# SECURITY EVENT RECOGNITION FOR VISUAL SURVEILLANCE

Wentong Liao[a], Chun Yang[b], Michael Ying Yang[c], Bodo Rosenhahn[a]

[a] Institute for Information Processing (TNT), Leibniz University Hannover, Germany
[b] Institute of Photogrammetry and GeoInformation (IPI), Leibniz University Hannover, Germany
[c] Scene Understanding Group, University of Twente, Netherlands
liao@tnt.uni-hannovder.de, michael.yang@utwente.nl

**Commission II/5**

**KEY WORDS:** Computer Vision, Event Recognition, Convolutional Neural Network, Video Surveillance

## ABSTRACT:

With rapidly increasing deployment of surveillance cameras, the reliable methods for automatically analyzing the surveillance video and recognizing special events are demanded by different practical applications. This paper proposes a novel effective framework for security event analysis in surveillance videos. First, convolutional neural network (CNN) framework is used to detect objects of interest in the given videos. Second, the owners of the objects are recognized and monitored in real-time as well. If anyone moves any object, this person will be verified whether he/she is its owner. If not, this event will be further analyzed and distinguished between two different scenes: moving the object away or stealing it. To validate the proposed approach, a new video dataset consisting of various scenarios is constructed for more complex tasks. For comparison purpose, the experiments are also carried out on the benchmark databases related to the task on abandoned luggage detection. The experimental results show that the proposed approach outperforms the state-of-the-art methods and effective in recognizing complex security events.

## 1. INTRODUCTION

Security at public place has always been one of the most important social topics. With rapidly increasing deployment of surveillance cameras, the reliable methods for automatically analyzing the surveillance videos and recognizing special events are demanded by different practical applications, such as security monitoring (Collins et al., 2000, Liao et al., 2015a), traffic controlling (Wang et al., 2009, Liao et al., 2015b), etc. Due to their large market and practical impact, much attention has been drawn in both computer vision and photogrammetry communities for decades. The task of security event analysis and detection refers to suspicious object detection and anomaly detection in given videos.

Since the object type of category occurring in surveillance scene is unexpected, traditional methods ignore the object type and use foreground/background extraction techniques to identify static foregrounds regions as suspicious object candidates. However, object type provides very important information for video event analysis. For instance, a black luggage is more suspicious than a pink wallet which has been left on the floor in an airport hall. Only detecting static items is insufficient to deeply and correctly analyze such complicated circumstance. The main reason that the previous works only focus on abandoned/left-luggage detection is the imperfect object detector which can only detect limited kinds of object categories with unsatisfied accuracy. In recent years, convolutional neural networks (CNNs) are driving advances in computer vision, such as image classification (Krizhevsky et al., 2012), detection (Girshick et al., 2014, Ren et al., 2015, Liu et al., 2016, Girshick et al., 2016), semantic segmentation (Long et al., 2015, Mustikovela et al., 2016), pose estimation (Toshev and Szegedy, 2014, Krull et al., 2015). CNNs have shown remarkable performance in the large-scale visual recognition challenge (ILSVRC2012) (Russakovsky et al., 2015). The success of CNNs is attributed to their ability to learn rich feature representations as opposed to hand-designed features used in traditional image classification methods. Therefore, it is a good choice to use deep learning methods to detect object type in the task of security event recognition.

Our goal in this work is to detect abandoned objects and then analyze the latter events related to them: its owner is taking it, or someone else is moving it to somewhere, or stealing it? These three security events are the most often occurring circumstances in our daily life. In this paper, CNN framework is used for object detection and verification. Because the previous works only focus on left object detection, appropriate benchmark dataset is missing for more complicated tasks. Therefore, we construct a new video event dataset: Security Event Recognition Dataset(SERD)[1] containing various scenarios within real-world environment. We evaluate our method on the benchmark PETS2006[2] and PETS2007[3]. A new dataset called ABODA provided by Lin et al. (Lin et al., 2015) is also used for further test. The results show that our algorithm outperforms the state-of-the-art methods for abandoned object detection inference. Besides our framework are evaluated on our dataset SERD for further more complicated tasks. Quantitative and qualitative comparisons with ground truth show that the proposed framework is effective for security event detection.

To summarize, our contributions are:

- We propose a novel framework which not only detects the abandoned object but also labels its owner and analyzes the event of a person interacting with the object.
- We utilize CNN for detection and verification tasks.
- A new video event dataset is provided especially for the task of security event detection.

This paper is structured as follows: related work is discussed in Sec. 2. The proposed framework is discussed in detail in Sec. 3.

[1] SERD will be publicly available on authors' homepage
[2] http://www.cvg.reading.ac.uk/PETS2006/data.html
[3] http://www.cvg.reading.ac.uk/PETS2007/data.html

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
doi:10.5194/isprs-annals-IV-1-W1-19-2017

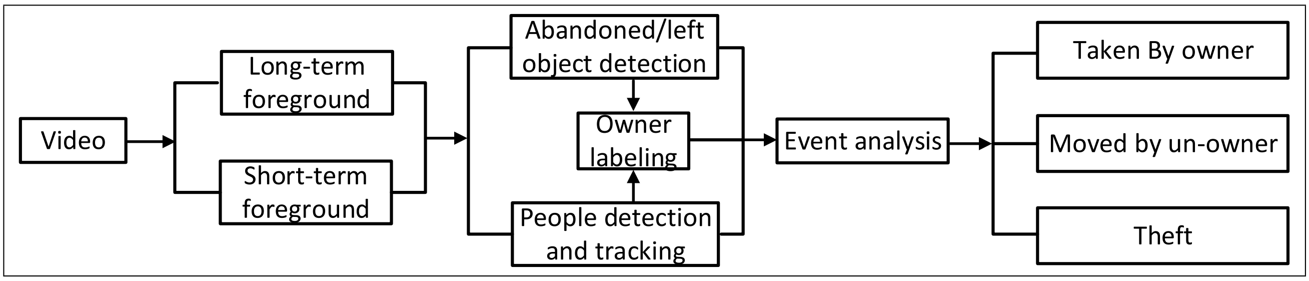


Figure 1: Flowchart of our framework.

Experimental results of the proposed framework are shown and analyzed in Sec. 4. Finally, conclusion in Sec. 5. summarizes this paper.

## 2. RELATED WORK

Security event recognition can be deemed as a special topic of activity analysis in video which has been one of the most popular topics for decades in computer vision. However, most of the attention focus on human motion/activity recognition and abnormal event detection (Wang et al., 2009, Ji et al., 2013, Wang et al., 2015, Simonyan and Zisserman, 2014). As a practical application topic, security event recognition attracted much less effort from researchers. And even most of the existing works for this topic only focus on the shallow task of detecting abandoned luggage (Porikli et al., 2007, Fan and Pankanti, 2011, Liao et al., 2008, Evangelio et al., 2011, Fan et al., 2013, Lin et al., 2015). They learn a robuster background model and then identify static foreground objects by subtraction. Their methods have some limitations in practice:

- First, pure foreground/background extraction model is very sensitive to illumination changing.
- Second, it is hard to divide individual foreground objects in crowded scene.
- Third, object category is ignored which is however very important for security event analysis.
- Fourth, background objects are also important components in some public scenes (such as retailer shop and lab), which have not drawn attention in their previous works.
- Last but not the least, to my best knowledge, all of the previous works don't care what will happen to the abandoned objects, for instance, who will move them or take them away. Such activity recognition is also crucial task for analyzing surveillance video.

To handle temporary occlusion in finding the owner, (Lin et al., 2015) used a back-tracing verification strategy. However, the verification is triggered only when there is no moving foreground object within the object's neighbor region of predefined radius. This method is inappropriate in practice. In addition, tracking person or object provide abundant information for further semantic analysis. Therefor, we also track persons to get their trajectories as (Tian et al., 2011, Fan et al., 2013, Liao et al., 2008) did. And we apply re-identification (Re-id) methods for person/object verification to solve the problems that they have encountered such as occlusion and imperfect tracking.

In this paper, we propose a framework to analyze complex security events in surveillance video of public scene. First, abandoned object is detected and its owner is also identified. Then the latter events happening on this object are analyzed. Different alarm is triggered if this object is moved by a un-owner. An overlook of our framework is illustrated in Fig. 1.

## 3. METHODOLOGY

Our framework is described by the key components of person and object detection, ownership labeling and security event analysis. In the following subsections, each component is discussed in detail.

### 3.1 Background Model

Static is the most obvious character of abandoned object. Thus, we also apply dual-background model to detect static region as previous works. The background is divided into long-term which is used for detecting static foreground objects, and short-term one for moving objects.

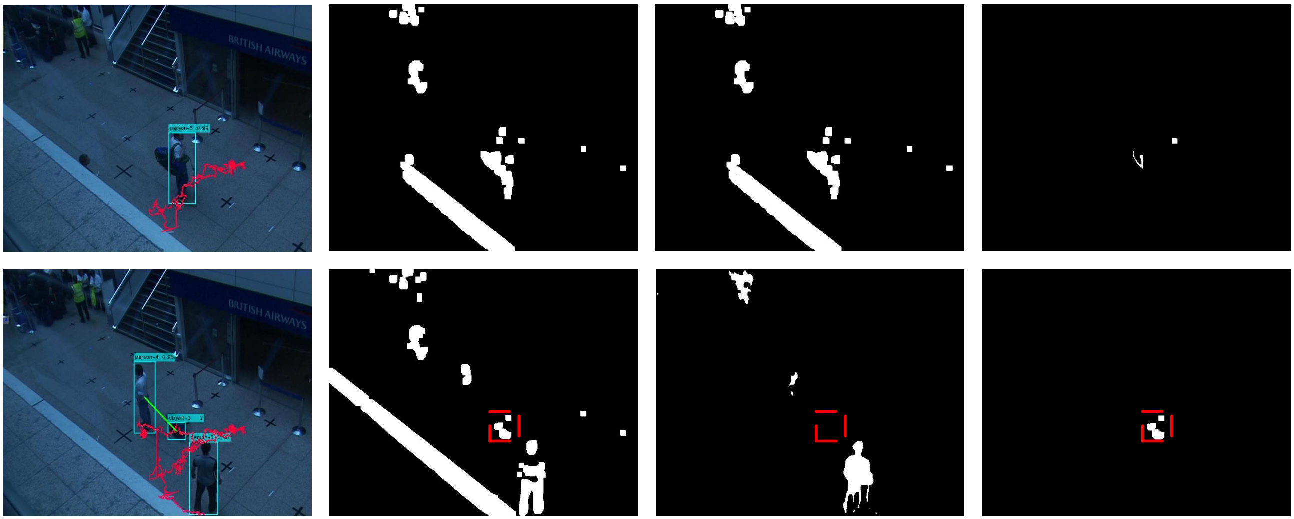
Long-term background model at time point $t$ is $\mathbf{BG}_L^t$ and the short-term one is $\mathbf{BG}_S^t$. We denote $\mathbf{F}_L^t$ as binary foreground image obtained via $\mathbf{BG}_L^t$ and $\mathbf{F}_S^t$ via $\mathbf{BG}_L^t$, as shown in Fig. 2(b) and 2(c) respectively.

We use the background model proposed in (Russell and Gong, 2006) in our framework. Here, 20 frames of each 50th frame are sampled for learning long-term background model and each 3th frame for short-term background model. With frame rate of 25Hz, the long-term background completely updates in each 40 seconds and the short-term background updates each 2 seconds.

### 3.2 Person and Object Detection

In recent years, deep learning based algorithms have shown great power in object detection and classification tasks (Russakovsky et al., 2015, Krizhevsky et al., 2012, Girshick et al., 2014, Ren et al., 2015). Thus, the algorithm faster region proposal convolution neural network (FrRCNN) is applied due to its "real-time" capability and high accuracy in this work.

We divide the objects of interest into background objects and foreground objects. Firstly, the FrRCNN is used to detect objects from the learned initial long-term background RGB image. These detected objects are registered in $\mathbf{BO} = \{BO_1, \dots, BO_{N_B}\}$, which indicates that these objects' belong to the background. To detect abandoned object and recognize security events, only the static objects are interested. Therefore, an $XOR$ operation is conducted between $\mathbf{F}_L^t$ and $\mathbf{F}_S^t$ to get the static foreground regions, as shown in Fig. 2(d). Then, FrRCNN is applied to detect objects within those regions. For instance, Fig. 2(d) shows the



(a) RGB image (b) Long-term foreground (c) Short-term foreground (d) Static foreground

Figure 2: An example of static foreground detection in PETS 2007 dataset. The time point in the second row is 270 frames after the one in the first row. (a) shows the person/object detection (bounding box) and owner labeling (green line). The red lines are the tracking traces of detected person. (b) and (c) are the foregrounds which are extracted from the long-/short-term model respectively. (d) shows the static foreground. The place of detected bag of interest is indicated by red bounding boxes.

foreground regions and the detected left bag is shown in Fig. 2(a). The FrRCNN is only used within the foreground regions instead of the whole image to reduce computation, which is important for real-time application. Here, 30 proposals are generated by FrRCNN instead of 300 proposals in the original work. All the detected objects are static objects and denoted as $\mathbf{SO} = \{SO_1, \ldots, SO_{N_O}\}$. $SO_i$ encodes the information of object category, bounding box, and its feature which will be discussed in Sec. 3.4. Note that, each $SO_i$ is checked in $\mathbf{BO}$ based on their bounding box and object type. The one which already exists in $\mathbf{BO}$ is canceled to avoid the misunderstanding of background objects as abandoned objects.

Persons are detected by FrRCNN based on the long-term foreground model $\mathbf{F}_L^t$ and denoted as $\mathbf{P} = \{P_1, ...P_{N_p}\}$. Subsequently, the real-time tracking algorithm proposed by Bewley et al. (Bewley et al., 2016) is utilized in our framework for tracking. The tracing information of each person is denoted as $T_i$.

### 3.3 Abandoning Detection and Ownership Labeling

Owner of an objects is an important information to make sure whether an object is abandoned or just left provisionally. It is also the crucial cue to analyze security events, such as theft. Thus, to identify the owner, we compute the average distance between $SO_i$ and each person's trace $T_i$ over time. The person with smallest distance to $SO_i$ is labeled as the owner and denoted as $OP_i$ (an example is shown in Fig. 2(a)). Because the shot-term background is updated in each 2 seconds, only the section of each trace from $T_i^{t-2s}$ to $T_i^t$ is considered, whereas $t$ is the time point of $SO_i$ being detected. A concrete example is shown in Fig. 7(b).

It is costly but unnecessary to watch all objects occurring in the surveillance scene. Security events of public scenes relates to abandoned objects mostly. Therefore, abandonment should be detected reliably. The based rule for abandoned object detection are originally defined by PETS2006. From temporal aspect, if an object is unattended move his bag in 30 seconds, the bag is declared as an abandonment. From the spatial aspect, an object is defines as abandonment if there is not owner within 3 meters. However, in practice the owner may stay in the scene for a very long time without touching his object. For instance, in the public rest area of a library, a student who wants a break put his bag on a table and then go to a vending machine for a while. This case satisfies the rules for abandonment, but the bag is not abandoned. A concrete example is shown in Fig. 7(b).
Besides, the spatial rule requires high quality calibration of cameras. Therefor, the rules for abandonment detection are modified to fit the practice application better as follows:

1) $OP_i$ is tracked going out of the surveillance scene, i.e. its trace is extending to the edge area of given scene.

2) If $OP_i$'s trace does not reach the edge area but it disperses from the scene longer than consecutive $T = 30$ seconds and $SO_i$ is still there, then $SO_i$ is labeled as abandoned object.

### 3.4 Security Event Analysis

To judge if an object is taken by its owner, moved or stolen by others, the person and object must be verified. Deep feature representations learned by CNN also show great effectiveness in the task of person Re-id (Li et al., 2014, Ahmed et al., 2015, Xiao et al., 2016). Here, the approach proposed by Xiao et al. (Xiao et al., 2016) is used to extract deep features for person and object verification.

To reduce unnecessary computation, only the objects which have been moved and the persons who are involved in the events are verified. When any object is being moved, the region indicated by its bounding box will be shown in the short-term foreground image $\mathbf{F}_S^t$. Therefore, the object whose bounding box involves foreground over a threshold of its area is counted as a possible moving object. Then this region is extracted as an input of FrRCNN to classify its object category. If the newly classified object category changes or its bounding box varies over a threshold, object $SO_i$ or $BO_i$ is recorded as moving/missed object $MO_i$. And the person who is now closest to it is labeled as candidate $CP_i$ for this event. Next, $CP_i$ needs to be verified if it is the

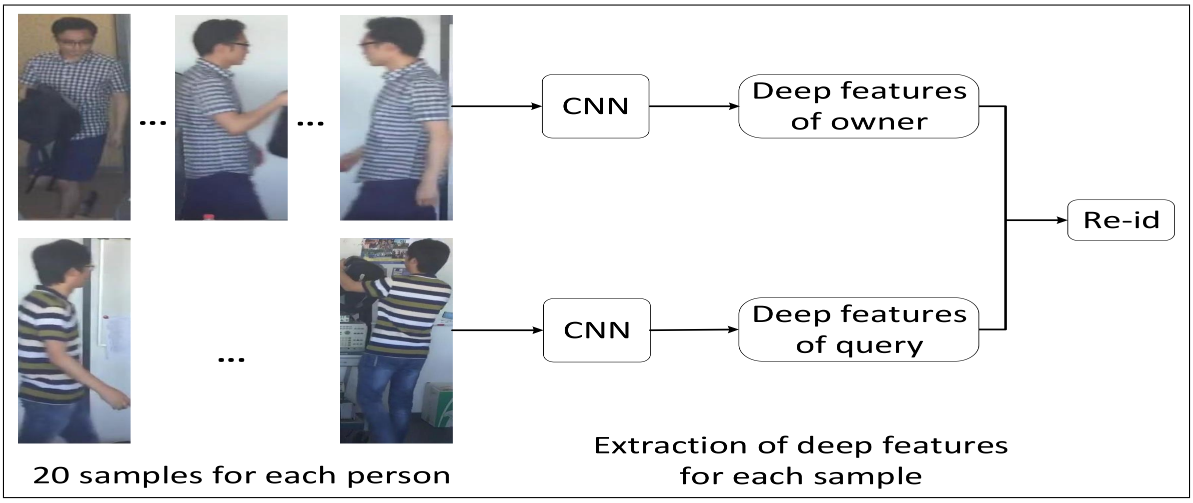

...
...
...
CNN
CNN
Deep features of owner
Deep features of query
Re-id
20 samples for each person
Extraction of deep features for each sample

Figure 3: Flowchart for person verification. 20 samples for the owner and the un-owner are extracted from the video sequence respectively. Then a CNN model which is specially trained for verification task is used to extract deep representation of each sample. Finally the person re-identification is done by comparing the deep features between the owner and the query person.

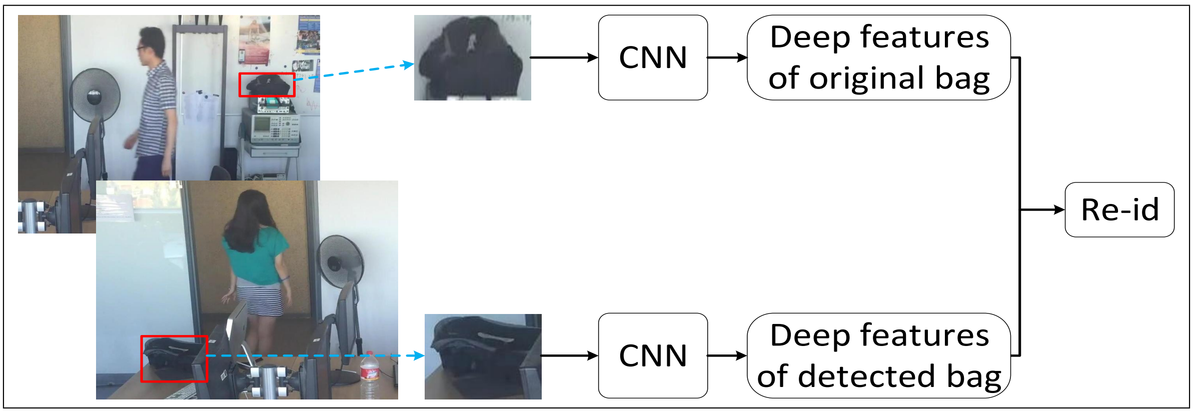


Figure 4: Flowchart for object verification. The process is very similar as person re-identification. But only one sample is extracted for each object when they are static.

owner of $MO_i$. If $MO_i$ is registered in $BO_i$, $CP_i$ is labeled as suspect because the background objects belong to the scene under surveillance. If $MO_i$ is from $SO_i$, a progress of people Re-id is carried out as follows.

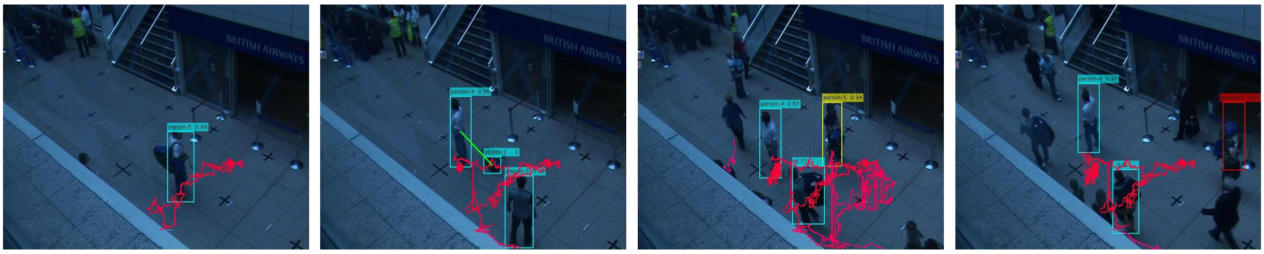
The pose and view angle of a person influence the verification results crucially. For example, two pictures which are captured from a man front and rear respectively are easily identified as two different persons. To enhance the Re-id accuracy, 20 samples are taken for each person. When a person is labeled as owner $OP_i$ or candidate $CP_i$, 20 frames are picked out from his first appearance till present in uniformly time interval, and 20 samples of them are cropped out from them respectively. In this way, the appearance information of this person is captured as different as possible. Each sample from $CP_i$ is compared with each one from $OP_i$ using the CNN framework (Xiao et al., 2016). This process is illustrated in Fig. 3. Then a $20 \times 20$ confused matrix is obtained to interpret the similarity of this two sets of samples. $M_{nm}$ denotes the similarity between $n$-th sample of $OP_i$ and $m$-th sample of $CP_i$. The similarity score is formally calculated as:

$$S_i = \arg\max_m \frac{1}{20} \sum_{n=1}^{20} M_{nm}. \qquad (1)$$

If $S_i$ is greater than a threshold, $CP_i$ and $OP_i$ are considered as the same person. $CP_i$, $SO_i$ and $MO_i$ are canceled from the their lists respectively, because it is not necessary to pay attention on $SO_i$ any more. Otherwise, $CP_i$ keeps the label as candidate for further watch.

In the later video frames, each newly detected object $SO_j$ is compared with each $MO_i$: $SO_j$ and $MO_i$ are cropped out from the their corresponding RGB images respectively and put into the CNN framework (Xiao et al., 2016) to verify if $SO_j$ is $MO_i$. Fig. 4 illustrates this verification progress. If yes, $CP_i$ is recognized as moving the object to a new place. When $CP_i$ disperses from the surveillance scene, or it reaches a predefined regions, such as exist, $MO_i$ is not detected again. Then this event is recognized as stealing and $CP_i$ is the theft.

We use the ImageNet (Russakovsky et al., 2015) pretrained CNN models and fine tune with some examples from the aforementioned datasets. For Re-id task, the pretrained CNN models is provided by (Xiao et al., 2016) without fine tuning.



(a) Person is detected and tracked (b) Luggage is put down and labeled (c) Luggage is picked by un-owner (d) It is recognized as steal

Figure 5: An example of experimental results on dataset PETS2007 from camera 3. Green line connects the bag with its owner. The yellow bounding box indicates a un-owner moving the bag while the red one indicates the man as a theft.

| - | (Li et al., 2006) | (Fan et al., 2013) | (Tian et al., 2011) | (Lin et al., 2015) | ours |
|---|---|---|---|---|---|
| Precision | 0.75 | 0.95 | 0.85 | 1.0 | 1.0 |
| Recall | 1.0 | 0.8 | 0.8 | 1.0 | 1.0 |

Table 1: Comparison of different methods on PETS2006 video dataset.

## 4. EXPERIMENTS

In this section, the performance of proposed framework is evaluated for security event recognition. In addition, the experimental results of abandoned luggage detection will also be compared with the-state-of-the-art methods.

### 4.1 Dataset and Implementation Details

The experiments are carried out on the following datasets to evaluate the performance of our framework for detecting security events: abandoned object detection, recognition of objects being moved by owner or non-onwer, or stolen.

1) The PETS2006 dataset consists of seven sequences of various scenarios. Beside the third one, each of the others includes an abandoning event.

2) The PETS2007 dataset comprises eight sequences captured from a crowded public scene and contains 3 scenarios: loitering, theft and abandoning object.

3) ABODA is proposed in (Lin et al., 2015) and more challenging for abandoned object detection. It has 11 sequences labeled with various scenarios as listed in Tab. 2.

4) The SERD video dataset is constructed by us for further evaluation of proposed framework for security event recognition. It comprises 3 sequences with more complex scenarios (such as theft) within a real-world environment. Two of them are captured from a student lab and the other one is taken from a public rest area of a university library.

On PETS2007/2007 datasets, only the sequences from camera 3 are used in our experiments.

### 4.2 Experimental Results

Our method is evaluated for detecting abandoned object on the benchmark dataset PETS2006. The experimental results are compared with the ones given by the state-of-the-art methods (Li et al., 2006, Fan et al., 2013, Tian et al., 2011, Lin et al., 2015). From the comparison in Tab. 1 we can see that, our results are same as the one form (Lin et al., 2015), but outperforms others. Furthermore, our method labels the owner of each abandoned object correctly.

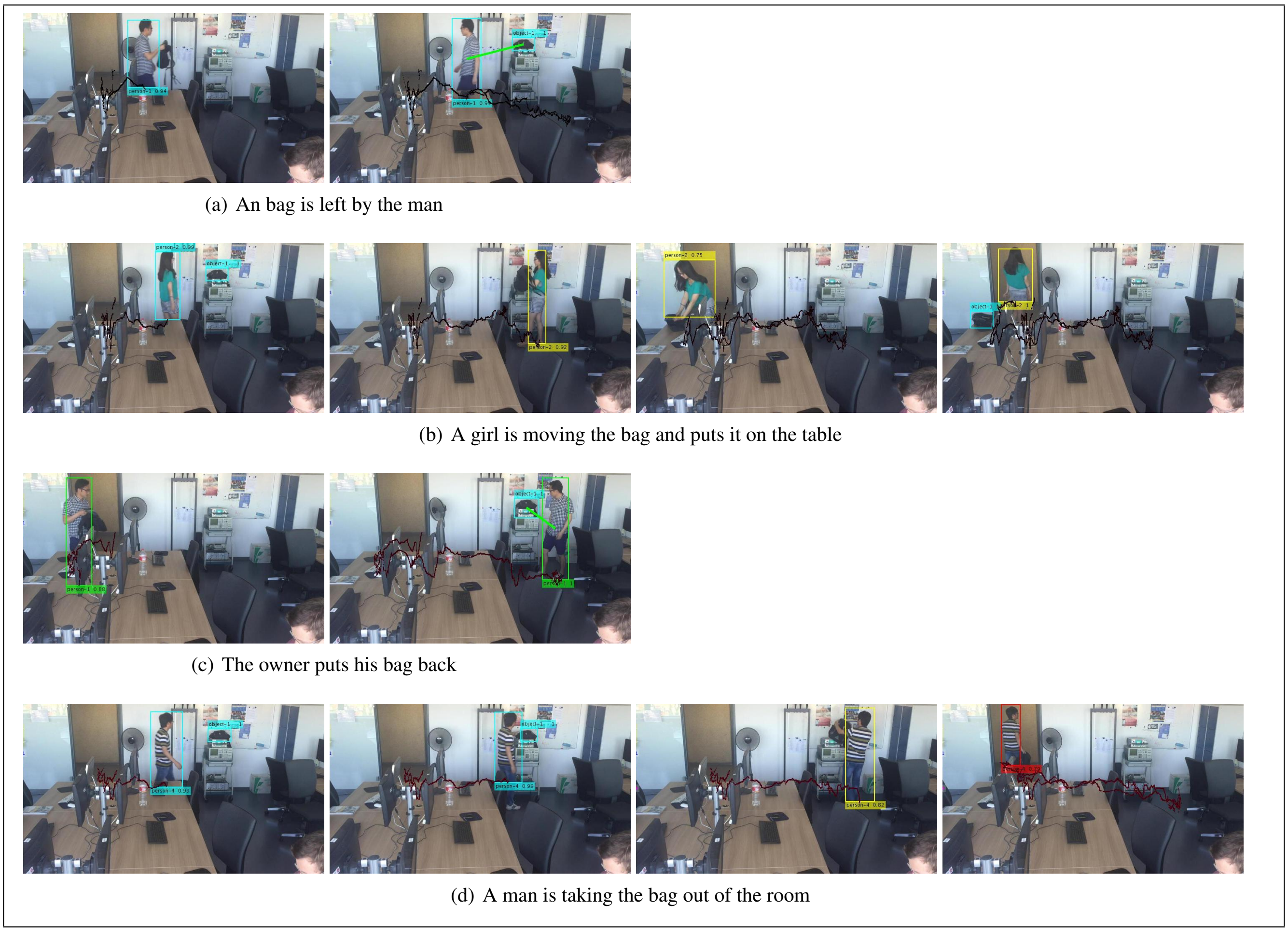
| - | - | (Lin et al., 2015) | | ours | | - |
|---|---|---|---|---|---|---|
| Video | GT | TP | FP | TP | FP | Scenerio |
| V1 | 1 | 1 | 0 | 1 | 0 | Outdoor |
| V2 | 1 | 1 | 0 | 1 | 0 | Outdoor |
| V3 | 1 | 1 | 0 | 1 | 0 | Outdoor |
| V4 | 1 | 1 | 0 | 1 | 0 | Outdoor |
| V5 | 1 | 1 | 0 | 1 | 0 | In Night |
| V6 | 2 | 2 | 0 | 2 | 1 | Light Switching |
| V7 | 1 | 1 | 1 | 1 | 0 | Light Switching |
| V8 | 1 | 1 | 1 | 1 | 1 | Light Switching |
| V9 | 1 | 1 | 0 | 1 | 0 | Indoor |
| V10 | 1 | 1 | 0 | 1 | 1 | Indoor |
| V11 | 1 | 1 | 3 | 1 | 1 | Crowded Scene |

Table 2: Comparison of different methods on ABODA. GT, TP, FP means ground truth, true positive and false positive respectively.

For further comparison, we conduct experiment on ABODA dataset and compare the results with the method (Lin et al., 2015). The experimental results are listed in Tab. 2, which shows that our method achieve comparable performance as (Lin et al., 2015) and outperforms it in crowded scene. That is because our method is based on object detection, which separates individual objects in crowded scene. In the scenario of light switching, both of our methods have made false positive detection. Illumination changing is really a challenging problem. Our method also successfully find the owner of each abandoned object on this dataset.

In the next step, we evaluate the performance of the proposed approach for analyzing complicate security events: object is taken by its owner, moved or stolen by a un-owner, which is the main goal of this work.

On dataset PETS2007, our method correctly detects abandoned object, labels owner and recognizes the theft in the 5th and 6th videos, and no false positive result is generated. However, false alarms about theft have been triggered in the 3rd and 4th videos. It is because in each of the scenarios, the owner places her/his bag on the ground, and then a familiar person of the owner picks the



(a) An bag is left by the man

(b) A girl is moving the bag and puts it on the table

(c) The owner puts his bag back

(d) A man is taking the bag out of the room

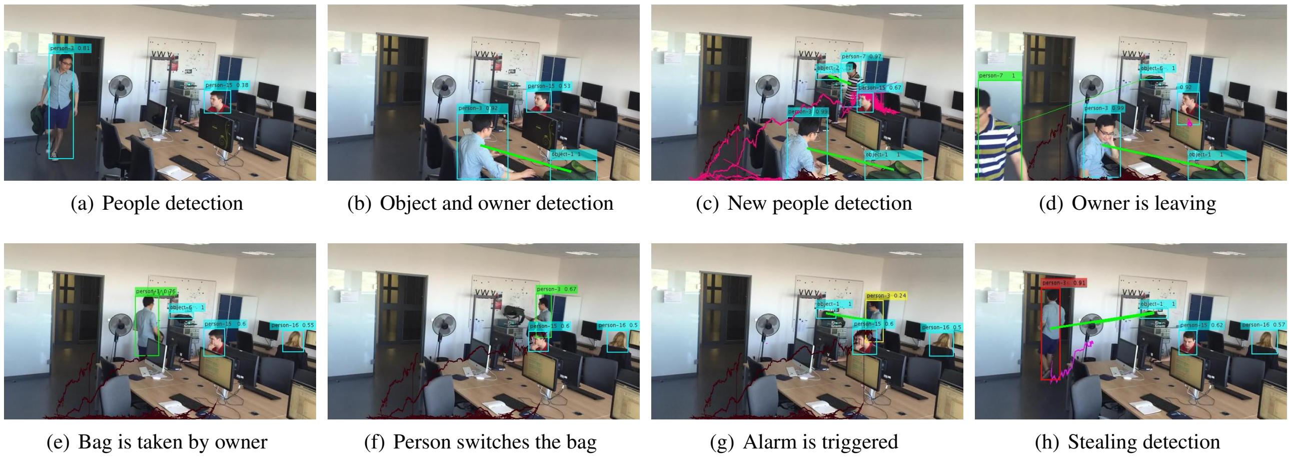
Figure 6: An example of typical experimental results on SERD. In this scene, series of events happens around a bag. The green bounding box means that the man is the owner and he is allowed to move the bag. (a) A bag is left in the room and the man is labeled as its owner (indicates with the green connecting line). (b) A girl comes to move the bag to a table. Her activity is alarm in yellow, because she is not the bag's owner but the bag is still within surveillance region. (c) The owner comes back to move his bag to another place. He is indicated in green because he is verified as the bag's owner. (d) Another man comes to take the bag away. A yellow alarm is caused when he is taking the bag but still in the room. Then the alarm turn in red when he goes out the room.

bag up and walks out of the scene. Our method does not proceed the semantic analysis of familiar/known person to the owner.

Finally, we validate proposed method on our own dataset. Fig. 6 illustrate the whole process of a series of events about an abandoned object. In the beginning, a person comes into the student lab, put his bag on the oscilloscope and then leaves the room. The object is recognized as an abandoned object and he is labeled as its owner (Fig.6(a)). Subsequently, a girl comes to pick the bag , which triggers an alarm by our algorithm. Then she put the bag on a table and leaves the room. Because the bag is detected again before the girl leaving the scene, this event is recognized as ”moved by un-owner” (Fig. 6(b)). Next, the owner comes back again and moves his bag to the original place. Since he is the owner of this bag, it is recognized as an allowable activity. After the bag is detected again on the oscilloscope, he is labeled again as the owner. When he is going to leave the room, the bag is recognized as an abandoned object again (Fig. 6(c)). Finally, another man comes to take the bag out of the lab. When he picks up the bag, an alarm for ”moved by un-owner” is caused. When he is detected to go out of the lab, the alarm for ”theft” is triggered (Fig. 6(d)). All security events are correctly detected in this video by our method.

The second video is also from the same lab but in different angle and scenarios as shown in Fig. 7. Person A comes into the lab and put his bag on the table, and then he sits there for a long time. He is labeled as the owner of the bag, and the bag is not recognized as an abandoned object(Fig. 7(a) and (b)). Person B put his bag on the oscilloscope and goes out of the camera view. He is labeled as the owner of his bag and the bag is recognized as an abandoned object when he is going out of the scene. Subsequently, A takes his bag away, which is recognized as allowable. A exchanges his bag with the bag of B, which causes an alarm. Meanwhile, A is still labeled as the owner of his own bag. When he is detected going out of the room, an alarm of stealing is triggered. Each event is correctly recognized and no false alarm is triggered by our method in this video.

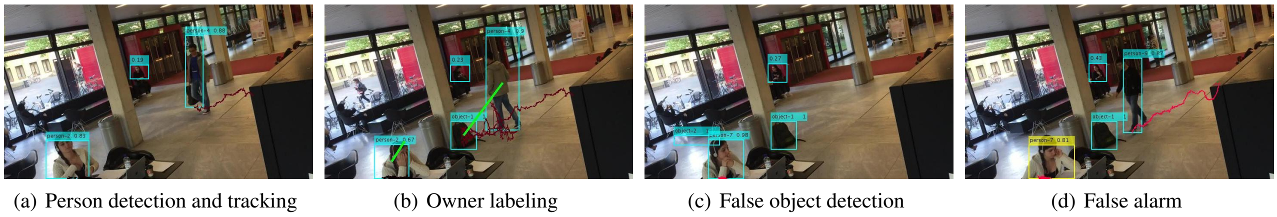
In the third video which is over a public rest area of a university library, our method doesn't perform well. The falsely detected events are shown in Fig. 8. Because of illumination changing in partial regions, objects are falsely detected which don't exist actually. For example, the slow sun light changing causes uneven illumination changing in the down-left part of the image. Then this part is recognized as some foreground objects. The girl is labeled as the owner shown in Fig. 8(b). Another falsely detected



(a) People detection (b) Object and owner detection (c) New people detection (d) Owner is leaving

(e) Bag is taken by owner (f) Person switches the bag (g) Alarm is triggered (h) Stealing detection

Figure 7: An example of experimental results on SERD. A man comes into the room (a). Then he left his bag on the table and begins to work(b). He is labeled as the owner but the bag is not labeled as abandoned object. Another man left his before the white board (c) and left the scene (d). He is labeled as the owner and the bag is labeled as abandoned object. The owner takes his bag away without alarm (e). He switches the bag by his (f), and a warning is issued that the bag does not belongs to him and he is labeled as the owner of the substitute objects (g). He is recognized as a theft when he is leaving (h).



(a) Person detection and tracking (b) Owner labeling (c) False object detection (d) False alarm

Figure 8: An example of experimental resutls on SERD of library scene.

| Event | Abandoning | | | Moved by owner | | |
|---|---|---|---|---|---|---|
| Scene | GT | TP | FP | GT | TP | FP |
| Lab1 | 1 | 1 | 0 | 1 | 1 | 0 |
| Lab2 | 2 | 2 | 0 | 1 | 1 | 0 |
| Library | 1 | 1 | 2 | 0 | 0 | 1 |
| | Moved by un-owner | | | Theft | | |
| Lab1 | 1 | 1 | 0 | 1 | 1 | 0 |
| Lab2 | 0 | 0 | 0 | 1 | 1 | 0 |
| Library | 1 | 1 | 1 | 1 | 1 | 0 |

Table 3: Experimental results on our video dataset SERD.

object in Fig. 8(c) is also recognized as abandoned object but not labeled belonging to the girl. When the light keeps changing, our method cannot detect the object anymore. Therefore, the girl is recognized taking it away because she is the closest person when this happens. A summary of the experimental results is shown in Tab. 3.

### 4.3 Real-Time Capability

The proposed system was developed using Matlab and ran in a DIGITS DevBox. Each frame with size $360 \times 240$ costs 0.12 seconds per average, i.e. computation speed is 8.33 fps. The most expensive computational cost of the framework is updating the dual-background models. Considering the motion speed of human beings is not so fast, if each 3 frames is taken as input to the framework, the proposed algorithm is scraped for real-time application without significantly decreasing the performance.

## 5. CONCLUSION

In this work, we propose a novel framework for security event recognition in surveillance videos which includes abandoned object detection and special event analysis. It is a significant extended application of state-of-the-art works which only focus on abandoned luggage detection. Different from previous works, our approach uses object detector, which benefits from the power of deep learning in visual tasks, instead of using foreground/background extraction for static item detection. The proposed approach outperforms the state-of-the-art methods for abandoned luggage detection. The effectiveness of our approach for more complex security event recognition has also been verified in various scenarios.

In the future, we will dedicate our effort to enable the algorithm to recognize more complex security events (such as familiar/known person recognition), improve the algorithm to accelerate the progressing speed for truly real-time application beyond update hardware, and make it more stable for dealing more challenging situations such as very crowded scenes.

## ACKNOWLEDGEMENTS

The work is funded by DFG (German Research Foundation) YA 351/2-1. The authors gratefully acknowledge the support.

## REFERENCES

Ahmed, E., Jones, M. and Marks, T. K., 2015. An improved deep learning architecture for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3908–3916.

Bewley, A., Ge, Z., Ott, L., Ramos, F. and Upcroft, B., 2016. Simple online and realtime tracking. arXiv preprint arXiv:1602.00763.

Collins, R. T., Lipton, A. J., Kanade, T., Fujiyoshi, H., Duggins, D., Tsin, Y., Tolliver, D., Enomoto, N., Hasegawa, O., Burt, P. et al., 2000. A system for video surveillance and monitoring. Technical report, Technical Report CMU-RI-TR-00-12, Robotics Institute, Carnegie Mellon University.

Evangelio, R. H., Senst, T. and Sikora, T., 2011. Detection of static objects for the task of video surveillance. In: IEEE Winter Conference on Applications of Computer Vision, pp. 534–540.

Fan, Q. and Pankanti, S., 2011. Modeling of temporarily static objects for robust abandoned object detection in urban surveillance. In: Advanced Video and Signal-Based Surveillance, pp. 36–41.

Fan, Q., Gabbur, P. and Pankanti, S., 2013. Relative attributes for large-scale abandoned object detection. In: International Conference on Computer Vision (ICCV), pp. 2736–2743.

Girshick, R. B., Donahue, J., Darrell, T. and Malik, J., 2016. Region-based convolutional networks for accurate object detection and segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 38(1), pp. 142–158.

Girshick, R., Donahue, J., Darrell, T. and Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition.

Ji, S., Xu, W., Yang, M. and Yu, K., 2013. 3d convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(1), pp. 221–231.

Krizhevsky, A., Sutskever, I. and Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp. 1097–1105.

Krull, A., Brachmann, E., Michel, F., Yang, M. Y., Gumhold, S. and Rother, C., 2015. Learning analysis-by-synthesis for 6d pose estimation in RGB-D images. In: International Conference on Computer Vision, pp. 954–962.

Li, L., Luo, R., Ma, R., Huang, W. and Leman, K., 2006. Evaluation of an ivs system for abandoned object detection on pets 2006 datasets. In: Proc. IEEE Workshop PETS, pp. 91–98.

Li, W., Zhao, R., Xiao, T. and Wang, X., 2014. Deepreid: Deep filter pairing neural network for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 152–159.

Liao, H.-H., Chang, J.-Y. and Chen, L.-G., 2008. A localized approach to abandoned luggage detection with foreground-mask sampling. In: Advanced Video and Signal Based Surveillance, pp. 132–139.

Liao, W., Rosenhahn, B. and Yang, M. Y., 2015a. Gaussian process for activity modeling and anomaly detection. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, ISPRS Geospatial Week, pp. 467–474.

Liao, W., Rosenhahn, B. and Yang, M. Y., 2015b. Video event recognition by combining HDP and gaussian process. In: International Conference on Computer Vision Workshop, pp. 166–174.

Lin, K., Chen, S.-C., Chen, C.-S., Lin, D.-T. and Hung, Y.-P., 2015. Abandoned object detection via temporal consistency modeling and back-tracing verification for visual surveillance. IEEE Transactions on Information Forensics and Security 10(7), pp. 1359–1370.

Liu, L., Lin, W., Wu, L., Yu, Y. and Yang, M. Y., 2016. Unsupervised deep domain adaptation for pedestrian detection. In: European Conference on Computer Vision Workshop on Crowd Understanding, pp. 676–691.

Long, J., Shelhamer, E. and Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.

Mustikovela, S. K., Yang, M. Y. and Rother, C., 2016. Can ground truth label propagation from video help semantic segmentation? In: European Conference on Computer Vision Workshop on Video Segmentation, pp. 804–820.

Porikli, F., Ivanov, Y. and Haga, T., 2007. Robust abandoned object detection using dual foregrounds. EURASIP Journal on Advances in Signal Processing 2008(1), pp. 1–11.

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

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C. and Fei-Fei, L., 2015. Imagenet large scale visual recognition challenge. International Journal of Computer Vision 115(3), pp. 211–252.

Russell, D. M. and Gong, S., 2006. Minimum cuts of a time-varying background. In: British Machine Vision Conference (BMVC), Vol. 6, Citeseer, pp. 809–818.

Simonyan, K. and Zisserman, A., 2014. Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems, pp. 568–576.

Tian, Y., Feris, R. S., Liu, H., Hampapur, A. and Sun, M.-T., 2011. Robust detection of abandoned and removed objects in complex surveillance videos. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 41(5), pp. 565–576.

Toshev, A. and Szegedy, C., 2014. Deeppose: Human pose estimation via deep neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1653–1660.

Wang, L., Qiao, Y. and Tang, X., 2015. Action recognition with trajectory-pooled deep-convolutional descriptors. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4305–4314.

Wang, X., Ma, X. and Grimson, W. E. L., 2009. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(3), pp. 539–555.

Xiao, T., Li, H., Ouyang, W. and Wang, X., 2016. Learning deep feature representations with domain guided dropout for person re-identification. arXiv preprint arXiv:1604.07528.