Abandoned Objects Detection in Video Sequences

A. Kukleva[1], V. Konushin[2], A. Konushin[1 3]
anna.kukleva@graphics.cs.msu.ru|vadim@tevian.ru|ktosh@graphics.cs.msu.ru
[1]Lomonosov Moscow State University, Moscow, Russia;
[2]Video Analysis Technologies, LLC, Moscow, Russia;
[3]NRU Higher School of Economics, Moscow, Russia

In this paper we handle a problem of abandoned objects detection in video sequences by using a deep convolutional neural network (CNN). At first, a background subtraction based detection [6] is used, which shows high recall and computational speed. Then CNN, which is trained on generated synthetic data, is used to filter out false positives and classify abandoned objects types. The proposed algorithm was tested on a privately collected video collection and showed good performance on long videos.

Keywords: *abandoned objects, object detection, neural network classifier.*

1. Introduction

Currently, the automatic video sequence analysis is a developing field due to the increase in safety requirements in public places. A city-scale video surveillance system includes the collection, processing and storage of videodata. In these systems hundrends of thousands of cameras are mounted in all public places, including staircase landings, courtyards, places of mass congestion of citizens, educational institutions, private enterprises, subways. The number of cameras is increasing every year, and it becomes impossible for operators to process incoming information in real time. Automatic video analysis tools can help personnel to focus only on dangerous or strange occurrences.

One of the main goals of automatic video surveillance is the detection of bags and luggage that were left without attendance in public places, such as airports. Such object are called ”abandoned”. It can also be noted that algorithms that solve this problem can be applied in other fields of video analysis as well: for example Fig. 1, monitoring the safety of valuable objects and when a parking lot is full.

Fig. 1. Examples of abandoned objects
Left – parked car, right – abandoned bag.

Currently existing algorithms, which are based on object tracking or background modeling, cannot ensure a sufficiently high accuracy under real conditions to provide a good reliability level. In recent years in many computer vision problems the rapid progress is demonstrated by methods, which are based on convolutional neural networks. But training a convolutional neural network requires a large amount of training data. It is important problem, since for some tasks it is very difficult to obtain a sufficient amount of training data. The detection of abandoned objects is an example of such task. If we provide enough training data for implementation of neural network, we can expect a significant increase in abandoned objects detection accuracy.

2. Related Works

Most of successful abandoned object detection algorithms developed in the previous years rely on background subtraction. Two different approaches can be distinguished.

The first family of works relies on using combination of existing background sustraction algorithms and other independant methods such as object tracking, cumulative mask formation for foreground objects, people detection [9–14].

The second family of works modifies background subtraction algorithms for solving this problem. For instance modifications can relate to alteration of one or several distribution of Gaussian mixture model or integration of a finite state machine into background model [1–5].

There are plenty of algorithms for classification of the founded objects: decision trees, linear classificator, SVM, a bag of words, etc. In recent years deep convolution neural networks demonstrate the best results in the classification problems.

The similar approach [12] to one described in this paper was published concurrently with this article in summer 2017. There are the following differences. First in [12] Gaussian Mixture Model is used for background modeling and static object detection while we employ the algorithm based on accumulation and analysis pixel-by-pixel masks. Second, we use different architecture of convolutional neural network for the classifier of abandoned objects and false detections.

3. Proposed algorithm

Our proposed algorithm consists of two main steps:

1. Detection of a potential abandoned object (hypotheses generation) with background subtraction algorithm.
2. Classification of the found hypotheses with neural network.

4. Hypotheses search algorithm

After analyzing the existing approaches for detecting static objects and those removed from a scene, we selected as a baseline the algorithm [6], which demonstrated a high

recall. The algorithm takes sparse video frames as input, about one frame per second, which allows it to work in realtime.

The output of the baseline algorithm is a set of static foreground regions (non-moving for several previous frames and different from background).

Several masks are generated to handle each frame:

- Frame difference accumulation mask – pixel-by-pixel difference between two neighbor frames.
- Background accumulation mask – pixel-by-pixel comparison of the gradient of a current frame and a background model.
- The intersection of the previous two masks is a stationary pixel mask, and a pixel is implanted into the background model upon reaching a certain threshold.

To obtain hypotheses on the stationary pixel mask, connected components are marked by bounding boxes. Fig. 2 shows baseline algorithm as a block diagram.

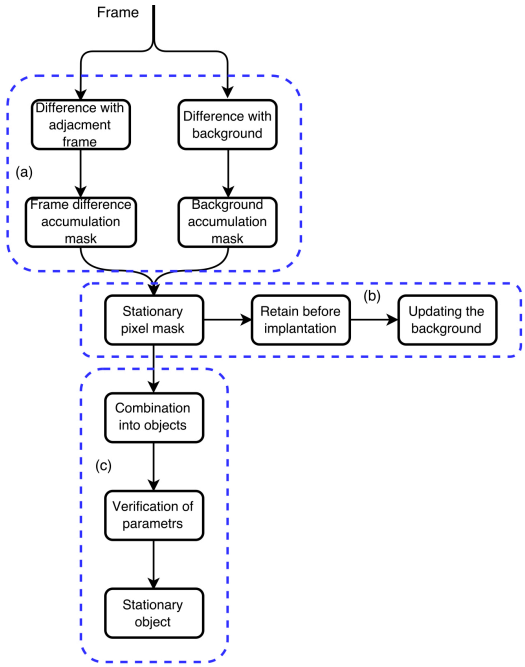

Fig. 2. Block diagram of the baseline algorithm.

An image gradient is used to reduce the impact of illumination intensity instead of the original image. The pixel difference is estimated considering the noise parameters of a particular video. These parameters are determined before the main algorithm starts working.

We propose to automate noise level estimation for each video.

We need to acquire the noise level in a video sequence from input images. We can assume that the noise is white – a stationary noise, the spectral components of which are evenly distributed over the entire range of frequencies involved, and we can minimize the problem up to the determination of the root-mean-square noise deviation.

We have selected the noise level estimation method on the basis of image blocks from [8]. It is easy to implement, requires a small amount of memory, and showed fairly high performance in the tests.

A frame sequence is selected from each video and then equally spaced geometrically rectangles with non-moving scenes were cut out. Since moving objects distort the evaluation, they should be absent in these areas. This is achieved by applying a background subtraction algorithm and generating an accumulating background mask.

$$AccMask_t(x) = \begin{cases} 0, & \text{foreground pixel;} \\ AccMask_{t-1}(x) + 1, & \text{otherwise.} \end{cases}$$

The Gaussian mixture model was chosen as background subtraction algorithm. To avoid random noise, a morphological erosion operator is applied to each background mask.

Every 30th iteration, a sliding window algorithm searches for the largest area of an accumulating background mask, which remains static at least 50 consecutive frames.

An image gradient is used to extract noise parameters. The mean value is calculated for the entire selected sequence of areas with 3x3 filter and the mean value of the root-mean-square noise deviation for N frames is calculated using the following formula:

$$f(x) = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\sigma_k^2},$$

where σ – Root-mean-square deviation of frame number k from mean.

5. Classification

5.1. Data preparation

To validate our approach in challenging realistic environments, we collected data from several cameras monitoring busy university hall, parking and subway with a lot of loitering people or cars within about 40 hours of video. All objects were manually classified in this data set. It contains 234 ground truth objects which is not enough for training deep neural network. As a result, we created synthetic data (Fig. 3). We have made the collection of background images, which reflects all kinds of conditions: cabinet, park, street, shop and others, and the collection of bags composed of bags with different colors and shapes. The average bag size corresponding to the real background conditions was determined manually for each background image.

For easy embedding of bags into background, it is necessary that its background is white, without noises and shadows. The image with a bag is binarized, the contour of an object is located, the alpha channel of a background around the object is equated to 0, the object is combined

with a randomly selected part of a background image. The size of the generated image depends on the manual mapping of the mean suitable size for the background and on the stretch factors of the chosen bag along the vertical and horizontal directions, the length of each side can range from 20 to 500 pixels.

Fig. 3. Left – real data
Right – synthetic data.

Thereby more then 15 thousand synthetic bags were produced.

5.2. Convolutional neural network

For hypotheses classification we used a CaffeNet, which is a replication of AlexNet [7], in which the order of normalization layers and poolingis is switched. It is pre-trained on ImageNet, a large image database, convolution layers capture well generalized features at lower levels and specific features for a specific task at higher levels of the architecture.

Two approaches are employed for training common convolutional neural network in this paper: fine-tuning the CaffeNet network pre-trained on Imagenet and training from scratch the same architecture. Mini-batch consists of 20 images, initial learning rate is 0.01 multiplying by 0.1 every 2500 iterations. Activation function is ReLU.

6. Experimental Results

Firstly, we compared the base method [6] with its modification consisting of the base algorithm and the block noise level estimation [8]. Results are in the table 1. P stands for precision and R stands for recall.

PETS2006	P	R
Base algorithm	1.00	1.00
Proposed algorithm	1.00	1.00

Table 1. Dataset PETS2006. Results.

As you can see from the table 1 both of the algorithms ideally detect all abandoned objects from PETS2006 [2].

For algorithm evaluation such number of videos and overall duration isn't enough. A proper dataset was collected and composed of videos with total duration of 40 hours. There are various background scenes such as university hall, classrooms, underground, parking places and etc. Further experimental estimation was carried out on this dataset.

Assembled dataset	P	R
Base algorithm	0.13	0.89
Proposed algorithm	0.09	0.96

Table 2. Assembled dataset. Results.

As seen in the table above (Table 2), precision was significantly decreased but this algorithm should only show high recall as it is followed by CNN classification.

Classification accuracy was measured on different subsets of the data. 2 classes[1] – bags and false alarms, 3 classes[2] – bags, cars and false alarms. $CaffeNet_0$ – trained from scratch, $CaffeNet_{ft}$ – fine-tuned.

Network	2 cl.[1]		3 cl.[2]	
	P	R	P	R
$CaffeNet_0$	0.91	0.95	0.72	0.86
$CaffeNet_{ft}$	0.93	0.95	0.77	0.87

Table 3. Classification result.

Fine-tuning shows better results (Table 3) than training from scratch. Apparently there is no such diversity of synthetic data as to ImageNet dataset.

Proposed method is a superposition of hypotheses search and images classification algorithms. The total recall of abandoned objects detection for long videos can be found in the table 4.

Network	2 cl.[1]		3 cl.[2]	
	P	R	P	R
$CaffeNet_0$	0.91	0.91	0.72	0.81
$CaffeNet_{ft}$	0.93	0.91	0.77	0.83

Table 4. Result of the whole algorithms.

7. Conclusion

In this work we proposed a method for abandoned object detection, which combines the hypotheses search algorithm and the neural network classifier. We employed algorithm [6] for detection of potential abandoned objects and implemented automated noise estimation in the baseline algorithm. We used convolutional neural network as a classifier. For training classifier we generated synthetic data. This approach demonstrates good results on long real videos where false alarms occur much more often than true positive abandoned objects. Fine-tuning has a slight advantage over training from scratch.

As future work, we will explore the use of different approach for background subtraction [10]. It can help to generate more qualitative hypotheses and decrease amount of ghosts. This algorithm can be integrated with people tracking algorithm [11] to define object owner.

8. References

[1] Bayona A., SanMiguel J.C., Martinez J.M. Stationary foreground detection using background subtraction and temporal difference in video surveillance. IEEE International Conference on Image Processing, 2010.

[2] Dataset Pets2006. http://www.cvg.reading.ac.uk/PETS2006/data.html

[3] Evangelio R.H., Sikora T. Complementary background models for the detection of static and moving objects in crowded environments. 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2011. – pp 71–76.

[4] Fan Q., Gabbur P., Pankanti S. Relative Attributes For Large-scale Abandoned Object Detection. ICCV2013, 2013, – pp 2736–2743.

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

[6] Kireev D., Konushin V. Stationary object detection. Conference "Lomonosov – 2016", 2016. (in Russian).

[7] Krizhevsky A., Sutskever I., Hinton G. ImageNet Classification with Deep Convolutional Networks. In NIPS, 2012.

[8] Lapshenkov E. Non-standard noise level estimation of digital image based on harmonic analysis. Computer optic, 2012, vol 36, №3. (in Russian).

[9] Miguel J.C.S., Martinez J.M. Robust unattended and stolen object detection by fusing simple algorithms. IEEE International Conference on Advanced Video and Signal Based Surveillance, 2008.

[10] Morozov F., Konushin A.S. Background subtraction using a convolutional neural network. Proceedings of the 26th International Conference on Computer Graphics and Vision GraphiCon'2016, pp. 445–447.

[11] Shalnov E., Konushin V., Konushin A. An improvement on an MCMC-based video tracking algorithm. Pattern Recognition and Image Analysis, Allen Press Inc. (United States), vol 25, pp. 532–540.

[12] Sidyakin S.V., Vishnyakov B.V. Real-time detection of abandoned bags using CNN. Proc. of SPIE Vol. 10334 103340J-2.

[13] Tian Y.L., Feris R., Liu H. Robust detection of abandoned and removed objects in complex surveillance videos. IEEE Transactions on Systems, Man, and Cybernetics, 2010. – 41(5) – pp 565–576.

[14] Wen J., Gong H., Zhang X., Hu W. Generative model for abandoned object detection. IEEE International Conference on Image Processing, 2009.