Journal of Physics:
Conference Series

PURPOSE-LED
PUBLISHING™

PAPER • **OPEN ACCESS**

Video Surveillance for Indoor Office Environment Based on Object-Level Anomaly Detection

To cite this article: Haotian Chen *et al* 2023 *J. Phys.: Conf. Ser.* **2504** 012029

View the article online for updates and enhancements.

You may also like

- FAIAD: Feature Adaptive-based Image Anomaly Detection
Qianqian Zhang, Hongyang Wei, Xusheng Du et al.
- An improved two-stream CNN method for abnormal behavior detection
Luo Sha, Yuan Zhiwen, Xu Kan et al.
- Research on Recognition Algorithm of Abnormal Behavior of Workers in Two-Stream Convolutional Network
Yinong Zhao, Ruoxi Li, Xuedong Zhang et al.

Enabling Electrochemistry
Inside and Outside the Lab

PalmSens

Discover our diverse range of Potentiostats engineered to support your research wherever it takes you

- Intuitive software for Windows and Android
- Proven in over 4000 scientific publications
- Single-channel and multi-channel
- Benchtop, portable and wearable

Discover More



This content was downloaded from IP address 162.105.91.221 on 01/11/2026 at 07:10

Video Surveillance for Indoor Office Environment Based on Object-Level Anomaly Detection

Haotian Chen[1,2*][0000-0001-6329-4091], **Chao Zhang**[1][0000-0002-1018-4144], **Zhouchen Lin**[1][0000-0003-1493-7569]

[1] School of Intelligence Science and Technology, Peking University, No.5 Yiheyuan Road Haidian District, Beijing, P.R.China 100871

[2] North King Information Technology Co., Ltd., 7th Floor, Qingzheng Building, No. 25, Xisanhuan North Road Haidian District, Beijing, P.R. China 100089

[*]Corresponding author's e-mail: archristy@pku.edu.cn

Abstract. Traditional methods of Abnormal Behavior Detection (ABD) process the surveillance video on frame-level, which ignores object-level abnormal behavior patterns. To address the problem, this paper presents Object-Level Anomaly Detection model (OLAD), which aims to model various normal behavior patterns of different objects and the normal interaction patterns between them. Specifically, OLAD introduces an encoding-embedding network to transform object-level information into the feature space. By integrating such information, OLAD processes both frame-level and object-level cues in the video for ABD. In addition, we construct our own dataset Northking2022 especially for office scenes because of the lack of public datasets for indoor office environments. Experimental results show that OLAD gains better performance on both the public benchmark and Northking2022.

1. Introduction

Abnormal Behavior Detection (ABD) task aims at automatically determining whether the behavior of individuals in surveillance videos is abnormal. Since abnormal behavior is often associated with suspicious behavior, abnormal behavior detection is critical to intelligent video surveillance systems. Machine learning methods in ABD can be divided into supervised and unsupervised categories.

Supervised methods attempt to model normal and abnormal samples through data annotation during the training phase, which requires predefined abnormal behaviors according to different scenarios [1-4]. However, supervised ABD methods are difficult to generalize. Due to the wide variety of human behaviors, the training data are not able to cover all the situations. In addition, in many scenarios, one cannot give a clear definition of all possible anomalies. These problems lead to unsupervised ABD methods that do not require labeled samples at all. Therefore, many scholars have turned to unsupervised methods to study ABD.

Compared with artificial features, deep features have stronger representation ability. Thus, many scholars propose unsupervised ABD methods based on deep learning. Sun et al. [5] used Growing Neural Gas (GNG) to cluster the training samples and judged the testing samples based on their distance to cluster centers. Hasan et al. [6] fitted the distribution of normal samples on a low-dimensional subspace/manifold using a convolutional auto-encoder (CAE), and then determined whether the testing sample can be reconstructed by the model. Convolutional 3D (C3D) can simultaneously capture spatial and temporal information. Nogas et al. [7] combined C3D with AE to

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

calculate the reconstruction error of video spatiotemporal blocks. Ravanbakhsh et al. [8] used U-Net to reduce the information loss caused by dimensionality reduction and realize cross-modal reconstruction of video frames and optical flow graphs.

We observe that all the above methods model frame-level features as a whole in the video. Nevertheless, abnormal behavior patterns of different objects are not the same. Moreover, some abnormal behavior patterns are contained in the interaction of different types of objects. Therefore, this paper proposes Object-Level Anomaly Detection model (OLAD), which aims to model the variety in the normal behavior patterns of different objects and the normal interaction patterns between different objects. Specifically, in order to distinguish different objects in an image, OLAD first adopts a pre-trained object segmentation method to acquire the semantic annotation of different objects. The semantic annotations and recognition probabilities will be used as the input of the ABD module in the form of a mask. Then, different object information needs to be integrated into the features, in order to guide the subsequent network to learn the normal behavior patterns of different objects and the normal interaction patterns between different objects. OLAD utilizes an encoding-embedding method to fuse semantic information and recognition probability into the feature map. Finally, a prediction scheme is used for training and testing OLAD.

In addition, we construct our own dataset Northking2022 for office scenes, because of the lack of public datasets for indoor office environments. This dataset is realistically captured from real office scenarios and its scale is equivalent to the Avenue dataset [9]. Experimental results show that our method achieves AUC=76.0% on Northking2022.

2. Dataset

Unsupervised ABD methods usually assume that objects in the training dataset have normal behavior patterns. Then they model normal behavioral patterns by training a model on the normal data. On the contrary, the test dataset contains both normal and abnormal behavior patterns. During the testing process, the detection of abnormal behavior is completed by comparing the difference between the output of the model and the real behavior.

To enable the best performance of the model, it is necessary to guarantee that the training and testing datasets are not collected from different camera perspectives. Traditional public datasets (such as CUHK Avenue [9]) are all outdoor scenes, and the models trained on them are difficult to be directly applied to indoor office environment. Therefore, this paper presents a realistic office-scene dataset, Northking2022.

Northking2022 contains 437 video clips, a total of 55551 frames. And each clip is about 7-9 seconds long. Among them, there are 41 indoor abnormal behavior clips, 6648 frames, consisted of 5 types: running, jumping, falling, smoking, and fighting. 241 normal behavior clips (40630 frames) are used for training. The remaining 52 and abnormal behavior fragments form the test dataset.

Table 1. Comparison between public benchmarks and our Northking2022

Dataset	Size	Anomaly
CUHK Avenue [9]	30,652 frames	12.5%
Ours (Northking2022)	55,551 frames	12.0%

The statistics of different datasets are shown in Table 1 for comparison.

3. Our approach

3.1. Object-Level Anomaly Detection (OLAD)

Figure 1 illustrates the detailed structure of OLAD. In general, OLAD follows a predictive approach for unsupervised anomaly detection. Similar to [10], the input of OLAD is t=4 adjacent frames in the video, and the output is the prediction result for the 5th frame. Actual frame's anomaly score is measured by the error between the predicted frame and it.

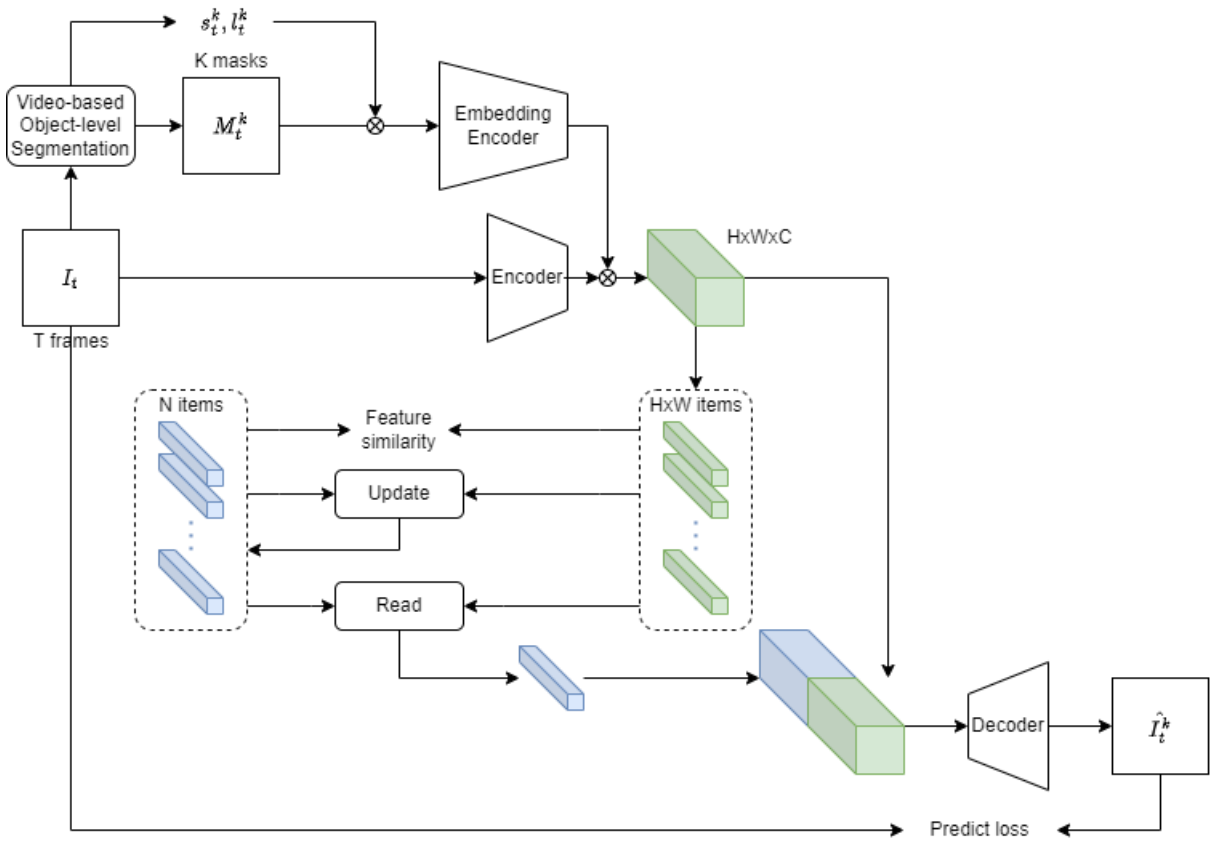


Figure 1. Detailed structure of our OLAD

OLAD processes the input video frames in two branches. One of them employ traditional frame-level unsupervised ABD methods (such as MNAD algorithm [10]), and the other introduces video-based object-level image segmentation methods (for example Mask2Former algorithm [13,14]). The MNAD model is mainly divided into three parts: encoder, decoder, and memory module.

Among them, the input image $I \in \mathcal{R}^{t\times h\times w\times 3}$ is encoded as the feature map $F_{\text{frame}} \in \mathcal{R}^{H\times W\times C}$ by a C3D Encoder. Here $h \times w$ and $H \times W$ are the size of frame and feature map, respectively. And C indicates the dimensional of features. The memory module stores N pattern features of size $f \in \mathcal{R}^{1\times 1\times C}$. The model calculates the similarity between each pixel on the feature map and each pattern feature, and then all pattern features are updated according to the similarity. At the same time, a weighted pattern feature from the memory module can also be "read" according to the similarity. Finally, these weighted pattern features will be concatenated with the original feature map and decoded into the predicted frame result. Detailed description can be found in [10].

Although MNAD explicitly identifies N patterns, these patterns do not involve object-level information in each frame and can only model frame-level behaviors, but not object-level ones. Therefore, the other branch of OLAD involves a video-based object-level image segmentation method to extract information of each object k, which consists of object category label $l_k \in \mathcal{L}$, object segmentation mask $M_k \in \mathcal{R}^{t\times h\times w\times 1}$, and confidences$_k \in \mathcal{R}$.

In order to integrate the object-level information, we design an encoding-embedding network based on C3D: Enc_Emb$\{\sum_k l_k \cdot s_k \cdot M_k\}$. This network transforms the input into the object-level feature map $F_{\text{object}} \in \mathcal{R}^{H\times W\times C}$, which is the same dimension as the traditional feature map. Finally, we fuse the two kinds of information to form the final feature map by an integrating function: $\text{Int}\{F_{\text{frame}}, F_{\text{object}}\} = F_{\text{frame}} * F_{\text{object}}$.

3.2. Loss function

In unsupervised ABD, the video frame itself is the supervised signal used to distinguish normal samples from abnormal ones. During training, the final loss function is calculated as the linear combination of prediction, feature compactness, and feature separateness losses [10].

3.3. Abnormal score

In order to quantitatively identify normal and abnormal frames in testing videos, it is necessary to assume that normal frames are similar to the patterns in the memory module. In other words, the video frames in the training process only contain normal patterns. Following [10], the min-max normalization of L2 distance and PSNR (peak signal-to-noise ratio) is adopted for anomaly score's calculation, where the L2 distance measures the gap between the actual observed frame features and the pattern features memorized by the model, and the PSNR measures the pixel-level gap between the actual observed frame and the predicted one from the model.

4. Experiments

4.1. Settings

As preprocessing, each video frame is normalized the color channel to the range of [-1, 1] and rescaled into 256×256. We set the parameters H, W, C to 32, 32 and 512, respectively. The memory number N of the feature map equals to 10. We adopt Adam [11] with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$ as the optimizer. The batch size is 4 and the epoch equals to 60. We choose the learning rate to be 2×10^{-4} with a cosine annealing strategy [12] as its decay. The model is trained using PyTorch. The program is running on a TITAN Xp.

To evaluate the results of ABD, during inference, we use average area under curve (AUC) [10].

4.2. Datasets

Our method is evaluated on Northking2022 along with CUHK Avenue [9], a public benchmark for ABD. Avenue consists of 47 abnormal events, (such as throwing stuff or running), which forms 16 training and 21 testing videos.

4.3. Comparison with the state-of-the-art

We compare the results of OLAD with SOTA, which is shown in Table 2. OLAD outperforms MNAD [10] with a 2.2% and 3.9% gain on Avenue and Northking2022, respectively. This is because MNAD models the frame in the video as a whole, without considering that abnormal behavior patterns of different objects are not the same. On the other hand, OLAD aims to model the differences in the normal behavior patterns of different objects and the normal interaction patterns between different objects, which boosts the ABD performance.

Table 2. OLAD compared with SOTA on Northking2022 and Avenue [9]. * indicates that the results are reproduced with their public code

Methods	AUC (%) on Avenue	AUC (%) on Northking2022
Unmasking [15]	80.6	-
ConvAE [16]	80.0	-
TSC [17]	80.6	-
StackRNN [17]	81.7	-
MemAE w/ memory module [18]	83.3	-
MNAD* [10]	84.3	72.1
OLAD	**86.5**	**76.0**

Table 3. Ablation studies for OLAD on Northking2022

Methods	AUC (%)
Baseline	72.1
OLAD (w/ fully-connect network as embedding)	71.6
OLAD (w/ addition as aggregation)	67.7
OLAD	**76.0**

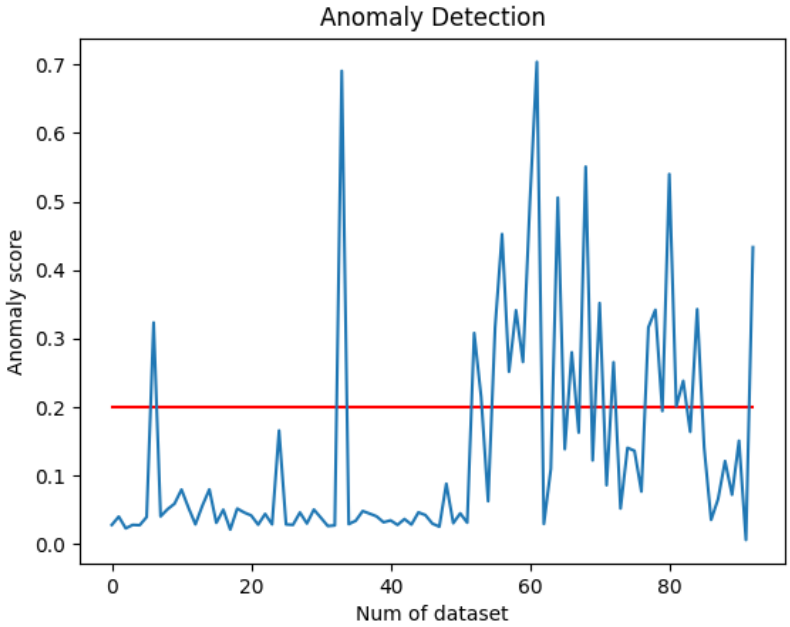


Figure 2. Qualitive Studies for OLAD in Surveillance Applications. "Num of dataset" means the clip number in the testing dataset of Northking2022. The first 50 of them are clips from normal scenes, and the rest are from abnormal scenes.

4.4. Ablation studies for OLAD

The results, where we tried different design choices on Northking2022, are demonstrated in Table 3. As a direct approach, object-level information could be integrated into the frame by a simple fully-connect network as embedding module. However, the second row of Table 3 shows that the integration realized in this way actually lower the ABD performance. On the contrary, OLAD introduces an encoding-embedding method that aggregates the object-level information through an encoder. As a result, the AUC is boosted by 4.4%.

The third row of Table 3 changes the aggregation operation from multiplication to addition. It is outperformed by the baseline and the one with fully-connect network and is inferior to our design choice for OLAD.

4.5. Qualitive Studies for OLAD in Surveillance Applications

OLAD outputs anomaly scores on frame level, which cannot be directly used in surveillance applications. Thus, anomaly scores on clip level are calculated by averaging the ones on all the frames within each clip. And a threshold is set to 0.2 to alert the surveillance system.

Figure 2 illustrates the results on the testing dataset of Northking2022. We can see that, only two of the normal clips wrongly raise the alert, and a majority of the abnormal ones are detected by our OLAD.

5. Conclusions

This paper proposes Object-Level Anomaly Detection model (OLAD) to address the problem of ignoring object-level abnormal behavior patterns in ABD. In order to process both frame-level and object-level cues in the video, OLAD contains an encoding-embedding network to transform object-level information into the feature space. By modeling various normal behavior patterns of different objects and the normal interaction patterns between the objects, OLAD gains better performance on both the public benchmark and our own dataset Northking2022 especially constructed for office scenes because of the lack of public datasets for indoor office environments.

Acknowledgments

This work is supported by the Beijing Postdoctoral Research Foundation (2022-ZZ-085), National Key R&D Program of China (2018AAA0100300), and National Nature Science Foundation of China (62071013 and 61671027).

References

[1] Sikandar T, Ghazali K H, Rabbi M F 2019 ATM crime detection using image processing integrated video surveillance: a systematic review Multimedia Systems 25(3) 229-251
[2] Wang M, Zhu Y, Sun Z, et. al. 2019 Abnormal behavior detection of atm surveillance videos based on pseudo-3d residual network 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA) 412-417
[3] Parab A, Nikam A, Mogaveera P, et. al. 2020 A new approach to detect anomalous behaviour in atms 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS) 774-777
[4] Goswami S, Goswami J, Kumar N 2015 Unusual event detection in low resolution video for enhancing ATM security 2015 2nd International Conference on Signal Processing and Integrated Networks (SPIN) 848-853
[5] Sun Q, Liu H, Harada T 2017 Online growing neural gas for anomaly detection in changing surveillance scenes Pattern Recognition 64 187-201
[6] Hasan M, Choi J, Neumann J, et. al. 2016 Learning temporal regularity in video sequences. Proceedings of the IEEE conference on computer vision and pattern recognition 733-742
[7] Nogas J, Khan S S, Mihailidis A 2020 Deepfall: non-invasive fall detection with deep spatio-temporal convolutional autoencoders Journal of Healthcare Informatics Research 4(1) 50-70
[8] Ravanbakhsh M, Sangineto E, Nabi M, et. al. 2019 Training adversarial discriminators for cross-channel abnormal event detection in crowds 2019 IEEE Winter Conference on Applications of Computer Vision (WACV) 1896-1904.
[9] Lu C, Shi J P, and Jia J A 2013 Abnormal event detection at 150 FPS in MATLAB 2013 IEEE International Conference on Computer Vision 2720-2727
[10] Park H, Noh J and Ham B 2022 Learning Memory-guided Normality for Anomaly Detection Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 14372-14381
[11] Kingma D P and Ba J 2015 Adam: A method for stochastic optimization International Conference on Learning Representations
[12] Loshchilov I and Hutter F 2017 SGDR: Stochastic gradient descent with warm restarts International Conference on Learning Representations
[13] Cheng B, Misra I, Schwing A G, et. al. 2022 Masked-attention Mask Transformer for Universal Image Segmentation Proceedings of the IEEE/CVF Conference on Computer Vision and

Pattern Recognition
[14] Cheng B, Schwing A G and Kirillov A 2021 Per-Pixel Classification is Not All You Need for Semantic Segmentation Conference on Neural Information Processing Systems
[15] Ionescu R T, Smeureanu S, Alexe B, et. al. 2017 Unmasking the abnormal events in video 2017 IEEE International Conference on Computer Vision
[16] Hasan M, Choi J, Neumann J, Roy-Chowdhury A K, et. al. 2016 Learning temporal regularity in video sequences Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
[17] Luo W X, Liu W, and Gao S H 2017 A revisit of sparse coding based anomaly detection in stacked RNN framework 2017 IEEE International Conference on Computer Vision
[18] Gong D, Liu L Q, Le V, et. al. 2019 Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection 2019 IEEE International Conference on Computer Vision