Global Information Guided Video Anomaly Detection

Hui Lv
huhrthui@njust.edu.cn
Nanjing University of Science and Technology
China

Chunyan Xu
cyx@njust.edu.cn
Nanjing University of Science and Technology
China

Zhen Cui
zhen.cui@njust.edu.cn
Nanjing University of Science and Technology
China

ABSTRACT

Video anomaly detection (VAD) is currently a challenging task due to the complexity of “anomaly” as well as the lack of labor-intensive temporal annotations. In this paper, we propose an end-to-end Global Information Guided (GIG) anomaly detection framework for anomaly detection using the video-level annotations (i.e., weak labels). We propose to first mine the global pattern cues by leveraging the weak labels in a GIG module. Then we build a spatial reasoning module to measure the relevance between vectors in spatial domain with the global cue vectors, and select the most related feature vectors for temporal anomaly detection. The experimental results on the CityScene challenge [1] demonstrate the effectiveness of our model.

CCS CONCEPTS
• Computing methodologies → Scene anomaly detection.

KEYWORDS

Anomaly detection, Weak supervision

ACM Reference Format:
Hui Lv, Chunyan Xu, and Zhen Cui. 2020. Global Information Guided Video Anomaly Detection. In Proceedings of ACM Multimedia Grand Challenge (Multimedia 2020). ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Detecting abnormal events in video sequences is a popular task due to the real-world applications such as surveillance and fault detection systems. Anomalies are often defined as behavioral or appearance patterns that do not conform to usual patterns [3, 6, 10]. And it is a time-and-labor-consuming job to manually identify anomalies in videos. Therefore, it is a pressing need to develop intelligent computer vision algorithms to analyze the large amount of raw video data and detect video anomalies automatically. So far the task is extremely hard. The challenges include insufficient annotated data due to the rare occurrences of anomalies, large inter/intra class variations, subjective definition of anomalous events, low resolution of surveillance videos, etc. Hence, the motion and appearance cues are vital for video anomaly detection. For example, the motion flow of road vehicles can be used to infer accidents. Also appearance information, e.g., fire and smoke, benefits a lot for detecting anomalies like explosion. In this paper, we are motivated to explore these motion pattern in temporal domain and appearance information in spatial domain.

The goal of a practical anomaly detection system is to timely signal an activity that deviates normal patterns and identify the time window of the occurring anomaly. We propose to leverage the video-level annotations for video anomaly detection under weak supervision. Previously, anomaly detectors tackle the problem in a two-stage manner [8, 11, 12]. In the first stage, the authors use well pre-trained models to extract high-level semantic feature vectors from the raw video clips, then they divide each video sequence into segments and group the clip-level feature vectors to segment-level. In the second stage, they build anomaly detection models by taking segment-level features as input and outputting the corresponding anomaly scores. The whole process is tedious that one must prepare specific semantic features first and read the saved feature files for anomaly detection.

In this paper, an end-to-end anomaly detection framework is designed to facilitate anomaly detection in videos with a high confidence and efficiency. We first utilize a backbone model to extract features from video-clips, upon which we build a Global Information Guided (GIG) module to localize anomalies in temporal domain. The whole network architecture is depicted in Figure 1. In the designed GIG module, we first leverage the video-level label to capture the global pattern cue in a video sequence. Then we enhance the spatio-temporal feature representation with the global information by channel-wise attention mechanism.

In addition, spatial information is largely neglected in previous approaches [8, 11, 12]. They extracted the semantic features after the global pooling operation from pre-trained model. In this way, the spatial resolution is directly reduced to $1 \times 1$, hence the appearance cue and relationship among objects are restrained largely. While these informations are of great important for detecting and distinguishing anomalies. For example, Robbery describes the action between people, Stealing usually happens between people and property, and if a gun is holding in one’s hand, it is easy to witness a Shooting event. Inspired by this, besides the motion semantics extracted by action classification model, we further investigate the spatial cues by exploring the relevance between spatial feature vectors and global pattern cue vector in a Spatial Reasoning (SR) Module. Then we select the most relevant spatial vectors for later anomaly classification and localization.

To tackle the weak supervision VAD, with only the video-level annotations available, we propose a novel Video-Segment objection function, namely VS Loss, for utilization of video-level annotations and expanding it to segment-level supervision. The whole function contains two components, which are Video-level Supervision...
Global Information Guided Module

Here, we explain the details of our Global Information Guided model for video anomaly detection. In this model, we propose to mine the Global Pattern Cue (GPC) within each video sequence and utilize the global information to enhance the anomaly representation. Given the extracted high-level semantic feature maps \( X = \{x_i^T\}_{i=1}^T \) of all segments, with resolution \( w, h, d \) of \( x \), we can obtain the global pattern cue by spatio-temporal dimension reduction operation. We apply the simple and effective Global MaxPooling – 3D function here as:

\[
g = \psi(X),
\]

where \( \psi \) denotes the dimension reduction operation, \( g \) is the vector of the GPC with resolution of \( \mathbb{R}^{1 \times 1 \times d} \).

Inspired by the great progress in channel attention [5, 7], we adopt a simple channel-wise attention operation to imply the GPC vector \( g \) as a guidance for representation enhancement, formally:

\[
\bar{X} = \sigma(g) \odot X + X,
\]

where \( \bar{X} \) represents the enhanced feature maps, \( \sigma \) is a sigmoid function for normalizing the weight scalars to \([0, 1]\) and \( \odot \) denotes the channel-wise multiplication operation. For preserving the original information, we leverage the skip connection to generate the final representation.

Further we propose to take the video-level annotation \( a \) as the supervision upon the GPC, which can also be viewed as an intermediate supervision. At First, we employ a fully connected function \( \phi_1 \) as the anomaly classification head to measure the anomaly status, using the GPC vector as input and outputting scores of \( 1 + C \) classes (one normal class and \( C \) anomaly class). Here we select to analyze the anomalous status of the GPC vector in a coarse level by detecting whether any anomaly instance exists in the video, rather than detecting which the anomaly class is, if exists. In detail, we take the maximum score among \( C \) anomaly classes of the GPC vector as the overall anomaly score \( S^a_g \):

\[
S^a_g = \max_{i=1}^{C} (\sigma(\phi_1(g_i))),
\]

where \( \sigma(\phi(g))_i \) denotes the anomaly score of \( i \)-th class. After the maximum operation, the global anomaly score becomes a two-channel vector, indicating the normal and abnormal probability respectively. Then we apply a Video–level Supervision with Binary Cross Entropy loss as:

\[
\ell^a_g = -(y^* \log S^a_g + (1 - y^*) \log (1 - S^a_g)).
\]

Here, \( y^* \) is set to 1 if the video contains any anomaly instance, otherwise 0.

Spatial Reasoning Module

After mining the global information, we further analyze the relationship between each spatio-temporal vector and the GPC vector to retrieve spatial cue vital for anomaly classification and localization. Previous methods [8, 11, 12] usually apply a global_pooling operation to integrate the spatial information, however in this way, the object appearance cues and the interaction between objects are largely neglected, which is of great importance for distinguishing the anomalies. For example, the instances of Robbery, Stealing and Shooting are prone to false detection. Robbery describes the event...
between at least two people, Stealing usually happens between people and property, and if a gun is holding in one’s hand, it is easy to witness a Shooting event.

In the designed SRM, we take the GPC vector $g$ and the enhanced spatio-temporal feature maps $\{\hat{x}\}_{i=1}^{T} \in \hat{X}$ as inputs. At the first step, we measure the spatial key by calculating the relation score between $g$ and $\hat{X}$, formally:

$$r_{i,j} = R(g, \hat{x}_{i,j}),$$

with $i, j$ denoting the row and column index, $r$ as the relation score, and $R$ as the Reasoning function. Here we choose the cos similarity as the relation measurement metric. At the second step, we choose the $k$ spatial vectors with the top-$k$ relation scores and further reduce the spatial dimension to 1 for obtaining the spatial pattern vector $\hat{x}_{k}$ of each segment, by aggregating the spatial vectors corresponding as:

$$\hat{x}_{k} = \sum_{i=1}^{w} \sum_{j=1}^{h} \text{top}_k(\hat{x}_{i,j}, r_{i,j}).$$

Then we apply an anomaly classification head $\phi_2$ for generating the pattern vectors of each segment as:

$$S_{k} = \sigma(\phi_2(\hat{x}_{k})).$$

Once we obtain the anomaly scores $\{S_{k}\}_{i=1}^{T}$ of $T$ segments of a video, we need to summarize the results from segment-level to video-level. In detail, we pick up the top $p$ maximum segment-level scores and average them to make a segment-consensus score as:

$$S_{\text{top}} = \frac{1}{p} \sum_{i=1}^{T} \text{top}_p(S_{k}).$$

We apply the Segment-level Supervision on the segment-consensus score. At first, we enable a supervision on the overall anomalous status as the GPC vector, formally:

$$S_{\text{g}} = \max_{i=1}^{T}(S_{\text{top}}^i).$$

$$\ell_{\text{g}} = -(y \log S_{\text{g}} + (1 - y) \log (1 - S_{\text{g}})),$$

Similar as the GPC vector, $S_{\text{g}}$ is the binary overall anomaly score and $\ell_{\text{g}}$ is the corresponding loss term. We further distinguish the anomaly instances by supervising the segment-consensus anomaly score under multi-class anomaly detection function as:

$$\ell_{k} = -(y \log S_{k} + (1 - y) \log (1 - S_{k})).$$

here, $y$ is the multi-hot anomaly label and $\ell_{k}$ is the anomaly classification loss. The Segment-level Supervision and Video-level Supervision make up our Video-Segment loss $\ell_{\text{vs}}$.

$$\ell_{\text{vs}} = \ell_{k} + \lambda_1 \ell_{\text{g}} + \lambda_2 \ell_{\text{g}},$$

with $\lambda_1, \lambda_2$ representing the balance weights for various terms. Further more, since the anomaly mostly occurs for a short period of time in real life, we also add a sparse constraint as in [12]:

$$\ell_{\text{sparse}} = \sum_{i=1}^{T} S_{\text{top}}^i.$$
Finally, we collect all the loss terms to get the overall supervision signal $\ell$, with $\lambda_3$ being the weight for the sparse term and $\ell_s$ from Eqn. 12. Formally:

$$
\ell = \ell_s + \lambda_3 \ell_{\text{sparse}}
= \ell_s + \lambda_1 \ell_s + \lambda_2 g_{\text{sparse}} + \lambda_3 \ell_{\text{sparse}}.
$$

4 EXPERIMENTS

4.1 Dataset

We conduct method on the CityScene challenge [1]. CitySCENE dataset consists of videos which cover 12 real-world anomalies related to public safety and city management, including Accident, Carrying, Crowd, Explosion, Fighting, Graffiti, Robbery, Shooting, Smoking, Stealing, Sweeping, and WalkingDog. The training set consists of 758 normal and 1319 anomalous videos. The videos in training set are trimmed manually and the labels are at video-level. Videos in testing set are untrimmed, with a total amount of 188. In addition, anomaly mostly occurs for a short period of time in testing videos.

4.2 Implementation Details

We adopt the Temporal Pyramid Network (TPN) [9], which is a state-of-the-art action classification network, as our backbone model. We use the resnet50 version of TPN and extract feature maps after the TPN module in [9] with the resolution of $R^{7x7x2048}$. We load the model pre-trained on kinetics400 as the default weights. In addition, anomaly mostly occurs for a short period of time in testing videos.

In this work, we propose an end-to-end Global Information Guided anomaly detection framework for anomaly classification and localization. It is the first end-to-end anomaly detection model under weak annotations, as far as we know. We leverage the video-level labels to mine the global pattern cue of the entire video sequences as first and enhance the learned representation with the global pattern information by channel-wise attention. Then we retrieve the spatial cue in each video segment by filtering the spatial vectors. The AUC of our method reaches up to 84.09%, superior than ActionLab (70.92%) and Orange-Control (31.65%). Although the AUC performance of our method is lower than the best approach, the gap is small. For task 2, our method wins the 3rd place in the challenge with MF1 of 52.33%, which exceeds the 4th method with a margin of 7%. In addition, our GIG model can run at almost 100 fps on a 2080Ti GPU. The fast speed, as well as the state-of-the-art performance shows the superiority of our GIG model.

5 CONCLUSION

This work was supported by the National Natural Science Foundation of China (Grants Nos.61772276).
REFERENCES

[1] [n.d.]. The CityScene Challenge. https://cityscene.github.io.

[2] John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* (2011).

[3] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. 2016. Learning temporal regularity in video sequences. In CVPR.

[4] Chengkun He, Jie Shao, and Juyu Sun. 2018. An anomaly-introduced learning method for abnormal event detection. *Multimedia Tools and Applications* (2018).

[5] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In CVPR.

[6] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. 2013. Anomaly detection and localization in crowded scenes. *TPAMI* (2013).

[7] Xiang Li, Wenhui Wang, Xiaolin Hu, and Jian Yang. 2019. Selective kernel networks. In CVPR.

[8] Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-world anomaly detection in surveillance videos. In CVPR.

[9] Ceyuan Yang, Yinghao Xu, Jiaping Shi, Bo Dai, and Bolei Zhou. 2020. Temporal Pyramid Network for Action Recognition. In CVPR.

[10] Bin Zhao, Li Fei-Fei, and Eric P Xing. 2011. Online detection of unusual events in videos via dynamic sparse coding. In CVPR.

[11] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. 2019. Graph Convolutional Label Noise Cleaner: Train a Plug-and-play Action Classifier for Anomaly Detection. In CVPR.

[12] Yi Zhu and Shawn Newsam. 2019. Motion-Aware Feature for Improved Video Anomaly Detection. *BMVC* (2019).