Modularized Framework with Category-Sensitive Abnormal Filter for City Anomaly Detection

Jie Wu
Sun Yat-sen University
wujie10558@gmail.com

Yingying Li
Department of Computer Vision Technology (VIS), Baidu Inc
liyingying05@baidu.com

Wei Zhang
Department of Computer Vision Technology (VIS), Baidu Inc
zhangwei99@baidu.com

Yi Wu
Energy Development Research Institute, China Southern Power Grid
wuyi2@csg.cn

Xiao Tan
Department of Computer Vision Technology (VIS), Baidu Inc
tanxiao01@baidu.com

Hongwu Zhang
Department of Computer Vision Technology (VIS), Baidu Inc
zhanghongwu@baidu.com

Shilei Wen
Department of Computer Vision Technology (VIS), Baidu Inc
wenshilei@baidu.com

Errui Ding
Department of Computer Vision Technology (VIS), Baidu Inc
dingerrui@baidu.com

Guanbin Li*
Sun Yat-sen University
liguanbin@mail.sysu.edu.cn

ABSTRACT

Anomaly detection in the city scenario is a fundamental computer vision task and plays a critical role in city management and public safety. Although it has attracted intense attention in recent years, it remains a very challenging problem due to the complexity of the city environment, the serious imbalance between normal and abnormal samples, and the ambiguity of the concept of abnormal behavior. In this paper, we propose a modularized framework to perform general and specific anomaly detection. A video segment extraction module is first employed to obtain the candidate video segments. Then an anomaly classification network is introduced to predict the abnormal score for each category. A category-sensitive abnormal filter is concatenated after the classification model to filter the abnormal event from the candidate video clips. It is helpful to alleviate the impact of the imbalance of abnormal categories in the test phase and obtain more accurate localization results. The experimental results reveal that our framework obtains a 66.41 MF1 in the test set of the CitySCENE Challenge 2020, which ranks first in the specific anomaly detection task.

KEYWORDS

Anomaly Detection, Temporal Segment Network, Temporal Shifting Module, Category-Sensitive Abnormal Filter

ACM Reference Format:
Jie Wu, Yingying Li, Wei Zhang, Yi Wu, Xiao Tan, Hongwu Zhang, Shilei Wen, Errui Ding, and Guanbin Li. 2020. Modularized Framework with Category-Sensitive Abnormal Filter for City Anomaly Detection. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA*. ACM, Seattle, USA, 5 pages. https://doi.org/10.1145/3394171.3416279

1 INTRODUCTION

With the increasing use of cameras on the road, the smart city center can collect large amounts of videos. These videos play an important role in urban management, public safety, traffic control, environmental protection and other aspects. One of the key tasks of these videos is to detect abnormal events such as traffic accidents, crimes or illegal activities. It is unrealistic to extract abnormal events from such large data by humans because abnormal events rarely occur compared with normal activities. Therefore, the research of abnormal event detection is of great significance, which is the focus of academia and industry. However, due to the complexity of anomalies in urban scenes and the diversity of backgrounds, imaging quality and motion, detecting various abnormal events in these realistic videos is a challenging problem.

A video anomaly detection challenge is held in ACM Multimedia 2020 to address the challenges of anomalous event detection in real-world scenario. This challenge publishes a new challenging dataset, namely CitySCENE. CitySCENE is a new video dataset that consists of more than 1000 untrimmed real-world videos, with 12 realistic anomalies related to public safety and city management, including accident, carrying, crowd, explosion, fighting, graffiti, robbery, shooting, smoking, stealing, sweeping, and walkingdog. This dataset and challenge serve as a common tool for researchers to benchmark their algorithms with each other and thus contribute to reproducible research.

*This work is done when Jie Wu was a research intern at Baidu. Corresponding author is Guanbin Li. This work was supported by the State Key Development Program under Grant No. 2016YFB1001004, the Guangdong Basic and Applied Basic Research Foundation under Grant No.2020B1515020048, National Natural Science Foundation of China under Grant No.61976250 and Grant No.61702565, the National High Level Talents Special Support Plan (Ten Thousand Talents Program), the Fundamental Research Funds for the Central Universities under Grant No.18lgpy63, and was also sponsored by CCF-Tencent Open Research Fund.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3416279

To deal with the above challenge, we design a modularized framework for city anomaly detection. The modularized framework contains a video segment extraction module, an anomaly classification module and a category-sensitive abnormal filter. The video segment extraction module first splits the video into several nonoverlapping video segments. Then an abnormal classification module utilizes the temporal segment network with temporal shift module to predict the abnormal score of the corresponding video segments. A category-sensitive abnormal filter strategy is employed to process the model outputs to obtain the final localization results.

The contributions of this work are summarized as follows:

• We design a modularized framework for city anomaly detection. This framework can address the general and specific anomaly detection tasks under a unified framework.

• A category-sensitive abnormal filter is designed to select the abnormal event from the candidate video clips. It contributes to alleviate the impact of the imbalance between abnormal categories and the normal ones at the testing stage and obtain more accurate localization results.

• Experimental results on CitySCENE Challenge 2020 demonstrate that our method outperforms other competition methods, and rank first in the specific anomaly detection task.

The source code of the two versions including general anomaly detection and specific anomaly detection has been made public at https://github.com/WuJie1010/CitySCENE2020-Anomaly-Detection.

2 RELATED WORK

As a most challenging task in the computer vision field, anomaly detection in the city scenario has been extensively studied in the last ten years [2–4, 6, 7, 10, 11, 14, 17–20]. Sultani *et al.* [14] collect the UCF-Crime [14] dataset that contains a series of videos from diverse categories in complicated surveillance scenarios. Sultani *et al.* [14] first propose the task of weakly supervised anomaly detection that merely resorts to video-level labels to learn anomaly. They simultaneously take advantage of both normal and abnormal videos to optimize the detection model. NVIDIA AI CITY CHALLENGE [8, 12, 13] have attracted considerable interests, which contributes to fine-grained anomaly detection in actual traffic accident scenarios and promoting the development of intelligent transportation.

In this paper, we design a modularized framework for city anomaly detection. This framework can address the general and specific anomaly detection tasks under a unified framework. Our proposed method achieves 66.41 mean F1 score and ranks the first place among all the participant teams in the specific anomaly detection task of the CitySCENE challenge.

3 METHODOLOGY

3.1 Problem Formulation

Task 1: General Anomaly Detection. General anomaly detection aims to predict frame-level probability to indicate the occurrence of abnormal events. Taking a video V as inputs, this task requires to output the abnormal probability $p^a(i)$ for each frame i. In this paper, we feed the particular video segment to the model and output the abnormal probability for this segment, and the frames within this segment share the same abnormal probability.

Task 2: Specific Anomaly Detection. The specific anomaly detection task aims to output a video segment $[j, k]$ (j and k indicate the start and end clip indices respectively) that semantically matches the specific abnormal event. It is worth noting that the training set of the CitySCENE is trimmed, while the testing set is untrimmed. This poses a difficulty for our task, i.e, how to use the abnormal concepts learned in the trimmed video to analyze the untrimmed video. In this work, we address the issue by designing a modularized framework. A video segment extraction module first splits the video into several non-overlapping video segments. Then an abnormal classification model is pre-trained and predicts the abnormal score of the corresponding video segments. A category-sensitive abnormal filter strategy is employed to process the model outputs to obtain the final localization results.

3.2 Modularized Framework

Video Segment Extraction. Following the common-used formulation, we represent a video V by N clips $\{V_1, V_2, ..., V_N\}$, each clip corresponds to a small chunk of sequential frames [16]. In our work, each video chunk contains 16 frames, and there is no intersection between each chunk.

Abnormal Classification Network. In our task, city abnormal events are divided into 12 categories, namely, accident, carrying, crowd, exploration, fighting, graffiti, robbery, shooting, smoking, steaming, sweeping, walking. We design a classification model to distinguish the abnormal category from the normal ones. Specifically, we take the temporal segment network [15] to perform the anomaly classification task. In order to speed up the prediction of the classification task, we introduced the temporal shift module (TSM) [9] that effectively model temporal concepts by moving feature mapping along the temporal dimension. It does not need additional calculation costs on the basis of two-dimensional (2D) convolution but has a strong ability of time modeling. In this paper, we leverage the bidirectional TSM combines past and future frames and current frames, which is suitable for high throughput offline video recognition. The partial shift strategy is used in the TSM to significantly bring down the memory movement cost. To balance the model capacity for spatial feature learning and temporal feature learning, the TSM is incorporated into a residual block. The residual shift can address the degraded spatial feature learning issue, as all the information in the original activation is still accessible after the temporal shift through identity mapping. In short, we construct the whole detection framework with ResNet-50 backbone and 8-frame input using no shift, 1/4 partial shift.

Category-Sensitive Abnormal Filter. During the training, the number of training samples of different abnormal categories has a significant impact on model training. We do not adopt methods like class-based weight or focal loss to adjust in the training stage. We design a category-sensitive abnormal filter that consists of a threshold lookup dictionary, which is pre-defined based on the number of different abnormal categories in the training set.

In the testing phase, we directly obtain the abnormal probability p_i^a of each frame by :

$$p_i^a = 1 - p_i^n, \tag{1}$$

where p_n denotes the output probability of the normal event category. For specific anomaly detection, we first obtain the probability

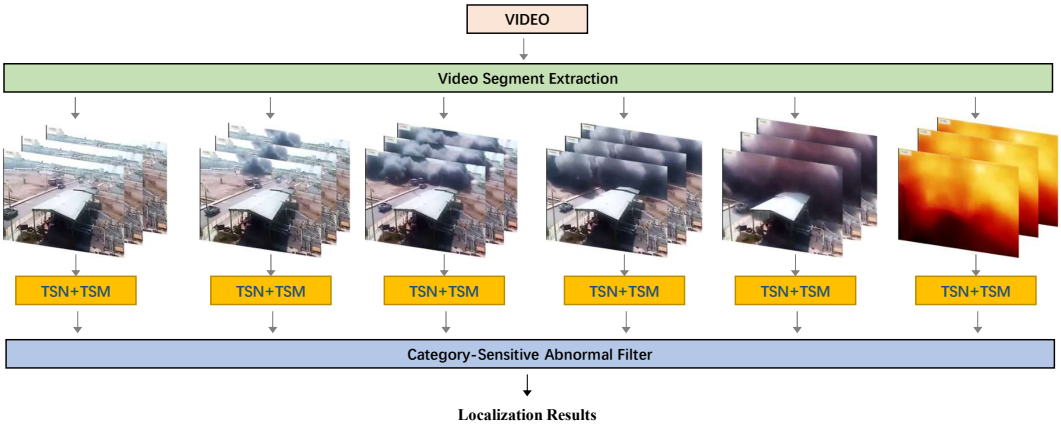

Figure 1: The architecture of the proposed modularized framework. The framework consists of three modules: video segment extraction, anomaly classification network (TSN+TSM) and category-sensitive abnormal filter.

of the abnormal category of each segment in the video (remove the normal event), then perform a softmax operation:

$$\alpha_c = \frac{p_i^c}{\sum_{c \in \mathbb{A}} p_i^c}, \tag{2}$$

where c and α_c denotes a specific category and the corresponding abnormal probability. $\mathbb{A}$ is the category space that contains 12 anomalies. The max category abnormal score is compared with the pre-defined threshold lookup dictionary $D(c)$ of class-sensitive abnormal filter to determine whether the video segment corresponds to the anomaly or filter it. Finally, the remaining video clips will be regarded as the localization result:

$$\begin{cases} c, & \max(\alpha_c) >= D(c) \\ Normal, & \max(\alpha_c) < D(c) \end{cases} \tag{3}$$

This category-sensitive abnormal filter contributes to alleviate the impact of the imbalance of abnormal categories at the testing stage and obtain more accurate localization results.

4 EXPERIMENTS

4.1 Datasets and Evaluation Metrics

Datasets. The training set in the CitySCENE consists of 758 normal and 1319 anomalous videos. The videos in the training set are trimmed manually and the annoation labels are at video-level. The testing set includes a set of untrimmed videos, where anomaly mostly occurs for a short period of time.

Evaluation Metrics.

Task 1: General Anomaly Detection. The frame-based ROC curve and AUC are adopted to validate the performance of the proposed method. An ROC space is defined by FPR and TPR as x and y axes, respectively, which depicts relative trade-offs between true positive (tp) and false positive (fp). The FPR and TPR are computed as following:

$$FPR = \frac{tp}{tp + fn}, \qquad TPR = \frac{fp}{fp + tn}. \tag{4}$$

Task 2: Specific Anomaly Detection. We use frame-based F1-score as evaluation metric to evaluate the performance of the method. For each predefined class c, we calculated the precision and recall from the results and get the F1-score. The class-based precision and recall is defined as:

$$precision^c = \frac{tp^c}{tp^c + fp^c}, \quad recall^c = \frac{tp^c}{tp^c + fn^c}. \tag{5}$$

The classbased F1 score is defined as:

$$F1^c = 2 * \frac{pre^c * rec^c}{pre^c + rec^c}. \tag{6}$$

Then the final macro-averaging F1-score is calculated by averaging the F1-score of each class to evaluate the total performance of methods on all classes:

$$MF1 = -\sum_{c \in C} \frac{F1^c}{N}, \tag{7}$$

where N is the number of the anomaly classes, i.e, 12.

Anomaly Category	Number	Anomaly Category	Number
Accident	57	Carrying	85
Crowd	68	Explosion	210
Fighting	264	Graffiti	103
Robbery	171	Shooting	75
Smoking	89	Stealing	30
Sweeping	98	Walkingdog	69
Normal events	758		

Table 1: Number of diffenent anomalies and normal event in the training set.

4.2 Implementation Details

We fine-tune our anomaly detection model from Kinetics [1] pre-trained weights and freeze the Batch Normalization [5] layers. We use 5-fold cross-validation to train the detection model. The final prediction result is the ensemble result of the prediction scores of the five models, that is to average the probability of the five predictions. In this paper, the threshold lookup dictionary of category-sensitive abnormal filter is defined as follows: accident:0.2, carrying: 0.2, crowd: 0.1, explosion:0.2, fighting:0.2, graffiti:0.1, robbery:

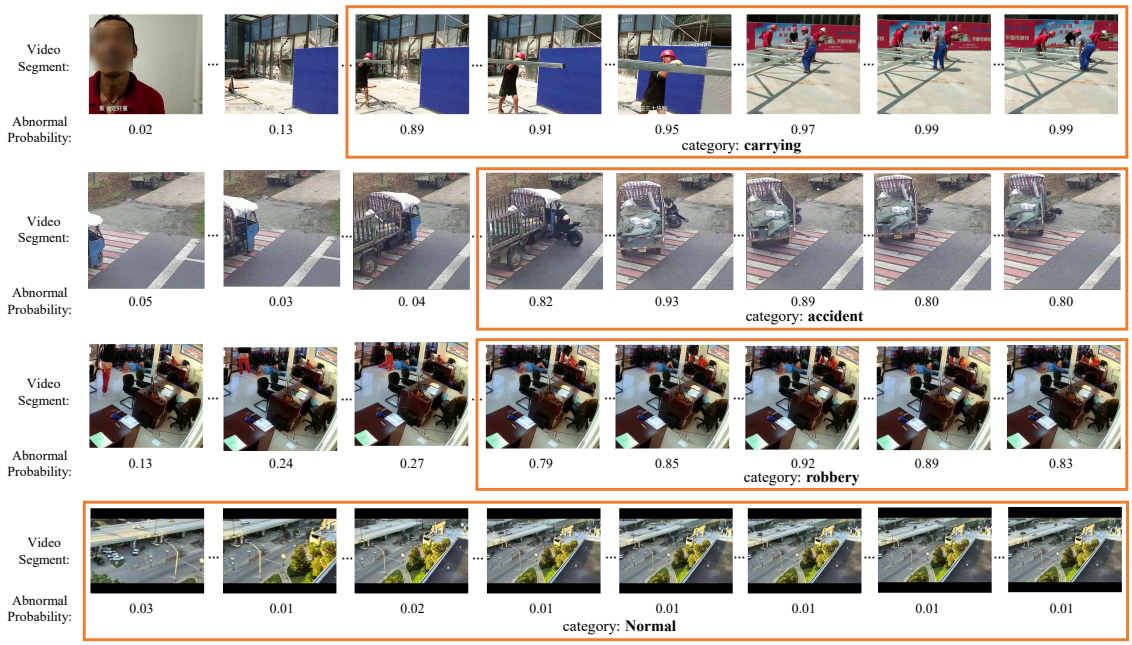

Figure 2: Anomaly detetcion and temporal localization results. The orange box represents the final localization result.

Rank	Team Name	AUC
1	BigFish	89.2
2	MonIIT	87.94
3	DeepBlueAI	86.85
4	**SYSU-BAIDU**	**86.52**
5	UHV	85.37
6	GOGOGO	84.09
7	ActionLab	70.92
8	Orange-Control	31.65

Table 2: The performance comparison (in %) of the state-of-the-art methods in the task 1: general anomaly detection.

Rank	Team Name	MF1
1	**SYSU-BAIDU**	**66.41**
2	BigFish	62.11
3	GOGOGO	52.33
4	MonIIT	45.52
5	Orange-Control	40.42
6	UHV	38.36
7	ActionLab	22.2

Table 3: The performance comparison (in %) of the state-of-the-art methods in the task 2: specific anomaly detection.

0.3, shooting: 0.3, smoking:0.3, stealing:0.2, sweeping: 0.2, walking-dog:0.2.

4.3 Comparison with the State-of-the-art

We evaluate our method on two required tasks on the CitySCENE testing set and compare it with several state-of-the-art methods in Table 2 and 3. As shown in Table 2, we achieve 86.52 AUC in the general anomaly detection task, which shows that our proposed method is competitive and rank fourth among all the contestants. The final leaderboard results for specific anomaly detection among all the teams are shown in Figure 3, we achieve 66.41 MF1 and rank the first place among all the participant teams. Specifically, our method boosts the MF1 from 62.11% to 66.41%, with an improvement of 4.3% compared with the second-ranked model.

4.4 Qualitative Visualizations

We illustrate four qualitative results in Figure 2 to show the anomaly detection and temporal localization results of the proposed algorithm. We observe that our algorithm can accurately predict the abnormal probability for each frame and localize the abnormal event. It demonstrates the effectiveness of the proposed algorithm.

5 CONCLUSIONS

We propose a modularized framework that resorts to address the anomaly detection tasks in the city scenario. The modularized framework consists of three modules: video segment extraction, abnormal classification network and category-sensitive abnormal filter. A video segment extraction module is employed to obtain the candidate video segments. Then the abnormal classification network is introduced to predict the abnormal score of each category. A category-sensitive abnormal filter is concatenated after the detection model to filter the abnormal event from the video candidate clips. Extensive experiments show that our approach establishes new state-of-the-art performance, i.e., 66.41 MF1 on the specific anomaly detection task in CitySCENE Challenge 2020.

REFERENCES

[1] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

[2] Yang Cong, Junsong Yuan, and Ji Liu. 2011. Sparse reconstruction cost for abnormal event detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3449–3456.

[3] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. 2016. Learning temporal regularity in video sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 733–742.

[4] Timothy Hospedales, Shaogang Gong, and Tao Xiang. 2009. A markov clustering topic model for mining behaviour in video. In *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 1165–1172.

[5] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).

[6] Jaechul Kim and Kristen Grauman. 2009. Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental updates. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2921–2928.

[7] Louis Kratz and Ko Nishino. 2009. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1446–1453.

[8] Yingying Li, Jie Wu, Xue Bai, Xipeng Yang, Xiao Tan, Guanbin Li, Shilei Wen, Hongwu Zhang, and Errui Ding. 2020. Multi-granularity tracking with modularlized components for unsupervised vehicles anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 586–587.

[9] Ji Lin, Chuang Gan, and Song Han. 2019. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE International Conference on Computer Vision*. 7083–7093.

[10] Weixin Luo, Wen Liu, and Shenghua Gao. 2017. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE International Conference on Computer Vision*. 341–349.

[11] Hajananth Nallaivarothayan, Clinton Fookes, Simon Denman, and Sridha Sridharan. 2014. An MRF based abnormal event detection approach using motion and appearance features. In *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 343–348.

[12] Milind Naphade, Ming-Ching Chang, Anuj Sharma, David C Anastasiu, Vamsi Jagarlamudi, Pranamesh Chakraborty, Tingting Huang, Shuo Wang, Ming-Yu Liu, Rama Chellappa, et al. 2018. The 2018 nvidia ai city challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 53–60.

[13] Milind Naphade, Zheng Tang, Ming-Ching Chang, David C Anastasiu, Anuj Sharma, Rama Chellappa, Shuo Wang, Pranamesh Chakraborty, Tingting Huang, Jenq-Neng Hwang, et al. 2019. The 2019 ai city challenge. In *CVPR Workshops*.

[14] Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6479–6488.

[15] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*. Springer, 20–36.

[16] Jie Wu, Guanbin Li, Si Liu, and Liang Lin. 2020. Tree-Structured Policy based Progressive Reinforcement Learning for Temporally Language Grounding in Video. *arXiv preprint arXiv:2001.06680* (2020).

[17] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. 2015. Learning deep representations of appearance and motion for anomalous event detection. *arXiv preprint arXiv:1510.01553* (2015).

[18] Jiangong Zhang, Laiyun Qing, and Jun Miao. 2019. Temporal Convolutional Network with Complementary Inner Bag Loss for Weakly Supervised Anomaly Detection. In *IEEE International Conference on Image Processing*. IEEE, 4030–4034.

[19] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1237–1246.

[20] Yi Zhu and Shawn Newsam. 2019. Motion-Aware Feature for Improved Video Anomaly Detection. *arXiv preprint arXiv:1907.10211* (2019).