Are Pre-training Multimodal Vision-Language Models Effective for Anomaly Detection from Surveillance Videos?

Kazuya Ueki

***Abstract*—This paper reports on a study that evaluates the effectiveness of using multimodal vision-language models for anomaly detection in surveillance videos. We conducted experiments using the UCF-Crime dataset, which includes 13 types of anomalies, such as arson, road accidents, and vandalism. Our findings demonstrate the capability of detecting targeted anomalies by inputting text descriptions within an unsupervised learning framework.**

I. Introduction

Anomaly detection in surveillance videos using machine learning is a research area of significant societal importance. However, there are challenges, such as achieving high accuracy in specific environments but facing reduced accuracy or failure to detect certain anomalies in real-world scenarios. This study proposes a system that aims to perform generic and precise anomaly detection in real-world environments through text-based specification of desired anomalies.

II. Multimodal Model-Based Anomaly Detection

The proposed anomaly detection system uses the multimodal vision-language model to detect anomalies based on specific context, as shown in Fig. 1. This system can be integrated with existing anomaly detection systems to enhance accuracy and identify more specific anomalies. In the future, we plan to include features informing users about the reasoning behind detected anomalies. This paper focuses solely on evaluations using the multimodal vision-language model, without considering integration with traditional systems.

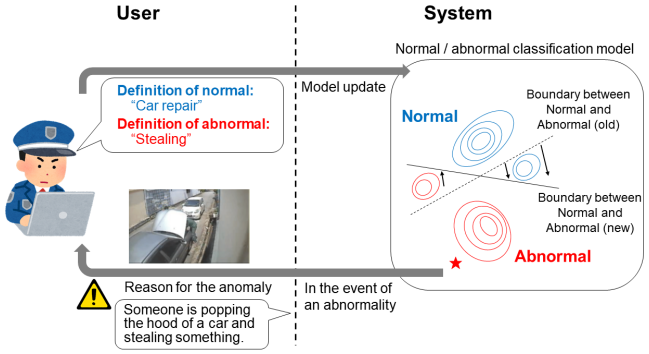

Fig. 1. Steps to create a face attribute model

III. Experiments

We performed experiments on the UCF-Crime dataset [1], which contains 13 anomalies, such as abuse, burglary, explosion, stealing, vandalism, etc., to determine whether anomalies can be successfully detected. Anomaly detection was performed using a multimodal model pre-trained in a contrastive learning framework called CLIP. Specifically, we used coca_ViT-L-14 (mscoco_finetuned_laion2b_s13b_b90k) provided by OpenCLIP[1], known for its high accuracy in video retrieval. The comparison results between conventional unsupervised methods and our multimodal vision-language model are presented in Table I. By varying the input prompts such as 'anomaly behavior,' 'dangerous activity,' or 'dangerous behavior,' we examined the accuracy of anomaly detection. While conventional methods extract rich video features such as C3D and I3D, the multimodal model offers an advantage in that it simplifies the process by allowing for the direct input of frame images extracted from the video into a pre-trained model to obtain results. We confirmed that this simple approach, relying on the multimodal model, yields detection performance comparable to that of the more complex conventional methods, with variations depending on the specific input text used.

We also assessed the ability to detect targeted anomalies by specifying particular text, as detailed in Table II. Our findings indicated that inputting specific text to define the anomalies of interest significantly facilitated the detection of these specific anomalies.

TABLE I
Frame-level AUC performance on UCF-Crime.

	Method	AUC
Conventional unsupervised approach	Sohrad et al.	0.5850
	Lu et al.	0.6551
	BODS	0.6826
	GODS	0.7046
	Prompt	**AUC**
Multimodal vision-language approach	'anomaly behavior'	0.5729
	'dangerous activity'	0.6192
	'dangerous behavior'	0.6725

TABLE II
Frame-level AUC performance for each anomaly on UCF-Crime.

Name	AUC	Name	AUC	Name	AUC
Abuse	0.7629	Explosion	0.8970	Shooting	0.5672
Arrest	0.7499	Fighting	0.7734	Shoplifting	0.7720
Arson	0.8682	Road Accidents	0.9480	Stealing	0.8734
Assault	0.8279	Robbery	0.7573	Vandalism	0.7748
Burglary	0.8482				

IV. Summary and Future Work

We therefore demonstrated the effectiveness of our proposed approach for anomaly detection using text as a key based on a multimodal vision-language model. In the future, we plan to enhance the accuracy of anomaly detection by integrating the proposed method with existing methods. Additionally, we aim to improve the system's functionality by providing explanations for detected anomalies.

References

[1] W. Sultani, C. Chen and M. Shah "Real-world Anomaly Detection in Surveillance Videos In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.6479–6488, 2018.

Kazuya Ueki is with School of Information Science, Meisei University, Tokyo, Japan (e-mail: kazuya.ueki@meisei-u.ac.jp).

[1]https://github.com/mlfoundations/open_clip