УДК 004.85

ПОШУК НЕБЕЗПЕЧНИХ ПОДІЙ У ВІДЕОАРХІВАХ СПОСТЕРЕЖЕННЯ ЗА ТЕКСТОВИМ ОПИСОМ НА ОСНОВІ МОДЕЛЕЙ ГЛИБОКОГО НАВЧАННЯ

Кит М.О.
e-mail: mykyta.kyt@nure.ua
Науковий керівник – канд. техн. наук, доц. Єсілевський В.С.
Харківський національний університет радіоелектроніки,
кафедра прикладної математики
м. Харків, Україна

This paper explores the application of vision-language alignment models for prompt-based retrieval of dangerous events in surveillance video archives. Traditional anomaly detection requires extensive labeled data and retraining for each scenario. The proposed approach leverages contrastive pretraining (CLIP, InternVideo2) and LLM-driven anomaly scoring (LAVAD, Holmes-VAD) to enable natural-language retrieval of hazardous events without fine-tuning. We analyze AnomalyCLIP's zero-shot anomaly recognition and Holmes-VAD's explainable detection on UCF-Crime, XD-Violence, and UCA benchmarks. A cascaded architecture combining lightweight CLIP-scoring with LLM-driven semantic verification is proposed for scalable, interpretable surveillance.

Сучасні системи відеоспостереження генерують величезні масиви даних, ручний перегляд яких є практично неможливим. Виявлення небезпечних подій – бійок, падінь, вторгнень у заборонені зони, залишених предметів – традиційно спирається на методи глибокого навчання, що потребують великих розмічених наборів даних і доменно-специфічного перенавчання для кожної нової сцени спостереження. Парадигма текстово-візуального вирівнювання (vision-language alignment) відкриває принципово інший підхід: оператор формулює текстовий запит природною мовою, що описує небезпечну ситуацію, а система автоматично здійснює пошук відповідних фрагментів у відеоархіві без додаткового перенавчання моделі [1].

Базовою технологією для реалізації такого підходу є моделі контрастивного вирівнювання візуальних та текстових представлень. Модель CLIP формує спільний латентний простір, у якому семантично близькі зображення та текстові описи мають високу косинусну подібність. Розширення CLIP на відеодані реалізується через побудову відеоенкодерів із часовою агрегацією. Модель InternVideo2 використовує прогресивне навчання з масковим відеомоделюванням, крос-модальним контрастивним навчанням та передбаченням наступного токена, що забезпечує формування узгодженого простору представлень для відео та тексту з підтримкою довготривалих часових залежностей [2].

Адаптацію CLIP-простору безпосередньо для задачі детекції аномалій реалізує модель AnomalyCLIP, представлена на ICLR 2024. Замість

класифікації об'єктів модель навчається розрізняти «нормальність» та «аномальність» сцени через спеціалізовані навчувані промпти, що перенаправляють увагу з семантики об'єктів на структурні відхилення від норми. Це забезпечує zero-shot детекцію аномалій у нових доменах без жодного прикладу з цільового набору даних. На бенчмарках UCF-Crime та XD-Violence AnomalyCLIP демонструє конкурентну точність порівняно з моделями, навченими у повністю керованому режимі [3].

Альтернативний підхід полягає у використанні великих мовних моделей як ядра для оцінки аномальності. Метод LAVAD (Language-driven Video Anomaly Detection), представлений на CVPR 2024, реалізує повністю training-free детекцію небезпечних подій. Відеокадри описуються текстовими підписами за допомогою VLM-каптіонера, підписи очищуються через крос-модальну подібність, після чого агрегуються у часовому вікні та подаються у LLM для генерації anomaly score. Цей підхід перевершує unsupervised та one-class методи на UCF-Crime без будь-якого навчання на цільових даних [4].

Подальшим розвитком є модель Holmes-VAD, яка поєднує мультимодальний LLM з часовим семплуванням для генерації не лише числового anomaly score, а й розгорнутого текстового пояснення виявленої аномалії природною мовою. Автори запропонували датасет VAD-Instruct50k із 50 000 пар “відео–пояснення” для instruction-tuning мультимодального LLM. Це принципово змінює парадигму взаємодії оператора з системою: замість абстрактного числового сповіщення оператор отримує повідомлення виду “виявлено бійку між двома особами біля входу о 14:23:07, один з учасників впав на землю” [5].

Важливим викликом залишається якість бенчмарків для текстово-візуального пошуку у відеоспостереженні. Датасет UCA (UCF-Crime Annotation), представлений на CVPR 2024, доповнює UCF-Crime деталізованими текстовими анотаціями на рівні речень із часовою прив'язкою – 23 542 речення для 1 854 відео. Експерименти засвідчили, що мультимодальні моделі, які добре працюють на загальних відео-мовних бенчмарках, значно деградують на surveillance-відео через низьку якість зображення, нетипові ракурси камер та контекстуальну неоднозначність подій [5].

Для масштабованого розгортання у реальних системах пропонується каскадна архітектура пошуку. На першому рівні легковагий CLIP-подібний скоринг (AnomalyCLIP або ViCLIP) швидко фільтрує відеоархів, виокремлюючи фрагменти з підвищеним anomaly score. На другому рівні LLM-верифікатор (Holmes-VAD або LAVAD) виконує детальний семантичний аналіз лише відфільтрованих фрагментів, генеруючи текстове пояснення та зіставляючи його з оригінальним текстовим запитом оператора. Такий дворівневий підхід дозволяє обробляти добові відеоархіви у сотні разів швидше за повний LLM-аналіз усього потоку, зберігаючи при цьому високу точність та інтерпретованість результатів.

Таким чином, текстово-візуальне вирівнювання трансформує парадигму виявлення небезпечних подій у відеоспостереженні – від навчання спеціалізованих детекторів до формулювання запитів природною мовою. Подальші дослідження у межах роботи зосереджені на підвищенні точності часового ґрунтування аномалій, покращенні крос-доменної генералізації на нові типи surveillance-сцен, а також на оцінці faithfulness текстових пояснень відносно реального візуального контенту.

Отже, текстово-візуальне вирівнювання трансформує виявлення небезпечних подій у відеоспостереженні, забезпечуючи перехід від ресурсомісткого навчання детекторів до гнучкого zero-shot пошуку природною мовою. Запропонована каскадна архітектура масштабовано обробляє відеоархіви, надаючи точні координати та текстові пояснення інцидентів, що суттєво знижує навантаження на операторів.

Подальші дослідження зосередяться на підвищенні точності часо-просторового ґрунтування (grounding) аномалій у незмонтованих відеопотоках. Також необхідне покращення крос-доменної генералізації для роботи на нових surveillance-сценах зі складними умовами. Нарешті, критичною є розробка метрик оцінки достовірності (faithfulness) згенерованих пояснень для уникнення «галюцинацій» мовних моделей у системах безпеки.

Список використаних джерел:

1. Yuan X., Zhang Z., Wang Z. et al. Towards Surveillance Video-and-Language Understanding: New Dataset, Baselines, and Challenges. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024)*. Seattle, 2024. P. 18527–18537.

2. Wang Y., Li K., Li Y. et al. InternVideo2: Scaling Foundation Models for Multimodal Video Understanding. *Proceedings of the European Conference on Computer Vision (ECCV 2024)*. Milan, 2024. P. 396–412.

3. Zhou Q., Pang G., Tian Y., He S., Chen J. AnomalyCLIP: Object-agnostic Prompt Learning for Zero-shot Anomaly Detection. *Proceedings of the International Conference on Learning Representations (ICLR 2024)*. Vienna, 2024.

4. Zanella L., Menapace W., Mancini M., Sebe N., Ricci E. Harnessing Large Language Models for Training-free Video Anomaly Detection. *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024)*. Seattle, 2024. P. 18527–18536.

5. Zhang H., Luo Y., Chen G. et al. Holmes-VAD: Towards Unbiased and Explainable Video Anomaly Detection via Multi-modal LLM. arXiv preprint. 2024. URL: https://arxiv.org/abs/2406.12235 (дата звернення: 05.03.2026).