Qeios

Peer Review

# Review of: "VADSK: Video Anomaly Detection with Structured Keywords"

**Fabio Ricardo Oliveira Bento**[1]

1. Coordenadoria de Eletrotécnica, Campus Guarapari, Universidade Federal do Espírito Santo, Brazil

**Dear Authors,**

I analyzed your paper, "VADSK: Video Anomaly Detection with Structured Keywords," which presents an approach to video surveillance anomaly detection using keywords and foundation models. The proposal of a **lightweight two-stage pipeline (induction and deduction)** is particularly interesting, especially as it targets edge devices and emphasizes interpretability.

Here are some strengths and areas for consideration:

**Strengths:**

1. **Novel Approach and Interpretability:** The idea of extracting keywords from descriptions generated by foundation models to represent frame features is creative. The use of TF-IDF-weighted keywords and the resulting encoding provides **inherent interpretability to the classification process**, a common challenge in many deep learning techniques. Figure 6 effectively illustrates how this interpretability manifests. This focus on transparency is crucial for building trust in surveillance systems.

2. **Real-Time Feasibility and Memory Efficiency:** The system's lack of temporal context and simple pipeline enable **near real-time inference**, a significant advantage for applications requiring rapid responses. Memory efficiency—by avoiding temporal information storage or parallel feature extraction—makes the system **suitable for resource-constrained devices**. Table 3 clearly highlights how VADSK uniquely combines these three qualities (interpretability, real-time performance, memory efficiency) compared to other reviewed SOTA methods.

3. **Use of Open Foundation Models:** The selection and use of open foundation models like Llama-3.2 Vision-Instruct-11B and MiniCPM-V-2_6-int4 enhance **transparency and reproducibility**. The quantization of Vision-Instruct-11B demonstrates practical concern for computational efficiency.



doi.org/10.32388/FW3LOG

4. **Clear Methodology Structure and Description:** Sections 3.1 (Induction) and 3.2 (Deduction) detail each pipeline stage, from description generation and corpus formation to TF-IDF calculation, keyword weighting, keyword encoding, and binary classification. Equations (6), (7), (8), and (9) provide the mathematical foundation for TF-IDF and weighting.

**Areas for Improvement and Considerations:**

1. **Performance on Complex Benchmarks**: While VADSK achieves comparable performance on ShanghaiTech (0.753 AUROC), it **still falls significantly behind top SOTA methods on UCSD Ped2 and CUHK Avenue**, where some approaches score above 90%. You suggest this may be due to the greater complexity of anomaly events in these datasets, indicating the method is better suited for simpler anomalies. A deeper discussion on the specific characteristics of anomalies in each dataset (e.g., pedestrians vs. bikes in UCSD Ped2; loitering, object throwing, running in CUHK Avenue; multiple scenarios in ShanghaiTech) and how the keyword-based approach may be inherently limited in capturing temporal or spatial complexity without additional context would be helpful.

2. **Ablation and Foundation Model Choice:** The ablation study in Section 4.4 compares Vision-Instruct-11B and Mini-CPM, showing a **clear trade-off between accuracy and inference speed**. Vision-Instruct-11B generally offers better AUROC on complex datasets (ShanghaiTech, CUHK Avenue), while Mini-CPM is significantly faster (~2s/frame vs. 5–6s/frame). The final choice depends on application requirements. Expanding on **why Vision-Instruct-11B's descriptions are more detailed** and capture subtler nuances (as mentioned) would strengthen the discussion. Additionally, exploring the potential impact of different or more specific prompts on description quality—and thus VADSK's performance—would be interesting for future work. The current prompt ("**You are a surveillance monitor... Describe the activities and objects...**") is reasonable, but testing the system's sensitivity to prompt variations could yield valuable insights.

3. **Keyword Selection and Weighting:** Using TF-IDF to identify and weight relevant keywords is a standard and suitable NLP technique. However, relying solely on normalized TF-IDF vector differences may limit the ability to capture complex term relationships or the importance of n-grams (though you opted for 1-grams). You describe the anomaly keyword selection as a "**simplified and explainable implementation**" compared to AnomalyRuler. Elaborating on this simplification's pros and cons would be useful. Section 4.2 details the TFidfVectorizer setup, which is very helpful. Discussing how the choice of parameters (e.g., limiting keywords to 100 or 200, adjusting min_df/max_df thresholds) affects

performance and interpretability (fewer keywords = more interpretability but potentially less accuracy?) would add depth.

4. **Classifier Training:** You use a simple feed-forward neural network and address class imbalance with weighted Binary Cross-Entropy loss. Five-fold cross-validation and early stopping are good practices to prevent overfitting. The dataset-specific batch sizes are mentioned; briefly explaining the rationale (e.g., based on dataset size or complexity) would be helpful. Specifying the exact feed-forward network architecture (number of hidden layers, neurons per layer, activation functions) in Section 3.2.2 would improve clarity.

5. **Future Work:** The suggestions in Section 5 are pertinent, particularly improving keyword generation/selection with advanced NLP techniques, dynamic updates, and domain-specific terms. Experimenting with different classifier architectures (e.g., beyond feed-forward to capture keyword interactions) is also valuable. Testing in specialized domains (industrial, healthcare, traffic) would further validate practical applicability.

**Summary:**

The paper presents a **promising and practical solution** for video anomaly detection, standing out for **interpretability and efficiency**—key qualities for real-world applications, especially in resource-limited scenarios. While quantitative performance on some benchmarks needs improvement to compete with top SOTA methods, the trade-offs are clear and well-justified.

Addressing the suggested areas—particularly a deeper analysis of the method's limitations on complex datasets and exploring sensitivity to keyword extraction parameters and prompts—would further enrich the work. The outlined future work is an excellent starting point.

I commend the authors for their research and the clear presentation of methodology and results.

**Sincerely,**

Fabio Ricardo Oliveira Bento.

## Declarations

**Potential competing interests:** No potential competing interests to declare.