FACULDADE ENGENHARIA

Departamento de
Informática

Leveraging AI for Video Understanding in Surveillance Scenarios

Project Proposal

Supervisor: João Neves (jcneves@di.ubi.pt)
Co-supervisor: João Pereira (joao.pereira@deepneuronic.com)

Objectives

Video surveillance can help authorities to maintain public order and security. However, the substantial increase in the number of security cameras present in public and private venues has not been matched by the number of human personnel required to monitor them. Considering that most of the events recorded by security cameras do not represent a threat to security, it is crucial to alleviate the waste of time and labour demanded of human operators by developing intelligent solutions that can assist them in maximising the product of their work.

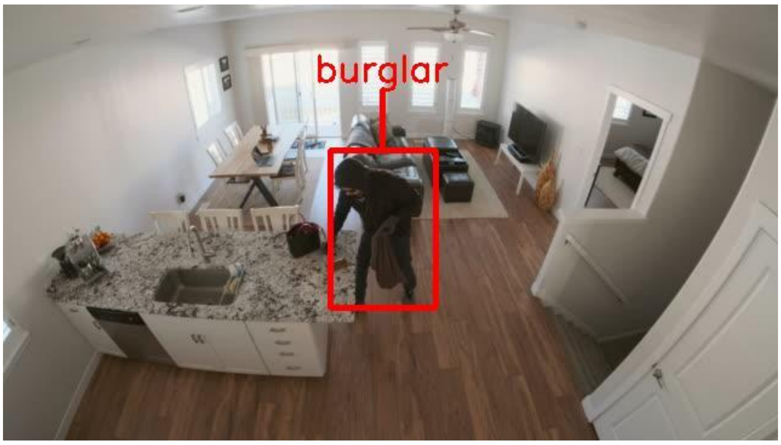

Figure 1: Results obtained from a SOTA Large Language Model. The model is capable of localizing and recognizing the type of action only based on a custom prompt. (source: iStock for image and https://llava.hliu.cc/ for bounding box detection).

Traditional AI methods applied in Surveillance contexts identify the presence or absence of anomalous events in security videos. However, these methods are limited because they lack crucial information that can be derived from more complex tasks such as Spatiotemporal Action Detection (e.g. identification of aggressor and victim, person counting, person re-identification), which consists of localising humans and classifying their actions from video data. Most Spatiotemporal Action Detection (STAD) methods report their performance in datasets depicting simple activities in typical environments (e.g., sports). This project envisages the development of a new dataset that allows the evaluation of STAD methods in Video Surveillance, a much more complex scenario. Additionally, the project intends the development of innovative evaluation strategies to assess the performance of predicting the localization and the type of actions using textual descriptions obtained from off-the-shelf video-based Large Language Models (e.g., LLaVA). The textual descriptions will be provided to the student, which is expected to focus on extending the traditional evaluation methods (e.g. accuracy, recall, bounding box IoU) with novel metrics that measure the descriptive power of AI in the context of Video Surveillance.

Tasks

T1: Study the state of the art on Spatiotemporal Action Detection methods and Video Surveillance datasets. (0.5 months)

T2: Acquisition of a dataset for Spatiotemporal Action Detection in Surveillance Videos. (2 months).

T3: Benchmarking of reference STAD methods in the collected dataset. (2 months).

T3: Development of novel indicators for assessing the quality of textual descriptions for Surveillance Videos. (2 months).

T4: Tests and debugging (0.5 months).

T5: Report writing (0.5 months).

Academic Prerequisites

- Interest in the field of Artificial Intelligence and Computer Vision.
- Proficient in Python or with an interest in learning.

Expected Results

- Dataset
- Computational Prototype
- Project Report