Towards a Usable Crime-Based Anomaly Detection Model

Daan Strijbosch
University of Twente
The Netherlands
d.strijbosch@student.utwente.nl

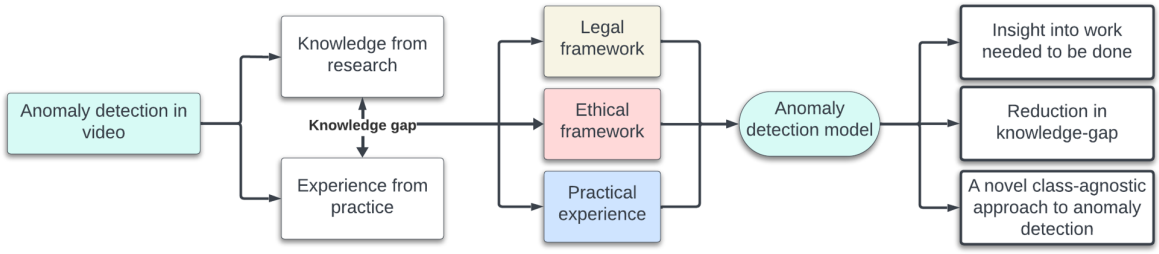


Figure 1: An overview of the work done and the end results of this paper.

ABSTRACT

With an ever-increasing amount of video footage being gathered, the interest in automated anomaly detection in video has seen a large increase. This paper aims to go beyond the technical implementation of a video anomaly detection model and look into what is needed to take this technology into practice. By describing the legal and ethical aspects of automated anomaly detection in video and gathering feedback from the Netherlands Forensic Institute, a novel anomaly detection methodology is created and tested. The proposed methodology uses I3D features to provide a class-agnostic method of detecting anomalies in video. Although performance is not at the level of the current state of the art, the agnostic nature as well as the analysis and discussion of relevant literature, provide an important step to bring these technologies into practice.

KEYWORDS

Video anomaly detection, I3D features, ethics, AI-law

1 INTRODUCTION

After its commercial introduction in 1949, video surveillance has seen a rise in popularity [20]. Currently, almost every public building has at least one camera that is used to monitor activity. However, this increase in popularity goes hand in hand with an increase in gathered video data. Current estimates indicate that more than one billion cameras are in use, which are estimated to collect multiple exabytes of data weekly [46]. This amount of data is impossible to monitor using only human analysts, leading to the concern of many critics [24].

1.1 Motivation

Because of this massive increase, a large push has been made in research to further automate the processing of CCTV footage [27]. This research is mostly focused on the technical implementation of detecting actions in video, resulting in a gap between research and practice [11, 43, 41]. The goal of this paper, is to take a step back and look at the automation of anomaly detection in its full scope, by incorporating the practical, ethical and legal groundwork that needs to be laid down before such technologies can become reality. The result of this is a class-agnostic anomaly detection model and the foundation of a framework that can be used to implement anomaly detection methods in practice.

For the purpose of this paper, an anomaly is defined as: 'An event that disrupts the expected flow of video imagery'. There are two parts to this definition: firstly, the event not the action which causes the event is labeled as the anomaly. For example, for an anomalous event like arson, the fire itself is the anomaly not the person lighting the fire. Secondly, what the natural flow of the video entails is subject for discussion on a case-by-case basis.

While this definition is valuable for the discussion to be had, it can not be used directly in the validation of the proposed methodology. The reason for this is that it can contradict with the ground-truth used in datasets. This happens when a dataset classifies the action rather than the change in scene as the anomaly. This will become relevant later when discussing the effectiveness of the methodology.

As for the ethical and legal aspects. In this work, both fields will be investigated in the context of a European Union centered view. The analysis will provide an overview of the legal landscape as well as the ethical considerations. The aim of this is to provide insight into the existing gap between research and a deployed product.

However, the largest driving factor behind this paper is improving the practicality of the solution, which is a topic that many parties are interested in. One of these parties is the Netherlands Forensic Institute (NFI) who will provide insights and guidance for the practical aspects of the methodology presented in this paper. Improving the practicality of a solution is twofold in topics like these; the dataset will need to mimic a real-life scenario as closely as possible and the solution should not be too specific in the actions it perceives as anomalous. This means that a dataset should be chosen which shows a lot of different actions but also contains actions that are relevant. For this reason, the UCF-Crime dataset was chosen, a

choice that will be further supported in Section 3.2 [42]. In addition to the UCF-Crime dataset, the ONFIRE and fight-detection dataset will also be used to complement the analysis of the methodology [1, 2]. To further improve on the solution, the NFI provided feedback at an early stage of this research which shaped the methodology.

1.2 Research questions

To provide guidance to the work in this paper, the following research questions will be answered. These are divided in main and sub-questions for each topic. The paper will cover two main topics: the technical implementation of an anomaly detection model and the improvement of the usability with the goal of deploying of such a model.

RQ1: How can a proposed system for anomaly detection be improved to make it more usable in a real-life situation?
SQ1.1: How can existing anomaly detection methods be improved to better fit a real-life anomaly detection use-case?
SQ1.2: What ethical considerations are important when designing a machine learning model for a sensitive topic such as video-based anomaly detection?
SQ1.3: What legal considerations need to be dealt with before being able to deploy and use an anomaly detection model?

RQ2: How can a combination of Inflated 3-Dimensional (I3D) features and various categorization methods be used for unsupervised or weakly-supervised anomaly detection?
SQ2.1: How can the L1 and L2 norm of a feature vector be used for a weakly-supervised anomaly detection model?
SQ2.2: Do different types of anomalies provide better performance in combination with a different feature type from the I3D feature set; RGB (Red Green Blue), Optical Flow and Combined features.

The start of answering research question 1 and its subquestions will be given in Section 2, the impact on the methodology and the dataset will be shown in Section 4.1 and a discussion in Section 7.1 will tie this together. Research question 2 will be answered more traditionally by showing the results of the validation which will be detailed in the methodology, Section 6. In Section 7.2 an overview of all research and sub questions and the answers to them will be presented.

1.3 Structure of this paper

The following Section, Section 2, will provide the analysis of the ethical and legal field since it will have an impact on the methodology and provides a foundation for the whole paper. Next, an overview of other work previously done in the field will be presented in Section 3. The implications of the ethical and legal analysis and the feedback provided by the NFI will be discussed in Section 4, together with the resulting proposed methodology. Section 5 will describe the metrics and visualizations that are the result of the methodology and will provide details on the implementation to make the methodology easier to reproduce. In the following Section, Section 6, the validation of the model will be presented which will then be discussed in Section 7. Next to this, the ethical and legal theory will be connected and recommendations for future work will be made in the same Section. Finally, Section 8 will provide concluding statements on the lessons learned.

2 LEGAL AND ETHICAL GROUNDWORK FOR AUTOMATED VIDEO ANALYSIS

As mentioned before, a clear understanding of the legal and ethical playing field of automated video surveillance is the key to making the correct choices in the design of a methodology. This Section will provide the foundation that will be used throughout the rest of this paper to facilitate these choices. Besides this, a lot of the information given cannot be processed in the methodology of a research paper as these include communication with users and the implementation of fail-safes when storing data. These parts will still be described in as much detail as possible here, later the impact on the methodology presented will be discussed in Section 4.1 and in Section 7.1 a more theoretical discussion will be presented. The legal groundwork looks into how legal systems view the implementation of AI from an EU-Centered viewpoint, using the EU AI Act and the GDPR as the basis for this Section [36, 14]. The ethical analysis will look into ethical literature surrounding AI, as well as more practical considerations.

2.1 Legal groundwork

2.1.1 EU AI Act. On the 9th of December 2023 the European council and parliament reached an agreement on the AI Act which has been in its proposal phase since April 2021 [23]. Although varying in severity, the AI Act has an impact on every AI product that will be released in the EU. This Section will go into how the AI Act could impact the solution proposed in this paper. The act differentiates between three levels of risk that an AI system brings with it: limited, high and unacceptable risk. The levels are explained below including a brief description on regulations related to each level [36]:

- Unacceptable risk: Unacceptable risk systems are systems that are considered a threat to humanity. These systems would include social scoring systems, systems aimed at cognitive manipulation and biometric systems in public spaces. Although some exceptions are made almost all of the systems that fall in this category will be banned.
- High risk: These are systems that have the potential to negatively impact safety and fundamental rights. This includes systems used in health, education, law enforcement and justice. These systems will have to be assessed before being put on the market as well as during the system's life cycle.
- Limited risk: This includes systems like AI generated art, video manipulation and other basic AI tools. These tools should provide the minimal amount of transparency that is needed for a user to make an informed decision.

Finally there are also AI systems that fall outside of these three categories and have even less risk like spam-filters or AI in video games. These fall outside of the described scope and will thus not be regulated by this act.

Looking at the descriptions for each category it seems most plausible that any proposed system to automatically processes video footage falls in at least the high risk category. An argument could be made that the system incorporates biometric aspects. However,

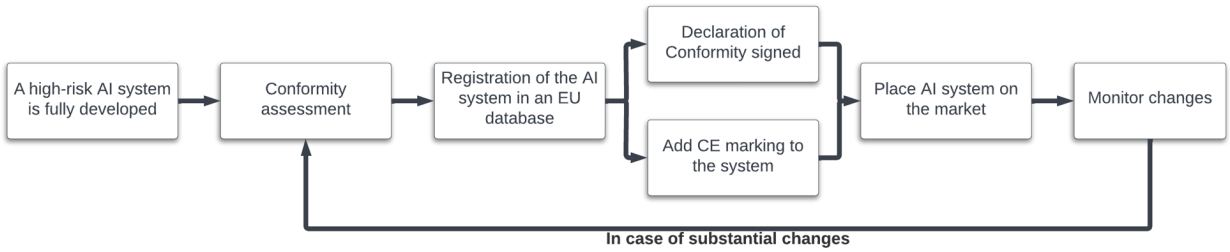


Figure 2: Steps to take for high-risk systems according to the EU AI Act.

there is no model training taking place in the proposed system. This means that biometric identifiers do not influence the final result but rather a whole scene is used. Furthermore, instead of determining who is performing the action, the system only aims to classify a whole scene as a moment of interest. This last point is contentious however, since it is not made clear whether a difference is made between determining the source of an action or the action itself. Still, it is most likely the system falls in the high risk category. The reason why it is never lower than high risk is because a system like the one proposed is always prone to misclassification of actions and in turn provides a risk to safety.

In Figure 2 the steps that a high risk system needs to take can be seen [9]. After development, the system will need to undergo the assessment and register itself in the EU database. This fundamental rights impact assessment (FRIA) has a couple of guidelines that need to be considered [36]:
"(a) a clear outline of the intended purpose for which the system will be used;
(b) a clear outline of the intended geographic and temporal scope of the system's use;
(c) categories of natural persons and groups likely to be affected by the use of the system;
(d) verification that the use of the system is compliant with relevant Union and national law on fundamental rights;
(e) the reasonably foreseeable impact on fundamental rights of putting the high-risk AI system into use;
(f) specific risks of harm likely to impact marginalised persons or vulnerable groups;
(g) the reasonably foreseeable adverse impact of the use of the system on the environment;
(h) a detailed plan as to how the harms and the negative impact on fundamental rights identified will be mitigated.
(i) the governance system the deployer will put in place, including human oversight, complaint-handling and redress."
Points A and B are both up to the deployer and where the system is going to be used. However, the other points would require a more thorough assessment of the system and an especially critical look into the data. On top of this assessment the deployers of the system need to notify the national supervisory authority that will be tasked with supervising and enforcing the EU AI Act in the phase of a system's life cycle that comes after its initial creation.

2.1.2 General Data Protection Regulation (GDPR). Another piece of relevant jurisdiction is found in the GDPR. The GDPR provides principles for the lawful processing of personal data. In the proposed system data is gathered and used at two moments, during development and during deployment.

In general the GDPR does not provide specifications for public datasets, also called open data, except when it includes personal data [14]. The definition given for personal data is as follows: "any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person" [13]. Since the UCF-Crime dataset contains non-anonymized footage of people they can be identified by physical features. According to the paper in which the dataset is introduced, the videos all come from YouTube and LiveLeak which is a gray area in terms of legality [42, 50, 30]. However, considering the popularity of the dataset it can be assumed that YouTube or LiveLeak have no issue with people using the data. Considering the GDPR however, it becomes more difficult to establish a sound legal basis for the system. For example, the following clause is listed in article 5 of the GDPR: "personal data shall be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes" [14]. Using video footage uploaded to YouTube and then incorporating it in an anomaly detection methodology can be seen as going beyond the initial purpose at the time of collection. However, other points relate more to limiting the usage and this is something that is clearly done. The data is used for certain parts of the methodology and is then not used in any other way, limiting the usage to only what is needed for the project at hand.

Looking further into the system's life cycle and the GDPR regulation. At one point the system will start collecting data and storing it for future use, since the system will always need a fail-safe. The recording of CCTV footage itself within the GDPR is already widely documented and as long as the subject on the video is aware this should pose no problem. Furthermore, the retrieved data should not be stored longer than needed which is quite ambiguous. This is something that is up to the user of the proposed system and how long they need to analyze the footage they retrieve. Using collected data for an AI model is also something that is addressed within the GDPR. There are two things that the system deployer needs to do:

- Inform the users how their data is used (so that the crime detection algorithm is used and what for).
- Make sure the users can ask for their data to be erased or rectified.

Besides this, as mentioned before, the data should be limited to the minimum amount needed and not stored longer than needed. Finally, the GDPR also states that someone has the right to not be subjected to automated decision making but due to the nature of the proposed system (including a human person that makes the final decision) there is no fully automated decision making taking place.

2.2 Ethical groundwork

The landscape of AI ethics can be very complex and has shown to form a high barrier of entry [4]. This Section aims to establish the groundwork to solve these ethical questions, whenever the proposed methodology would be taken beyond research. During development and after deployment are the two key points where ethics play a large role. Looking at key research in AI Ethics there are a multiple major themes prevalent in ethics, the first part of this Section will highlight the themes relevant to the proposed methodology and provide an analysis of those.

Starting with the themes relevant to the development phase. Although public datasets are used, it does not mean this phase is exempt from ethical problems. Firstly, the theme of privacy plays a large role during this phase [22]. Data stewardship, which is the management of the data during all stages (collection, processing and usage), is an important topic within this theme. Due to the dataset used being created outside the scope of this research, only the final two points can be addressed. To address these points it is important that the data is not used beyond its intended usage, which is research. This means that in the context of this paper the dataset is still used for its intended goal.

However, besides using the dataset for its intended use, the data that is used should also provide a (as close to) truthful depiction of the problem a model is trying to solve. This leads into the second theme; fairness, where bias plays an important role [22]. Research shows that bias in AI that is related to demographic diversity in a real-life situation, is often not the fault of the programmer but rather of the data used [10]. For this reason it is important that the data used for developing a model is as realistic as possible. A lot of research, including this paper, uses public datasets where this burden often, for better or worse, falls on the people who publish these datasets. This can create the problem that when asked about bias, a lot of researchers will not be sure if the dataset they used is actually unbiased. However, a problem is that there is not a lot of things a researcher can do since to be able to compare work the data has to stay unchanged. The goal of this paper however is to show the work beyond the research phase. At this stage biases that have a high probability of doing harm to (a group of) people, such as over-representation of a certain group of people in the dataset, should be known and circumvented to the best possible standard. This can be done by either changing the dataset by adding more videos to make up for any inconsistencies in the demographics of the subjects shown (so that the dataset represents the population in the location where the system will be used) or by providing extra human input during the usage of the system. Both options need to be analysed before deployment.

Bias is shown to overlap and be relevant to both phases but there are more themes in AI ethics that become relevant during deployment, transparency and accountability [22]. Transparency is the topic of making sure that people involved with the system, in the case of anomaly detection anyone on camera, can understand what is happening with the video data. A large part of this is monitored by laws highlighted in the last Section, but a part is also making sure that the algorithm itself is understandable [14, 26]. This means that anyone without any sophisticated knowledge should be able to understand the given explanation of how the algorithm works. When deploying it is thus important to make sure that this information is available to people. Furthermore, the topic of accountability focuses on making sure that there is someone who develop and deploy the system are aware of its impact and that someone can always be held responsible.

Although, these ethical themes and the awareness around them is important, research shows that the guidelines of AI can sometimes be misused to hide behind to avoid responsibility [15]. What this means is that developers may use it as a red herring to hide behind when actual laws are discussed to prove that they already pay attention to the ethical issues. For this reason it is always important to be transparent and take these ethical guidelines further by providing an open and transparent line of communication, both with users and with a 3rd party to keep the developer of such a system accountable.

3 RELATED WORK IN COMPUTER VISION FOR ANOMALY DETECTION IN VIDEO

This Section will provide an overview of the current state of the art in anomaly detection in video. Furthermore, it will also show various datasets on which the proposed methodology could be validated and provides an explanation for the choice of dataset.

3.1 Anomaly detection

Anomaly detection deals with the problem of making a binary distinction for each frame, group of frames, or whole video. This distinction is whether or not the frame shown is 'normal' which means that the video continues in an expected way or whether certain patterns deviate from the expected flow of a video. In the case of surveillance footage this means any event that presents a danger or aims to damage either individuals or infrastructure.

An example of this is shown in a 2020 paper where a combination of a pre-trained ResNet50 model and a bi-directional long short-term memory (BD LSTM) model, which is a type of recurrent neural network (RNN), was used [17, 18, 40, 47]. The ResNet50 model was used to extract features which were then passed to two stacked LSTM models. Since the model is bi-directional the output of the model is not only dependent on the current frame but also on the future frames. The choice for a LSTM was made because it is able to keep track of longer sequences than other types of RNNs. The performance shown using the ResNet50 as a feature extractor in combination with the multi-layer BD-LSTM was 85.53% accuracy on the UCF-Crime dataset.

(a) Example from the Fight Detection Surveillance dataset.

(b) Example from the ONFIRE dataset.

Figure 3: Example frames from the Fight Detection Surveillance [2] and ONFIRE [1] datasets.

Another example comes from a 2023 paper [25]. Here a method is shown where none of the labels provided by the UCF-Crime dataset are used. This method assigns a vector of the mean and standard deviation for each video. These vectors are then used to iteratively cluster the videos into two clusters of anomalous and normal videos. Afterwards, the normal videos are labeled and put aside. The anomalous videos are segmented and a hypothesis test is used to find the segment in which the anomaly happens in the anomalous videos. In this hypothesis test H0 is that a video segment is normal and a p value is calculated for each video. The assumption is then made that the videos are Gaussian distributed and various significance levels are tested. Using the labeled segments, a neural network is trained that achieves a AUC score of 80.65% on the UCF-Crime dataset.

Another method, as shown by Chen et al., is using two different branches for video description [8]. One branch generates video captions and extracts the text-features from those, the other branch extracts video-features from the video. These features are then fused and fed to an anomaly predictor. This predictor, in the form of a binary classifier, calculates an anomaly score for each snippet of video and uses a threshold to label the segments. This method achieves a 84.9% AUC score on the UCF-Crime dataset.

A final methodology uses bi-directional frame interpolation to find anomalies in video footage [12]. Given a frame F the frame and its neighbours F-1 and F+1, frame F is discarded and generated again using the neighbouring frames by warping with the interpolated optical flow. This generated frame F' is then compared to the actual frame F to see how much it differs. The more F' differs from F the higher the chance of an anomaly. This method was tested on the Ped2, Avenue and Campus datasets where it got a 98.9%, 89.7% and 75.0% accuracy respectively [32, 31, 52].

3.2 Datasets and literature in (crime-related) anomaly detection from surveillance videos

In the field of crime-related anomaly detection in surveillance videos, datasets are relatively sparse. Although datasets containing hours and hours of surveillance footage are widely available in the form of, for example, Shanghai Tech and the VIRAT dataset, the focus of these datasets is not providing a diverse array of relevant anomalies [52, 34]. The Shanghai Tech dataset is focused on crowd counting and the VIRAT dataset on various events like moving vehicles or people talking on a phone. If we narrow the search down to purely relevant anomaly related footage from surveillance videos the leading dataset that comes up is the UCF-Crime dataset [42]. The UCF-crime dataset consist of 1900 videos spanning 13 crime-related anomalies as well as normal videos (videos not containing an anomaly). Furthermore, the UCF-Crime dataset has been used in many different papers on anomaly detection making performance comparisons easily possible. Although the UCF-Crime dataset features a wide array of videos, this also means that the currently used categories are spread thin. This means that finding

anomaly-specific performance differences tends to be quite difficult. To deal with this, two different datasets are chosen which are hypothesized to be on the opposite end of the spectrum of being dependent on optical flow related features and RGB related features. The two datasets are the fight-detection dataset, consisting of 150 fighting and 150 normal videos, and the ONFIRE dataset, consisting of 219 fire-related videos and 103 normal videos [2, 1]. These will be used to answer the question if different anomalies work better with different feature sets.

In short, three datasets will be used. The UCF-Crime dataset will be tested on all methods and will be used for the main comparisons with other literature. The ONFIRE dataset and fight-detection dataset will be used to answer the questions of the impact of different feature sets.

4 METHODOLOGY: PROPOSED ANOMALY DETECTION FRAMEWORK

This Section will provide an overview of the methodology that will be used for the proposed anomaly detection model. It will start by providing an overview of the feedback from the NFI and its impacts, as well as the impact of the ethical and legal framework. Afterwards the implementation of the methodology will be described.

4.1 Feedback and implications from practice and literature

The groundwork laid in Section 2 and the feedback gathered during the initial phases of the project plays a key role in the proposed method. Due to this, only describing the methodology would not do this impact justice. This Section aims to explain how the feedback by both the NFI as well as the legal and ethical groundwork shaped the methodology.

The feedback from the NFI throughout the initial stages of the progress, led to the following key changes:

- Initially the methodology also included a classifier, this was shifted to a two-class problem (either a video is anomalous or not) due to the fact that the NFI noted that classification will always be done by a human.
- The scope changed from focusing only on crime, and thus removing categories which were not crimes (like road accidents), to also including these other anomalies. The reason for this is twofold, again exact definitions of activities are not part of the scope of a realistic product (and thus removing edge-cases imposes additional risk) and secondly, activities like road accidents still warrant intervention by emergency services.
- The methodology aims to be as class-neutral as possible. The goal of this is to mitigate the impact of an incomplete dataset (no dataset has every possible anomalous activity)

Besides the lessons learned from the feedback from the NFI, the following key points will also be incorporated into the design of the methodology based on the earlier described ethical and legal foundation.

- An action is not connected to a person automatically, the goal is to define a scene rather than an individuals actions as to reduce privacy concerns.
- Next to the feedback of the NFI, the choice to not temper with the dataset is further backed-up by the ethical ideas of being unbiased.
- The data is handled according to its intended use and the described methodology using this dataset is not intended for actual use outside of research.

4.2 Methodology overview

The goal of the methodology is to provide a video-level label which is either 'normal' or 'anomalous'. The key factor to the methodology proposed is that research found that feature magnitude could be related to anomalous behavior in video [25]. Theoretically, this means that in a video in which nothing 'unnatural' to the flow of a video happens, the feature magnitude should be stable. To test this, three tests will be used to distinguish between videos. In Figure 5 a general overview of the pipeline can be seen with each node having the corresponding Section next to it. Each method has the same general part, features are extracted from which the L1 and L2 norms (the feature magnitude) are calculated. From these L1 and L2 norms the mean and standard deviation is calculated. Afterwards, the three different methods diverge and will use the aforementioned values to classify the videos.

4.3 Feature extraction and video representation

The features that are used are I3D features which were introduced by Carreira et al. [6]. I3D features are spatio-temporal features extracted from videos using an Inflated 3D CNN, capturing both the spatial and temporal information in a video. These features have often been cited to provide the best performance for video classification tasks [8]. The batch size for the feature extraction is set to 16 frames. The end result of this extraction is three feature types for each video: the I3D features themselves and the two components of these features separately (RGB features and optical flow features). The experiments will be performed on all three of the feature sets.

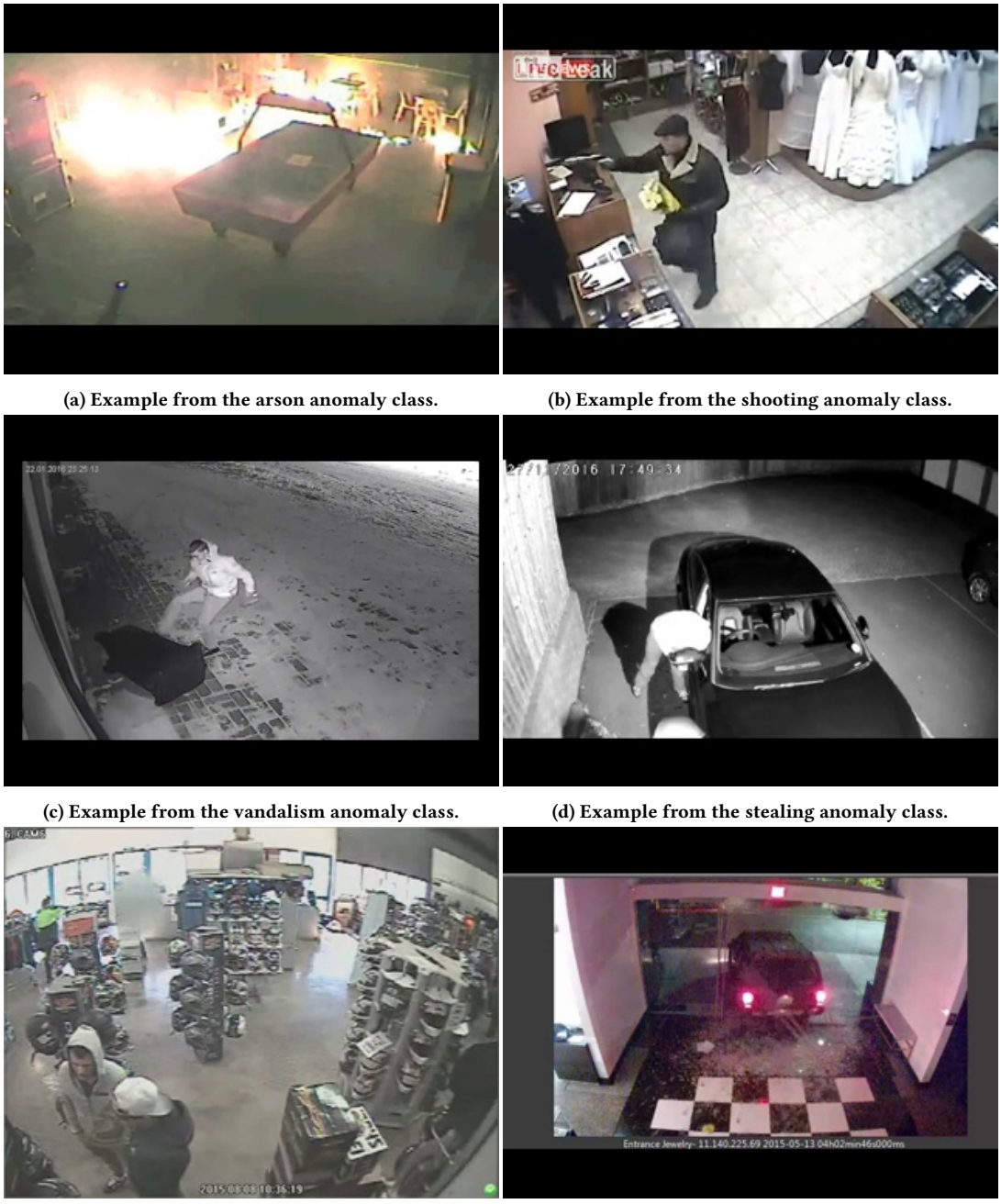
After this step each video now has three arrays of features of the size 1024 * (total number of frames / 16) for the RGB and optical flow features and 2048 * (total number of frames / 16) for the combined I3D features. Using these arrays the L1 and L2 norm can be calculated for each batch of 16 frames using the two formulas below respectively.

(1) $\sum_{i=1}^{n} |X_i|$

(2) $\sqrt{\sum_{i=1}^{n} X_i^2}$

The L1 Norm is the purest form of the feature magnitude and is the sum of all the absolute values in a feature matrix. The L2 norm is the square root of the sum of all the values in a feature matrix squared. This results in the L2 norm being more sensitive to outliers, due to the fact that squaring a value increases its impact on the total. Each video now has a list of L1 and L2 norms assigned with a length equal to the total number of frames / 16.

(1) $\mu_i = \frac{1}{m_i} \sum_{j=1}^{m_i} \left\| \mathbf{f}_{ij} \right\|_{1 \vee 2}$
Where $\left\| \mathbf{f}_{ij} \right\|_{1 \vee 2}$ is equal to the L1 or L2 norm of a vector and m_i is equal to the number of batches for the video



(a) Example from the arson anomaly class.

(b) Example from the shooting anomaly class.

(c) Example from the vandalism anomaly class.

(d) Example from the stealing anomaly class.

(e) Example from the shoplifting anomaly class.

(f) Example from the burglary anomaly class.

Figure 4: Example frames from the UCF-Crime dataset [42].

(2) $\sigma_i = \sqrt{\frac{1}{(m_i-1)} \sum_{j=1}^{m_i} \left(\left\| \mathbf{f}_{ij} \right\|_{1 \vee 2} - \mu_i \right)^2}$

Where $\left\| \mathbf{f}_{ij} \right\|_{1 \vee 2}$ is equal to the L1 or L2 norm, m_i is equal to the number of batches for the video and μ_i is the previously calculated mean

With each video labeled with a list of L1 and L2 norms, a mean and a standard deviation the methods diverge. The next Section describes each method.

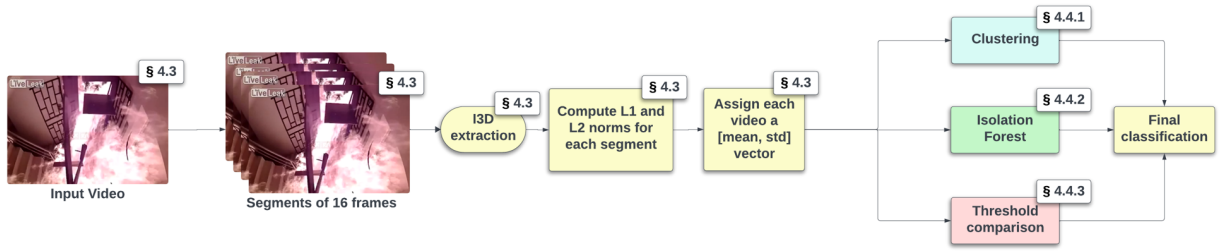


Figure 5: The pipeline of the proposed methodology.

4.4 Anomaly Detection from extracted features

4.4.1 Clustering. The first method used relies on clustering of the vector [mean, standard deviation], for each video. This clustering is done using the L2 norm since, as explained, the L2 norm is more susceptible to outliers providing better clustering performance. Five different clustering methods have been used, three regular and two iterative clustering methods. The three regular methods are the following:

- Kmeans clustering - Clusters data into K (in this case two, normal and anomalous) clusters by minimizing the sum of squared distances between data points and their respective cluster centroids
- Spectral clustering - Uses the eigenvalues of a similarity matrix constructed from the data to perform dimensionality reduction before clustering in fewer dimensions.
- Gaussian Mixture Model (GMM) Clustering - Models the data as a mixture of multiple Gaussian distributions, assigning probabilities to each data point belonging to each Gaussian component.

Besides this, for the Kmeans clustering and the spectral clustering methods an iterative method will also be tested similar to the work described in [25]. The iterative method follows the algorithm described in Algorithm 1.

Algorithm 1 Iterative Clustering.

```
normal cluster ← all datapoints
while len(normal cluster) / len(anomaly cluster) > 0.5 do
    Cluster the normal cluster into two different clusters
    Largest cluster -> normal
    Smaller cluster -> anomaly
    Calculate the ratio between the clusters
end while=0
```

Using the algorithm, the anomaly cluster is iteratively appended with the furthest outliers which in theory should improve clustering when the data is not clearly separated. The ratio between the two clusters is set to 0.5 due to the fact that the size of the classes are known for the test set. In other scenarios this parameter should be adapted accordingly.

4.4.2 Isolation Forest. The second method uses isolation forest, a machine learning algorithm used for anomaly detection. The foundation of isolation forest is based upon the idea that anomalous data points are few and different. It relies on two main concepts:

- Random partitioning: A random forest algorithm selects a random feature and a random value in the range of values of this feature to split the data.
- Isolation path: the isolation path is the amount of splits that a particular data point needs before it is completely isolated.

The intuition in this case is that anomalous data points would require less splits before being isolated. The process of partitioning and calculating the isolation path for each data point is done iteratively. The mean of these values is transformed into an anomaly score for each data point. A threshold, often expressed in a ratio of normal and anomalous data points is then used to get the top x% of data points with the highest anomaly score. In the case of the UCF-Crime dataset this value is set to 50% (logically, the highest value still possible) since the test set is equally divided in normal and anomalous videos. Similar to the clustering methods, this is also done on the L2 norms.

4.4.3 Video threshold. The final method relies on both the L1 and L2 norms for each video and the mean and standard deviation of these values. Given a video of, for example, 160 frames, there are 10 L1 and L2 norms of the patches used for I3D feature extraction. These patches have a mean and standard deviation calculated according to the aforementioned formulas.

The idea is that videos with a higher standard deviation in their L1 and L2 norms have more irregular occurrences which would increase the probability of an anomaly in these videos. For each video in the test set the standard deviation of the video is compared to a threshold, if the video scores higher than the threshold it is deemed to be an anomalous video and vice versa. The threshold score is gathered from the training set, where the performance of the method was optimized and the threshold value for which the accuracy was the highest is used for testing.

4.5 Comparison of I3D feature components

All three of the tests will be performed on the UCF-Crime dataset to provide insights into performance relative to the state of the art [42]. However, besides this the best performing method will also be tested on the ONFIRE and fight-detection dataset [1, 2]. This will result in a comparison between the three feature types present in I3D features, RGB, optical flow and the combination of the two.

5 EXPERIMENTAL SETUP

This Section provides details on the results that will be produced when the above methodology is used, and the metrics to summarize the quantitative results. Next to this, implementation details are given to make reproduction of the methodology easier.

5.1 Metrics

This subsection will provide an overview of the metrics used, their relevance and the formula to calculate them. As can be seen all these metrics are deterministic, something which might not be best when implementing the actual system. However, this is needed for proper comparisons of the methods. In Section 7.2 and 7.3 a non-deterministic method is highlighted.

- Accuracy: The proportion of correctly classified instances out of the total instances.

$$Acc. = \frac{TruePositive + TrueNegative}{(TotalPositives + TotalNegatives)}$$

- Balanced Accuracy: Aims to produce the same information as the accuracy but particularly useful in scenarios with imbalanced class distributions. It combines sensitivity and specificity to provide a measure of overall model performance that is less affected by class imbalance.

$$Balanced\ Acc. = \frac{1}{2}\left(\frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}} + \frac{\text{TrueNegative}}{\text{TrueNegative} + \text{FalsePositive}}\right)$$

- Recall: The proportion of actual positive instances that are correctly detected by the model. Its importance stems from the fact that, depending on the application, this value should be almost 100%. This depends on if the system is fully automated, where false-positives could be harmful, or still supervised where a false negative is much more harmful than having to manually relabel a false positive.

$$Recall = \frac{TruePositive}{(TruePositive + FalseNegative)}$$

- Precision: The proportion of predicted positive instances that are actually positive. Relevant since having too many false flags either makes manual work very difficult or automatic labeling too high-risk.

$$Precision = \frac{TruePositive}{(TruePositive + FalsePositive)}$$

- F1 score: The harmonic mean of the precision and recall, in cases like previously described having a high recall and low precision or vice versa can be more desirable. The F1 score provides a more balanced picture than the two other metrics on their own.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

- Area Under Curve: The area under the Receiver Operating Characteristic (ROC) curve, which plots the true positive and false positive rate at various thresholds. This score and the corresponding graph will be given for methods using a threshold value.

$$A = \int_a^b f(x)\,dx$$

Where f(x) is the TPR-FPR curve

The most important metric and the one which should be close to 100% for anomaly detection is recall. However, the recall is also a metric which can easily be influenced by changing the threshold value, cluster size or contamination. The accuracy is relevant but not most important since the importance of classifying a true negative is severely outweighed by the importance of correctly finding a true positive (an anomaly). The F1 score is most important when comparing methods since the method with the highest F1-score would perform best when the parameters are tuned to get a recall value of close to 100%.

5.2 Visualizations

Although the previously described methods metrics are the key result that can be used for comparing methods, more work has been done to make the results more intuitively understandable. A short list is presented to showcase some of these minor tests from which the results, in the forms of additional Figures, will be presented in Section 6.

- Visualization of the clusters was performed to better understand the performance
- A principal component analysis (PCA) of the feature vectors was performed. PCA maps high dimensional data (in this case 1024 or 2048 dimensions) to a two dimensional space while keeping the dimensions which distinguish the datapoints best.
- Visualization of the mean of the L1 and L2 norm over time was performed. This to support the third method, the video threshold.

5.3 Implementation details

This Subsection describes how the methodology was implemented. All of the code was written in Python. To extract the I3D features, the I3D feature extraction package was used, which is a freely distributable package developed by Hao Vy Phan [39]. The package uses a modified and pretrained ResNet50 model, where the kernel is transformed to handle 3D convolutions and the model is pretrained on the Kinetics-400 dataset [17, 21]. The package also requires the Pytorch and Torchvision packages [37, 33]. The extracted features are stored using the Numpy package as arrays with each video being stored in a separate folder with the same name as the video and containing a single 'Feature.npy' file [16]. For the calculation of the L1 and L2 norm, the corresponding Numpy methods from the linear algebra and the basic Numpy modules was used.

For the Isolation Forest component, from the Scikit-Learn package the ensemble module was used [38]. The Scikit-Learn clustering module was used for all three clustering methods. For the threshold tests no further packages other than Numpy were needed.

To calculate all the needed metrics the Scikit-Learn metrics module was used which contained easy calculations for these metrics when a list of ground truths and predictions is given. Finally, for

Clustering type	Feature	Accuracy	Recall	Precision	F1
Kmeans	*Comb.*	51,38	-	-	-
Kmeans	*Flow*	51,38	-	-	-
Kmeans	*RGB*	51,38	-	-	-
Kmeans (iterative)	*Comb.*	54,83	60,71	52,15	56,11
Kmeans (iterative)	*Flow*	53,00	58,57	51,25	54,67
Kmeans (iterative)	*RGB*	56,21	62,86	53,99	58,09
Spectral	*Comb.*	51,38	-	-	-
Spectral	*Flow*	51,38	-	-	-
Spectral	*RGB*	51,38	-	-	-
Spectral (iterative)	*Comb.*	53,10	65,71	51,11	57,8
Spectral (iterative)	*Flow*	51,38	65,00	49,73	56,35
Spectral (iterative)	*RGB*	54,48	**69,28**	51,25	59,51
GMM	*Comb.*	58,62	68,57	55,81	**61,54**
GMM	*Flow*	53,10	65,00	51,12	57,23
GMM	*RGB*	**62,07**	35,00	**72,06**	47,11

Table 1: Results on the UCF-Crime dataset when using clustering.

Method	Feature	Accuracy	Recall	Precision	F1
Isolation Forest	*Comb.*	47,93	47,86	46,21	47,02
Isolation Forest	*Flow*	47,59	45,00	46,65	45,32
Isolation Forest	*RGB*	54,32	54,32	48,63	49,65

Table 2: Results of isolation forest method.

Method	Feature	Threshold value (std.)	Accuracy	Recall	Precision	F1
L1 + L2 std	*Comb.*	40 + 1,2	65,17	82,86	60,10	69,67
L1 + L2 std	*Flow*	34 + 1,3	57,24	66,43	54,71	60,00
L1 + L2 std	*RGB*	15 + 0,8	**68,62**	**83,57**	**63,24**	**72,00**

Table 3: Results on the UCF-Crime dataset when using threshold values.

Method	Feature	Threshold value (std.)	Accuracy	Recall	Precision	F1	AUC
L1	*RGB*	15	57,59	95,00	53,41	68,39	74,00
L2	*RGB*	0,8	68,28	83,57	62,90	71,78	73,00

Table 4: Ablation study on the threshold values method.

the visualizations, Scikit-Learn was used to provide the ROC-curve which was then visualized using the Matplotlib package [19]. Matplotlib was also used to visualize the clustering methods and the L1 and L2 norm over time.

For the ONFIRE and fight-detection dataset no train and test split was given. This train and test split was randomly made using the train test split function from the Scikit-Learn model selection module using random state '42'. These splits were then transferred to different folders for easier access later using built-in features in Python 3.11. Access to folders throughout the methodology was also done using the built-in OS features from Python 3.11.

This results in the following list of requirements:

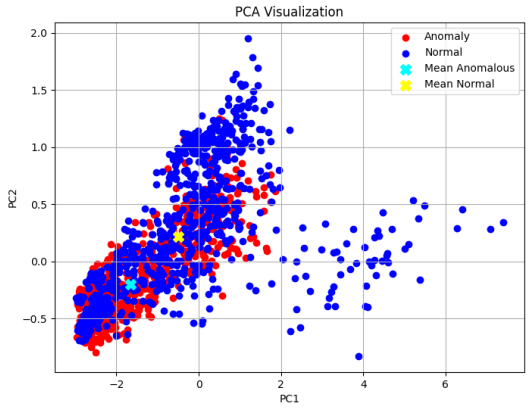
- Python - version 3.11
- I3DFeatureExtraction - version 0.3.3
- Pytorch - version 2.3.1
- Torchvision - version 0.18.1
- Numpy - version 1.23.4
- Scikit-Learn - version 1.4.2
- Matplotlib - version 3.62

6 RESULTS

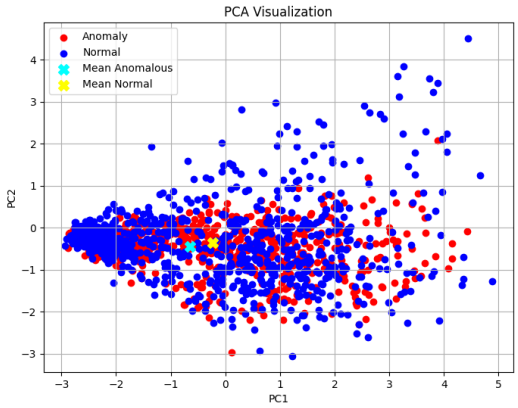
This Section will show the results of the previously described methodology. Starting with the validation on the test set of the three proposed methods and then showing the other visualizations of the extracted values.

6.1 Anomaly detection performance

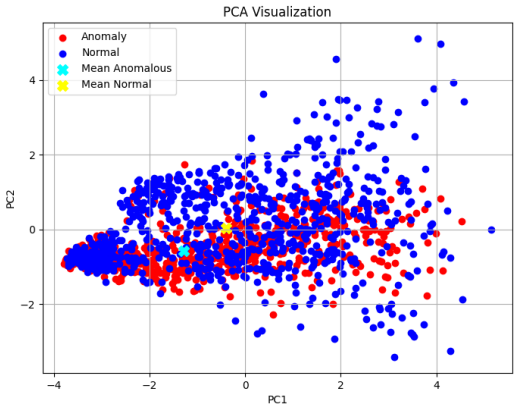
6.1.1 Clustering. Table 1 shows the results for the clustering method. As can be seen clustering using a Gaussian Mixture Model (GMM) works best while the non-iterative methods of both Kmeans and



(a) PCA analysis of the RGB features.



(b) PCA analysis of the optical flow features.



(c) PCA analysis of the combined features.

Figure 6: PCA analyses on the UCF-Crime training-set for all three feature types.

Spectral clustering simply reduce the whole dataset to one cluster (for this reason the recall, precision and F1 score is not shown since they cannot be calculated or are 0 without any positive labels). Looking specifically at GMM, the combined feature set has the highest F1 score of 61,54% and although the accuracy is slightly lower than when the RGB featureset was used, 58.62% vs. 62,07%, the higher F1-score is more important as explained before.

6.1.2 Isolation forest. In Table 2 the results of performing Isolation Forest can be seen, it is clear that the RGB features slightly outperform the other features, with an F1-score of 49,65%. However, none of these come close to the results of the other methods.

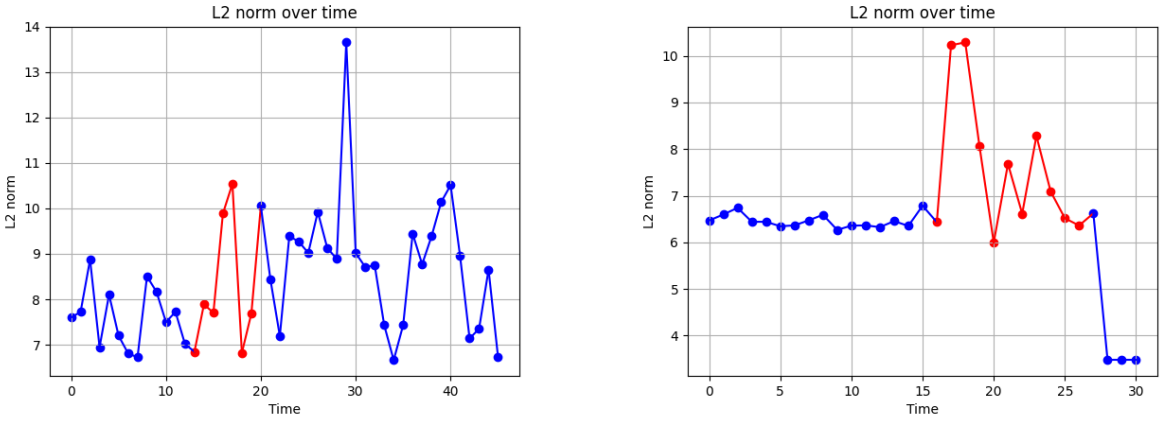
6.1.3 Video threshold. In Table 3 the results for the threshold methodology can be seen. The RGB features again perform best with an F1 score of 72,00%. In Table 4 the performance of the two norms can be seen separately. In Figure 8, the corresponding ROC-curves can be seen. Both methods show a reduction in the F1 score when used independently with the L2 norm reducing by 0,22% and the L1 norm by 3,61%.

6.2 Visualization

In Figure 6 three PCA analyses can be seen of the different feature types. As can be seen, the RGB features have the highest distance between the mean of the normal and of the anomalous segments of video.

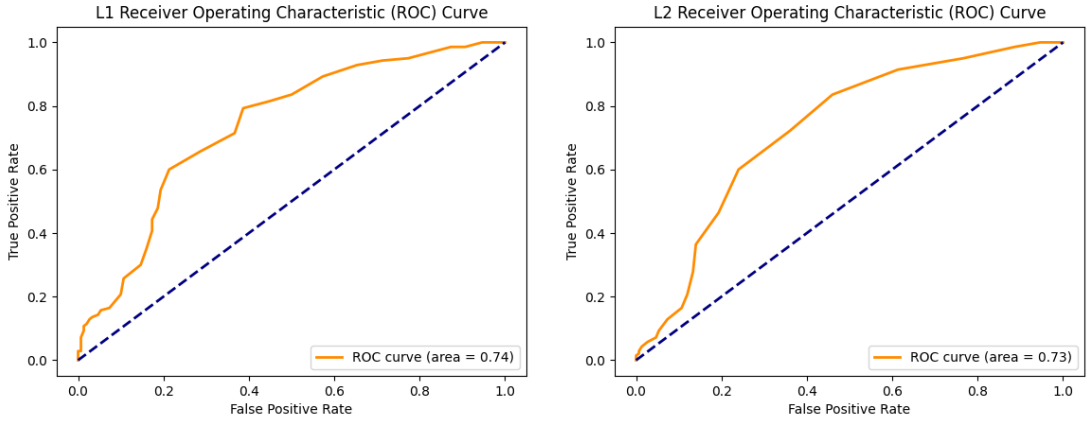
in Figure 7 the L2 norms calculated from the RGB features for each segment in a video can be seen over time. Each dot indicates a 16 frame batch and the red colors indicate where the anomaly takes place. As can be seen, in certain cases this indication works very well while in other cases the L2 norm spikes at a different point in time.

6.2.1 Performance comparison. Looking at all the results presented, it becomes clear that using a threshold value for the mean L1 and L2 norm of a video provides the highest accuracy of 68,62% and a F1 score of 72,00%. The Isolation forest method and clustering, especially the non-iterative Kmeans and Spectral methods, lack behind with GMM clustering being the closest. The RGB features



(a) Video: Arson009. **(b) Video: Explosion025.**

Figure 7: Two examples of the L2 norm over time for RGB features.



(a) ROC Curve for the L1 threshold. **(b) ROC Curve for the L2 threshold.**

Figure 8: ROC curves for both threshold values on the UCF-Crime RGB features.

clearly outperform all the other features for almost all of the presented methods except for GMM where although the accuracy for the RGB features is higher, the F1 score of the combined and optical flow features are 14,43% and 10,12% higher than RGB features respectively.

Looking at the PCA visualization presented in Figure 6 the reason for the better performance of the RGB features becomes clear. Although both feature sets have a large overlap and creating perfect clusters for either is impossible, the mean of all segments when using RGB features differs more than when using optical flow or combined features. These clusters also provide another explanation for the poor performance of the clustering methods which at best are 10,46% behind in F1 score when comparing GMM on combined features with threshold values on RGB features.

When looking at the graph of the L2 norm over time, it is clear that not for every video the increase, or even change, in L2 norm is an indication of an anomaly. When thinking about RGB features, it makes sense that a video where an explosion happens would show a larger change than a video where a pickpocket occurs. However, what is also an issue with methods like these, is that the dataset does not perfectly fit the expected video under which the methodology was hypothesized to work. The UCF-Crime dataset features videos with intros and multiple cameras, the fact that there is then a large difference between the L2 norms make sense when a scene shifts, just like when an explosion would happen for example. In a real life scenario the camera would be pointed in the same direction 24 hours a day and thus finding a pattern in the L2 norm becomes easier in such a controlled environment. This could lead to devising a certain upper and lower bound which is deemed 'normal' for the L2 norm. Instead of relying on videos longer than the batch number and finding the mean and standard deviation, each batch

Paper	Unsupervised (US) / Weakly-supervised (WS) / Fully supervised (FS)	AUC
Wang et al. [48]	US	70.46%
I3D features - L1 + L2 std. threshold	WS	74.00%
Zhang et al. [51]	WS	78.66%
Liang et al. [29]	WS	84.52%
Zhong et al. [53]	FS	74.60%
Chang et al. [7]	FS	84.62%
Tan et al. [44]	FS	86.71%

Table 5: Comparison of proposed method to state of the art.

could be examined independently against this threshold. A segment-based approach like this was shortly tested but resulted in almost all videos being marked as anomalous because, as can be seen in Figure 7, normal segments also show spikes in the L2 norm.

What also reduces the effectiveness of a segment based approach on a dataset like UCF-Crime, is that the exact timestamp of an anomaly can be ambiguous. In the arson example shown, the L2 norm spikes after the anomaly happens but it could be that at this point the fire is biggest. The human activity has already been done but the anomaly would still be picked up by the system and handled correspondingly. This is also why a clear definition of an anomaly needs to be defined before such datasets are created, as was discussed previously and will be discussed later in Section 7.1. All these reasons lead to a more general approach where a whole video is labeled according to a mean works best.

6.3 State of the art comparison

In Table 5 the comparison between the best performing method presented in this paper, and the state of the art can be found. Almost all research uses the AUC metric to measure performance meaning this is the metric that will be compared. Looking at the state of the art, the best performing fully supervised methods still have a slight performance advantage over the best weakly-supervised with 86,71% AUC versus 84,52% AUC for weakly-supervised methods [29, 44]. The gap to the unsupervised methods is a lot bigger where the AUC is 74,00% [48]. The method presented in this paper is still behind the other weakly-supervised methods but slightly above the unsupervised method. It could be possible that the L1 and L2 norm may not be the best indication for an anomaly. However, it has to be said that the threshold method presented, although it is labeled as weakly supervised, could be adapted to be fully unsupervised when having more data present and a stationary camera is used. Then instead of relying on a training set to define these thresholds, historical data of a camera could be used. This would make it competitively performing when compared to the unsupervised methods. Also, as was hypothesized before when comparing these results, having more data on one single camera and viewpoint can improve the proposed method considerably.

6.4 Feature type comparison

In Table 6 and 7 the results using the threshold methodology can be seen on the fight-detection and ONFIRE dataset respectively [2, 1]. Before looking at the results, there is one key-difference that had to be made for the fight-detection dataset. Instead of using the standard deviation and comparing that to the threshold, the mean of the L1 and L2 norm was used. The reason for this is that the videos in this dataset are often only 1 or 2 seconds long meaning the standard deviation is of no use.

Looking at the results, two things become apparent. Firstly, when comparing the F1 scores of all the features, the ONFIRE dataset scores as intended with a F1 score of 72,36% for the best feature set, versus the 72% of the UCF-Crime dataset. The fight-detection dataset shows slightly worse results with the highest F1 score being 66,67%. However, the main reason these datasets were chosen was

Method	Feature	Threshold value (std.)	Balanced accuracy	Recall	Precision	F1
L1 + L2 mean	*Comb.*	0,30 + 0,03	48,00	**90,00**	60,50	**72,36**
L1 + L2 mean	*Flow*	0,33 + 0,15	48,25	82,50	60,55	69,84
L1 + L2 mean	*RGB*	0,30 + 0,08	**49,75**	87,50	**61,40**	72,16

Table 6: Results on the ONFIRE dataset when using threshold values.

Method	Feature	Threshold value	Accuracy	Recall	Precision	F1
L1 + L2 mean	*Comb.*	0,28 + 21,68	60,00	73,33	57,89	64,71
L1 + L2 mean	*Flow*	0,30 + 14,10	65,00	70,00	63,64	66,67
L1 + L2 mean	*RGB*	0,27 + 15,03	58,33	70,00	56,76	62,69

Table 7: Results on the fight-detection dataset.

to show differences in different anomalies for different feature sets. These can clearly be seen in the fight-detection dataset where the optical flow now performs better by a 1,96% margin in the F1 score to the next feature set. Although for the ONFIRE dataset the optical flow performs worst, the combined features still outperform the RGB features by a difference of 0,20% in F1 score. A marginal difference but not completely expected. This leads to the conclusion that there are significant differences when focusing on different anomalies, however, these are not always as expected. It is important to note that the ONFIRE dataset does not only contain clear fires however but also has a lot of smoke coming from forests for example. This could skew the results away from the RGB features compared to when a dataset only contains clear fires. Furthermore, if a fire has already started at the beginning of a video, this also means the video does not show as much change when compared to a fire being shown from being lit until fading out.

7 DISCUSSION

This Section will discuss the main outcomes of this paper, with the goal of connecting the initial analysis of the ethical and legal implications to the proposed methodology.

7.1 Connecting theory to practice

In Section 2 the most important legal and ethical literature pieces and topics were discussed. The initial discussion regarding known factors like the dataset and an early analysis was also done. This Section will, after having implemented and validated the full methodology, provide a step-by-step rundown of what future work could do with this knowledge and relate it to the datasets and methodology proposed in this paper.

7.1.1 Step 1: Data collection. The topic of data collection was discussed in both the legal and ethical literature. In the legal part the GDPR played a key role and the main point in collecting training data was keeping it for its intended purpose [14]. Many datasets, including the UCF-Crime, ONFire and fight-detection dataset, have the purpose of being used in research. Furthermore, the UCF-Crime dataset is also a rather grey area due to the fact that the data is collected from YouTube or LiveLeak [50, 30]. This is due to two reasons: the first reason is that it is impossible to know if each video was gathered and uploaded with consent. The second problem is the right to be forgotten. There could be a scenario someone wants the YouTube video of them gone, but is either unaware of the dataset also including the video or they are aware and the dataset has to be changed resulting in problems for researchers.

The ethical issue of bias was also discussed before. While it is possible to manually check all videos and analyse potential bias and augment a dataset to reduce the existing bias, this is very labour intensive. Furthermore, bias is also country-specific since ethnic representations differ around the world. Many of the datasets online however, gather data from all over the world. Although for research purposes this provides a diverse set of data, this could run into the problem of not entirely representing a specific country's population. The same can be said for the UCF-Crime dataset.

Combining the issue of legal data collection and bias makes it very difficult to work with public datasets gathered by a 3rd party outside of a research context. Although there are companies like OpenAI and their GPT models which rely on public data, video footage is a much more delicate theme not only in law but also in ethics [35]. For this reasons none of the datasets used in this paper can be recommended when bringing anomaly detection solutions into practice.

For future work, the foundations of ethical data gathering should be further explored. Research should delve into topics of informed consent and lawful data gathering resulting in a dataset which does not exist in all these grey areas. Furthermore, research is also being done on the topic of AI-generated data which could also provide a solution to the problems described (as long as this AI is unbiased and the instructions given are clear) [49].

Finally, from a purely practical point of view, the issue regarding the definitions of an anomaly has to be tackled here. A working definition was given in Section 1.1 but this definition was already shown to be problematic in Section 6.2.1 when datasets do not adhere to this. When labeling data the anomaly should be labeled consistently, be it the action or the change in scenery, and the natural flow of a video needs to be described clearly (these are the two parts of the earlier given definition).

7.1.2 Step 2: Developing and training. In the scenario where step 1 is completed and results in an unbiased and legal dataset, the next step would be developing a model. Again, the theme of bias plays a key role because even though a dataset might be 'balanced', a model or methodology could still work for the wrong reasons. For example, when using RGB features that are dependent on color, problems can arise when certain skin or clothing colors are over-represented in one type of category (even when the dataset as a whole is balanced). The methodology proposed in this paper could still be used as there is no inherent bias since there is no training taking place, but monitoring false positive anomalies should be done to see why these frames are considered anomalous.

The theme of transparency should also be dealt with during this phase. A large field of research, explainable ML/AI, is finding ways to keep these new technologies transparent and understandable for the general public [5]. This was not in the scope of this paper but should be a part of future research in anomaly detection.

Furthermore, as mentioned before in Section 4.2, all metrics used for research purposes in this paper are deterministic in nature. For an actual implementation, a likelihood ratio would better serve the potential users of this technology. In Section 7.3 ideas for this will be discussed. Finally, the proposed technology should be created in close correspondence with the EU according to the AI Act but also other stakeholders as to close the gap between research and practice [36].

7.1.3 Step 3: Validation. The model should be tested more extensively than what is currently the norm in research. This paper showed a dataset-agnostic approach to finding anomalies and recommends it when trying to adapt these technologies to the real world. Although the downsides were shown in performance, the upside can be found in the validation. This validation can easily take place over multiple datasets and in turn aim to prevent bias.

Furthermore, besides validating the methodology on datasets as was shown in this paper, this validation should also be done on the data gathered by the camera(s) which will be used. During this step it is important to find the key parts in which the method might lack.

At these points it is important to always allow room for human intervention, an idea further backed up by the feedback from the NFI.

7.1.4 Step 4: Deployment. Before actual deployment the system has to be assessed by the according to the EU AI Act [36]. This process is fully guided and in the previous steps the needed work to pass this assessment has already been done. This assessment has been described in Section 2. Assuming the assessment is passed, this is the moment when the technology will get the most attention from the general public and it is also the moment in which the guidelines to deal with data and ethical issues need to be fully planned out. People should be informed about how their data is handled and what the model does through publicly available information sources.

7.2 Answering the research questions

Here, the posed research questions are answered based on the analysis of the ethical and legal landscape and the implementation of the proposed methodology.

- *SQ1.1: How can existing anomaly detection methods be improved to better fit a real-life anomaly detection use-case?* This can be done by using a weakly or unsupervised method which does not rely on having every possible anomaly available. For example, by using general methods that rely on the characteristics of an anomaly like breaking an expected pattern rather than the action itself.
- *SQ1.2: What ethical considerations are important when designing a machine learning model for a sensitive topic such as video-based anomaly detection?* At the various stages in a product's life-cycle different ethical topics are important. Especially the topic of bias, explainability and the proper management of data (not going beyond its intended usage) are important. Furthermore, awareness is not enough and there should be accountability when these ethical risks are not dealt with accordingly.
- *SQ1.3: What legal considerations need to be dealt with before being able to deploy and use an anomaly detection model?* In the EU there are two main pieces of legislation, the EU AI Act, which focuses on AI implementations, and the GDPR which aims to defend data against being exploited. They are both relevant at all stages of the proces, the GDPR should be used to guide the usage of data and the EU AI Act should guide the development and monitoring of every AI system.
- *RQ1: How can a proposed system for anomaly detection be improved to make it more usable in a real-life situation?* Anomaly detection models not only pose a significant risk, due to the nature of anomalous behaviour, they are also very hard to make into a practical solution. During development of these models multiple experts on ethics and legal cases need to be present in combination with experts from the field. A joint-effort needs to be made to protect all stakeholders and users to bring these methods beyond research. Furthermore, instead of the work shown here that used deterministic metrics for comparison and research purposes, the final system should not be purely black and white but instead use a likelihood measurement.
- *SQ2.1: How can the L1 and L2 norm of a feature vector be used for a weakly-supervised anomaly detection model?* The L1 and L2 norm can be used in various ways like clustering of tuples of the mean and standard deviation of these norms, assigning a threshold value to them or using isolation forest to group them.
- *SQ2.2: Do different types of anomalies provide better performance in combination with a different feature type from the I3D feature set; RGB, Optical Flow and Combined features.* The results on the different datasets show that anomalies with a large change in the image, like fire-related anomalies, respond better to RGB and combined features. Features with more intricate movement, like fighting (a movement which can easily be mistaken for normal behaviour) respond better to optical flow features which show a change in movement.
- *RQ2: How can a combination of I3D features and various categorization methods be used for unsupervised or weakly-supervised anomaly detection?* The usage of a threshold values provided the best score for unsupervised and weakly-supervised methods using I3D features with a F1 score of 72,00%.

7.3 Future work

Although the performance is not quite enough at this stage, the correlation between the L1/L2 norm and an anomaly in footage is evident. Future work could focus on a multitude of aspects of this correlation.

Research into using pseudo-labels from unsupervised methods has proven to be effective already [25]. These labels were then used to train a model which can be used instead of the unsupervised methods. However, these methods could be combined with what was presented in this paper. For example, although the threshold method is not fully unsupervised as it had to rely on the training set due to the videos being taken with different cameras. A dataset like ShanghaiTech, with longer videos from the same viewpoint, could use the upper and lower bound discussed in Section 6.2.1 to make this method fully unsupervised [52]. These could then be used to generate the pseudo-labels and train a model similar to earlier work done.

Other work has also shown that motion-aware features provide significant benefit when detecting anomalies [54]. I3D features were used for this paper but C3D features were shown to sometimes even outperform the I3D feature set [45]. These features could also be tested with the aforementioned methods.

Furthermore, as discussed in Section 7.1, when working on getting a proposed system beyond research, ideas should be gathered on how to provide an anomaly score rather than a black and white distinction as is also shown in other research [28]. Looking at how this would fit into the proposed methodology of this paper, a possibility would be to make the threshold values a sliding scale instead of a hard divider. Thresholds at certain intervals could get a percentage of certainty and when using both the L1 and L2 norm this could be translated into an anomaly score (either by multiplication of the initial gradients or some other metric).

Finally, although in this paper a foundation for this was presented, a better framework to guide future research on the topics

of ethical and legal issues is needed. These frameworks exist in a basic form but are either too broad or do not relate specifically to anomaly detection [3].

8 CONCLUSION

This paper provided a foundation to put the work done in anomaly detection theory to practice. Feedback from the Netherlands Forensic Institute (NFI) served as the starting point for a deeper look into what current research is missing to more closely resemble real life use cases. By looking at the most relevant literature in the legal and ethical domains and by analysing feedback from the NFI, a class-agnostic methodology was created. The choice for this methodology aimed to both cover new ground in research, as well as provide a potential solution to many of the problems that arise when taking solutions like these beyond research.

Using I3D features and their L1 and L2 norm as a foundation, different methods were used to differentiate between normal and abnormal videos. The method that performed best assigned a threshold value to the standard deviation of these norms. Although the validation of the methods on the UCF-Crime dataset showed that there is still a long way to go, a foundation for future work using these methods has been provided. The reasoning as to why the performance is not as high as one would hope is also given and tests on other datasets further strengthen these points.

Furthermore, besides these technical details a discussion on how other research in the field could take its next step towards being put into practice was also given. This discussion also connected the work done in this paper to the lessons learned from practice, and provided the start of a framework. This framework can serve as a foundation for future work in the field of practical video anomaly detection solutions.

REFERENCES

[1] Gincheva A. *ONFIRE Dataset: Harmonizing Decades of Wildland Fire Data*. 2023. DOI: https://doi.org/10.5194/egusphere-egu24-8496,2024.

[2] Seymanur Akti, Gözde Ayse Tataroglu, and Hazim Kemal Ekenel. "Vision-based Fight Detection from Surveillance Cameras". In: *CoRR* abs/2002.04355 (2020). arXiv: 2002.04355. URL: https://arxiv.org/abs/2002.04355.

[3] Mona Ashok et al. "Ethical framework for Artificial Intelligence and Digital technologies". In: *International Journal of Information Management* 62 (2022), p. 102433. ISSN: 0268-4012. DOI: https://doi.org/10.1016/j.ijinfomgt.2021.102433. URL: https://www.sciencedirect.com/science/article/pii/S0268401221001262.

[4] Jacqui Ayling and Adriane Chapman. "Putting AI ethics to work: are the tools fit for purpose?" In: *AI and Ethics* 2.3 (Sept. 2021), pp. 405–429. ISSN: 2730-5961. DOI: 10.1007/s43681-021-00084-x. URL: http://dx.doi.org/10.1007/s43681-021-00084-x.

[5] Vaishak Belle and Ioannis Papantonis. "Principles and practice of explainable machine learning". en. In: *Front. Big Data* 4 (July 2021), p. 688969.

[6] João Carreira and Andrew Zisserman. "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset". In: *CoRR* abs/1705.07750 (2017). arXiv: 1705.07750. URL: http://arxiv.org/abs/1705.07750.

[7] Shuning Chang et al. "Contrastive Attention for Video Anomaly Detection". In: *IEEE Transactions on Multimedia* 24 (2022), pp. 4067–4076. DOI: 10.1109/TMM.2021.3112814.

[8] Weiling Chen et al. "TEVAD: Improved video anomaly detection with captions". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2023, pp. 5549–5559. DOI: 10.1109/CVPRW59228.2023.00587.

[9] European Commision. *Shaping Europe's digital future: AI Act*. 2022. URL: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

[10] Bo Cowgill et al. *Biased Programmers? Or Biased Data? A Field Experiment in Operationalizing AI Ethics*. 2020. arXiv: 2012.02394 [econ.GN].

[11] Bruno Degardin and Hugo Proença. "Iterative weak/self-supervised classification framework for abnormal events detection". In: *Pattern Recognition Letters* 145 (2021), pp. 50–57. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2021.01.031. URL: https://www.sciencedirect.com/science/article/pii/S0167865521000507.

[12] Hanqiu Deng et al. "Bi-directional Frame Interpolation for Unsupervised Video Anomaly Detection". In: *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2023, pp. 2633–2642. DOI: 10.1109/WACV56688.2023.00266.

[13] EU. *Protecting data and opening data*. 2018. URL: https://data.europa.eu/en/publications/datastories/protecting-data-and-opening-data.

[14] European Parliament and Council of the European Union. *Regulation (EU) 2016/679 of the European Parliament and of the Council*. of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). May 4, 2016. URL: https://data.europa.eu/eli/reg/2016/679/oj (visited on 04/13/2023).

[15] Thilo Hagendorff. "The Ethics of AI Ethics: An Evaluation of Guidelines". In: *Minds and Machines* 30.1 (Feb. 2020), pp. 99–120. ISSN: 1572-8641. DOI: 10.1007/s11023-020-09517-8. URL: http://dx.doi.org/10.1007/s11023-020-09517-8.

[16] Charles R. Harris et al. "Array programming with NumPy". In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2.

[17] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *CoRR* abs/1512.03385 (2015). arXiv: 1512.03385. URL: http://arxiv.org/abs/1512.03385.

[18] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

[19] J. D. Hunter. "Matplotlib: A 2D Graphics Environment". In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[20] Bryan Johnston. *A brief history of surveillance cameras*. July 2023. URL: https://www.deepsentinel.com/blogs/home-security/history-of-surveillance-cameras/.

[21] Will Kay et al. "The Kinetics Human Action Video Dataset". In: *arXiv preprint arXiv:1705.06950* (2017).

[22] Emre Kazim and Adriano Soares Koshiyama. "A high-level overview of AI ethics". In: *Patterns* 2.9 (2021), p. 100314. ISSN: 2666-3899. DOI: https://doi.org/10.1016/j.patter.2021.100314. URL: https://www.sciencedirect.com/science/article/pii/S2666389921001574.

[23] Mauritz Kop. *European AI Alliance - EU artificial intelligence act: The European approach to ai*. URL: https://futurium.ec.europa.eu/en/european-ai-alliance/document/eu-artificial-intelligence-act-european-approach-ai?language=pt-pt.

[24] Nancy G La Vigne et al. 2011.

[25] Anas Al-lahham et al. *A Coarse-to-Fine Pseudo-Labeling (C2FPL) Framework for Unsupervised Video Anomaly Detection*. 2023. arXiv: 2310.17650 [cs.CV].

[26] Stefan Larsson and Fredrik Heintz. "Transparency in artificial intelligence". en. In: *Internet Pol. Rev.* 9.2 (May 2020).

[27] Tony Lawson, Robert Rogerson, and Malcolm Barnacle. "A comparison between the cost effectiveness of CCTV and improved street lighting as a means of crime reduction". In: *Computers, Environment and Urban Systems* 68 (Mar. 2018), pp. 17–25. DOI: 10.1016/j.compenvurbsys.2017.09.008. URL: https://doi.org/10.1016/j.compenvurbsys.2017.09.008.

[28] Tianyu Li et al. "Anomaly Scoring for Prediction-Based Anomaly Detection in Time Series". In: *2020 IEEE Aerospace Conference*. 2020, pp. 1–7. DOI: 10.1109/AERO47225.2020.9172442.

[29] Weijie Liang, Jianming Zhang, and Yongzhao Zhan. "Weakly supervised video anomaly detection based on spatial–temporal feature fusion enhancement". en. In: *Signal Image Video Process.* 18.2 (Mar. 2024), pp. 1111–1118.

[30] LiveLeak. *LiveLeak - Redefine the Media*. Accessed: 2024-07-27. 2024. URL: https://www.liveleak.com.

[31] Cewu Lu, Jianping Shi, and Jiaya Jia. "Abnormal Event Detection at 150 FPS in MATLAB". In: *2013 IEEE International Conference on Computer Vision*. 2013, pp. 2720–2727. DOI: 10.1109/ICCV.2013.338.

[32] V. Mahadevan et al. "Anomaly detection in crowded scenes". In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2010, pp. 1975–1981. DOI: 10.1109/CVPR.2010.5539872.

[33] Simon Marcel and Yannick Rodriguez. *Torchvision the Machine-Vision Package of Torch*. https://github.com/pytorch/vision. 2010.

[34] Sangmin Oh et al. "A large-scale benchmark dataset for event recognition in surveillance video". In: *CVPR 2011*. 2011, pp. 3153–3160. DOI: 10.1109/CVPR.2011.5995586.

[35] OpenAI. *GPT-4: Generative Pre-trained Transformer 4*. Accessed: 2024-07-27. 2024. URL: https://www.openai.com.

[36] European Parliament. *EU AI Act: first regulation on artificial intelligence*. Dec. 2023. URL: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

[37] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*. Ed. by H. Wallach et al. 2019, pp. 8024–8035.

[38] Fabian Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830. URL: http://jmlr.org/papers/v12/pedregosa11a.html.

[39] Hao Vy Phan. *i3dFeatureExtraction*. Available at https://pypi.org/project/i3dFeatureExtraction/, version 0.3.3. (Visited on 03/04/2024).

[40] David E. Rumelhart and James L. McClelland. "Learning Internal Representations by Error Propagation". In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*. 1987, pp. 318–362.

[41] Strijbosch. "Proposing an ontology for human-related crime recognition in videos". In: (2022).

[42] Waqas Sultani, Chen Chen, and Mubarak Shah. "Real-world Anomaly Detection in Surveillance Videos". In: *CoRR* abs/1801.04264 (2018). arXiv: 1801.04264. URL: http://arxiv.org/abs/1801.04264.

[43] Weijun Tan, Qi Yao, and Jingfeng Liu. "Overlooked Video Classification in Weakly Supervised Video Anomaly Detection". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*. Jan. 2024, pp. 202–210.

[44] Weijun Tan, Qi Yao, and Jingfeng Liu. "Overlooked Video Classification in Weakly Supervised Video Anomaly Detection". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*. Jan. 2024, pp. 202–210.

[45] Du Tran et al. "C3D: Generic Features for Video Analysis". In: *CoRR* abs/1412.0767 (2014). arXiv: 1412.0767. URL: http://arxiv.org/abs/1412.0767.

[46] Hannu Turtiainen, Andrei Costin, and Timo Hämäläinen. "CCTV-Exposure: System for Measuring User's Privacy Exposure to CCTV Cameras". In: *Business Modeling and Software Design*. Ed. by Boris Shishkov. Cham: Springer International Publishing, 2022, pp. 289–298. ISBN: 978-3-031-11510-3.

[47] Waseem Ullah et al. "CNN Features with Bi-Directional LSTM for Real-Time Anomaly Detection in Surveillance Networks". In: *Multimedia Tools and Applications* 80 (May 2021). DOI: 10.1007/s11042-020-09406-3.

[48] Jue Wang and Anoop Cherian. "GODS: Generalized One-class Discriminative Subspaces for Anomaly Detection". In: *CoRR* abs/1908.05884 (2019). arXiv: 1908.05884. URL: http://arxiv.org/abs/1908.05884.

[49] Zuhao Yang et al. *AI-Generated Images as Data Source: The Dawn of Synthetic Era*. 2023. arXiv: 2310.01830 [cs.CV].

[50] YouTube. *YouTube - Broadcast Yourself*. Accessed: 2024-07-27. 2024. URL: https://www.youtube.com.

[51] Jiangong Zhang, Laiyun Qing, and Jun Miao. "Temporal Convolutional Network with Complementary Inner Bag Loss for Weakly Supervised Anomaly Detection". In: *2019 IEEE International Conference on Image Processing (ICIP)*. 2019, pp. 4030–4034. DOI: 10.1109/ICIP.2019.8803657.

[52] Yingying Zhang et al. "Single-Image Crowd Counting via Multi-Column Convolutional Neural Network". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 589–597. DOI: 10.1109/CVPR.2016.70.

[53] Jia-Xing Zhong et al. "Graph Convolutional Label Noise Cleaner: Train a Plug-and-play Action Classifier for Anomaly Detection". In: *CoRR* abs/1903.07256 (2019). arXiv: 1903.07256. URL: http://arxiv.org/abs/1903.07256.

[54] Yi Zhu and Shawn D. Newsam. "Motion-Aware Feature for Improved Video Anomaly Detection". In: *CoRR* abs/1907.10211 (2019). arXiv: 1907.10211. URL: http://arxiv.org/abs/1907.10211.