Proposing an ontology for human-related crime recognition in videos

DAAN STRIJBOSCH, University of Twente, The Netherlands

Video surveillance has been around for a very long time and, throughout the years of its existence, it has seen a large growth in popularity. However, due to this increase in produced footage, manual monitoring is not always an option. This paper introduces a framework that groups crime and provides a classification of different relevant aspects, such as location and demeanor of the suspect. This so-called ontology of crime, in the form of a semantic tree, is then used to build a model upon the existing ResNet50 model. The proposed model achieves an accuracy of 37.7% compared to 34.4% accuracy of the regular ResNet50 model. Furthermore, an implementation where multiple frames are used to classify one instance of crime is shown. As well as an implementation of a threshold, that filters out low quality frames during the classification process.

Additional Key Words and Phrases: Computer vision, ontology design, scene-recognition

1 INTRODUCTION

Closed-circuit television (CCTV), also known as video surveillance, is a concept that has been around for many years. The idea of using cameras as a surveillance system was born in the year 1947 and has since then seen many different applications ranging from street monitoring to behavioural analysis[16]. Because of this rise in popularity, the amount of footage that is being gathered every day is also rising to extreme levels. It is estimated that there are 200,000,000 cameras currently operating that, altogether, produce more than 10 million GB of video footage per week[8]. And although many experts still see the merits of CCTV, more critical voices have also raised their concerns about its effectiveness. Manually analyzing the footage that is gathered by these CCTVs is a strenuous task and research shows that the operators of these CCTVs often cannot keep up with both monitoring the footage and all other related tasks. This is problematic since research also indicates that the biggest factor in the effectiveness of CCTV is the operator monitoring the video footage.[7, 26]

All of this goes to show that the current way of working with surveillance footage is becoming outdated quickly. Because of this many calls have been made to reduce the amount of CCTV and instead opt for other solutions [9]. However, on the other side of the spectrum, research is being done to modernize CCTV. This new area of research focuses on the development of solutions to alleviate the workload from the people that monitor the surveillance footage. These solutions aim to summarize videos or detect anomalies using different machine learning methods[18, 19]. But although these new ideas all show improved effectiveness, the amount of research currently available is little. Furthermore, the research that is done often varies largely in scope and goal.

Crime can come in many different forms, from assault to burglary, but current research often focuses only on one type of crime, such as shoplifting or weapon detection, and fails to touch upon the different types of crime[18, 19]. The reason for the narrow scope of many of these papers is that there is a lack of a solid foundation. Where the human eye can easily detect and categorize crimes up to a certain level, a clear framework for machine learning applications does not exist

The goal of this paper is to create a clear foundation for future research regarding crime detection on CCTV using machine learning. It will do so by laying out the different types of crime and designing an ontology to detect these crimes. By first describing the different types of crime and finding the similarities and dissimilarities between, them this paper will group types of crime. Afterwards it will develop an ontology that describes what these types of crimes look like and how they can be systematically detected. The ontology will include the different types of physical identifiers that are characteristic to a certain crime such as body movement and location. This ontology will then be implemented and tested to verify its quality when used in real life scenarios.

1.1 Aim

Although research has been done on the topic of machine learning for crime detection, a clear foundation for this is still missing. The small amount of research papers published create their own methods and techniques to focus on their particular section of crime [11, 18]. On top of this, most research focuses more on the machine learning than the crime meaning that many identification methods are inconsistent among different papers. This paper will analyze different types of crime and look into how these can be grouped and detected. From this analysis an ontology will be created that will include what certain types of crime look like and how they can be systematically detected. This analysis and ontology will then be put to the test on a provided dataset which will show the effectiveness of using this ontology for future research. This leads to the following research question (**RQ**):

RQ. Can we propose an ontology on which to base a hierarchical classification approach for human-related crime recognition in videos?

To answer the main research question the following sub-questions (**SQ**) will be used:

SQ1. What types of crime exist and what physical characteristics can categorize them?

SQ2. How can the different categories found in SQ1 be used to create an ontology of crime?

SQ3. Does the proposed ontology improve the performance of crime detection using machine learning?

TScIT 37, July 8, 2022, Enschede, The Netherlands

1.2 Contribution

By answering the previously described research question, we aim to provide an ontology for crime detection that is well tested and scientifically relevant. This ontology could then serve as the foundation of future research into the field of crime detection or could provide a baseline for other fields which require something similar. The framework makes use of both the physical location and the physical attributes, such as movement and posture, of a criminal activity to find ways to classify these activities. Furthermore, once fed to an algorithm, the framework improves on both the speed and accuracy of the classification process.

1.3 Organization of this paper

The paper will be structured as follows. First the related work will be laid out and analyzed 2. This section will be subdivided into three different parts of related work all relevant to the research; computer vision, crime categorization and ontology design. With this related work clear the next section will be the methodology 3. This methodology will go into details on how the proposed ontology will be designed and provide the reasoning behind the choices made. This proposed methodology will then be tested. In the experiments section the setup of the experiment, like the dataset, validation methods and testing methods, will be explained and then the results will be shown 4. The results will be discussed in the discussion section which will try to interpret the results 5. Finally, a conclusion can be made which serves as a summary of the findings of this paper and will lead into the future work in this area of research 6.

2 RELATED WORK

This section will go over some of the related work in the fields of computer vision and crime categorization.

2.1 Computer vision

Starting with computer vision, the concept of getting computers to gain understanding from an image or a video has been around since the 1960s, when L. Roberts described the idea of gaining 3D info from 2D objects. Ever since then, the first attempts have been made to gain this information and process it for various use cases[22]. The first attempts focused on extracting edges and then trying to deduce a 3d-object from there [5]. These early efforts often resulted in vague outlines and, while results were noteworthy, they were nowhere near good enough for any practical use cases. Since then much has changed but it is not until the year 2001, when real-time face recognition became possible, that we can see possible applications for crime detection [24]. Even now, the created algorithm performs well, however, since 2001 a lot has changed.

This brings us to the modern, and for us relevant, research into computer vision. Most research currently done uses a type of neural networks known as convolutional neural networks (CNN) [8, 11, 12, 15]. A CNN is a type of neural network that is mainly used in image and video processing to find patterns and make sense out of the image or video it is presented with. Due to its popularity, the research done into these CNNs is extensive and many useful applications can be found.

An example of the use of a CNN is found in research done in 2014 by M. Malekar [12]. In the paper, an algorithm is devised that uses a CNN to extract important information out of videos and create a summary from this information. The summary created by this algorithm only missed 2-5 frames of suspicious footage in videos ranging from 85-439 frames duration.

Work by Singh et al. [20] showcases a CNN used to track contour displacement to identify shoplifting. The work in this paper is based on the OpenCV library, a library that is also used in many other papers with similar goals [2, 14]. OpenCV is an example of one of the open source computer vision libraries that can assist researchers with topics such as scene-recognition or image classification.

2.2 Crime categorization

Besides work in computer vision, it is also important to establish a foundation in crime categorization. When looking at crime classification, one of the most important pieces of literature is the Crime Classification Manual written by J.E Douglas[3]. The manual serves as a cohesive list of all different types of crime and explains the definition and characteristics of them. Over the years, the crime classification manual has grown to be one of the most, if not the most, influential book for crime classification in criminal cases.

The Crime Classification Manual provides an extensive list on types of crime and definitions and can prove useful in our research on the types of crime. However, it does not provide a foundation for the design of our ontology since there is no extensive documentation on visual characteristics of the crime itself. Other research on this topic is scarce and while some efforts have been made to classify and detect crime [6, 25], none provide an in-depth framework for future work. This lack of research does, however, highlight the importance of this paper.

2.3 Ontology

The final important piece of preliminary research, is to look at ontology design. Looking at the literature, three different approaches have been found with regards to ontology design for image/video analysis.

The first approach is showcased in the paper by W. Fang et al. [4]. In the paper, a knowledge graph is used to identify hazards on construction sites. The paper defines certain entities such as people, equipment and materials. These entities are then extracted from the video footage and, using bounding boxes, the spatial-relationship between these entities is checked. From these spatial-relationships tables are made specifying if, for example, the bounding box of someone's helmet is in the bounding box of that person, meaning they are wearing the helmet. This extracted knowledge is then compared to a set of rules that specify safety hazards. While promising, the amount of examples given were scarce, as is also acknowledged in the limitations sections, meaning that the approach is not yet extensively tested.

The second approach that was found is using a hierarchical classification approach. This approach is used in a paper by E. Martínez et al. [23]. The paper focuses on defining a hierarchical approach that can classify different food related scenes. In practice this means that different levels are created that differ in abstraction with the

most abstract level first and the least abstract level last. In the case of food related scenes this resulted in level 1 relating to the physical activity, eating, preparing or acquiring, and level 2 relating to the environment. Experimenting with this hierarchical classification shows an increase in weighted accuracy over general algorithms. A limitation mentioned is that it is difficult to classify scenes when there are less characteristic differences over other locations.

The third approach is closely related to the second approach and is shown in a paper by D. Cavaliere et al. [1]. The paper focuses on designing an ontology to analyse unmanned aerial vehicle (UAV) video content. The approach proposed here also uses different layers but not necessarily in the form of a semantic tree. The layers used in this case relate to the understanding of the sensor data with level 0 being the raw data and level 3 being the situation. Just like in the previous paper there is a step by step approach where, in this case, the scene is interpreted first, then the activity/event and then the situation. However, unlike the previous paper, there is no semantic tree that is used as a guide. Rather, multiple ontologies are combined that interpret each layer separately and pass that information on to the next layer. Although examples are shown, the paper does not feature extensive validation but rather, aims to act as a guideline for future projects. Furthermore, this approach is quite convoluted since there is a lot of cross-referencing between different ontologies and ways to interpret them.

3 BUILDING A HUMAN-RELATED CRIME ONTOLOGY

This section describes the methodology that will be used as a foundation for the later experiments. The project will start by performing literature research, focused on finding what different types of crimes can be found and how they can be grouped. Afterwards, literature research will be performed to see how crime can be detected manually, so, how a human would systematically categorize crime. And how crime can be detected with the use of computer vision, answering the question which algorithm will be used to verify the ontology. Finally, the research done will be verified by performing tests on the previously designed ontology.

3.1 Classification approach

Before we can actually start designing an ontology, it is important to perform research into crime that will serve as the groundwork for the rest of the research. In the related work section, some sources regarding crime classification were already mentioned [3, 6]. For this part of the research we will use the Crime Classification Manual by Douglas et al. as a source for all different types of crime [3]. The manual has been used as a foundation for countless other research papers and serves as the standard for language and classification of the criminal justice system. Using this manual, first the relevant types of crime for our research will be selected. The definition of what is deemed relevant, together with the selection procedure, will be described in the corresponding subsection. After this selection has been made, the types of crime that have been selected will be mapped to the crime that is available in the dataset which will later be used to perform the experiments.

3.1.1 Category selection.

Before starting the selection process it is important to clearly identify

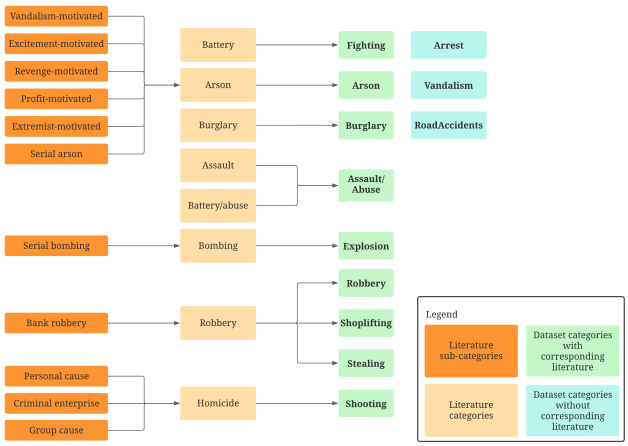

Fig. 1. Mapping of crime literature to the dataset

the selection criteria. Since the aim of the research is to create a method that can differentiate between types of crime, all categories listed that do not have any physical identifiers will have to be left out. This means that, for example, motivational aspects of crime, which are labelled in the classification manual, will be left out, as they do not provide added value to the ontology. This leaves us with all crimes that are:

- Visually identifiable
- Distinguishable from other selected crimes

Furthermore, some videos will feature more than one instance of a crime or even multiple crimes. In these cases, the approach will be to label each crime individually instead of looking at the video as one instance.

From the categories found in the Crime Classification Manual, the following meet the set requirements [3]:
Robbery, burglary, assault, battery/abuse, solo homicide (a collection of all single suspect homicides), group homicide (a collection of all multiple suspects homicide), arson, crime concealment, riots/civil disturbance, bombing and sexual assault

3.1.2 Mapping to the dataset.

Now that the types of crime have been selected, it is important to map them to the 13 categories provided in the dataset [21]. Starting with the easy to map categories, like for example *assault*, which is both a category in the Crime Classification Manual and in the dataset. Other categories are linked to categories that are closest related to them based on visual identifiers. Fig. 1 shows the mapping of the manual's categories on the left to the datasets' categories on the right.

As can be seen in the figure, some categories that have been previously listed have been omitted; sexual assault and crime concealment. In the case of sexual assault this is because the dataset has nothing that is closely related, meaning that this is not something that can be worked with further from here. In the case of crime concealment because it was previously explained that multiple instances of crime in the same video will be separated. This makes

it difficult to classify something as crime concealment when the previous crime is not taken into account. This assures that crime will be detected properly and makes it is easier to differentiate when a video contains multiple crimes.

Additionally, in some cases the category found in the dataset does not have a counterpart in the crime classification manual. This is the case for the following categories: *arrest, road accidents, shooting, shoplifting* and *vandalism*. In the case of *arrest* the reason is clear, this category is not a crime but rather an action performed by police. Because it is likely that in video footage of crime an (attempt to) arrest will be seen, it is still important for the design of the ontology so that the algorithm does not classify it as, for example, regular fighting. In the case of *road accidents* it is the same, this is not a crime but rather an accident. However, this category is suitable to be used since it is still closely related to *assault*. *Shooting* is also not a crime which is classified individually in the manual but which will be kept separate. The reason for this is because in the case of *shooting* this gives us the chance to identify a firearm and see how the classifier handles these situations.

Shoplifting, stealing and *vandalism* are three categories in the dataset that are not separately classified in the manual. In the case of *stealing* the category will be kept separate from *robbery* because there is an identifiable difference according to literature [17]. The difference is that *robbery* uses force or intimidation where stealing/theft does not. Besides *stealing* and *robbery* the final category in the same scope of crime is *shoplifting*. In the mapping process we also kept this category separate since there is also a distinguishable difference. Similar to *stealing*, *shoplifting* does not use force or intimidation but the location does change opposed to stealing. *Shoplifting* happens when an individual takes something from a shop whereas stealing is a more general term. As with *shooting*, keeping these categories separate allows us to fully test the algorithm when two crimes are very closely related. Finally, in the case of *vandalism* it is listed in the book as a subcategory or motivation of other crimes but not separately. For this reason this relation is also not shown in the mapping process. However, for our ontology design and future research we will keep it as a category and define it as follows: 'conduct that damages others' property'. The reason for this is that this definition is both clearly separate from other categories and also allows for visual detection.

3.2 Ontology design

After having completed this mapping process, we now have a complete picture of crime in literature and the crime in our dataset. Using the findings discussed in the previous sub-chapter an ontology can be designed of the 13 categories presented. The ontology will be based on literature on the category itself, the explanation given in the previous sub-chapter and the mapping presented including the closely related crimes which are not in the dataset.

Firstly, before an ontology can be made, it is important to establish what it will look like. In the related works section three different approaches have been shown. Next to this, during the mapping process, two different defining factors for certain categories have already been found: use of a weapon (in the case of *shooting* vs. *homicide*) and location of the crime (in the case of *shoplifting* vs. stealing). Looking at the literature presented the best option is to design a hierarchical approach, presented in the paper about classifying food related scenes [23]. The reason for this is that using spatial-relationships, while possibly more accurate when looking at weapon use, does not allow the incorporation of location or other visual characteristics[4]. The other approach mentioned which also showcases a hierarchical classifier focuses on multiple different types of sensor data. Since the only sensor that will used is a camera, there is not enough data to properly interpret this result is also not viable for our use case [1].

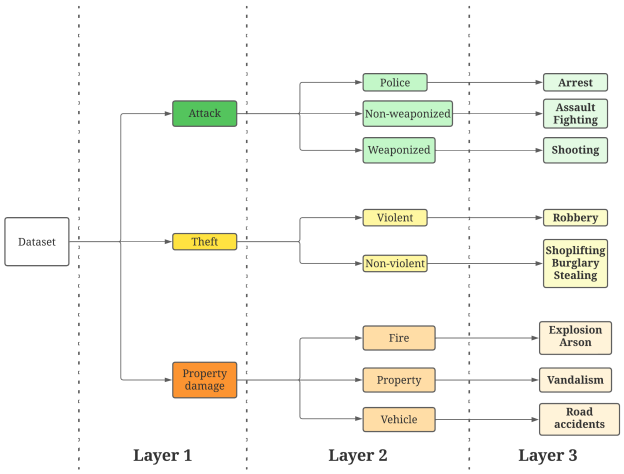

Fig. 2. Proposed semantic tree for classification

For the design of the semantic tree a combination of the Crime Classification Manual, dataset inspection and dictionary definitions has been used [3]. Using these methods every category has been linked to certain keywords between which overlap was found and the semantic tree was designed. The proposed semantic tree can be found in Fig. 2.

4 EXPERIMENTAL FRAMEWORK

This section will show the setup and the results of the performed experiments, which are used to validate the previously made classification approach. The section contains an explanation of the dataset, validation metrics used, the design of the experiment and the results from the experiments.

4.1 The dataset

To perform the experiments, a dataset has been provided. As mentioned previously, the dataset that will be used is the UCF-crime dataset [21]. The dataset features 13 different categories of crime containing in total 950 videos and information on these videos. Table 1 shows the different categories and the amount of videos in these categories. Besides the name and category, some videos in the dataset included frame level annotations. These annotations specify for each frame whether it is anomalous or normal. In Fig. 3 example frames for some of the included categories can be seen.

Table 1. Overview of the provided dataset

Category	No. of videos	Category	No. of videos
Abuse	50	Robbery	150
Arrest	50	Shooting	50
Arson	50	Shoplifting	50
Assault	50	Stealing	100
Burglary	100	Vandalism	50
Explosion	50	Fighting	50
Road Accidents	150	**Total**	950

Table 2. Amount of frames used during experimentation

Category	No. of frames	Category	No. of frames
Arrest	570	Robbery	1520
Arson	570	Shooting	560
Assault	1130	Shoplifting	530
Burglary	1070	Stealing	1010
Explosion	580	Vandalism	500
Road Accidents	1520	Fighting	590
		Total	**10.150**

4.1.1 Dataset preparation. Before starting the experiments, some work had to be done on preparing the dataset. The classification approach considered in this paper is based on still images and not on video footage. This means that in order to be able to test the approach, frames have to be extracted from the videos. Furthermore, only anomalous frames are useful since the research is not about finding anomalous frames but rather about classifying them. Previously it was mentioned that only certain videos had the frame level annotations, meaning for a large list of videos no anomalous frames could be extracted. Luckily, the MPVIR research group researched anomaly recognition in videos and for this they used the UCF crime dataset [13]. The results of their research were frame level annotations for all videos. Thankfully, after getting into contact with the research group, they shared their results meaning that for every video the anomalous frames could now be extracted. Instead of extracting all of the anomalous frames, around 200 per event, 10 frames spaced out evenly over the event were extracted from each anomalous instance. Some videos contained multiple instances of the same event, meaning from these videos 20 frames were extracted. This process resulted in 10.150 frames from 1015 events which will be used to train, validate and test the proposed approach. In table 2 the total amount of frames per video can be seen. Note that for this table the assault and abuse class have been merged into assault, as explained in the *Building a Human-related crime ontology* section 3.

4.2 Metrics

To evaluate the approach, four different metrics will be calculated for each model: accuracy, precision, recall and F1-score. Accuracy is the most straightforward metric, it is the ratio of correctly categorized videos to the total amount of videos. The precision determines the amount of correctly categorized types of crime in the pool of identified crime. The recall determines how many times the algorithm correctly identified crime and how many times it could have identified it. The F1-score, also known as balanced F-score, shows the balance between the recall and the precision. The calculations will be done according to the following formulas:

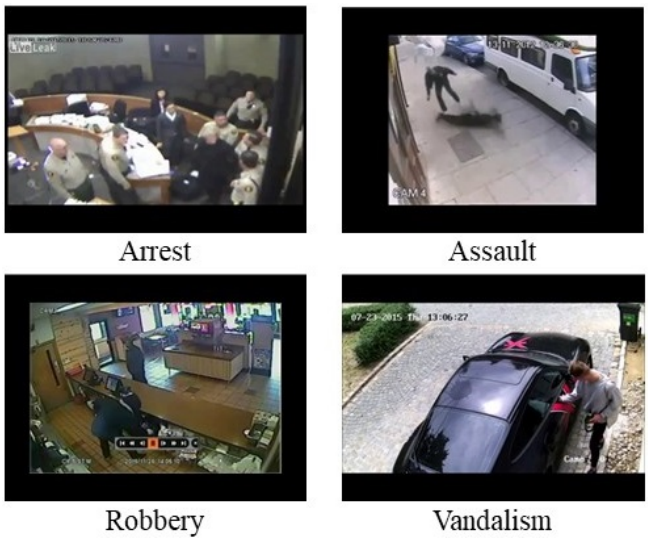

Fig. 3. Example frames from the UCF-crime dataset

$$Accuracy = \frac{TruePositive + TrueNegative}{(TruePositive + FalsePositive + TrueNegative + FalseNegative)}$$

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

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

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

For all 4 metrics, the *weighted* and *macro* variants will be calculated. Both variants calculate the specific metric for each class. The difference between the two however, is that the macro score takes the arithmetic mean of these calculations. The weighted score aims to normalize class imbalances by assigning a different weight to a certain class when calculating the metric.

The reason why solely using accuracy is not enough is because the dataset used is not balanced. The biggest classes (*Road accidents* and *robbery*) both take up roughly 15% of the dataset. This could lead to a situation where only classifying these two classes on a 50/50 basis could lead to a relatively high accuracy. By looking at multiple metrics we can compare the different results and get a better picture of the actual performance.

Next to this, for every category, the accuracy will be calculated and transformed into a confusion matrix. These matrices will be used to investigate categorical differences. For this goal only using the accuracy is enough since every class will be looked at independently.

Table 3. Macro metrics for the ResNet50 and Semantic model

	ResNet50				Semantic			
Threshold	**Precision**	**Recall**	**F1-score**	**Accuracy**	**Precision**	**Recall**	**F1-score**	**Accuracy**
0	0.278	0.268	0.26	0.311	0.226	0.279	0.245	0.338
0.3	0.269	0.265	0.262	0.305	0.249	0.297	0.260	0.358
0.5	0.240	0.244	0.241	0.291	**0.293**	**0.324**	**0.290**	**0.377**
0.7	0.289	**0.299**	**0.291**	**0.344**	0.270	0.305	0.274	0.357
0.9	**0.297**	0.285	0.287	0.331	0.259	0.285	0.262	0.344

Table 4. Weighted metrics for the ResNet50 and Semantic model

	ResNet50				Semantic			
Threshold	**Precision**	**Recall**	**F1-score**	**Accuracy**	**Precision**	**Recall**	**F1-score**	**Accuracy**
0	0.317	0.311	0.309	0.268	0.277	0.338	0.299	0.279
0.3	0.311	0.305	0.304	0.265	0.304	0.358	0.316	0.297
0.5	0.289	0.291	0.289	0.244	**0.339**	**0.377**	**0.339**	**0.324**
0.7	0.342	**0.344**	**0.341**	**0.299**	0.318	0.358	0.324	0.305
0.9	**0.346**	0.331	0.335	0.285	0.311	0.344	0.316	0.284

4.3 Experimental setup

To perform the experiments, two different models had to be implemented. Firstly, a regular non-changed convolutional neural network was used and secondly the same model with the semantic tree model built on top of it. From now on, the model using the semantic tree will be referred to as the semantic model. The regular model of choice is the ResNet50 model since it shows promising results in similar use cases, both when used as an action classifier and when using it as a foundation for another model [10].

To built the semantic model, it was trained on each layer of the tree separately. To get to the final prediction the joint probability of each layer of the tree will be calculated and combined. As an example, to get the probability of the image relating to *shoplifting* the calculation would look as follows:

$$P(shoplifting, x) =$$
$$P(dataset, x|Theft, x) * P(Theft, x|Nonviolent)$$
$$*P(Nonviolent, x|Shoplifting, x)$$

As mentioned before, 10 frames were extracted from every anomaly. Instead of classifying each separately the combined probability of all frames corresponding to an event will be used. This will result in one classification per anomalous event so in total 1.011 classifications, the same amount as the amount of events in the 950 videos. On top of this, experiments will be done defining a certain threshold for frames used in the combined probability. What this means is that a frame is only used in the final prediction whenever the highest probability is above a defined threshold. By doing this there is the potential of having more accurate classifications since frames that are too uncertain are not used to classify a video.

This will result in the following results which will be shown in the next section:

- A comparison of the baseline ResNet 50 and the semantic model based on the metrics: accuracy, precision, recall and F1-score

Table 5. Accuracy per category for best performing models

Category	ResNet50	Semantic model
Arrest	0	**0.12**
Arson	0.62	**0.88**
Assault	**0.22**	0.17
Burglary	0.38	0.38
Explosion	**0.25**	0.12
Fighting	**0.29**	0
Road accidents	0.5	**0.67**
Robbery	**0.39**	0.35
Shooting	0	0
Shooting	0.44	0.44
Stealing	0.5	**0.62**
Vandalism	0	**0.14**

- A comparison of the thresholds 0.3, 0.5, 0.7 and 0.9 for both of the models
- A comparison of the accuracy for each category derived from the highest performing threshold for both models
- A comparison of the accuracy for each layer and each node in the semantic tree

4.4 Results

The obtained, macro and unweighted, accuracy, precision, recall and F1-score for both the baseline ResNet50 and semantic model, as well as the different threshold levels, can be found in Table 3 and Table 4 respectively. Looking at the results, it can be seen that in terms of accuracy, the semantic model always outperforms the regular ResNet50 model. Looking at the other metrics, however, it becomes clear that for these it is not always the case. For each threshold level it can be seen that the balance shifts between the two models. As expected due to the quality of some frames, both models perform

better when using a threshold for categorization as opposed to no threshold. Looking at the scores for the best performing threshold levels the semantic model does outperform the ResNet50 model marginally.

In Table 5 the two highest performing threshold levels, 0.7 for the ResNet50 model and 0.5 for the semantic model, can be seen in more detail. Looking at the accuracy of the two models, it can be seen that there is no clear winner between which model categorizes the most categories correct. However, what is interesting to note is that the top 3 categories for the semantic model all have a higher accuracy than the best category for the ResNet50 model. The same can be said for the lower categories. Whereas the semantic model has 6 categories below 0.2 accuracy, the ResNet50 model has only 3. This means that it seems like the semantic model is doing significantly better on some categories while completely dropping the accuracy on some.

Taking a look at the confusion matrix of the semantic model in Fig. 4 this hypothesis can be confirmed. However, looking deeper into the confusion matrix shows another interesting fact. Looking at the categories with accuracy below 0.2, it can be seen that categories closely related in the semantic tree are often mistaken for these categories. Some examples are the *explosion* category where *arson* and *road accidents* are the two leading predictions, both relating to 'property damage, fire' and 'property damage' respectively in the semantic tree. Likewise, *shoplifting* has an accuracy of 0.44 but when we also count *robbery* then the accuracy increases to 0.88. Finally, when looking further into classes for which *robbery* is predicted it can be seen that the 3 highest classes are *shoplifting*, *vandalism* and *fighting*. While these three classes, except for *shoplifting*, are all in different branches of the semantic tree it is apparent how this classification came to be. *Fighting* and *vandalism* both are inherently violent and it can be difficult to classify whether someone's intent for the violence is to steal, fight or destroy.

Comparing the semantic model's confusion matrix to the ResNet50 model's confusion matrix in Fig. 5, it can be seen that the ResNet50 model is much more spread out. Whereas the semantic model is more focused and predictions are closer to the actual category, the ResNet50 model is less clear.

Finally, we can look at the accuracy per layer and also the accuracy per node in Fig. 6. From this analysis something interesting can be seen. As can be seen, the second layer of the model clearly outperforms the other two layers. However, the third layer also outperforms the first layer by quite a margin. Looking at the accuracy for each node in the third layer. It can be seen that the node relating to classes such as *explosion* and *arson*, from which arson was clearly categorized more often, does not necessarily have a low accuracy. The reason for this can be attributed to multiple reasons which will be discussed in the discussion section 5.

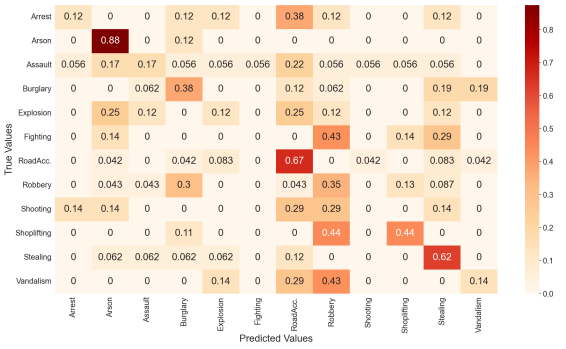

Fig. 4. Confusion matrix for the semantic model with threshold 0.5

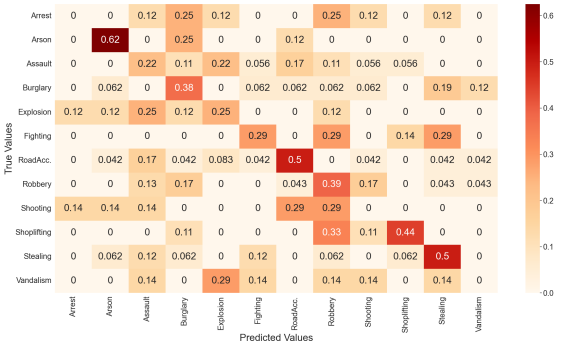

Fig. 5. Confusion matrix for the ResNet50 model with threshold 0.7

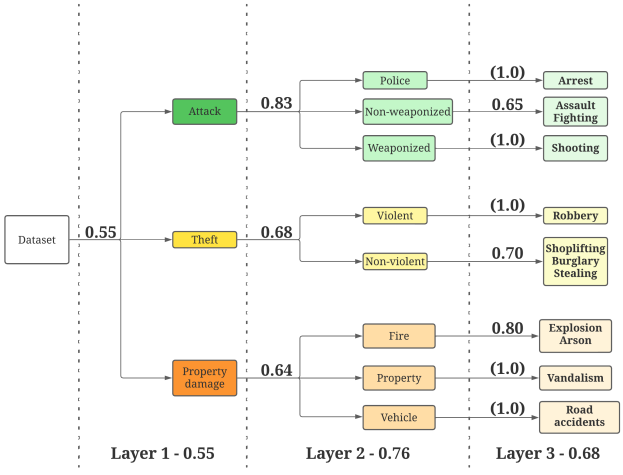

Fig. 6. Accuracy for each layer and node of the semantic tree

5 DISCUSSION

The research aimed to provide a classification approach based on an ontology of crime that improves regular methods of computer vision. Looking at the results it can be seen that the semantic model which was eventually chosen, provides slight benefits over its ResNet50 counterpart in terms of recall, precision and accuracy. Next to this by looking at the confusion matrices it becomes clear that the marginal improvements shown are not the full picture. Comparing the two, results in the conclusion that the semantic model is often much closer to the actual result than the ResNet50 model. Classes that are closely related, such as *shoplifting* and *robbery*, are much more likely

to be mistaken for each other when using the proposed semantic model. Comparing this to the ResNet50 model, it can be seen that results are more spread out and less interpretable.

However, what also becomes clear, is that the performance of both models on the dataset is far from ideal. Both models provide under 40% accuracy when testing them on the UCF crime dataset [21]. The reason for this can be attributed to two factors: Firstly, as previously mentioned, the process of extracting the frames from the video was difficult and required some concessions. From each anomaly, 10 frames were extracted but among these frames were; blurry, distanced or even empty frames. Although this problem was alleviated somewhat by using a threshold, a concept that has proven its worth, this also obscured the training and validation phase of the process. However, this is also part of using real world footage, even if the unclear frames could be filtered and the accuracy would improve this would not provide an honest and clear picture.

Secondly, in the final dataset only 10.150 images were available, spread over 12 different classes. On top of this, each incident had 10 images linked to it, meaning that, in the end, only 1015 scenarios were trained, validated en tested upon. With an average of less than 1000 images over 100 scenarios per class training and testing, a proper classifier results in noticeably worse quality. Furthermore, the data was unevenly divided, leaving a lot of classes with only 50 videos and thus not even 1000 images.

Although the performance was not as high as expected, this does not mean that the results should be discredited. The aim of the research was not to create the perfect classifier but, to see if the approach proposed showed improvement in the results. This improvement has been shown and, looking at other research proposing the same, the improvements shown are as expected [23]. Although the improvements in terms of metrics are not extremely high, the difference in the confusion matrix is clear. The semantic model often predicts a lot closer to the correct class, showing an improvement in classification.

Finally, in the results section it was shown that two classes which were often confused, meaning that they both belonged to the same node but one was clearly favored during the classification process, did not necessarily belong to a node with a low accuracy. Looking at the example of the *arson* and *explosion* category, with an accuracy of 0.88 and 0.12 respectively, the classes were related to a node with an accuracy of 0.80. The reason why arson is clearly favored can be attributed to many things. The most likely option is that pictures belonging to arson performed significantly better on the earlier layers. Meaning that once a picture got to the stage of being classified as 'property damage relating to fire' the final classification, whether it is *explosion* or *arson*, often turned out in the favour of *arson*. This idea is also backed up by the fact that *explosion* is predicted a lot less often than *arson*.

6 CONCLUSION

In this paper, a hierarchical classification approach, based on a semantic tree, for classification of crime-related activities was introduced. The semantic tree was based on literature research and fit to the provided UCF crime dataset [21]. The following contributions are presented:

- A semantic tree based on literature research. The semantic tree breaks down the main defining factors of crime and groups crime accordingly into different semantic levels. Using the tree improves the understanding of what exactly a crime constitutes and can be expanded upon in the field of crime or used as a baseline for other types of images.
- Using the semantic tree, a classification approach has been proposed using the different layers of the semantic tree. Each image is passed through the different layers and using the combined probability of each layer's prediction, the final classification is made. The proposed semantic tree can both be adopted to new classification problems or could be expanded upon in the area of crime.
- Experimentation on the model has shown that using multiple frames per video can be beneficial. Furthermore, using a threshold to filter out frames, on which the model is too uncertain, has also shown to improve performance.

The approach shown in this research is one that can both be expanded on in the area of crime, as well as adapted to new situations. The workflow shown has proven itself to be effective. Both in creating a better understanding of what the physiological features of a certain category entail and in implementing this knowledge into a computer vision model.

7 FUTURE WORK

Looking at the contributions presented in this paper, future work could focus on two different aspects:

Firstly, further testing and development on the proposed semantic tree. As mentioned in the discussion section, the baseline accuracy of the ResNet50 model was not as high as in other research. In this research, only one model was used on a, not yet ideal, dataset. Using the same dataset, different models could be tested such as VGG19 and InceptionV3. Additionally, models could be evaluated using K-fold cross validation, something which was not managed within the time-frame of this research. Next to this, more research could be done into better preparing the UCF crime dataset for future image classification related work. By cleaning the dataset, it could possibly lead to improved results for both models shown. Future research could also look into creating a comparison between the other two methods mentioned in the related work section of this paper 2. Furthermore 5, it was shown that the large difference between closely related classes, such as *explosion* and *arson*, does not necessarily lead to the conclusion that the third layer performs worse. It was discussed that either a restructuring of the semantic tree, where these classes are positioned elsewhere, or an analysis of the quality of the frames belonging to these classes, could lead to better results.

Secondly, besides further testing and developing the proposed methods, the semantic tree model could be adapted to different areas of research. As shown, the model has the ability to improve performance and on top of this is often predicting closer to the actual class than the regular model. By adapting the semantic model to as many different areas, a better idea of what does and does not work could be created. This could lead to improvements to the current semantic tree proposed by researching what exactly is the best way to divide and label categories.

REFERENCES

[1] Danilo Cavaliere, Vincenzo Loia, and Sabrina Senatore. 2019. Towards an Ontology Design Pattern for UAV Video Content Analysis. *IEEE Access* 7 (2019), 105342–105353. https://doi.org/10.1109/ACCESS.2019.2932442

[2] G. Chandan, Ayush Jain, Harsh Jain, and Mohana. 2018. Real Time Object Detection and Tracking Using Deep Learning and OpenCV. In *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*. IEEE. https://doi.org/10.1109/icirca.2018.8597266

[3] John E. Douglas, Ann Wolbert Burgess, Allen G. Burgess, and Robert K. Ressler. 2013. *Crime Classification Manual: A standard system for investigating and classifying violent crime* (3 ed.). Wiley.

[4] Weili Fang, Ling Ma, Peter E.D. Love, Hanbin Luo, Lieyun Ding, and Ao Zhou. 2020. Knowledge graph for identifying hazards on construction sites: Integrating computer vision with ontology. *Automation in Construction* 119 (2020), 103310. https://doi.org/10.1016/j.autcon.2020.103310

[5] M.A. Fischler and R.A. Elschlager. 1973. The Representation and Matching of Pictorial Structures. *IEEE Trans. Comput.* C-22, 1 (Jan. 1973), 67–92. https://doi.org/10.1109/t-c.1973.223602

[6] Kadari Kishore Kumar and Husnabad Venkateswara Reddy. 2022. Crime activities prediction system in video surveillance by an optimized deep learning framework. *Concurrency and Computation: Practice and Experience* 34, 11 (Feb. 2022). https://doi.org/10.1002/cpe.6852

[7] Nancy G La Vigne, Samantha S Lowry, Joshua A Markman, and Allison M Dwyer. 2011. .

[8] Po Kong Lai, Marc Decombas, Kelvin Moutet, and Robert Laganiere. 2016. Video summarization of surveillance cameras. In *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE. https://doi.org/10.1109/avss.2016.7738018

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

[10] Wentao Ma and Shuang Liang. 2020. Human-Object Relation Network For Action Recognition In Still Images. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. https://doi.org/10.1109/icme46284.2020.9102933

[11] John MacIntyre, Ilias Maglogiannis, Lazaros Iliadis, and Elias Pimenidis (Eds.). 2019. *Artificial Intelligence Applications and Innovations*. Springer International Publishing. https://doi.org/10.1007/978-3-030-19823-7

[12] Mrunal Malekar. 2021. Detecting Criminal Activities of Surveillance Videos using Deep Learning. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology* (Feb. 2021), 188–193. https://doi.org/10.32628/cseit217111

[13] Ramna Maqsood, Usama Bajwa, Gulshan Saleem, Rana Raza, and Muhammad Anwar. 2021. Anomaly recognition from surveillance videos using 3D convolution neural network. *Multimedia Tools and Applications* 80 (05 2021). https://doi.org/10.1007/s11042-021-10570-3

[14] Mauricio Marengoni and Denise Stringhini. 2011. High Level Computer Vision Using OpenCV. In *2011 24th SIBGRAPI Conference on Graphics, Patterns, and Images Tutorials*. IEEE. https://doi.org/10.1109/sibgrapi-t.2011.11

[15] Estefania Talavera Martinez, Maria Leyva-Vallina, Md. Mostafa Kamal Sarker, Domenec Puig, Nicolai Petkov, and Petia Radeva. 2020. Hierarchical Approach to Classify Food Scenes in Egocentric Photo-Streams. *IEEE Journal of Biomedical and Health Informatics* 24, 3 (March 2020), 866–877. https://doi.org/10.1109/jbhi.2019.2922390

[16] Clive Norris, Mike McCahill, and David Wood. 2002. The Growth of CCTV: a global perspective on the international diffusion of video surveillance in publicly accessible space. *Surveillance &amp Society* 2, 2/3 (Sept. 2002). https://doi.org/10.24908/ss.v2i2/3.3369

[17] Rebecca Pirius. 2017. Differences between theft and robbery. https://www.nolo.com/legal-encyclopedia/differences-between-theft-robbery.html

[18] Syed Atif Ali Shah, Mahmoud Ahmad Al-Khasawneh, and M. Irfan Uddin. 2021. Review of Weapon Detection Techniques within the Scope of Street-Crimes. In *2021 2nd International Conference on Smart Computing and Electronic Enterprise (ICSCEE)*. IEEE. https://doi.org/10.1109/icscee50312.2021.9498007

[19] Samit Shirsat, Aakash Naik, Darshan Tamse, Jaysingh Yadav, Pratiksha Shetgaonkar, and Shailendra Aswale. 2019. Proposed System for Criminal Detection and Recognition on CCTV Data Using Cloud and Machine Learning. In *2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN)*. IEEE. https://doi.org/10.1109/vitecon.2019.8899441

[20] Kartikeya Singh, Deepak Arora, and Puneet Sharma. 2020. Identification of Shoplifting Theft Activity Through Contour Displacement Using OpenCV. In *Computational Methods and Data Engineering*. Springer Singapore, 441–450. https://doi.org/10.1007/978-981-15-6876-3_34

[21] Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-world Anomaly Detection in Surveillance Videos. https://doi.org/10.48550/ARXIV.1801.04264

[22] Richard Szeliski. 2011. *Computer Vision*. Springer London. https://doi.org/10.1007/978-1-84882-935-0

[23] Estefanía Talavera, Maria Leyva-Vallina, Md. Mostafa Kamal Sarker, Domenec Puig, Nicolai Petkov, and Petia Radeva. 2019. Hierarchical approach to classify food scenes in egocentric photo-streams. *CoRR* abs/1905.04097 (2019). arXiv:1905.04097 http://arxiv.org/abs/1905.04097

[24] P. Viola and M. Jones. [n.d.]. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*. IEEE Comput. Soc. https://doi.org/10.1109/cvpr.2001.990517

[25] Pengfei Wang, Yu Fan, Shuzi Niu, Ze Yang, Yongfeng Zhang, and Jiafeng Guo. 2019. Hierarchical Matching Network for Crime Classification. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM. https://doi.org/10.1145/3331184.3331223

[26] Paul Wilson and Helene Wells. 2007. What do the watchers watch? an Australian case study of CCTV monitoring. *Humanities Social Sciences Papers* (Jan. 2007).