Analyzing descriptive captions for crime recognition in surveillance footage

VINCENT GUSTAV ALBERTSSON, University of Twente, The Netherlands

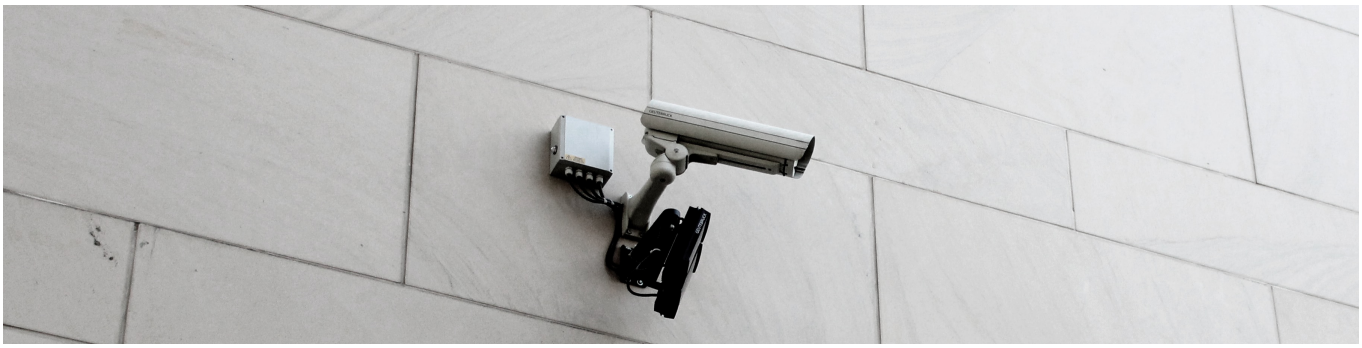

Fig. 1. A surveillance camera in Bonn, Germany, 2018.

Crime recognition in surveillance footage is a complicated and challenging task that can benefit from computer vision and machine learning techniques. With an immense amount of footage being recorded by security cameras all the time, automated video analysis techniques can help balance the workload. This paper proposes a novel approach that extracts descriptive captions from surveillance footage for crime recognition to support crime investigation and prevention. A pipeline is adapted to extract captions, process them, and train a model for crime recognition with the HR-Crime dataset. The aim is to evaluate the effectiveness of captions for crime recognition and compare the results with state-of-the-art methods. The findings contribute to the body of research on crime detection by highlighting the potential of this approach, and providing insights into its limitations and further work that can be done.

Additional Key Words and Phrases: Crime recognition, surveillance footage, captioning, forensics, neural networks

1 INTRODUCTION

Computer vision is an interdisciplinary field that integrates machine learning and artificial intelligence techniques to analyze and interpret visual information [12]. This typically involves processing visual inputs, such as images or videos, and extracting information from them to classify objects, situations, or activities, thereby yielding meaningful insights.

In the context of crime recognition, computer vision holds significant potential [1]. Computer vision techniques can be used to analyze the vast amounts of footage captured by surveillance cameras daily, a task that would be unfeasible using manual methods. Furthermore, it opens the door to proactive crime detection and prevention [14].

Traditional surveillance cameras passively record footage that can be used as evidence in retrospect but do little to prevent the actual occurrence of crimes. Live monitoring of camera feeds by security operators incurs high costs, leading to the integration of motion detection into surveillance systems. These systems provide cameras with the ability to send alerts or start recording only when motion is detected. While this approach is beneficial in static scenes where movement is rare, it does not suffice in busy scenes like streets, stores, airports, and other locations with consistent movement.

Building on this concept, anomaly detection systems can still function in busy environments because they don't rely on motion detection. These systems create a model of the baseline, or 'normal' activity using typical video footage of an area. They can then classify situations that deviate significantly from the baseline as anomalous, such as a person loitering or sudden crowds in otherwise quiet areas. However, anomaly detection suffers from a lack of granularity as classification is binary. Situations are either normal or anomalous, with no context or specificity as to what is happening. Anomaly detection systems are also usually confined to single-scene classification because the model is trained on a specific location, and footage from another location will always be anomalous in the context of what it was trained on [13].

More advanced techniques utilize optical flow to extract useful features such as the movement or trajectory of objects [5]. Recent research extracts poses of subjects in surveillance footage to estimate skeleton trajectories, classifying specific criminal activities with promising results [8]. However, these techniques carry comparatively high computational costs, and contextual information along with potential objects, such as weapons, are often overlooked.

This paper proposes a novel approach to crime recognition using narrative flow, achieved by extracting captions from surveillance videos. This approach not only utilizes contextual information from the extracted captions, but could also leverage temporal clues, as multiple captions are extracted from a single video through subsampling. With further enhancements, this method could compete with state-of-the-art techniques while reducing computational load.

TScIT 39, July 7, 2023, Enschede, The Netherlands

1.1 Related Work

In recent years, there has been a substantial improvement in the field of crime recognition due to the proliferation of artificial intelligence and machine learning techniques [9] [11].

A recent paper utilized descriptive captions extracted from images as a feature for emotion recognition [4]. These captions were processed through a pipeline and used to train an emotion classifying model . This paper adapts the pipeline of the emotion recognition paper to the task of crime recognition. This new pipeline was used to train a model with the graph isomorphism network (GIN) architecture [16] and videos from the HR-Crime dataset for the purpose of crime recognition.

There are a limited number of available annotated datasets for surveillance footage, especially in the case of those containing abnormal activity. With this consideration and its past use in crime recognition work, this paper uses the HR-Crime dataset. The HR-Crime dataset is a curated subset of UCF-Crime [15], consisting of human-related surveillance footage belonging to one of thirteen anomalous categories or normal activity [2]. These categories, which include but are not limited to assault, burglary, and robbery, feature full-length videos from a variety of settings to enhance the dataset's versatility for multi-scene anomaly detection.

The HR-Crime dataset has been used for crime recognition recently by exploring the utility of skeleton trajectories [8]. By tracking the poses of human subjects and their movement over time it was possible to extract skeleton trajectories and thereby use the motion associated with anomaly categories to train the model. An accuracy of 0.49 was achieved with this approach, showing the potential and versatility of the HR-Crime dataset for research in crime recognition.

Image caption generation has seen massive leaps in recent years, with more and more techniques, architectures, and datasets being available [11]. However, despite these advancements, there remains a substantial room for improvement before image captioning techniques can truly rival human performance. This presents a challenge for the proposed approach, as it relies on the accuracy and quality of the extracted captions.

1.2 Research Questions

To address the research gap and assess the proposed approach, this paper will answer the following main research question (RQ).

- How effective are captions for crime recognition compared to other state-of-the-art approaches?

The following sub-RQs (SRQ) will be answered to support the main RQ and address possible further improvements.

- SRQ1: How does the imbalanced nature of the dataset affect accuracy?
- SRQ2: How can multiple frames be used for classification and how does it affect accuracy?

2 METHODOLOGY

This research paper first answers the two SRQs to answer the main RQ. The proposed approach includes data selection, feature extraction, and evaluation.

2.1 Dataset

The HR-Crime dataset was used for this research paper. The dataset includes anomalous and normal footage, with the anomalous videos being categorized into 13 classes which are: Abuse, Arrest, Arson, Assault, Burglary, Explosion, Fighting, Road Accidents, Robbery, Shoting, Shoplifting, Stealing, and Vandalism. This study focuses on the recognition of anomalous activities, hence the normal footage was omitted for the purposes of this research. This exclusion halves the available data and makes the resulting model unable to classify normal footage accurately.

Due to the size of the dataset and the limited processing power and time available, subsampling was used. Rather than using all the frames of a video for feature extraction, 10 equidistant frames were selected from each video. These frames were then passed through the caption extraction pipeline and used for training and testing. The dataset was divided into a training and a testing set, with an 80:20 split respectively.

The following chart highlights the number of videos for each category to give an overview of the size of the dataset and its distribution. From Figure 2 the dataset is seemingly imbalanced with some classes consisting of 150 videos while others consist of only 50 videos. The table in *Appendix A* shows the dataset composition in more detail.

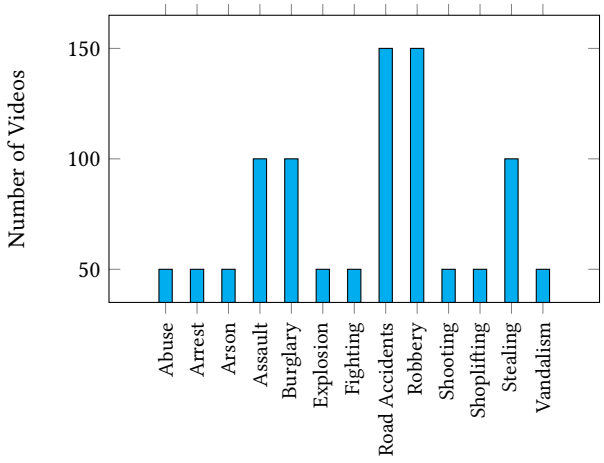

Fig. 2. Distribution of Videos across Anomalous Classes in the HR-Crime Dataset

2.1.1 Dataset Balancing. As mentioned, 10 frames are sampled equidistantly from each video. For instance, with 10 videos, the dataset would encompass 100 frames. It is important to note that the number of videos in each class varies, which may cause bias toward certain classes. To address this, a combination of over- and under-sampling was used. Classes with 50 videos were duplicated to have 100 videos, classes with 100 videos were left unchanged, and classes with 150 videos had random videos removed until they had 100 videos. Therefore, the modified dataset comprises 1300 videos, with 100 videos in each of the 13 classes.

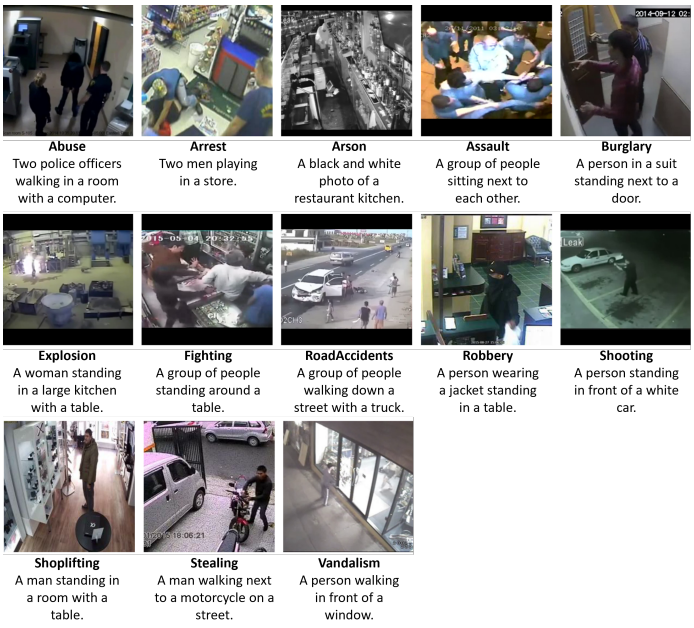

Fig. 3. Example Frames from each Class with Ground-Truth Labels and Generated Captions.

2.2 Captioning

The ExpansionNet v2 architecture is used to train image captioning models [7]. For this study, the pretrained model included with the ExpansionNet v2 paper was employed. This model was used to extract descriptive captions from the sampled frames in the HR-Crime dataset. An overview of the captions generated for each class with this method can be seen in *Figure 3*.

2.3 Pipeline

The pipeline used for this study is adapted from the paper on emotion recognition [4]. A detailed figure of the pipeline can be seen in *Appendix B*.

(1) Subsampling: 10 equidistant frames are sampled from each video.
(2) Caption Extraction: For each sampled frame, a caption is extracted and stored alongside the ground-truth label associated with the video.
(3) Caption Processing: Each caption is lemmatized and processed to transition from its raw form to one devoid of redundant stop words, as per the method outlined in the emotion recognition paper [4].
(4) Graph Generation: The co-occurrence matrices, in conjunction with GloVe embeddings [10], are used to generate a graph representation of the information extracted and processed in the previous steps.
(5) Model Training and Testing: The model is trained with the generated graphs using the GIN architecture [16].
(6) Evaluation: The model is evaluated using various metrics to analyze its performance in crime recognition.

The pipeline largely parallels the original [4], with the exception of the subsampling step since videos, rather than images, are used as the initial dataset. While SenticNet was utilized in the original emotion recognition paper during the graph generation step, it was omitted in this research due to its primary focus on emotion and polarity recognition from textual information which differs from the use case at hand.

2.4 Evaluation

The performance of the proposed approach was assessed using common classification metrics, which include:

- Accuracy, precision, recall, F1-score
- Confusion matrix with heatmap
- ROC curves for each class
- Average ROC AUC score

In addition, the metrics were computed based on the classification results of single frames as well as by grouping them by video as an aggregate. With the aggregate classification approach, frames from a single video were grouped together, and their separate classifications were averaged. This method's aim was to assess whether the temporal sequence of events could be considered to enhance the model's classification performance.

3 RESULTS

The proposed approach was evaluated using the metrics outlined in the evaluation section. This was done under three conditions: using the dataset before balancing with single frame classification, using the dataset after balancing with single frame classification, and finally, using the dataset after balancing with multi-frame classification.

3.1 Dataset Before Balancing

The initial training of the model was conducted with an imbalanced dataset that included 1000 videos. These videos were unevenly distributed across the classes, with some classes represented by as few as 50 videos as seen in Figure 2, while others were represented by as many as 150 videos. The model trained with the imbalanced dataset achieved an accuracy of 0.09, precision of 0.29, recall of 0.10, and an F1-score of 0.06.

The confusion matrix reveals a tendency of the model to frequently classify various frames as belonging to the Arrest class, regardless of their true class. The underlying cause of this bias for the Arrest class is not immediately evident and could be tied to various limitations inherent in the adopted approach, which are discussed later in the Discussion section.

While the heatmap did provide some insight into the overall performance of the model, the disparity across the classes became more evident upon analyzing the Receiver Operating Characteristic (ROC) curves.

The ROC curves were computed for each class and plotted to yield the area under curve (AUC) which is useful for gauging the performance of classifiers. The AUC ranges from 0.5 (equivalent to a random guess) to 1 (perfect classification). In this case, the classes Shooting and Vandalism exhibited AUCs close to 0.5, indicating that the model's ability to classify these categories comes close

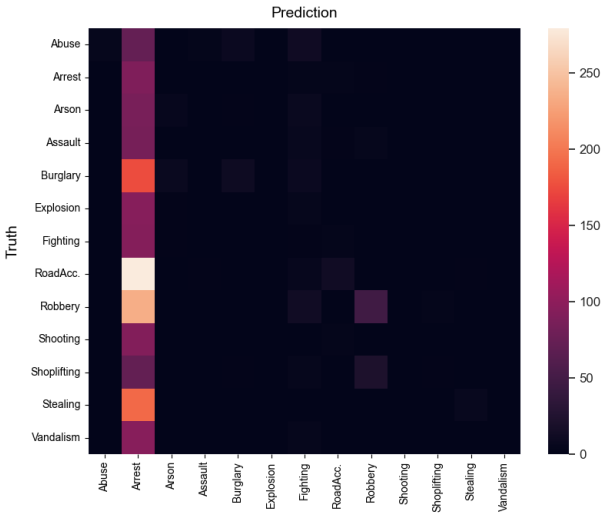

Fig. 4. Confusion Matrix with Dataset before Balancing

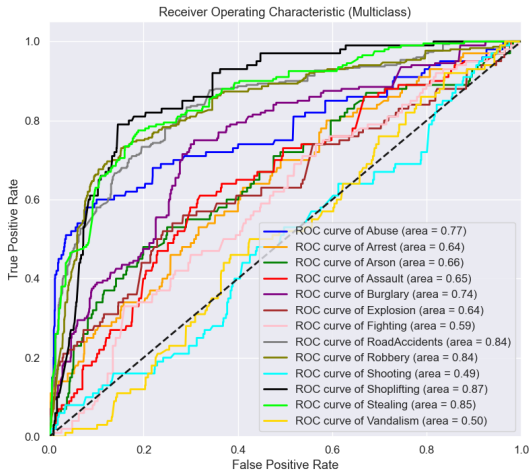

Fig. 5. ROC Curves for each Class with Dataset before Balancing

to randomly guesseing. However, more promising results were obtained for Shoplifting, Stealing, Robbery, and Road Accidents, which demonstrated AUCs of 0.87, 0.85, 0.84, and 0.84, respectively. These results suggest that the model was able to reasonably distinguish these categories from others.

It is noteworthy, though, that the classes Robbery and Road Accidents had 150 videos each which is significantly more than other classes. This could have made the model biased toward better recognition of these categories, as the increased availability of samples might have enhanced the learning capability of the model for these classes.

The average ROC AUC across all classes was 0.70, which indicates that the model is able to perform better than a guess, but significant improvements can still be made in the model's performance.

3.2 Dataset After Balancing

To address the imbalance in the dataset, a combination of over- and undersampling was used so that all classes have the same number of videos. The model trained with the balanced dataset achieved an accuracy of 0.17, precision of 0.31, recall of 0.17, and an F1-score of 0.13. These results show a substantial improvement over the imbalanced dataset and demonstrate that the dataset imbalance has a significant effect on the classification performance for this task.

The confusion matrix reveals that the model is no longer biased toward Arrest, but is now biased toward Assault and, to a lesser extent, toward Vandalism. The bias is less substantial, but further work is required to address this issue and improve the model's performance.

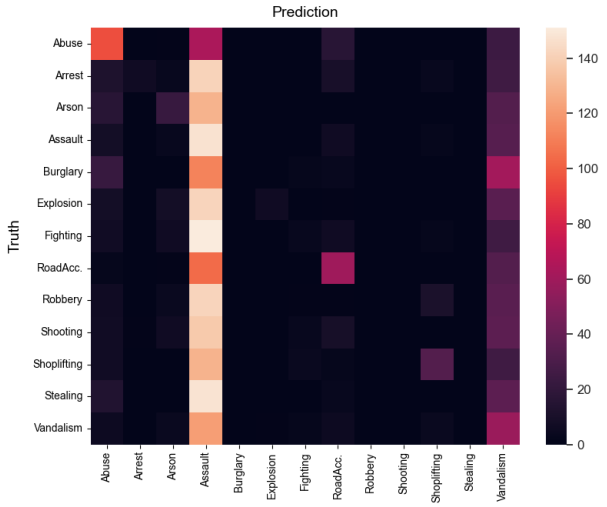

Fig. 6. Confusion Matrix with Dataset after Balancing

The average ROC AUC score increased to 0.73, a modest improvement over the results with the imbalanced dataset. More encouragingly, all of the classes perform better than a random guess. The classes with the best AUC were Shoplifting, Abuse, RoadAccidents, and Arson, with values of 0.92, 0.83, 0.81, and 0.80 respectively. The Shoplifting class stands out for its performance with the highest AUC among all classes, showing how the model was capable of classifying that category very accurately.

In Appendix C, examples of images from the Shoplifting class with their corresponding extracted captions are shown. It is noticeable that all of the images in this class depict indoor settings, which aligns with the nature of the crime. The captions frequently include words such as 'Store' or 'Room'. This consistency in the setting and the presence of distinctive keywords may have contributed to the high performance of the classification for this particular class.

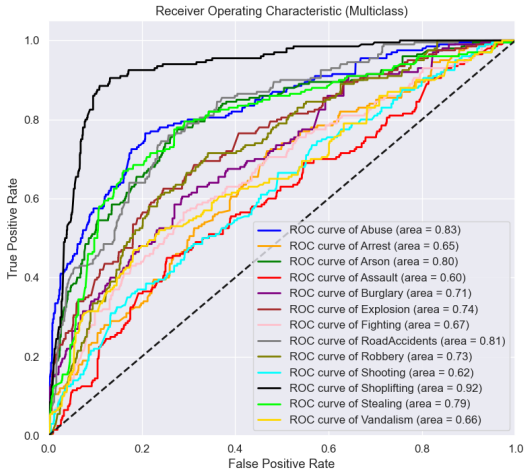

Fig. 7. ROC Curves for each Class with Dataset after Balancing

Appendix D showcases examples of images from the worst performing class, Assault, along with their extracted captions. In contrast to the Shoplifting class, videos and captions in the Assault class exhibit more variability. They encompass scenes from both indoor and outdoor settings. The diverse scenes within this class, coupled with captions that may not be distinct from those in other classes, pose a challenge for the model. This variation and potential similarity in captions across classes could lead to a bias towards certain classes during classification.

It is important to note that this approach considers the general location, number of people, and main subjects based on the extracted captions. However, the low quality of the videos, combined with the loss of information through training the model solely on the extracted captions, makes it challenging for the model to distinguish between different classes accurately.

3.3 Multi-frame Classification

Classification was conducted using multiple frames from the same video to determine if the added information would improve performance and whether or not the model was able to utilize temporal information.

The model achieved an accuracy of 0.17, precision of 0.22, recall of 0.17, and an F1-score of 0.13. Performing classification based on the average of 10 frames, as opposed to a single frame, does not seem to significantly change the model's performance, with the exception of a slightly lower precision. This suggests that the current approach does not effectively utilize the information from multiple frames nor does it utilize temporal information. It is worth noting that the original pipeline was designed for separate images and was not altered to use multiple frames during training or classification.

3.4 Comparison with State of the Art

The best accuracy achieved with the approach in this paper was 0.17, while the approach using skeleton trajectories was able to achieve an accuracy of 0.49 [2]. This indicates that the current approach is not yet competitive with the state of the art. However, it shows promise and could yield improved results if the identified shortcomings are addressed.

4 DISCUSSION

The results of this paper illustrate the utility of caption extraction for crime recognition in surveillance videos. Various experiments were conducted to test the proposed approach and understand its limitations.

4.1 Temporal Information

The proposed approach uses multiple frames from a single video, but treats each frame as an isolated instance rather than part of a sequence of events. Table 1 shows the comparison of results between image and video based classification. The performance metrics were unchanged except for precision which went down using the video based model. These results show that temporal information was not utilized in the model.

	Accuracy	Precision	Recall	F1-score
Image Based (Unbalanced)	0.09	0.29	0.10	0.06
Image Based (Balanced)	0.17	0.31	0.17	0.13
Video Based (Balanced)	0.17	0.22	0.17	0.13

Table 1. Results between Image Based and Video Based Classification

Crimes often have a temporal aspect, so the sequence of events plays a big role in the type of crime that can be detected. Without being able to use this information, it is not surprising that the model is unable to achieve a satisfactory performance, especially considering the nature of the captions used for training.

Upon inspecting the most common words found in extracted captions for each class, the issue of similarity becomes evident.

Abuse	Stand (236)	Room (207)	Person (207)	Group (156)	Street (147)
Arrest	Car (279)	Street (262)	Stand (209)	Group (196)	Park (114)
Arson	White (316)	Black (278)	Photo (274)	Street (208)	Person (192)
Assault	Street (278)	Group (246)	Walk (214)	Black (189)	Stand (164)
Burglary	Stand (205)	Table (203)	White (202)	Black (190)	Room (186)
Explosion	Street (250)	Group (198)	Car (195)	White (186)	Black (177)
Fighting	Group (359)	Street (226)	Stand (161)	Black (155)	White (155)
RoadAccidents	Street (565)	Car (442)	City (224)	Park (201)	Group (180)
Robbery	Store (346)	Stand (267)	Person (202)	Table (132)	Group (129)
Shooting	White (257)	Street (227)	Car (214)	Black (201)	Photo (159)
Shoplifting	Store (505)	Stand (243)	Table (221)	Group (195)	Sit (168)
Stealing	Car (502)	Park (363)	Street (289)	White (260)	Black (216)

Table 2. Top 5 Words in Extracted Captions per Class (with Frequency)

There is a considerable overlap in the captions between different classes and their most frequent words. Most classes had the same words in their first or second most frequently occurring words. The Abuse class had the word 'Room' as one of its most frequent words which does not occur as much in other classes, possibly contributing to its improved performance over other classes. This issue is also noticeable in Figure 3 and Appendix E, where the captions can be

seen to be similar between categories and almost identical between frames of a single video. This similarity between captions could make it difficult for the model to accurately recognize crimes without incorporating temporal cues.

Leveraging this information could be the key to improving the performance of the model. Adapting the pipeline to use a recurrent neural network (RNN) or long short-term memory (LSTM) during the training phase could allow it to use temporal information from multiple frames of a single video.

4.2 Subsampling

The use of subsampling in this study involved extracting 10 frames from each video in the dataset at regular intervals. A large amount of information was lost, with lots of it being potentially critical to capturing the context in each situation.

The subsampling method applied in this study would result in 10 frames, or less than 1% of the frames in the video being used to represent the whole. As the dataset is comprised of real surveillance footage, most of the videos are a minute or longer. This subsampling approach was used due to time and hardware limitations, but could be expanded to use more frames in the future.

The main issue of subsampling is that anomalous events are sparse in the videos, meaning that omitting a few frames could result in missing the critical moments in the footage.

4.3 Bias in Classification

The confusion matrix, even after balancing the dataset, reveals an inclination toward certain classes with many incorrect predictions. This indicates a potential limitation in the model's ability to distinguish between classes.

This bias could be due to the way the dataset was balanced, with the use of Synthetic Minority Oversampling Technique (SMOTE) [3] [6] being a possible improvement rather than the random over- and under-sampling used. The difference between SMOTE and random over-sampling, is that SMOTE adds new synthesized data points from existing data in the minority classes rather than just duplicating existing data. These techniques can help to overcome bias so that the model has better performance on minority classes. The bias could also be due to the captions themselves, with certain classes being more easily identifiable or others having captions that are more indistinguishable. It could also be a result of the dataset, where most of the videos contain lots of normal activity dotted with short periods of anomalous activity. The image quality is also poor, with low resolution and blurry footage in some cases that makes it harder for the algorithm to recognize what is happening.

5 CONCLUSIONS

The findings of this study answer the primary RQ: "How effective are captions for crime recognition compared to other state-of-the-art approaches?" revealing that the current approach is not competitive yet, but does demonstrate potential if more research is done and limitations are addressed.

The SRQ1 "How does the imbalanced nature of the dataset affect accuracy?" and SRQ2 "How can multiple frames be used for classification and how does it affect accuracy?" revealed important points. The effect of the imbalanced dataset was shown to be substantial on the classification performance of the model. In terms of using multiple frames, it was clear that the current pipeline does not utilize temporal information and the performance remained largely unchanged whether single or multiple frames were used. It is clear that there are several limitations such as the inability of the model to handle temporal sequences, the loss of information with subsampling, and a potential bias in the model toward certain classes.

In conclusion, this paper contributes to the body of research by exploring the use of captions for crime recognition in surveillance footage. It adapts a pipeline used successfully for emotion recognition and highlights limitations in applying this method to the task of crime recognition. Although the current approach does not compete with state-of-the-art methods, it contributes to the collective effort in aiding policing and stopping crime.

5.1 Future Work

Researchers conducting similar work should aim to address the aforementioned limitations. Specifically, incorporating temporal information, altering the subsampling technique to capture significant frames or extracting more frames in total, and refining dataset balancing techniques to address bias. Furthermore, improvements could be made to the caption extraction to ensure that the captions are more representative of what is happening in the frame.

REFERENCES

[1] Mahmoud Al-Faris, John Chiverton, David Ndzi, and Ahmed Isam Ahmed. 2020. A Review on Computer Vision-Based Methods for Human Action Recognition. *Journal of Imaging* 6 (6 2020). Issue 6. https://doi.org/10.3390/JIMAGING6060046

[2] Kayleigh Boekhoudt, Alina Matei, Maya Aghaei, and Estefanía Talavera. 2021. HR-Crime: Human-Related Anomaly Detection in Surveillance Videos. (2021). https://doi.org/10.34894/IRRDJE

[3] Prima Bouchaib and Mohammed Bouhorma. 2021. TRANSFER LEARNING AND SMOTE ALGORITHM FOR IMAGE-BASED MALWARE CLASSIFICATION. *Proceedings of the 4th International Conference on Networking, Information Systems Security* (4 2021). https://doi.org/10.1145/3454127.3457631

[4] Willams de Lima Costa, Estefania Talavera Martínez, Lucas Silva Figueiredo, and Veronica Teichrieb. 2023. High-level context representation for emotion recognition in images. (2023).

[5] Prasad D. Garje, M. S. Nagmode, and Kiran C. Davakhar. 2018. Optical Flow Based Violence Detection in Video Surveillance. *2018 International Conference On Advances in Communication and Computing Technology, ICACCT 2018* (11 2018), 208–212. https://doi.org/10.1109/ICACCT.2018.8529501

[6] Mamta Ghalan and Rajesh Kumar Aggarwal. 2022. Cross-Validated Ensemble ARDenseNet for Human Activity Recognition. *2022 IEEE Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)* (2022). https://doi.org/10.1109/IATMSI56455.2022.10119280

[7] Jia Cheng Hu, Roberto Cavicchioli, and Alessandro Capotondi. 2022. ExpansionNet v2: Block Static Expansion in fast end to end training for Image Captioning. (8 2022). https://arxiv.org/abs/2208.06551v3

[8] Alina-Daniela Matei, Estefanía Talavera, and Maya Aghaei. 2022. Crime scene classification from skeletal trajectory analysis in surveillance settings. (2022).

[9] Azhee Wria Muhamada and Aree A. Mohammed. 2022. Review on recent Computer Vision Methods for Human Action Recognition. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal* 10 (2 2022), 361–379. Issue 4. https://doi.org/10.14201/ADCAIJ2021104361379

[10] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* (2014), 1532–1543. https://doi.org/10.3115/V1/D14-1162

[11] Vikas Pogadadanda, Shafeullah Shaik, Gogula Venkata Sai Neeraj, Hima Varshini Siralam, Iwin Thanakumar Joseph S, and K. B.V.Brahma Rao. 2023. Abnormal Activity Recognition on Surveillance: A Review. *2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)* (2023), 1072–1077. https://doi.org/10.1109/ICAIS56108.2023.10073703

[12] Imane Rahil, Walid Bouarifi, and Mustapha Oujaoura. 2022. A Review of Computer Vision Techniques for Video Violence Detection and intelligent video surveillance systems. *International Journal of Advanced Trends in Computer Science and Engineering* 11 (11 2022), 62–70. Issue 2. https://doi.org/10.30534/IJATCSE/2022/051122022

[13] Bharathkumar Ramachandra, Michael J Jones, and Ranga Raju Vatsavai. 2020. A Survey of Single-Scene Video Anomaly Detection. (2020).

[14] Neil Shah, Nandish Bhagat, and Manan Shah. 2021. Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention. *Visual Computing for Industry, Biomedicine, and Art* 4 (12 2021), 1–14. Issue 1. https://doi.org/10.1186/S42492-021-00075-Z/FIGURES/3

[15] Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-world Anomaly Detection in Surveillance Videos. (2018). http://crcv.ucf.edu/projects/real-world/

[16] Keyulu Xu, Stefanie Jegelka, Weihua Hu, and Jure Leskovec. 2018. How Powerful are Graph Neural Networks? *7th International Conference on Learning Representations, ICLR 2019* (10 2018). https://arxiv.org/abs/1810.00826v3

A APPENDIX - DATASET DISTRIBUTION TABLE

	Abuse	Arrest	Arson	Assault	Burglary	Explosion	Fighting	Road Accidents	Robbery	Shooting	Shoplifting	Stealing	Vandalism	**Total train**	**Total test**	**Total**
Videos	50	50	50	100	100	50	50	150	150	50	50	100	50	800	200	1000
Extracted frames	500	500	500	1000	1000	500	500	1500	1500	500	500	1000	500	8000	2000	10000

Table 3. Distribution of Videos and Extracted Frames across Anomalous Classes in the HR-Crime Dataset for Training and Testing, before balancing.

B APPENDIX - PIPELINE OVERVIEW

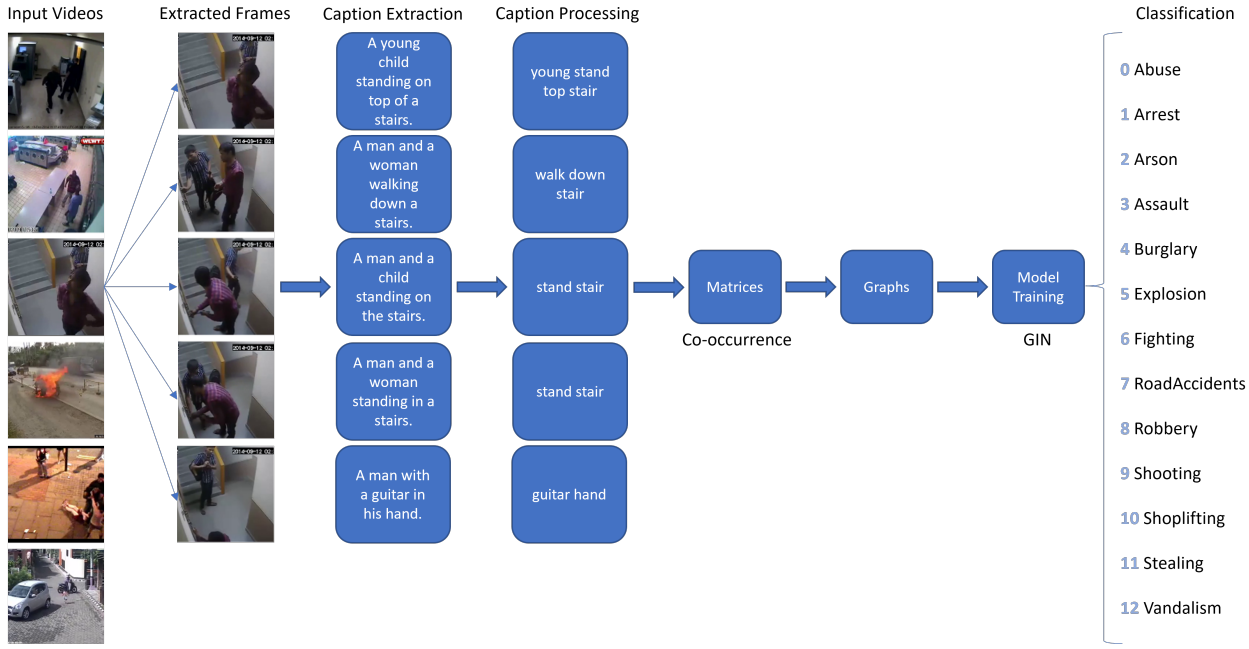

Fig. 8. Detailed Overview of the Steps Involved in the Image Extraction and Training Pipeline.

C APPENDIX - EXAMPLES FROM BEST PERFORMING CLASS

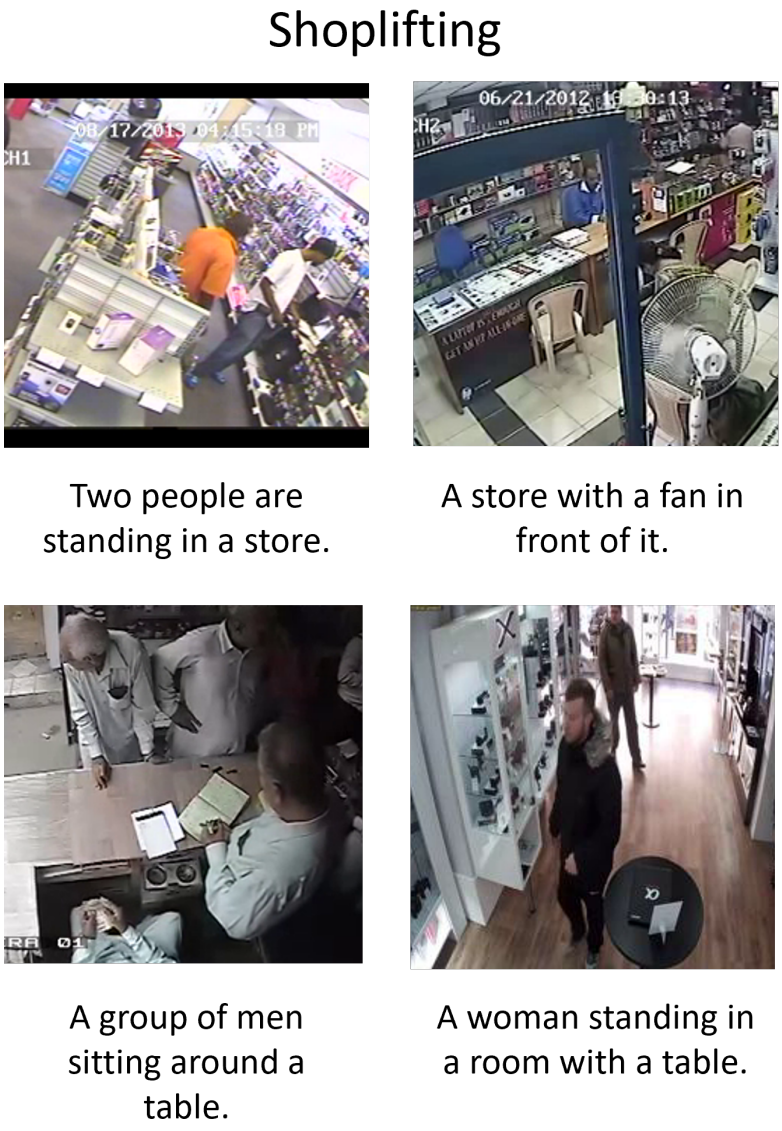

Fig. 9. Frames and Extracted Captions from the Shoplifting Class

D APPENDIX - EXAMPLES FROM WORST PERFORMING CLASS

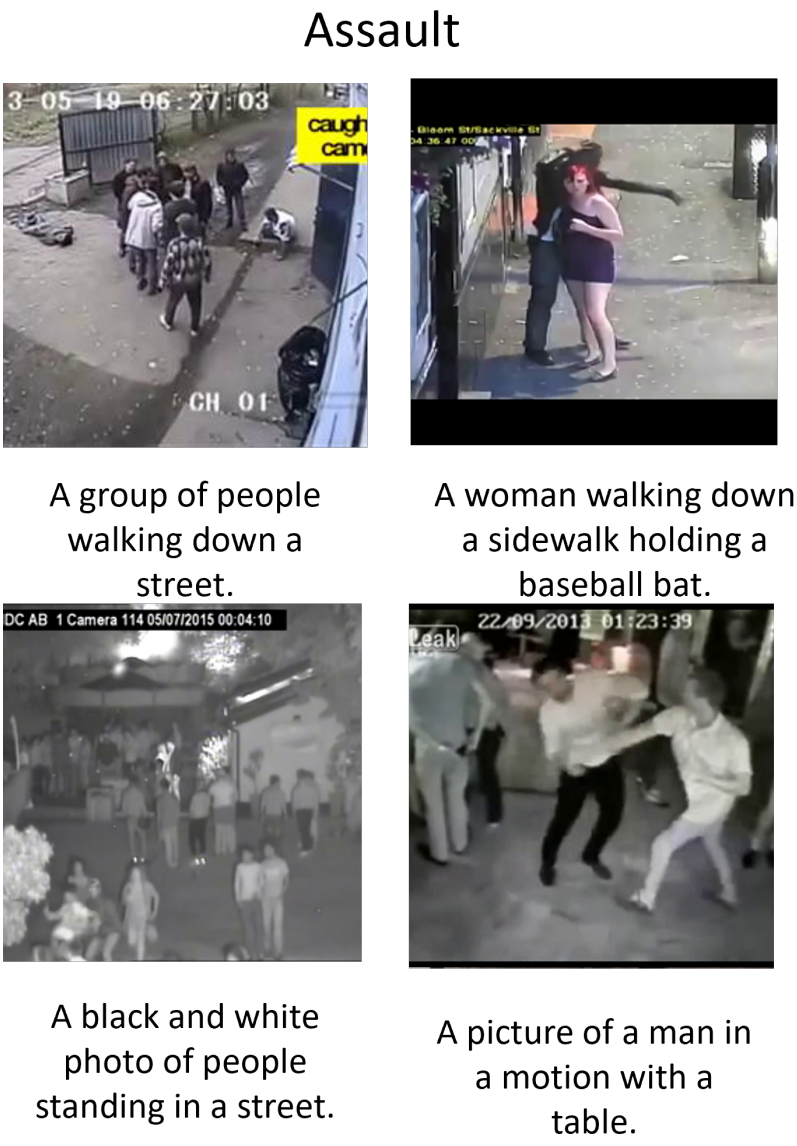

Fig. 10. Frames and Extracted Captions from the Assault Class

E APPENDIX - EXAMPLES OF CAPTIONS FROM A SINGLE VIDEO

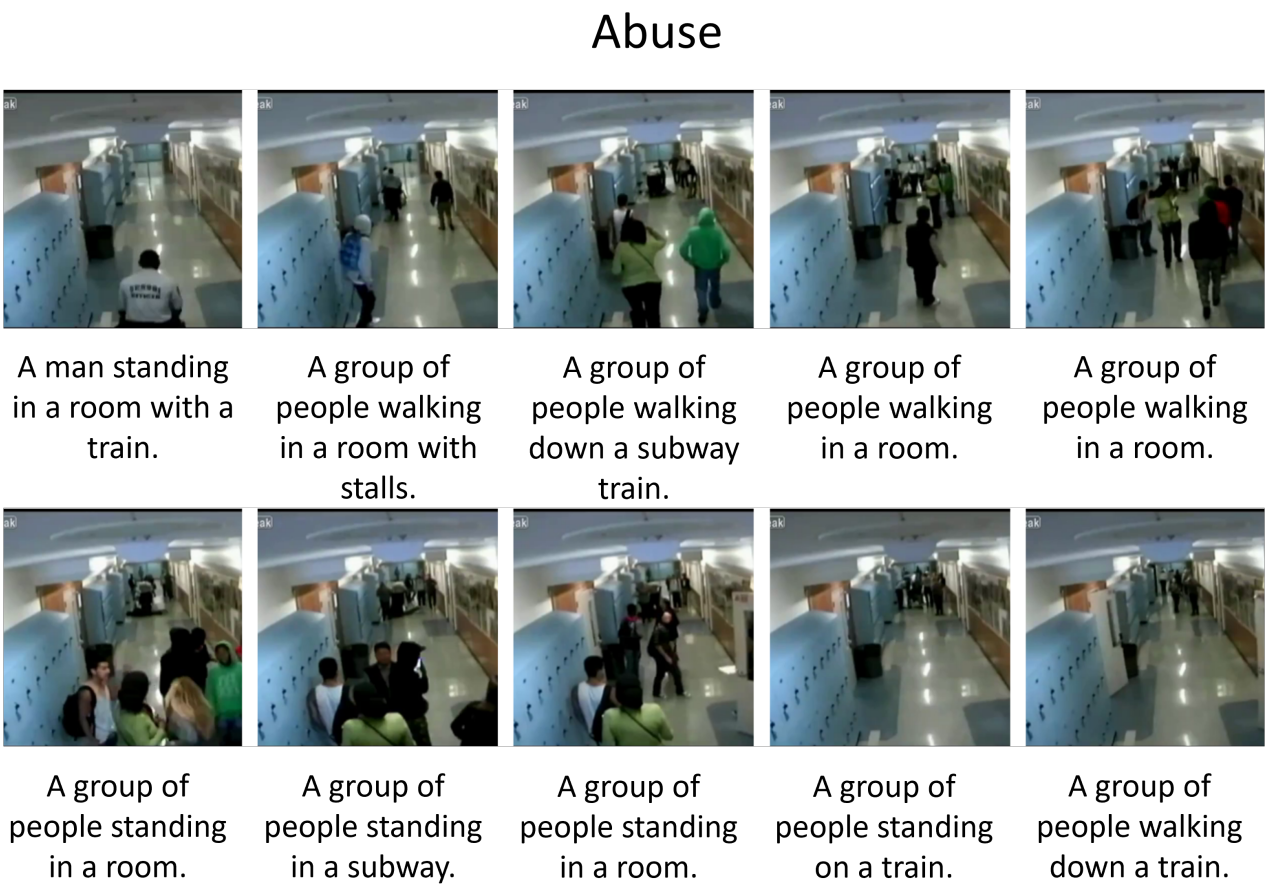

Fig. 11. Frames and Extracted Captions from a Single Video in the Abuse Class