Violence Detection: A Serious-Gaming Approach

Derkjan Elzinga[1], Stan Ruessink[1], Giuseppe Cascavilla[2], Damian Tamburri[2], Francesco Leotta[3],
Massimo Mecella[3] and Willem-Jan Van Den Heuvel[1]

[1]*JADS - Tilburg University, The Netherlands*
[2]*JADS - Technical University Eindhoven, The Netherlands*
[3]*Sapienza - University of Rome, Italy*

Keywords: AI, Convolutional Neural Network, Anomaly Behavior, Video Games, Cyber-Physical Space Protection.

Abstract: Widespread use of IoT, like surveillance cameras, raises privacy concerns in citizens' lives. However, limited studies explore AI-based automatic recognition of criminal incidents due to a lack of real data, constrained by legal and privacy regulations, preventing effective training and testing of deep learning models. To address dataset limitations, we propose using generative technology and virtual gaming data, such as the Grand Theft Auto (GTA-V) platform. However, it's unclear if synthetic data accurately mirrors real-world videos for effective deep learning model performance. This research aims to explore the potential of identifying criminal scenarios using deep learning models based on gaming data. We propose a deep-learning violence detection framework using virtual gaming data. The 3-stage deep learning model focuses on person identification and violence activity recognition. We introduce a new dataset for supervised training and find virtual persons closely resembling real-world individuals. Our research demonstrates a 15% higher accuracy in identifying violent scenarios compared to three established real-world datasets, showcasing the effectiveness of a serious gaming approach.

1 INTRODUCTION

A large part of our life takes place in public spaces. We define public spaces as circumscribed physical areas located either in the open air or municipal institutions and government buildings belonging to civic spaces in a community (Persson et al., 2002). To guarantee safety in such spaces, it is essential to detect threatening incidents proactively (Rest et al., 2014; Cascavilla et al., 2021). The number of surveillance cameras is rapidly increasing to improve security in public spaces (Flight et al., 2022). Despite the increase in the number of surveillance cameras, their effectiveness remains questionable What is more, manual surveillance is tedious, expensive, and time-consuming for Law-enforcement agencies. Therefore, a need emerges for smarter intelligence systems - based on artificial intelligence for the active detection of crime via surveillance cameras - to accompany classical approaches that require a human to monitor the multiple video screens and to identify criminal activity (Bouma et al., 2014).

At the same time, due to the different possible person movements, high dimensions of video data, different motion speeds, and different color videos, precise criminal incident recognition is still a challenging task. In the past decade, several machine learning-based methods have been developed for identifying criminal incidents. However, these handcrafted-based approaches are not usable in practice since performances reduce with different camera positions and monitoring of large groups of people. In addition, the approaches have a large computational time and require real data from real-world scenarios, with Data augmentation typically used to cope with limited available data (Alex et al., 2012).

Next to these technical concerns, privacy is also an important factor hampering training these models. The datasets required are often restricted and unavailable (Paulin and Ivasic-Kos, 2023).

This paper offers another solution: to generate data through the use of serious-gaming. A deep learning network can be trained based on virtual gaming data. The use of virtual gaming data is altogether a new approach. Therefore, limited data and academic literature are available on this topic. One of

Elzinga, D., Ruessink, S., Cascavilla, G., Tamburri, D., Leotta, F., Mecella, M. and Van Den Heuvel, W.
Violence Detection: A Serious-Gaming Approach.
DOI: 10.5220/0012762300003767

In *Proceedings of the 21st International Conference on Security and Cryptography (SECRYPT 2024)*, pages 163-174
ISBN: 978-989-758-709-2; ISSN: 2184-7711

the main concerns of the training of virtual gaming data is whether virtual data is similar enough to the real world to make an effective application possible.

Our work proposes a framework to train deep learning networks for violent behavior detection (Cascavilla et al., 2023) based on virtual gaming data. This work's contribution is twofold: (1) an extension to the existing academic literature for the detection of criminal incidents using a deep learning network and (2) using virtual reality to train deep learning networks that can ultimately be used in real situations. Moreover, we introduce a new self-created and curated dataset, GTA-V Fight, that allows supervised training of deep learning network models (datasets and code are available in the online appendix (Appendix, 2022)). For this, the following research question (RQ) is formulated:
"How can criminal incidents be automatically detected using virtual gaming data?"

To answer the research question, the following sub-questions (sRQ) have been formulated:

- To what extent can people be recognized in virtual gaming data?
- To what extent can virtual gaming data be used to improve the training of machine learning techniques on real data?
- What scenario complexity is more accurately identifiable?

The rest of the paper is organized as follows. In Section 2, we provide an overview of the literature currently available online. Section 3 presents the research methodology used to answer our research questions. In Section 5, we illustrate our results. While in Section 6, we discuss the outcome of our approach and the related limitations. Finally, Section 8 concludes the paper.

2 RELATED WORK

For years, law enforcement (LEAs) used surveillance cameras, relying on manual monitoring and time-consuming video searches. In smart cities today, Artificial Intelligence (AI) transforms this process, enabling intelligent video surveillance to classify normal and abnormal activities for more efficient law enforcement.

The core of these crime-fighting technologies is video analytics, a subset of Artificial Intelligence (AI), which helps LEAs improve initial threat assessment and real-time response. In recent years, several machine-learning (ML) methods have been developed for video surveillance. These techniques utilize methods that analyze audio, video, and images from video surveillance cameras to detect objects and activities automatically. For example, algorithms have been developed to detect abandoned objects (Li et al., 2010), theft (Chuang et al., 2009), crowd behavior (Nguyen et al., 2005; Mahmoodi and Salajeghe, 2019) and violent activities (Oliver et al., 2000; Goya et al., 2009; Ullah et al., 2019; Zhou et al., 2018).

Early violence detection in videos proposed using flame and blood detection, assessing motion intensity, and recognizing characteristic sounds (Nam J. et al., 1998). Additionally, a suggested approach involves joint audio-visual data representation for detecting violent scenes, emphasizing strong multi-modal cues by statistically revealing joint patterns after combining audio and visual features (Derbas and Quénot, 2014; Zhou et al., 2018).

The key to using machine learning techniques is to extract features that represent violent activity. Many of these techniques can be classified into two categories: handcrafted feature-based approaches and deep learning-based approaches.

In (Nievas et al., 2011), a fight detection system using Bag-of-Words (BoW) framework with STIP and SIFT descriptors achieved 90% accuracy. Similarly, (Nadeem et al., 2019) used GTA-V to classify weapon-based violence with BoVW, reaching 0.88 precision for cold weapons and 0.74 for hot weapons. In contrast, our proposed approach focuses on human identification and recognizing different fight events, reaching a higher accuracy.

One approach to real-time violence detection considers statistics of how flow-vector magnitudes change over time by using the Violent Flows (ViF) descriptor (Hassner et al., 2012). They obtained an accuracy of approximately 82% and outperformed the existing techniques by only using magnitudes of the optical-flow fields. However, performance decreased significantly in a non-crowd behavior dataset. The Oriented ViF (OViF) followed (Gao et al., 2016), obtaining an accuracy of 88% and 87.5% for the two different datasets. However, these results only apply in calm and normal situations. Using videos of crowded scenarios, the accuracy dropped significantly.

Although many handcrafted features-based approaches have been proposed to determine violent activities better using machine learning field, progress still faces different challenges, including monitoring large numbers of people and their activities, different camera positions, complex tracking algorithms, etc. Therefore, resorting to deep learning-based approaches is a natural option (Sargana et al., 2017).

One of the most popular types of deep neural networks is known as Convolutional Neural Networks (CNN). A CNN convolves learned features with input data and uses 2-dimensional (2D) or 3-dimensional (3D) convolutional layers, making this architecture well suited to process 2D or 3D data, such as images or videos. CNN eliminates the need for manual feature extraction. This automated feature extraction makes deep learning models highly accurate for machine learning tasks such as object or activity classification.

CNN has demonstrated great success on various tasks (Wu et al., 2015). Several studies have shown that CNN has higher accuracy and better results for various machine learning techniques, such as behavior recognition and security (Sajjad et al., 2019a; Batchuluun et al., 2017; Sajjad et al., 2019b), object tracking and activity recognition (Ullah et al., 2018; Lee and Kim, 2019). However, not much research has been performed on the automatic detection of violent activities based on deep learning models. Serrano et al. (Serrano et al., 2018) proposed an approach that used Hough forests with 2D CNN to detect violent activities. The approach demonstrated superiority over different handcrafted feature approaches for this recognition task and obtained 99% accuracy. Ullah et al. (Ullah et al., 2019) proposed a violence detection system using spatiotemporal features with 3D CNN. The 3D CNN model from Ullah et al. is a fine-tuning of the original model that was developed in 2015 (Tran et al., 2015).

The model was able to achieve an accuracy of 98% to 99% accuracy in the detection of violent activities. The 3D CNN approach from Ullah outperforms handcrafted-based approaches and state-of-the-art deep learning approaches for different benchmark datasets.

Recently, several publications used Grand Theft Auto-V (GTA-V) and other video game images to train and test deep learning models. These trained models were used for autonomous driving cars (Chen et al., 2015; Filipowicz et al., 2017).

There are three widely used publicly available datasets for violence detection. These are the movies fight dataset, hockey fight dataset (Nievas et al., 2011), and violent crowd dataset (Hassner et al., 2012). Using these datasets means that the studies can be compared and evaluated. A summary of the studies, including the aforementioned studies, is shown in table 1.

As large datasets are not available to train and evaluate deep-learning CNN technology, this study uses virtual gaming data to solve that problem (Martinez et al., 2017). Considering the limitations of the existing recognition techniques and inspired by the performance of the CNN studies, this work will study the possibility of learning a 3D CNN model to predict violent activities accurately based on virtual gaming data.

3 RESEARCH DESIGN

Here, we outline our research methodology for addressing our specific and general research questions, covering data collection, preparation, and our proposed deep learning framework with evaluation metrics and experiments.

Our methodology follows the Cross Industry Standard Process for Data Mining (CRISP-DM) for developing the deep learning framework (Shearer, 2000). We merged the *data collection* with the *data understanding* phase, and given the research nature of the project, there is no *deployment phase*.

3.1 Data Collection and Understanding

This project aims to train a deep-learning framework using virtual gaming data, specifically leveraging the realism offered by GTA-V. Previous studies utilized virtual sources like Udacity (Heylen et al., 2018), The Open Racing Car Simulator (Chen et al., 2015), and Half-Life (Geoffrey et al., 2007) for machine learning model training, but they lacked realism. Academic literature (Chen et al., 2015; Filipowicz et al., 2017), on the other hand, demonstrated that GTA-V provides a more realistic environment for model training, with authentic representations of crime and violent scenarios (Martinez et al., 2017). This choice enables data collection without real-world constraints and allows diverse scenario testing, reflecting the complexity of the real world.

GTA-V Fight Dataset: This paper collected a GTA-V Fight dataset for people fight pose estimation by exploiting the realistic video game GTA-V. The videos were collected from YouTube videos and self-created videos from the video game. The collected videos represent different types of scenes and scenarios. For example, the videos feature different body poses, in several scenarios at varying conditions and viewpoints. The use of different videos ensures that no introduction biases arise for particular scenes or behaviors. The videos in the dataset were labeled as fight and non-fight. A number of examples are shown in Fig. 1. The videos were stored as MP4 files. We decided to use recordings of GTA-V due to the fact that PlayStation did not allow us to use the PlayStation development platform for scientific purposes.

Table 1: Summary of violence detection methods tested on the three well-known datasets: movies fight dataset, hockey fight dataset, and violence crowd dataset.

Methods	Datasets Accuracies (%)		
	Movies dataset	Hockey fight dataset	Violent crowd dataset
STIP, SIFT, BoW (Nievas et al., 2011)	-	87.5	88
ViF (Hassner et al., 2012)	-	82.9	81.3
OViF,AdaBoost,SVM (Gao et al., 2016)	-	87.5	88
Motion Blobs, Random Forests (Gracia et al., 2015)	97.7	79.3	-
Fisher vectors (Bilinski and Brémond, 2016)	99.5	93.7	96.4
sHOT (Rabiee et al., 2018)	-	-	82.2
Hough Forests, 2D CNN (Serrano et al., 2018)	99	94.6	-
3D CNN (Ullah et al., 2019)	99.9	96	98

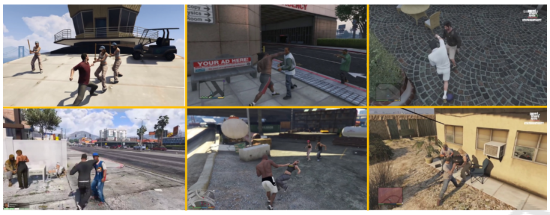

Figure 1: Examples randomly selected from the GTA-V Fight dataset exhibiting its variety in viewpoints, scenarios and number of people.

The dataset contains 250 short videos with different durations. In the dataset, 125 videos are labeled as fight and 125 videos as non-fight. The videos have an average resolution of 1280x720 pixels and a frame rate of 25 frames per second.

Evaluation Datasets: The results of the GTA-V fight dataset were assessed by comparing them with three well-studied datasets for violence recognition. The first dataset is the movie fight dataset. The movie fight dataset was introduced in (Nievas et al., 2011) and was designed for assessing fight detection. The dataset consists of 200 videos in which person-on-person fight videos were extracted from action movies. The videos have an average resolution of 360x250 pixels and a frame rate of 25 frames per second. The second dataset is the hockey fight dataset. This dataset was also introduced in (Nievas et al., 2011) for assessing fight detection. The hockey fight dataset consists of 1000 videos of action from hockey games of the National Hockey League. The dataset was divided into two groups, 500 fight videos and 500 non-fight videos. The videos have a resolution of 360x288 pixels and a frame rate of 25 frames per second. The non-fight videos are also related to the hockey ground environment. The third dataset is the violent crowd dataset. The violent crowd dataset was introduced in (Hassner et al., 2012) and was designed for violence detection and violence classification tasks. This dataset contains 246 videos taken from YouTube, divided into two categories: 123 violent videos and 123 non-violent videos. The videos have a resolution of 320x240 pixels and an average frame rate of 25 frames per second. A number of example frames of these datasets are shown in Fig. 2.

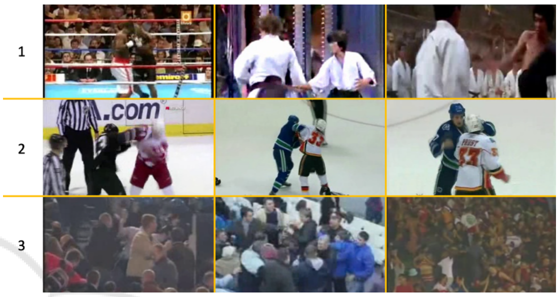

Figure 2: A number of examples frames randomly selected from: (1) movies fight dataset, (2) hockey fight dataset and (3) violent crowd dataset.

A detailed description of the datasets is given in table 2. Table 2 contains the number of total videos in the dataset, the number of violent and non-violent videos, the resolution of the videos, and the frame rate of the videos. All videos and data are packaged and are publicly available on the GitHub page[1].

3.2 Data Preparation

The raw GTA-V Fight dataset was pre-processed to be used for training and testing purposes. The preprocessing steps consisted of image extraction, various data augmentation techniques, and splitting up the dataset. In the following, we provide a detailed description of the preprocessing steps.

Image Extraction: The input layer of the 3D CNN network expects sequence of frames as inputs. The GTA-V Fight dataset consists of videos and must therefore be converted into sequences of frames. The frames of each video are extracted with a rate of 25 frames per second. Then the video frames are resized

[1] https://github.com/srues2/ViolenceDetectionUsingGTAV.git

Table 2: Detailed description of the used datasets.

Datasets	Videos	Resolution	Violent Scenes		Non – Violent Scenes	
			# Videos	Frame rate	# Videos	Frame Rate
GTA-V	250	1280x720	125	25	125	25
Movies Fight	200	320x250	100	25	100	29.97
Hockey Fight	1000	360x288	500	25	500	25
Violent Crowd	246	320x240	123	25	123	25

Table 3: Parameters traditional data augmentation.

Rotation_range	Horizontal flipping	Zoom_range	Shear_range	Height_shift_range	Width_shift_range	Fill_mode
40	True	0.2	0.2	0.2	0.2	Nearest

into 112x112 pixels. The next step is to convert the frames of each video to sequences of 16 frames. During the generation of these sequences is taken into account an 8-frame overlap between two sequences. The main advantage is that there is less information loss between two sequences. In addition, more sequence data has been generated, which leads to more data regularization.

Data Augmentation: Data augmentation is applied on the sequences of frames to regularize the data and to prevent overfitting in the model. Two different data augmentation techniques have been applied to the data: a mix of seven traditional augmentation techniques and style augmentation. Style augmentation is a new form of augmentation technique and Jackson et al. (Jackson et al., 2018) have shown that a combination of traditional augmentation techniques and style improve network performance. Hyperparameter search has determined the optimal values for the ratio of unaugmented to augmented images and the strength of the style transformer. A ratio of 2:1 appears to be optimal. For both augmentation techniques, the same data augmentation technique has been applied to all images in one sequence. As a result, the images in a sequence contain the same data augmentation transformation, and the images between sequences contain different data augmentation transformations.

The *Keras ImageDataGenerator* was applied for the mix of seven traditional data augmentation techniques. The traditional augmentation techniques consist of horizontal flipping, rotations, zooming, erasing, shearing, conversion to grayscale and random perturbations of hue, saturation, brightness and contrast. The preprocessing parameters are randomly chosen and are shown in table 3.

Style augmentation is a new form of data augmentation based on a random style transformer. Style augmentation randomizes texture, contrast and color, while preserving the shape and semantic content. Jackson et al. (Jackson et al., 2018) style augmenter was used to apply style augmentation. A number of examples to which style augmentation was applied to a random frame of the GTA-V Fight dataset are shown in Fig. 3.

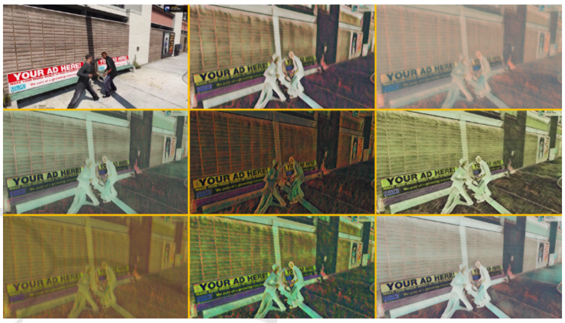

Figure 3: Style augmentation applied to a frame of the GTA-V Fight dataset. The original frame is shown on the top left.

Dataset Preparation: Ultimately, all GTA-V Fight video data was split into sequences of 16 frames with an eight frame overlap between the frames. Next, the sequence data was divided into three splits: a training set (75%), a validation set (12.5%), and a test set (12.5%). A stratified split was applied to ensure the same distribution of the data over the number of occurrences of violent situations in the dataset. When splitting the sequence data, it was taken into account that sequences from the same video were in the same subset.

3.3 Modeling

In general, violence detection models can be classified into two machine learning techniques: handcrafted feature-based approaches and deep learning-based approaches. The handcrafted features approach is based on the expert-designed feature detectors and descriptors. However, handcrafted feature-based approaches are not used in practice due to performance reduction related to the usage of different camera positions and monitoring of large groups of people. Hence deep, learning-based approaches is a natural option.

As already discussed a small number of studies have been done in automatic recognizing violent activities based on deep learning models due to the shortage of data available and the legal and privacy regulations. To avoid this problem, we investigated a deep learning model that was trained on virtual gaming data.

Inspired by Ullah's research results and the possibility of comparing the results with other studies on the three evaluation datasets, we decided to use the deep learning framework proposed in (Ullah et al., 2019).

3.4 Deep Learning Framework

The proposed deep learning framework is based on the three-staged end-to-end framework of Ullah's research (Ullah et al., 2019). The deep learning framework will be divided into two parts: person identification and violence activity identification. The first part of the framework is to identify persons from the input surveillance videos. A MobileNet-SSD CNN model performs person identification. When a person is identified, the images are passed to a 3D CNN model. The second part of the framework is to identify violence scenarios. This model is trained on virtual gaming data and extracts spatiotemporal features. These features are fed to the Softmax output layer of the model and predict whether or not there is a violent scene in the video. This part of the framework can determine to what extent virtual gaming data can be used to improve the training of machine learning techniques on real data. A visual representation of the deep learning framework is shown in figure 6 available in the online appendix (Appendix, 2022).

3.5 Person Identification

The incoming surveillance sequences were first assessed by a MobileNet-SSD CNN model (Howard et al., 2017). This model was originally designed for object detection, fine-grain classification, face attributes, and large-scale geolocation. In this paper, the model from (Howard et al., 2017) is used for person identification. If people were identified in the input surveillance sequences, these sequences with person identification were forwarded to the 3D CNN Network. So only the videos in which people occur were forwarded to the 3D CNN Network and not the videos in which no people occur. This means that the 3D CNN Network does not have to process unimportant sequence frames. The results of Ullah's study (Ullah et al., 2019) showed that this MobileNet-SSD CNN model helps the system optimize latency and size. MobileNets are built primarily from depthwise separable convolutions to detect objects instead of regular convolutions. The MobileNet provided the classification of the input sequences and the SSD version was used to locate the multibox detector. Together MobileNet and MobileNet-SSD performed the person identification. Some examples of person identification in the GTA-V Fight dataset are shown in Fig. 4.

3.6 Violence Identification

Inspired by the performance of Ullah's (Ullah et al., 2019) network, we decided to use his 3D CNN network to determine the performance of the newly created GTA-V dataset. The network consists of eight convolutional layers, five pooling layers, two fully connected layers, and a Softmax output layer. The network architecture is shown in Fig. 6. Each convolutional layer has 3x3x3 kernel size with stride 1x1x1. All pooling layers have 2x2x2 kernel size with stride 2x2x2 except for the first pooling layer with a kernel size of 1x2x2 and stride 1x2x2. The number of filters for each convolutional layer differs per layer. The first and second convolutional layers have 64 filters, the third and fourth convolutional layers have 128 filters, the fifth and sixth convolutional layers have 256 filters and the other convolutional layers have 512 filters. Stochastic gradient descent with a mini-batch size of 16 was used to update the parameters, with a learning rate of 0.001. Dropout was used in the fully connected layers with a rate of 0.5. Each fully connected layer has 4096 output units. The Softmax layer contains two outputs because there were two classes in the dataset: fight and non-fight scenarios. The model was trained for over 40 epochs. Initially, the 3D CNN network received a sequence of 16 frames as an input size of 1280x720 pixels. To avoid overfitting and achieve effective learning, all frames from the original input sequence were resized to crops of 3x16x112x112. Then the sequence of frames passed through the network and the network acted as a generic feature extractor. The network learned to extract features while training. The convolutional layers were made up of a bank of filters whose weights were learned during the training. The pooling and fully connected layers were employed to reduce the learned number of parameters and the size of the image feature descriptor. In fact, all layers generated image feature descriptors for a sequence of frames inputs can be classified by the Softmax layer of the 3D CNN network. Generally, the top activation layers contain larger receptive fields that learn high level and global features, while the bottom activation layers contain smaller receptive fields that are more sensitive towards patterns, such as shapes, edges and

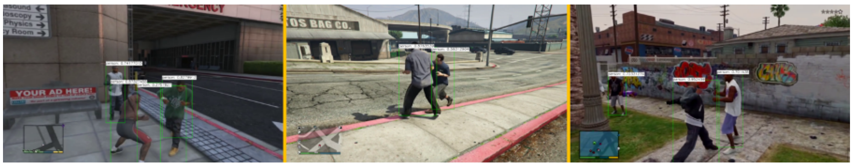

Figure 4: A number of example frames of the GTA-V Fight dataset to which person identification is applied using MobileNet-SSD CNN model.

corners. At the end of the network, the Softmax layer will predict an output label as a fight or non-fight.

4 EVALUATION

We performed several experiments with the aim of learning a 3D CNN model to predict violent activities based on video gaming data. This section provides an overview of the experiments and the order in which they were performed.

4.1 Experiments

The first experiment investigated whether people can be recognized in the GTA-V dataset. To perform this experiment, the MobileNet-SSD CNN model was used for person identification. The model was also performed on the three evaluation datasets. Next, the performance of the GTA-V dataset was compared with the performance of the evaluation datasets to determine to what extent people in virtual gaming data are realistic. All videos of the datasets contained people.

The second part of the experiment was to train the 3D CNN model on the GTA-V Fight dataset. Then, the model was tested on the GTA-V Fight dataset. Than the performance of the model have been compared with the performances of the model of the three evaluation datasets (Ullah et al., 2019).

The third part of the experiment involved training the 3D CNN model on the GTA-V Fight dataset. The best-performed model was then tested on the three evaluation datasets. The results of this part of the experiment can help us determine how generalizable the trained model is on other datasets.

4.2 Evaluation Metrics

As described above, the first experiment determined how many people were identified in the different datasets. To determine how often people were identified in the scenarios, we used the *accuracy* metrics for evaluation. Accuracy is the proportion of true results among the total number of videos examined. Since all videos in the datasets contain persons, the accuracy is suitable for comparing the results of the different datasets.

For the second and third experiment, the dataset was split into a train (75%), validation (12.5%), and test (12.5%) dataset. The trained 3D CNN model can be used to give a prediction of the label of every item in the test dataset. Subsequently, the predicted label can be compared with the actual label. A confusion matrix is suitable for comparing predicted values with the actual values. The confusion matrix contains the following variables:

- True Positives (TP): The predicted label is positive and the actual label is true.
- False Positives (FP): The predicted label is positive and the actual label is false.
- True Negatives (TN): The predicted label is negative and the actual label is false.
- False Negatives (FN): The predicted label is negative and the actual label true.

Due to these confusion matrix variables, different performance metrics can be calculated to assist in evaluating the performance of the model. Since this model had a binary classification task, the following performance metrics have been used to evaluate the experiments performed:

1. **Accuracy**
   The proportion of the total number of predictions that were correct.
   $$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$
2. **Precision**
   The proportion of positive cases that were correctly identified.
   $$Precision = \frac{TP}{TP+FP}$$
3. **Recall**
   The proportion of actual positive cases that were correctly identified.
   $$Recall = \frac{TP}{TP+FN}$$

4. **AUC - ROC Curve**
   AUC - ROC curve is a performance measurement for classification problems. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the 3D CNN model is capable of distinguishing between classes. The ROC curve is plotted with True Positive Rate (TPR), also called recall, against the False Positive Rate (FPR) where TPR is on the y-axis and FPR is on the x-axis.

$$FPR = \frac{FP}{TN + FP}$$

5 RESULTS

After the pre-processing steps, the dataset contained 4988 sequences of 16 frames. The input sequences consisted of 2458 fight labels and 2530 non-fight labels. Three experiments are conducted to evaluate the performance of the proposed method to detect violent activity. The first experiment used the MobileNet-SSD CNN network to identify people in the GTA-V fight dataset and evaluation datasets. The accuracy of identifying a person in a video per dataset is shown in table 4. In the experiment, approximately 96% of the videos in the GTA-V Fight dataset were people identified. This means that in only 4% of the input videos, no people were identified, while the videos contained people. In the movies fight dataset, violent crowd dataset and the hockey fight dataset, people were identified by the network with an accuracy of 98%, 88.2% and 80.9%, respectively. Compared to the accuracy of the GTA-V Fight dataset, this means that the network identified people more often in the virtual gaming dataset than in the realistic hockey fight dataset and violence in the crowd dataset.

Table 4: Accuracy percentage of the MobileNet-SSD CNN model on the used datasets.

Dataset	Accuracy person identified (%)
GTA-V Fight dataset	95.6
Movies fight dataset	98.0
Violent crowd dataset	88.2
Hockey fight dataset	80.9

The second experiment trained the 3D CNN model on the GTA-V fight dataset. Next, the trained model was tested on the GTA-V fight dataset. The results of the experiment are shown in table 5. The trained 3D CNN model scored a performance of 89% accuracy in identifying violence on the GTA-V Fight dataset. The table also shows the performances of the model on the three evaluation datasets. In this case, the model is both trained and tested on the same dataset. The performances of the three evaluation datasets are known from Ullah's (Ullah et al., 2019) research. When the model was trained on the movies fight dataset, the model had a performance of 99.9% accuracy in identifying violence on its own dataset. The violence crowd and hockey fight dataset had an accuracy of 98% and 96%, respectively. Table 5 shows that the GTA-V Fight dataset has approximately an 8% lower accuracy in violent activity identification compared to the three evaluation datasets.

Table 5: Accuracy percentage of the 3D CNN model on the used datasets.

Dataset	Accuracy violence activity identified (%)	AUC
GTA-V Fight dataset	89	0.962
Movies fight dataset	99.9	0.997
Violent crowd dataset	98	0.980
Hockey fight dataset	96	0.970

The Receiver Operating Characteristic (ROC) curve of the GTA-V Fight dataset is shown in Fig. 5. The ROC curve is constructed by plotting the true positive rate against the false-positive rate. The figure is zoomed-in because all true positive rates below 0.4 had a false positive rate of approximately 0, and all false positives above 0.4 had a true position rate of approximately 1. The figure shows that the curve bends to the top-left corner. Further, the performance of the 3D CNN model on the GTA-V Fight dataset was determined by calculating the precision, recall and Area Under Curve (AUC). The precision and recall with AUC values are shown in table 6. The performance values of the three evaluation datasets are derived from Ullah's research (Ullah et al., 2019). The GTA-V Fight dataset had a precision value of 0.841. This means that of all input sequences classified as violence, 84% of these sequences were actually violent situations. The precision value of the evaluation datasets is about 10% higher. The recall value of the GTA-V fight dataset is 0.948 and is slightly lower than the recall value of the evaluation datasets. The recall value means that, of all input sequences that were actually violence, 95% of these sequences were classified as violence. The last column of table 6 shows the AUC values. The GTA-V fight dataset had an AUC value of 0.962 and is approximately equal to the AUC value of the evaluation datasets.

Table 6: Precision, Recall, and AUC values of the used datasets.

Dataset	Precision	Recall	AUC
GTA-V Fight dataset	0.841	0.948	0.962
Movies fight dataset	1.0	1.0	0.997
Violent crowd dataset	0.982	0.988	0.980
Hockey fight dataset	0.960	0.967	0.970

Table 7: Accuracy percentage of the 3D CNN model on the used datasets.

Trained model	**Tested models accuracy (%)**		
	Movies Fight	**Violent Crowd**	**Hockey Fight**
GTA-V Fight dataset	65	68	72
Movies fight dataset	-	54	63
Violence crowd dataset	65	-	47
Hockey fight dataset	49	52	-

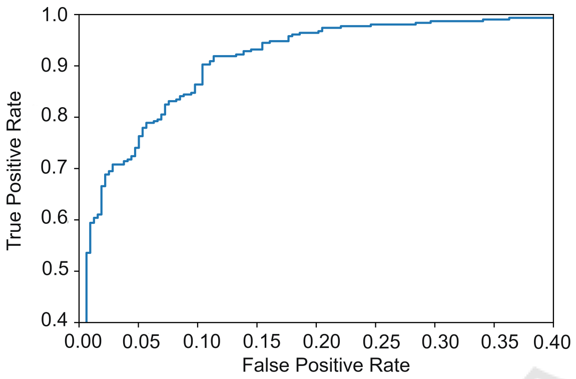

Figure 5: ROC curve on GTA-V Fight dataset.

The third experiment was to test the best trained 3D CNN model, based on GTA-V Fight dataset, on the three evaluation datasets. The results of this experiment are shown in table 7. In Ullah's research, the proposed 3D CNN model was also trained on one of the evaluation datasets and tested on the other two evaluation datasets. When we trained the model on the GTA-V Fight dataset and tested on the evaluation datasets, we had an accuracy of 65%, 68% and 72%, respectively. The trained model had higher accuracy in the identification of violent activities in the hockey fight dataset compared to the violence crowd dataset. In addition, the accuracy percentages of the model were lower if it was trained by an evaluation dataset and subsequently tested on other evaluation datasets. The evaluation datasets had an average accuracy value of 55%, while the model based on the GTA-V Fight dataset had an average accuracy value of approximately 70%.

6 DISCUSSION

This research proposed a framework to train a deep learning violent detection network using virtual gaming data. The virtual gaming dataset is a newly created dataset based on collected and self-created GTA-V videos. The proposed framework is based on a three-staged end-to-end framework (Ullah et al., 2019). This allowed us to compare and evaluate their results with the performance of the GTA-V Fight dataset. First, it was examined whether the virtual persons were similar enough to persons in the real world. Second, it was examined to what extent video gaming data can be used to improve the training in virtual gaming data. Measuring these two aspects of the deep learning framework makes it possible to address our sRQs.

To What Extent Can People Be Recognized in Virtual Gaming Data? In this research, a MobileNet-SSD CNN network was used for person identification. The network had an accuracy of 95.6% in identifying persons in the GTA-V Fight dataset. When this accuracy rate was compared with the accuracy percentages of the evaluation datasets, the accuracy of the virtual gaming dataset was equivalent or higher. This means that the network identified persons as often or more in the virtual gaming dataset than in the realistic evaluation datasets. This result is consistent with other studies on virtual to real transfer learning (Bak et al., 2018; Karttunen et al., 2019; Hoffmann et al., 2019). We can therefore answer our first RQ claiming that persons from GTA-V are just as realistic as persons in the real world.

To What Extent Can Virtual Gaming Data Be Used to Improve the Training of Machine Learning Techniques on Real Data? The second part of the deep learning framework contained a 3D CNN model to identify violent scenarios. The model was trained on virtual gaming data. When the trained model was tested on the GTA-V dataset, it had an accuracy of 89%. The dataset had an 8% lower accuracy in violent activity identification compared to the evaluation datasets. A possible explanation for this difference is the difficulty of the datasets. In the evaluation datasets, the non-violent scenarios are very clear, while the GTA-V dataset contains scenes that resemble violent situations. However, the classification accuracy is typically not enough information to measure the performance of the 3D CNN model.In addition, it is essential to look at other evaluation metrics, such as the recall and precision metrics (see table 6). The goal of this product is for the right actor to be alerted if a violent situation occurs. For example, if a brawl occurs, the police need to intervene. Therefore, there must be as few false negatives as possible, so this means a high recall value. The recall value of the GTA-V fight dataset was approximately 95% and corresponds to the evaluation datasets' results. Of all input sequences that were actually violence, 95% of these sequences were classified as violence. To get a more interpretable assessment, it is also essential to look at the precision metric of the model. The police do not have to be alerted continuously, while there is no violent situation. This means that the number of false positives must be as low as possible, so a high precision value. The GTA-V dataset had a precision value of 84%. The precision value of the evaluation datasets is about 10% higher. Of all input sequences classified as violence, 84% of these sequences were actually violent situations. This means that the model trained on the GTA-V dataset is less good at classifying non-violent situations than the evaluation datasets. Both evaluation metrics showed the performance of the model that has been trained and tested on its own dataset. However, the aim of this research was to investigate how generalizable a model is that was trained on virtual gaming data and tested in the real world.

Therefore, in the third experiment the best trained 3D CNN model, based on the GTA-V Fight dataset, was tested on the three evaluation datasets (see table 7). The performance of the GTA-V dataset model was compared with research also based on the three datasets (Ullah et al., 2019). The GTA-V trained model had an average accuracy of approximately 70%, while the evaluation datasets had an average accuracy value of 55%. This means that the virtual gaming dataset is generalizable compared to a model trained on the evaluation datasets. The virtual gaming dataset had an average 15% higher accuracy than the evaluation datasets. Therefore, virtual gaming data is a solution to improve the training of machine learning techniques on real data.

What Scenario Complexity Is More Accurately Identifiable? The third experiment also focused on the performance of the datasets separately. When the models were trained on their own dataset and subsequently tested on the violence in movies dataset, the GTA-V dataset was about as good at classifying violence. However, if the datasets were trained on their own datasets and then tested on the violent crowd and hockey fight dataset, it is striking that the GTA-V dataset had an 20% higher accuracy. An explanation is the latter two contain more persons in the video clips. This makes that the model trained on the virtual gaming dataset can identify videos with few persons in a brawl about as well or slightly better than models trained on real-world datasets. In addition, the trained virtual gaming model is better at identifying violent scenarios with many persons than models trained on real-world datasets. This result answered our third sub-question. The the model trained on the virtual gaming dataset is better at identifying complex violent situations than existing models in the real world.

Since this technique is relatively new, no further results on the generalizability of virtual gaming data models in violent scenarios were found. In addition, a striking observation is that within the field of identification of violent situations automatically, models are mainly trained and tested on the same datasets. For example, of the eight studies on violence detection methods in table 1, only one study (Ullah et al., 2019) published their generalizability of the model on the various evaluation datasets. The other studies only published the performance of the model where the model is trained and tested on the same dataset. In itself, these results about the performance of a model are interesting, but this concludes nothing about the generalization of the model in the real world. These models are only useful and can be deployed in the real world if they are generalizable enough.

We can conclude our discussion by answering to our main RQ:

"How Can Criminal Incidents Be Automatically Detected Using Virtual Gaming Data?".

We demonstrated through our experiments virtual gaming data can be extremely useful to train models to detect criminal behaviours and incidents. The proposed deep learning framework showed an average 15% higher accuracy than real world datasets. In addition, the GTA-V dataset was much better at identifying violence in complex busy situations than the evaluation datasets. The performances of the experiments show that video gaming data can offer an alternative way to compile large datasets for direct training or augmenting real-world datasets.

7 LIMITATIONS

Our research replicated Ullah's framework to compare GTA-V dataset results. Constraints included resizing all images to 112x112 pixels for the 3D CNN model input. This fixed size minimizes deformation, improving classification accuracy. Evaluation datasets with 320x250-pixel frames limit information loss, close to the 3D CNN model's input resolution. Original GTA-V frames (1280x720 pixels) may result in higher accuracy due to potential information preservation. The second constraint involves using Ullah's research's input sequence length, set at 16 for comparison. This parameter significantly impacts model effectiveness; a length of 16 may be challenging to train due to information overload, while a length of 4 can lead to overfitting. An optimal choice, like a length of 8, may enhance accuracy in video-based identification. Another limitation is that the GTA-V Fight videos are mainly focused on a surveillance camera perspective. The videos were viewed at a fixed-site high-angle from above.

Finally, the trained framework is limited to the movements of GTA-V characters. The model's performance can possibly increase if the movements of people from other video games are also used.

8 CONCLUSION AND FUTURE WORK

In this research, we proposed a violence prediction framework using virtual gaming data and introduced a new dataset, GTA-V Fight, for training deep learning models. Furthermore, virtual persons closely resembled real-world individuals. The person identification network performed as well or better on virtual data than realistic datasets. In addition, violence detection, though 8% less accurate on virtual data, demonstrated a 15% higher average accuracy compared to real-world datasets. The GTA-V dataset excelled in identifying violence in complex situations. These findings suggest video gaming data as an alternative for large, diverse datasets in training or augmenting real-world models.

Future work may be performed to improve the methodological limitations mentioned above. This means that research can be done into adjusting the input size of the frames and the length of the input sequences, as these variables can positively influence the performance of the model (Hashemi, 2019; Li et al., 2019).

The model in this study was trained using only virtual gaming data. It was found that combining real images with synthetically generated ones improves performance in the case of object recognition (Movshovitz-Attias et al., 2016). Future studies may investigate whether a mix of the evaluation datasets and the virtual gaming data can improve the current performance of the violence identification framework.

Finally, this research is the first step in using virtual gaming data to identify criminal incidents. Various models have already been developed for identifying criminal activities (see table 1). In the future, we intend to train existing models with virtual gaming data to compare the performance. The code our approach and a sample dataset of GTA-V violent and non-violent videos is available in the online appendix (Appendix, 2022).

REFERENCES

Alex, K., Ilya, S., and Hg, E. (2012). Imagenet classification with deep convolutional neural networks. *Proceedings of NIPS, IEEE, Neural Information Processing System Foundation*, pages 1097–1105.

Appendix (2022). Violence detection: A serious-gaming approach https://figshare.com/s/d794f30aa0865f70d92d.

Bak, S., Carr, P., and Lalonde, J. (2018). Domain adaptation through synthesis for unsupervised person re-identification. In *ECCV*.

Batchuluun, G., Kim, J., Hong, H., et al. (2017). Fuzzy system based human behavior recognition by combining behavior prediction and recognition. *Expert Syst. Appl.*, 81:108–133.

Bilinski, P. and Brémond, F. (2016). Human violence recognition and detection in surveillance videos. *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 30–36.

Bouma, H., Baan, J., Burghouts, G., et al. (2014). Automatic detection of suspicious behavior of pickpockets with track-based features in a shopping mall. In *Security and Defense*, volume 9253.

Cascavilla, G., Tamburri, D. A., Leotta, F., Mecella, M., and Van Den Heuvel, W. (2023). Counter-terrorism in cyber–physical spaces: Best practices and technologies from the state of the art. *Information and Software Technology*, 161:107260.

Cascavilla, G., Tamburri, D. A., and Van Den Heuvel, W.-J. (2021). Cybercrime threat intelligence: A systematic multi-vocal literature review. *Computers & Security*, 105:102258.

Chen, C., Seff, A., Kornhauser, A., and Xiao, J. (2015). Deepdriving: Learning affordance for direct perception in autonomous driving. *CoRR*.

Chuang, C., Hsieh, J., et al. (2009). Carried object detection using ratio histogram and its application to suspicious event analysis. *IEEE Trans. Circuits Syst. Video Techn.*, 19:911–916.

Derbas, N. and Quénot, G. (2014). Joint audio-visual words for violent scenes detection in movies. In *ICMR*.

Filipowicz, A., Liu, J., and Kornhauser, A. (2017). Learning to recognize distance to stop signs using the virtual world of grand theft auto 5. *Transportation Research Record*.

Flight, S., Klein Kranenburg, L., and Straaten, G. v. (2022). Cameratoezicht door gemeenten.

Gao, Y. et al. (2016). Violence detection using oriented violent flows. *Image Vis. Comput.*, 48-49:37–41.

Geoffrey, R., Andrew, J., and Paul, C. (2007). Ovvv: Using virtual worlds to design and evaluate surveillance systems. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8.

Goya, K., Zhang, X., Kitayama, K., and Nagayama, I. (2009). A method for automatic detection of crimes for public security by using motion analysis. *2009 Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pages 736–741.

Gracia, I., Suarez, O., Garcia, G., and Kim, T. (2015). Fast fight detection. *PLoS ONE*, 10.

Hashemi, M. (2019). Enlarging smaller images before inputting into convolutional neural network: zero-padding vs. interpolation. *Journal of Big Data*, 6:1–13.

Hassner, T., Itcher, Y., et al. (2012). Violent flows: Real-time detection of violent crowd behavior. *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–6.

Heylen, J., Iven, S., Brabandere, B. d., Oramas, J., Gool, L. v., and Tuytelaars, T. (2018). From pixels to actions: Learning to drive a car with deep neural networks. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 606–615.

Hoffmann, D., Tzionas, D., Black, M., and Tang, S. (2019). Learning to train with synthetic humans. In *GCPR*.

Howard, A., Zhu, M., et al. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *ArXiv*, abs/1704.04861.

Jackson, P., Abarghouei, A., Bonner, S., Breckon, T., and Obara, B. (2018). Style augmentation: Data augmentation via style randomization. In *CVPR Workshops*.

Karttunen, J., Kanervisto, A., Hautamäki, V., and Kyrki, V. (2019). From video game to real robot: The transfer between action spaces. *ArXiv*, abs/1905.00741.

Lee, S. and Kim, E. (2019). Multiple object tracking via feature pyramid siamese networks. *IEEE Access*, 7:8181–8194.

Li, X., Zhang, C., and Zhang, D. (2010). Abandoned objects detection using double illumination invariant foreground masks. In *2010 20th International Conference on Pattern Recognition*, pages 436–439. IEEE.

Li, Y., Yin, G., Hou, S., Cui, J., and Huang, Z. (2019). Spatiotemporal feature extraction for pedestrian re-identification. In *Wireless Algorithms, Systems, and Applications: 14th International Conference*.

Mahmoodi, J. and Salajeghe, A. (2019). A classification method based on optical flow for violence detection. *Expert Syst. Appl.*, 127:121–127.

Martinez, M., Sitawarin, C., et al. (2017). Beyond grand theft auto v for training, testing and enhancing deep learning in self driving cars. *ArXiv*, abs/1712.01397.

Movshovitz-Attias, Y., Kanade, T., and Sheikh, Y. (2016). How useful is photo-realistic rendering for visual learning? *ArXiv*, abs/1603.08152.

Nadeem, M. S., Franqueira, V. N. L., Kurugollu, F., and Zhai, X. (2019). Wvd: A new synthetic dataset for video-based violence detection. In Bramer, M. and Petridis, M., editors, *Artificial Intelligence XXXVI*, pages 158–164.

Nam J., Alghoniemy, M. et al. (1998). Audio-visual content-based violent scene characterization. *Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No.98CB36269)*, 1:353–357 vol.1.

Nguyen, N., Phung, D., et al. (2005). Learning and detecting activities from movement trajectories using the hierarchical hidden markov model. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2:955–960 vol. 2.

Nievas, E. et al. (2011). Violence detection in video using computer vision techniques. In *CAIP*.

Oliver, N., Rosario, B., and Pentland, A. (2000). A bayesian computer vision system for modeling human interactions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22:831–843.

Paulin, G. and Ivasic-Kos, M. (2023). Review and analysis of synthetic dataset generation methods and techniques for application in computer vision. *Artificial Intelligence Review*, 56(9):9221–9265.

Persson, P., Espinoza, F., Fagerberg, P., S, A., and Cöster, R. (2002). Geonotes: A location-based information system for public spaces. In *in Kristina Höök, David Benyon and Alan Munro (eds), Readings in Social Navigation of Information Space*, pages 151–173.

Rabiee, H. et al. (2018). Detection and localization of crowd behavior using a novel tracklet-based model. *International Journal of Machine Learning and Cybernetics*, 9:1999–2010.

Rest, J. v., Roelofs, M., and Nunen, A. v. (2014). Afwijkend gedrag maatschappelijk verantwoord waarnemen van gedrag in context van veiligheid. In *TNO 2014 R10987*. TNO.

Sajjad, M., Khan, S., et al. (2019a). Cnn-based anti-spoofing two-tier multi-factor authentication system. *Pattern Recognit. Lett.*, 126:123–131.

Sajjad, M., Nasir, M., Ullah, F., Muhammad, K., Sangaiah, A., and Baik, S. (2019b). Raspberry pi assisted facial expression recognition framework for smart security in law-enforcement services. *Inf. Sci.*, 479:416–431.

Sargana, A., Angelov, P., and Habib, Z. (2017). A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition. *Applied Sciences*, 7:110.

Serrano, I., Déniz, O., Espinosa-Aranda, J., and Bueno, G. (2018). Fight recognition in video using hough forests and 2d convolutional neural network. *IEEE Transactions on Image Processing*, 27:4787–4797.

Shearer, C. (2000). The crisp-dm model: The new blueprint for data mining. *Journal of Data Warehousing*, 5(4).

Tran, D., Bourdev, L., et al. (2015). Learning spatiotemporal features with 3d convolutional networks. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497.

Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M., and Baik, S. (2018). Action recognition in video sequences using deep bi-directional lstm with cnn features. *IEEE Access*, 6:1155–1166.

Ullah, F., Ullah, A., Muhammad, K., Haq, I., and Baik, S. (2019). Violence detection using spatiotemporal features with 3d convolutional neural network. *Sensors (Basel, Switzerland)*, 19.

Wu, Z., Wang, X., et al. (2015). Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In *MM '15*.

Zhou, P., Ding, Q., Luo, H., and Hou, X. (2018). Violence detection in surveillance video using low-level features. *PLoS ONE*, 13.