Jheronimus Academy of Data Science

Violence Detection Using Grand Theft Auto V

Master Thesis

Stan Ruessink, BSc

Supervisors:
Damian Tamburri, JADS
Eric Postma, JADS
Giuseppe Cascavilla, Sapienza

Version 1.0

's-Hertogenbosch, May 2020

Abstract

in progress...

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A large part of our public life takes place in public spaces. Public spaces are usually located in the open air, but also municipal institutions and government buildings belong to public spaces. More than a thousand small and large events occur every year in such public areas in the Netherlands. Safety is one of the most critical aspects of public spaces. To guarantee safety, it is essential to detect threatening incidents proactively [1]. At the moment, public order and safety are generally guaranteed by an integrated approach of organizers, municipalities, law enforcement officers, fire brigades, and medical services. Surveillance cameras are a popular tool with these chain partners. The number of surveillance cameras is rapidly increasing to improve security in public spaces [2]. Despite the increase in the number of surveillance cameras, its effectiveness is questionable. Manual surveillance seems tedious and time-consuming. The systems require a human to monitor the multiple video screens and to identify criminal activity. However, only a selection of the video streams can be observed, and the effectiveness of predictive behavior is not clear [3]. Automatic detection in surveillance cameras can help to prevent criminal incidents on time and handle the vast amount of data.

The automatic criminal incident detection of surveillance cameras is a challenging field in real-world actions and has advanced rapidly over the last few years. Due to the different possible person movements, high dimensions of video data, different motion speed, different color videos, precise criminal incident recognition is still a big challenging task. In the past decade, several machine learning-based methods have been developed for identifying criminal incidents. For example, Chuang et al. [4] developed a system to identify theft. The system was able to identify theft in low-quality surveillance videos. Another example is Kumar and Bhatnagar [5], who developed a system to detect the behavior of large groups of people. These types of systems showed high performance in detecting suspicious objects and activities. However, these handcrafted based approaches are not used in practice. The reason is that these approaches' performance reduces with different camera positions and monitoring of large groups of people. In addition, the approaches had a large computational time. Recently, the use of neural networks and deep learning algorithms have shown significant progress in the automatic detection of objects and activities. Currently, network models are used for behavioral recognition, object tracking, and activity recognition. However, only a limited number of models are used for identifying violent activities since a limited number of datasets are available to train network models. Data augmentation techniques are used to cope with the limited available data [6]. Besides, pre-trained networks are trained for a similar task.

Another solution is to generate data through the use of virtual gaming data. For example, a deep learning network has been trained, based on virtual gaming data, to apply in self-driving cars [7, 8]. In addition, a network has been developed to detect abnormal behavior in crowded scenarios [9]. The use of virtual gaming data is a new technique. Therefore, limited data and academic literature are available on this topic. One of the main concerns of the training of virtual gaming data is whether the virtual data is close enough to the real world to make an effective application possible. Inspired by the deep learning-based approaches and the new solution to generate data with virtual games, this thesis proposes a framework to train a deep learning network based on

virtual gaming data. This work's contribution is twofold: (1) an extension to the existing literature for the detection of criminal incidents using a deep learning network and (2) using virtual reality to train deep learning networks than can ultimately be used in real situations. Also, we introduce a new self-created dataset, GTA-V Fight, that allows supervised training of deep learning network models. For this, the following research question is formulated:
"How can criminal incidents be automatically detected using virtual gaming data?"

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

- To what extent can people be recognized in virtual game data?
- To what extent can virtual game 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. Section 2 explores the related work to this area. Section 3 introduces the proposed approach of the framework. Section 4 provides our results, and section 5 discusses the results and some future work. Finally, section 6 concludes the paper.

Chapter 2

Related work

In recent years, several machine learning methods have been developed for video surveillance. Machine learning for video surveillance utilizes methods that analyze audio, video and images from video surveillance cameras in order to detect objects and activities automatically. For example, algorithms have been developed to detect abandoned objects [10], theft [4], smoke and fire [11], crowd behavior [12, 13] and violent activities [14, 15, 16, 17]. Li et al. [10] created a model to detect and recognize abandoned objects with the use of a Gaussian mixture model and Support Vector Machine (SVM). Results showed that the used method could detect very small abandoned objects within low-quality surveillance videos, and it is robust to the varying illuminations and dynamic background. Chuang et al. [4] developed a system to recognize theft using a Forward-backward ratio histogram and finite state machine. The system has detected 96% of the cases of theft and proved that the model is robust, accurate, and powerful in carried object detection. Seebamrungsat et al. [11] proposed a fire detection system based on light detection. This system uses HSV and YCbCr color models to separate colors from each other and colors from the background. The differences between the generated frames of the models are analyzed and predict the presence and growth of smoke and fire. The overall accuracy of the system has been greater than 90%. Kumar and Bhatnagar [5] developed a crowd behavior detection system using a hybrid tracking model and integrated features enabled neural network. The crowd behavior detection system estimates the direction of the movement of objects as well as their activity. The performance of the developed system obtained an accuracy of 95%.

With respect to the detection of violent activities, Nam et al. [18] is one of the first proposals for violence detection in videos. This study proposed to recognize violent scenes in videos using flame and blood detection and capture the degree of motion, as well as the characteristic sounds of violent events. Further, Derbas and Quénot [19] proposed an audio-visual data representation for violent scenes detection. They proposed a feature that provides strong multi-modal audio and visual cues by first joining the audio and the visual features and then revealing the joint multi-modal patterns statistically. However, audio-based methods are always restricted since the absence of the audio channel in many surveillance cameras [17]. Since the audio channel is often absent in surveillance cameras, violent detection from surveillance videos is essentially a task of activity recognition. Several machine learning techniques and methods have been developed to detect violent activities. Using machine learning techniques, the key point is to extract features that represent the violent activity. Many of these techniques can be classified into two categories: handcrafted feature-based approaches or deep learning-based approaches.

2.1 Handcrafted feature-based approaches

Features designed by humans are defined as handcrafted features. The handcrafted features approach is based on the expert-designed feature detectors and descriptors such as Histogram of Oriented Gradients (HOG), Hidden Markov Models (HMM), Scale-Invariant Feature Transform

(SIFT), Violent Flows (ViF), Gaussian Mixture Models (GMM) and Space Time Interest Points (STIP).

Some of these handcrafted features-based approaches are used in violent activity recognition. Nievas et al. [20] developed a fight detection system based on a Bag-of-Words (BoW) framework using the STIP and SIFT descriptors. Experiments showed that the BoW approach could accurately recognize fight activities with approximately 90% accuracy. However, the results differed between the type of dataset. For the one dataset, they concluded that the accuracy was insensitive to the choice of descriptor and vocabulary size. In contrast, the accuracy of another dataset depended on the choice of the descriptor, with SIFT dramatically outperforming STIP. Hassner et al. [21] described an approach to real-time detection of violence in crowded scenes. The detection system considered statistics of how flow-vector magnitudes change over time by using the ViF descriptor. The ViF descriptors are classified as either violent or non-violent scenes using a Support Vector Machine (SVM). They obtained an accuracy of approximately 82% and outperformed the existing techniques by relying on magnitudes of the optical-flow fields alone. However, the ViF-based approach performance decreased significantly in a non-crowd behavior dataset. Based on ViF descriptor, Gao et al. [22] developed the Oriented ViF (OViF). The OViF features describe the changes of motion magnitudes based on the statistics of motion orientations. The approach used AdaBoost as feature extraction and linear SVM for classification. They obtained an accuracy of 88% and 87.5% for the two different datasets and achieved improved performance over the existing ViF approaches. However, these results only apply in calm and normal situations. When videos of crowded scenarios were used, the accuracy dropped significantly. Furthermore, Gracia et al. [23] used motion blobs and random forests for the detection of fight and non-fight activities. In this approach, blobs of movement were first detected, and then different features were used to characterize the fights. The approach makes no assumptions on the number of individuals body part detection or salient point tracking. Results showed that the approach has a significantly faster computational time than existing approaches, as mentioned above. However, the approach did not outperform the approaches considered. The accuracy is ranged from 70% to 95% depending on the type of dataset. Bilinski and Brémond [24] used the sliding window approach and improved the Fisher vector approach to detect violence. The approach employed local features and spatio-temporal positions. Although the base of the approach is in the temporal sliding window, the authors sped up the detection of violence for a range of frames by using a summed-area table. The advantage of this area table is that no temporal segments need to be calculated, which leads to a more accurate and faster performance compared to existing approaches. They obtained accuracies of approximately 96%. Moreover, Rabiee et al. [25] used two descriptors to detect and localize abnormal behaviors in crowded scenes. They proposed the simplified Histogram of Oriented Traclet (sHOT) model, which contains both orientation and magnitude information in a single feature. They combined sHOT with a Dense Optical Flow (DOF) to detect abnormal behavior in a crowd. This abnormal behavior descriptor obtained an accuracy of 82.2%. They concluded that the descriptor could deal with abnormality detection in various crowd densities, and evaluate on medium level crowd and dense crowd scenarios. However, the method may not be satisfying as person detector-based methods, since they are detecting the crowd behaviors rather than individual behaviors.

Handcrafted feature-based approaches are still widely used due to some bottlenecks such as computational complexity of deep learning-based approaches for violent activity recognition. Although many handcrafted features-based approaches have been proposed to better determine violent activities in the field of machine learning, 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 [26].

2.2 Deep learning-based approaches

Recently, there has been a growing trend of learning robust feature representations from raw data with deep neural networks. Deep learning is a subset of machine learning methods based on Artificial Neural Networks (ANN) that use recently developed training techniques to train their models. These networks are basically an abstract representation of data points. The high-level representation consists of multiple layers for processing the networks that are used to reach higher complexity. The different layers can learn different abstraction levels of the data using the input of previous layers until they reach a final layer. The final layer makes the final decision for the class. In fact, several features are learned at each layer of hierarchy in the network [27].

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 2D or 3D convolutional layers, making this architecture well suited to processing 2D or 3D data, such as images or videos. CNN eliminates the need for manual feature extraction. This means that it is not necessary to identify features used to classify images. The CNN works by extracting features directly from the data. The relevant features are not pre-trained and will be learned while the network trains on a collection of data. This automated feature extraction makes deep learning models highly accurate for machine learning tasks such as object or activity classification. The success of training deep learning networks is dependent on the existence of large datasets for training and evaluation [9]. It takes a significant amount of human time and effort to build large, with ground truth labels, datasets which is extremely expensive. Building a good dataset for detecting activities should be diverse, capture all possible aspects of the problem, and precisely annotated.

CNN has demonstrated great success on various tasks [28]. Several studies have shown that CNN has higher accuracy and better results for various machine learning techniques, such as behavior recognition and security [29, 30, 31], object tracking and activity recognition [32, 33]. However, not much research has been done into the automatic detection of violent activities based on deep learning models. Serrano et al. [34] 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. [16] 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 [35]. The approach from Ullah consists of two steps. First, persons are detected in a video using a light-weight CNN model to reduce and overcome the voluminous processing of useless frames. Second, a sequence of frames with detected persons is passed to a 3D CNN, where the spatiotemporal features of these frames are extracted and fed to the Softmax classifier. This classifier predicts whether a violent activity occurred in the input video. The model was finally 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.

As described above, a small number of studies have been done in automatic recognizing violent activities based on deep learning models. This because little data is available due to legal and privacy regulations. A solution to the problem of generating datasets is through the use of virtual gaming data.

2.3 Training with virtual gaming data

Virtual gaming data provides the opportunity to readily create a scenario, capture the labeled data of those scenarios and effectively investigate neighboring variances of those scenarios [36]. Recently, several publications used GTA-V and other video game images to train and test deep learning models. These trained models were used for autonomous driving cars [7, 8]. For example, Filipowicz et al.[8] used GTA-V data to detect the distance to stop signs from an in-game image. In addition to the use of virtual reality for self-driving cars, there is one study that used virtual reality in detecting abnormal behavior in crowded scenes using density heatmaps and optical flow[9]. The first results of this study are positive, but there is still much improvement possible. For example, investigating other deep learning techniques, using other virtual gaming data, improving detection accuracies and evaluating other datasets.

All aforementioned violent detection models have in common that the models have been tested on three well-known datasets for violence detection. The three widely used publicly available datasets for violence detection are movies fight dataset [20], hockey fight dataset [20] and violent crowd dataset [21]. The advantage of the consistent use of these datasets is that the results of the studies can be compared and evaluated. A summary of the studies is shown in table 1.

Methods	**Datasets Accuracies (%)**		
	Movies dataset	**Hockey fight dataset**	**Violent crowd dataset**
STIP, SIFT, BoW [20]	-	87.5	88
ViF [21]	-	82.9	81.3
OViF,AdaBoost,SVM [22]	-	87.5	88
Motion Blobs, Random Forests [23]	97.7	79.3	-
Fisher vectors [24]	99.5	93.7	96.4
sHOT [25]	-	-	82.2
Hough Forests, 2D CNN [34]	99	94.6	-
3D CNN [16]	99.9	96	98

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

Based on the literature above, it can be concluded that the development of automatic detection of violent activity systems is not improving because there is a lack of large datasets for training and evaluation. Thus, the latest and successful deep-learning CNN technology cannot be tested and evaluated for the automatic detection of violent activities. To avoid this problem, it is useful to investigate whether there are other ways to generate datasets. A possible solution is using virtual gaming data. 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. A more detailed description of the research is described in the ‘Methodology’ section.

Chapter 3

Methodology

In this section, the method that is used to answer the sub-questions of the research question defined in the earlier section will be described in detail. First, an overview of the deep-learning framework will be given in section 3.1. Then, a detailed explanation of all steps in the deep-learning framework will be given in the consequent sections. Finally, section 3.6 and section 3.7 will provide a detailed description of the evaluation metrics and the experiments performed.

3.1 Deep learning framework

In general, this section will discuss how a violent activity will be detected based on a deep learning framework using virtual gaming data. The proposed framework is based on the three-staged end-to-end framework of Ullah's research[16]. The virtual gaming data is a new created dataset based on collected and self-created GTA-V videos. 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 using the 3D CNN model. This model is trained on virtual gaming data and extract spatiotemporal features. The spatiotemporal features are fed to the Softmax output layer of the 3D CNN model and predict whether or not there is a violent scene in the video. When a violent scenario is discovered, an alert could be sent to the nearest security department or a police station. A visual representation of the deep learning framework is shown in figure 1. More information about the used datasets, models, and experiments are further discussed in detail in the sub-sections below.

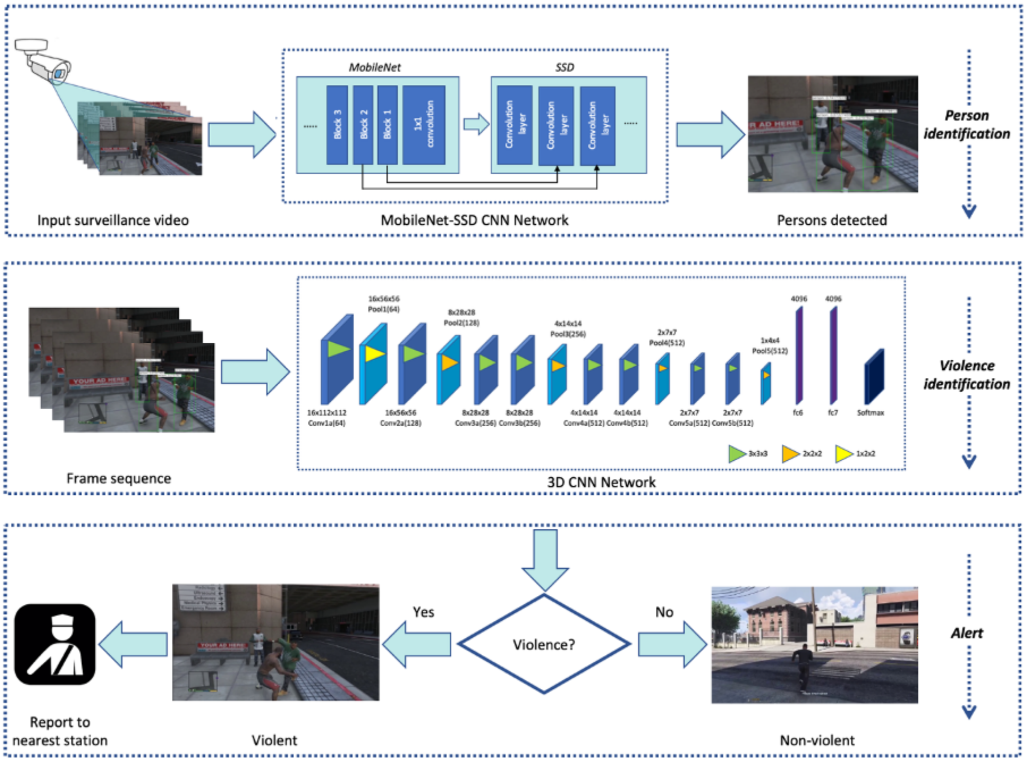

Figure 1: Deep learning framework of the violent detection method.

3.2 Datasets

3.2.1 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 figure 2. The videos were stored as MP4 files.

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.

3.2.2 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 movies fight dataset. The movies fight dataset was introduced in [20] 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 [20] 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

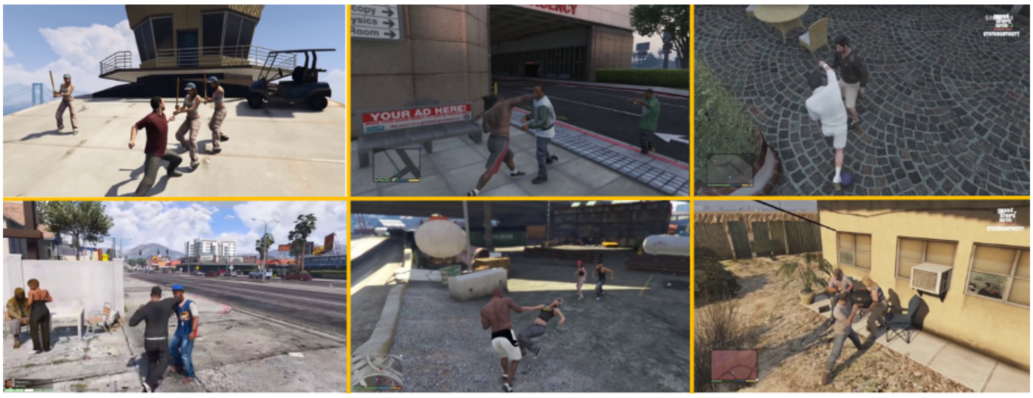

Figure 2: Examples randomly selected from the GTA-V Fight dataset exhibiting its variety in viewpoints, scenarios and number of people.

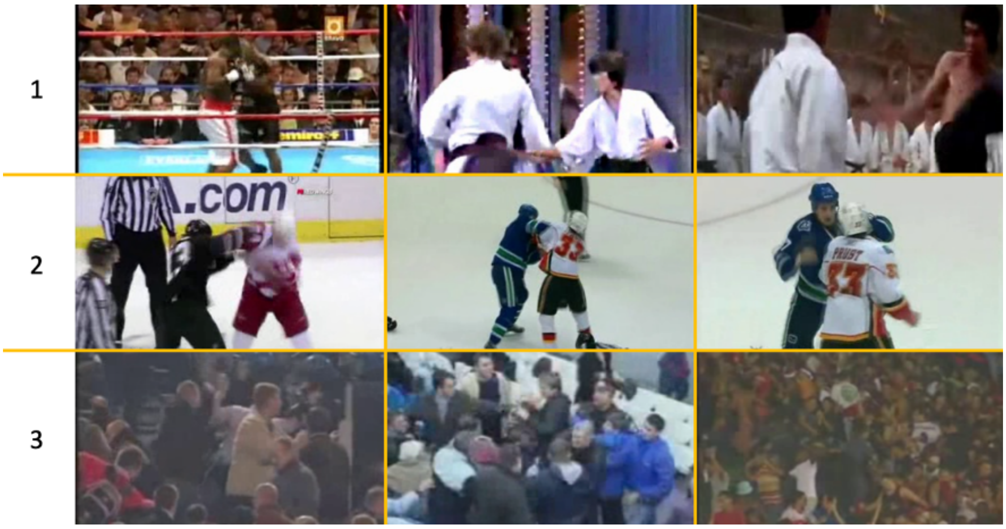

Figure 3: A number of examples frames randomly selected from: (1) movies fight dataset, (2) hockey fight dataset and (3) violent crowd dataset.

dataset is the violent crowd dataset. The violent crowd dataset was introduced in [21] 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 figure 3. A detailed description of the datasets is given in table 2.

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 2: Detailed description of the used datasets

3.3 Data preparation and usage

Preprocessing was done on the GTA-V Fight dataset. The preprocessing steps consisted of image extraction, various data augmentation techniques, and splitting up the dataset. The preprocessing steps are described in the sections below.

3.3.1 Image extraction

The input layer of the C3D 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 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 is generated which leads to more data regularization.

3.3.2 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. [X21] 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 augmentations 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.

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

Table 3: Parameters traditional data augmentation

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. [37] 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 figure 4.

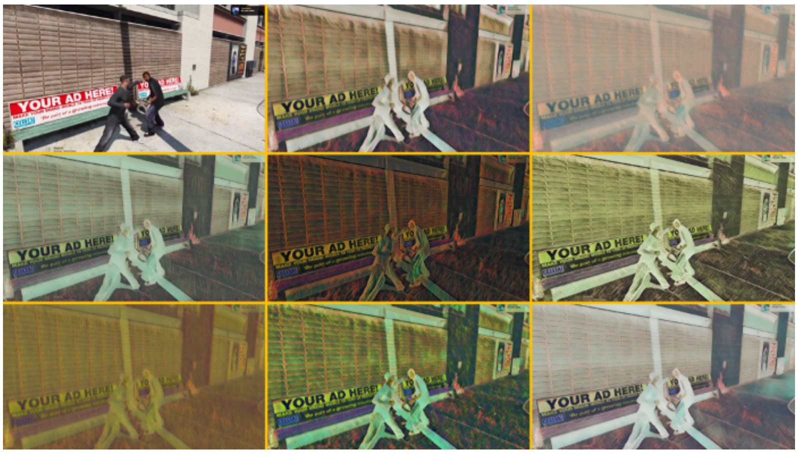

Figure 4: Style augmentation applied to a frame of the GTA-V Fight dataset. The original frame is shown on the top left.

3.3.3 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 based on the number of occurrences of fight situations. When splitting the sequence data, it was taken into account that sequences from the same video were in the same subset.

3.4 Person identification

The incoming surveillance sequences were first assessed by a MobileNet-SSD CNN model [38]. This model was originally designed for object detection, fine-grain classification, face attributes and large-scale geolocation. In this thesis, this model was used for person identification. If people were identified in the input surveillance sequences, these sequences with person identification were forwarded to the C3D Network, which is explained in section 3.4. So only the videos in which people occur were forwarded to the C3D Network and not the videos in which no people occur. This means that unimportant sequence frames do not have to be processed by the C3D Network. The results of Ullah's study [16] 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 they performed person identification. Some examples of person identification in the GTA-V Fight dataset are shown in figure 5.

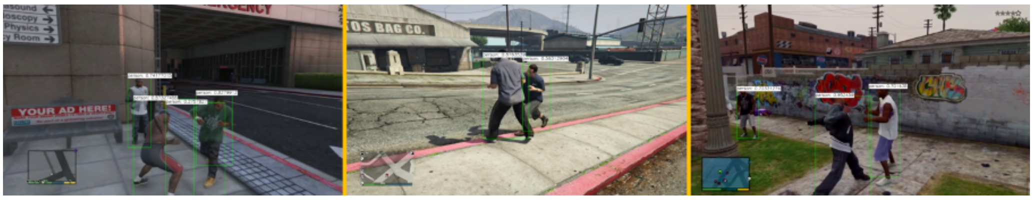

Figure 5: A number of example frames of the GTA-V Fight dataset to which person identification is applied using MobileNet-SSD CNN model.

3.5 Violence identification

Inspired by the performance of Ullah's[16] network, we decided to use his C3D 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 figure 1. 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 second 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 over 40 epochs.

Initially, the C3D 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 number of parameters to be learned and the size of the image feature descriptor. In fact, all layers generated image feature descriptors for a sequence of frames inputs that can be classified by the Softmax layer of the C3D 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 corners. At the end of the network, the Softmax layer will predict an output label as a fight or non-fight.

3.6 Experiments

Several experiments were performed during this thesis with the aim of learning a 3D CNN model to predict violence activities based on video game data. This section gives an overview of the experiments and the order in which they were performed.

The first sub-question is to what extent people can be recognized in virtual gaming data. This shows how realistic people in virtual videos are compared to people in real data. Therefore, 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 MobileNet-SSD CNN model was processed on the GTA-V dataset and the performances were stored. 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 sub-question is to what extent video gaming data can be used to improve the training of a 3D CNN network on real data. The sub-question was answered with the second and third experiment. The second 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 and the performance was stored. The performance of the model was compared with the performances of the model of the three evaluation datasets. The performances of the three evaluation datasets were known from Ullah's [16] research. By comparing the performances, the accuracy of the models on the different datasets was determined. With these results, it could be argued how well the model can recognize violence scenarios.

The third experiment is to train the 3D CNN model on the GTA-V Fight dataset. The best-performed model was then tested on the three evaluation datasets. The performance of each evaluation dataset was compared and stored. Ullah's research [16] also showed how the trained evaluation sets performed when tested on each other. From these results, it could be determined how generalizable the trained model is on other datasets. Finally, this experiment can answer the research question of whether virtual game data can be used to recognize real fight situations automatically.

3.7 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, the accuracy metrics was used 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.

As described above, for the second and third experiment, the dataset was split into a train, validation, and test 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.
   $$FPR = \frac{FP}{TN + FP}$$

Chapter 4

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 a 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 crowd dataset.

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

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

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 had 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 [x18] 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.

The Receiver Operating Characteristic (ROC) curve of the GTA-V Fight dataset is shown in figure 6. The ROC curve is constructed by plotting the true positive rate against the false-positive rate. 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 [X18]. The GTA-V Fight dataset had a precision value of 0.841. This means that of all input sequences

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

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

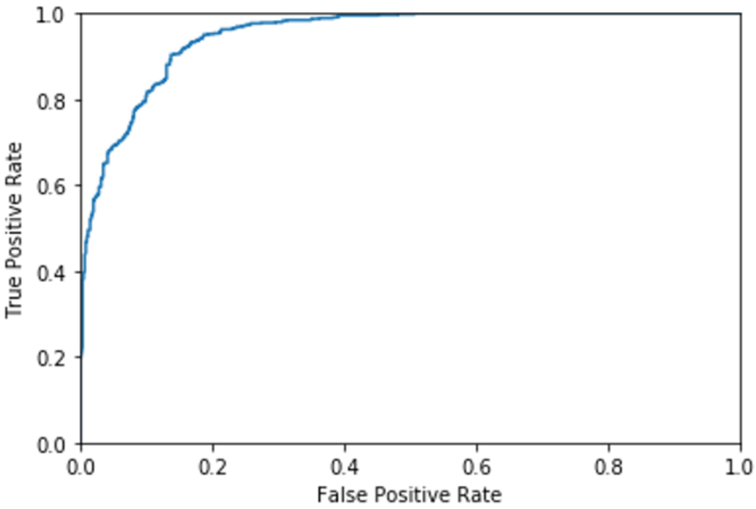

Figure 6: ROC curve on GTA-V Fight dataset.

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.

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

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. The performance of this experiments is also shown in the table. 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. Further, 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%.

Trained model	**Tested models accuracy (%)**		
	Violence in movies	**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	-

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

Chapter 5

Discussion

in progress...

Chapter 6

Conclusion

in progress...

Bibliography

[1] J. van Rest, M. Roelofs, and A. van Nunen. Afwijkend gedrag maatschappelijk verantwoord waarnemen van gedrag in context van veiligheid. In *TNO 2014 R10987*. TNO, 2014.

[2] Autoriteit Persoonsgegevens. Cameratoezicht. In *Beleidsregels voor de toepassing van bepalingen uit de Wet bescherming persoonsgegevens en de Wet politiegegevens*, 2016.

[3] H. Bouma, J. Baan, G. Burghouts, P. Eendebak, J. Huis, J. Dijk, and J. Rest. Automatic detection of suspicious behavior of pickpockets with track-based features in a shopping mall. In *Security and Defense*, volume 9253, 09 2014.

[4] C.H. Chuang, J.W. Hsieh, L.W. Tsai, S.Y. Chen, and K.C. Fan. Carried object detection using ratio histogram and its application to suspicious event analysis. *IEEE Trans. Circuits Syst. Video Techn.*, 19:911–916, 06 2009.

[5] M. Kumar and C. Bhatnagar. Crowd behavior recognition using hybrid tracking model and genetic algorithm enabled neural network. *International Journal of Computational Intelligence Systems*, 10:234, 01 2017.

[6] K. Alex, S. Ilya, and E. Hg. Imagenet classification with deep convolutional neural networks. *Proceedings of NIPS, IEEE, Neural Information Processing System Foundation*, pages 1097–1105, 01 2012.

[7] C. Chen, A. Seff, A. Kornhauser, and J. Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. *CoRR*, 05 2015.

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

[9] L. Lazaridis, A. Dimou, and P. Daras. Abnormal behavior detection in crowded scenes using density heatmaps and optical flow. *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2060–2064, 2018.

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

[11] J. Seebamrungsat, S. Praising, and P. Riyamongkol. Fire detection in the buildings using image processing. *2014 Third ICT International Student Project Conference (ICT-ISPC)*, pages 95–98, 2014.

[12] N.T. Nguyen, D.Q. Phung, S. Venkatesh, and H.H. Bui. 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, 2005.

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

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

[15] K. Goya, X. Zhang, K. Kitayama, and I. Nagayama. 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, 2009.

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

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

[18] Nam J., M. Alghoniemy, and A.H. Tewfik. Audio-visual content-based violent scene characterization. *Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No.98CB36269)*, 1:353–357 vol.1, 1998.

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

[20] E.B. Nievas, O. Déniz-Suárez, G.B. García, and R. Sukthankar. Violence detection in video using computer vision techniques. In *CAIP*, 2011.

[21] T. Hassner, Y. Itcher, and O. Kliper-Gross. Violent flows: Real-time detection of violent crowd behavior. *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–6, 2012.

[22] Y. Gao, H. Liu, X. Sun, C. Wang, and Y.C. Liu. Violence detection using oriented violent flows. *Image Vis. Comput.*, 48-49:37–41, 2016.

[23] I.S. Gracia, O.D. Suarez, G.B. Garcia, and T.K. Kim. Fast fight detection. *PLoS ONE*, 10, 2015.

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

[25] H.R. Rabiee, H. Mousavi, M. Nabi, and M. Ravanbakhsh. Detection and localization of crowd behavior using a novel tracklet-based model. *International Journal of Machine Learning and Cybernetics*, 9:1999–2010, 2018.

[26] A.B. Sargana, P. Angelov, and Z. Habib. A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition. *Applied Sciences*, 7:110, 01 2017.

[27] K. Pogorelov, O. Ostroukhova, M. Jeppsson, H. Espeland, C. Griwodz, T. de Lange, D. Johansen, M. Riegler, and P. Halvorsen. Deep learning and hand-crafted feature based approaches for polyp detection in medical videos. *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*, pages 381–386, 2018.

[28] Z. Wu, X. Wang, Y. Jiang, H. Ye, and X. Xue. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In *MM '15*, 2015.

[29] M. Sajjad, S.H. Khan, T. Hussain, K. Muhammad, A.K. Sangaiah, A. Castiglione, C. Esposito, and S.W. Baik. Cnn-based anti-spoofing two-tier multi-factor authentication system. *Pattern Recognit. Lett.*, 126:123–131, 2019.

[30] G. Batchuluun, J.H. Kim, H.G. Hong, J.K. Kang, and K.R. Park. Fuzzy system based human behavior recognition by combining behavior prediction and recognition. *Expert Syst. Appl.*, 81:108–133, 2017.

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

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

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

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

[35] D. Tran, L.D. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015.

[36] M. Martinez, C. Sitawarin, K. Finch, L. Meincke, A. Yablonski, and A.L. Kornhauser. Beyond grand theft auto v for training, testing and enhancing deep learning in self driving cars. *ArXiv*, abs/1712.01397, 2017.

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

[38] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *ArXiv*, abs/1704.04861, 2017.

Appendix A

My First Appendix

appendix...