Doctoral (PhD) dissertation

Viktor Huszár
2023 February

NATIONAL UNIVERSITY OF PUBLIC SERVICE

Doctoral School of Military Engineering

Viktor Huszár:

Possibilities and challenges of the defensive use of artificial intelligence and computer vision based technologies and applications in the defence sector

Doctoral (PhD) dissertation

Supervisors:

Dr. Imre Négyesi

. . . . . .

Dr. Csaba Krasznay

. . . . . .

Budapest, 2023 February

Declaration

I hereby certify that the Ph.D. thesis entitled “Possibilities and challenges of the defensive use of artificial intelligence and computer vision-based technologies and applications in the defence sector” is the result of my own research and investigation, except where otherwise stated. Where other sources of information have been used, they have been duly acknowledged. This research has not already been accepted for any degree and is also not being concurrently submitted for any other degree.

Viktor Huszár

2023 February

Acknowledgments

I am thankful to all those who have supported me on my journey to write my doctoral dissertation. I would especially like to thank my supervisors, Dr. Imre Négyesi and Dr. Csaba Krasznay for their trust in me and all the contribution they have made to this paper. Their valuable comments and continuous support and understanding mean a lot to me.

I would also like to thank Professor Zsolt Haig and Professor Zoltán Szenes for all their professional encouragement and feedback, as my research journey started a decade ago with my master's studies, and they treated me as professors should treat the odd ones like myself: they never gave up on me. Thank you for all your unforgettable and interesting lectures.

I would like to thank the Director of the Doctoral School of Military Engineering, Professor József Padányi. We had beautiful arguments and his true contribution was guiding me to be a team player with my supervisors. I hope he will be proud of me and the way I represented the Doctoral School. I would like to thank all members of the Doctoral School for their help and support throughout these amazing years of study. I would like to also thank my colleagues, especially Szilvia Blascsák, as her contributions and constant pushing meant a lot to end up writing this paper, her detailed professionalism is highly appreciated.

Last but not least, I would like to thank my family, my wife Enikő Huszár, my three wonderful children, and my friends as well for their patience, for their motivation, and for their unconditional caring support in achieving the scientific goal of my life.

Contents

Chapter 1

Introduction

1.1 Preface

Artificial Intelligence (AI) and Computer Vision (CV) technologies have the potential to revolutionize the defence sector, providing new opportunities for military operations, situational awareness, and decision-making. This doctoral dissertation explores the complex relationship between AI and CV technologies and their potential applications in defence, taking a comprehensive approach to the examination of the current state of these technologies and the associated challenges and limitations.

The research for this dissertation was conducted through a combination of a literature review and case studies, providing a thorough examination of available information and data. This analysis provides an in-depth understanding of the impact that AI and CV technologies could have on the defence sector and the wider implications for society.

One of the key findings of this research is the significant potential for AI and CV technologies to enhance the defence capabilities of a nation. However, this potential is not without its challenges. Ethical and moral dilemmas, security risks, and the possibility of misuse are some of the many challenges that must be addressed to ensure the responsible deployment of these technologies in defence.

This doctoral dissertation serves as a valuable resource for individuals and organizations working in the defence sector, AI and CV researchers, and policymakers who play a role in shaping the future of these technologies. It provides insights and recommendations to help guide the responsible development and deployment of AI and CV technologies in defence, taking into account the ethical and security implications that arise from their use.

Briefly, this doctoral dissertation provides a comprehensive examination of the relationship between AI and CV technologies and their potential use in defence. By highlighting the opportunities and challenges of these technologies, it offers valuable insights

and guidance for their responsible development and deployment in the defence sector, helping to ensure that these technologies are used in a way that benefits society and protects against potential harm.

1.2 Introduction

The concept of cybersecurity has added a new dimension to the challenges and threats addressed by traditional security measures. With the advent of the internet, new technologies such as Artificial Intelligence and computer vision have emerged, which have the potential to revolutionize various industries. The deep learning capabilities of these technologies bring with them a host of new challenges for the military and IT sectors.

One such challenge is the analysis of real-time computer images, which can now be done in a cost-effective manner through the use of AI systems. These systems can be trained using artificial, synthetic data, allowing for the analysis of large amounts of data in a short amount of time. Additionally, AI and computer vision technologies have made it possible to develop new and innovative military applications, such as autonomous weapons systems.

The field of information technology is constantly evolving and transforming. As early as 1995, scholars such as Christensen et al. [8] highlighted the importance of disruptive innovations that have the potential to fundamentally change the way companies and governments operate. Floppy disks, CDs, and the internet are just a few examples of such innovations that have had a profound impact on the world.

However, on the ubiquitous information network a revolutionary innovation appeared based on AI and computer vision in [9]. Most people identify AI with self-driving automotive industrial developments, but it has a wide range of applicability. Combining AI with computer vision and machine learning can result in a disruptive technology that changes the way the world works in economic, legal, and, above all, IT terms [10]. It will have a significant impact on IT systems, but less attention has been paid to the opportunities and threats for military and defence administration. Military science similarly has faced new challenges at the end of the 20^{th} century with the emergence of the Internet.

The potential applications of the AI and computer vision obtained data raise numerous military technical scientific and legal issues. Military information is most of the time classified, therefore obtaining data is difficult, legally challenging, and expensive. Despite certain smart solutions, where encrypted data could be obtained from uncontrolled resources [11], the data would be still subject to human error and would need secondary evaluation and validation. Data for military applications are even more interesting, as data security must be a priority in defence administration and in the daily communica-

tion of authorities.

For military applications, using AI and computer vision inherits several scientific problems depending on the source of data, encryption algorithms, and AI modelling structure. The computing power for real-time results can be another technical challenge, where scarce resources can make computations costly. The need for centralized data storage and an uncontrolled security management system should be explored for the efficient use of resources. The problem extends to the artificial isolation of such systems and the military risks of ”awaking” machine learning or programmed artificial intelligence. Science should investigate how data security, data integration, and isolation of the artificial intelligence decision-making environment and the framework for authorization levels can be achieved in such an automated distributed network-based military environment. Currently, due to centralized data storage and central control, breakable data communication between departments is a major problem. The most important aspect in terms of the degree of vulnerability is the activity of the user or the organization and - closely related to this - the value of their data. Particularly popular targets of attackers are financial institutions and organizations dealing with state or official secrets. Using manually obtained data can also conflict with the use of the personal data it processes and the related data processing operations. GDPR and other regulatory policies, therefore, limit the source of data usage [12].

In the case of analysis based on visual data, the data structure changes: digital images are represented in the form of feature descriptors. Shape recognition uses these feature descriptions to create object classes. Some applications in computer vision need to create three-dimensional models based on images or videos. This requires processing, analysis, and recognition, both of which require high computing power [13], and currently, most of the image analysis methodologies are often slow and do not operate in real-time. However, the power resource can be problematic, so several technologies can be combined to be strategically synergic for the best performance [14]. Eventually, competitive machine vision solutions will be the ones with reduced expensive hardware and resource requirements. In developing countries, test self-driving tracks and 5th Generation mobile network (5G) exemplary tests show how expensive the implementation of innovative machine vision-based R&D results is.

Artificial Intelligence (AI) and computer vision technology have been shaping the future of global societies, too. These disruptive technologies are game changers for the armed forces and law enforcement, as they can reduce human error and improve the efficiency of decision-making. It is often mentioned, that AI has the potential to become greater than the internet itself. The world’s leading militaries in Russia, China, and the United States have expressed their interest in applying and combining the technologies for improvements and advancements in military and law enforcement areas. Despite

the widespread use of biometric recognition (face recognition) of AI, there is still a lot of space for research on the divergent military applicability of emerging technologies. Therefore, comprehensive research was required to identify the potential of computer vision-based technologies and to determine best-use case scenarios and military and law enforcement application purposes.

1.3 Artificial Intelligence and security related applications

Artificial Intelligence was first coined by John McCarthy, also known as the father of AI, at a Dartmouth conference in 1956 [15]. Since then, AI has been continuously proven to be beneficial in several domains including human assistance systems. Figure 1.1 shows various important inventions over the years based on AI that opened ways to multiple diverse areas of research. In July 2018, where Daniel Faggella, the head of Research and CEO of Emerj, addressed the Interpol–United Nations (UNICRI) Global Meeting on the Opportunities and Risks of Artificial Intelligence and Robotics for Law Enforcement. This was the beginning of discussions regarding Artificial Intelligence in policing, security, and law enforcement.

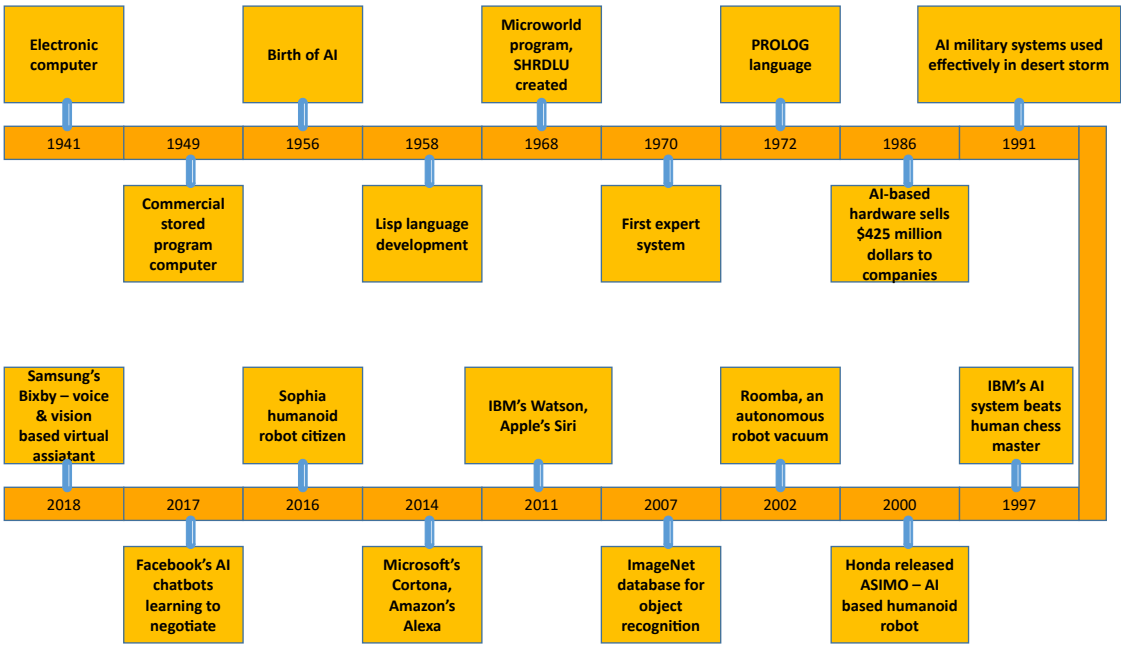

Figure 1.1: Timeline of major Artificial Intelligence events (original illustration by Viktor Huszár).

Kevin McCaney notes in Law Enforcement Using Analytical Tools to Predict Crime, that law enforcement agencies are starting to rely on predictive analytics software in an-

ticipating and preventing crime. Until recently, such technology was mostly used by competitive enterprises. An example would be the IBM Blue Crush (Criminal Reduction Utilizing Statistical History) software in use by the Memphis, Tennessee Police Department to "analyze crime and arrest data, and combine it with weather forecasts, economic factors, and information on events such as paydays and concerts to create predictive models" [16]

The criminal justice system is a crucial component of society, responsible for maintaining law and order, investigating and prosecuting crimes, and providing justice and punishment to those who have committed crimes. This system is comprised of three main agencies: police, courts, and correctional facilities. The role of these agencies is to work together to enforce the law, protect the public, and hold those who have committed crimes accountable.

In order to perform their functions effectively, these agencies must have access to cutting-edge technology and expert analysis. One of the most important areas of focus in this regard is the use of Artificial Intelligence (AI) programs to enhance the precision of crime analysis and scenario simulation. This is where the National Intelligence Model (NIM) comes in. Developed in the UK, NIM is designed to improve and assist intelligence-led police operations.

The National Intelligence Model consists of nine individual elements, each of which plays a critical role in improving the efficiency and effectiveness of law enforcement activities [17].

- Crime Pattern (number/relations) focuses on analyzing the number and relationships of crimes committed in a given area. This information is then used to better understand the types of crimes being committed, which can inform law enforcement efforts to prevent and respond to these crimes.

- Criminal surveys are another important component of the National Intelligence Model. This element focuses on gathering data on crime and criminal activity in a given area, which can be used to develop a deeper understanding of criminal behaviour. This information can then be used to create targeted and effective responses to criminal activity.

- Demographic/ Social Trend Analysis is another important element of the National Intelligence Model. This element focuses on analyzing demographic and social trends in a given area, which can be used to identify areas that may be at higher risk of criminal activity. This information can then be used to inform law enforcement efforts to prevent crime in these areas.

- Profiling criminal operations is another critical element of the National Intelligence Model. This element focuses on identifying the characteristics of criminal operations, including the actors involved, the types of crimes being committed, and the methods used to commit these crimes. This information can then be used to create targeted and effective responses to criminal activity.

- Network analysis (actors who make up such networks) is another important element of the National Intelligence Model. This element focuses on analyzing the actors who make up criminal networks, including the relationships between these actors and the types of crimes being committed. This information can then be used to disrupt criminal networks and prevent crime from being committed.

- Risk analysis is another critical component of the National Intelligence Model. This element focuses on identifying the risks associated with criminal activity, including the likelihood of a crime being committed and the potential consequences of these crimes. This information can then be used to inform law enforcement efforts to prevent crime and protect the public.

- target profile analysis is another important element of the National Intelligence Model. This element focuses on identifying the characteristics of criminal targets, including the types of crimes being committed and the methods used to commit these crimes. This information can then be used to create targeted and effective responses to criminal activity.

- Operational Intelligence assessment is another important element of the National Intelligence Model. This element focuses on assessing the effectiveness of law enforcement operations, including the strategies and tactics used to respond to criminal activity. This information can then be used to refine and improve law enforcement efforts, ensuring that they are as effective as possible.

- Results analysis is the final element of the National Intelligence Model. This element focuses on evaluating the results of law enforcement activities, including the effectiveness of various strategies and tactics. This information can then be used to inform future law enforcement efforts, ensuring that they are as effective as possible.

The use of Artificial Intelligence (AI) and robotics in policing have become increasingly popular due to their potential to improve crime prevention, investigation, and response. However, the same features that make AI and robotics appealing to law enforcement can also be used by criminals and terrorist groups to carry out malicious attacks.

This has been highlighted in a report by the United Nations Interregional Crime and Justice Research Institute (UNICRI), which concludes that AI and robotics could be both a tool for good and a weapon for evil in policing. The report identified three main areas of attack, each with its unique set of challenges and implications for the criminal justice system [18]. The report identified three main areas of attack:

- Digital attacks, such as automated spear phishing, automated discovery, and exploitation of cyber vulnerabilities. These types of attacks rely on the use of AI to automate the process of finding and exploiting cyber vulnerabilities, allowing criminals to gain access to sensitive information, such as login credentials and financial information.
- Political attacks, such as the spread of fake news or media to generate confusion, conflict, or the use of face-swapping (deep fake) and spoofing tools to manipulate video and create trust issues in political figures. These types of attacks can have a significant impact on the public's trust in political leaders, as well as the validity of evidence in court. Furthermore, the use of deep fake technology can lead to the creation of false narratives, making it difficult to differentiate between what is real and what is not.
- Physical attacks, such as facial recognition capabilities in armed drones or drones smuggling contraband. These types of attacks can have serious consequences, such as the potential loss of life or the smuggling of illegal goods into the country. The report also highlights the potential of AI to be used to subvert another AI system by poisoning data sets, which could result in incorrect decisions being made by law enforcement agencies. [19].

Therefore, the UNICRI report highlights the need for caution in the use of AI and robotics in policing and the criminal justice system. While these technologies have the potential to improve the efficiency and accuracy of law enforcement, they also present new challenges and risks that must be considered. The report suggests that law enforcement agencies must work closely with experts in AI and cybersecurity to mitigate these risks and ensure that AI and robotics are used for the public good in policing.

1.4 AI in military applications

The application of Artificial Intelligence (AI) and Computer Vision in the military and defence sector has gained significant attention in recent years. In the literature, there is a growing body of research exploring the use of these technologies for various military and defence applications.

1.4.1 Target recognition and classification

The application of AI and Computer Vision technologies for target recognition and classification in the military and defence sector is an important area of research and development. These technologies can be used to automatically identify and classify targets from surveillance images, which can help to speed up the target recognition process and support decision-making. For example, deep learning algorithms can be trained to recognize and classify different types of weapons, vehicles, and other targets based on their visual appearance.

1.4.2 Autonomous systems

The development of autonomous systems is another key area of focus in the literature on AI and Computer Vision applications in the military and defence sector. These systems can be designed to operate without human intervention, reducing the risk to human personnel and supporting operations in dangerous environments. For example, autonomous vehicles, unmanned aerial vehicles (UAVs), and other systems can be equipped with AI and Computer Vision technologies to enhance their situational awareness and decision-making capabilities. This can include object detection and tracking, target recognition and classification, and other capabilities that can help to enhance their performance and reduce the risk to human personnel.

1.4.3 Weapon systems

AI and Computer Vision technologies are also being used to enhance weapon systems in the military and defence sectors. This can include target acquisition and tracking systems, which can improve the accuracy and effectiveness of weapon systems. For example, AI and Computer Vision technologies can be used to automatically identify and track targets, allowing weapon systems to engage them more accurately and effectively. This can help to reduce the risk to human personnel and improve the effectiveness of military and defence operations.

1.5 Computer Vision using AI for public safety

Computer vision has been making great strides in recent years, with applications ranging from facial recognition to autonomous vehicles. One area where computer vision is having a major impact is in the field of surveillance and security. With more and more video cameras being deployed in public spaces, the amount of recorded footage that must be an-

alyzed by operators is becoming overwhelming. To address this problem, much research is being conducted to develop systems that can automatically analyze this footage.

The goal of this research is to reduce the mechanical load on the operators of these cameras. By automating the process of analyzing the footage, the operators will be freed up to focus on other important tasks, such as responding to potential threats. In addition, automatic analysis of footage will allow for faster and more accurate identification of potential threats, as well as more efficient use of resources.

The deployment of video cameras in public spaces is becoming increasingly common, with cameras being placed in educational institutions, parks, shopping malls, traffic lights, highways, and even at country borders. The goal of these cameras is to ensure safety, by providing a record of events that can be used to identify potential threats and respond to them in a timely manner.

The development of systems that can automatically analyze footage from these cameras is a critical area of research, as it will have a major impact on the efficiency and effectiveness of surveillance and security systems. By reducing the workload of operators, these systems will be able to provide more accurate and timely information, which will in turn help to ensure the safety of the public. Some important applications related to surveillance and security where computer vision and AI technologies can be extremely useful are shown in Figure 1.2. These applications are discussed in the following.

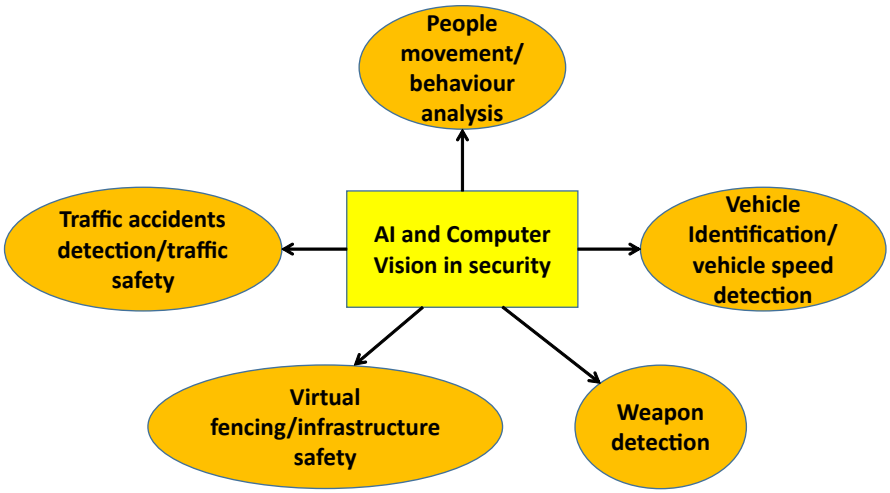

Figure 1.2: Applications of AI and computer vision in the fields of video surveillance and security monitoring (original illustration by Viktor Huszár).

1.5.1 Object detection and localization

The use of facial image data for face detection and person identification has become ubiquitous in today's society, being utilized in both smartphones and offices. With the

advent of AI and modern methods for face detection, these systems are now able to accurately identify individuals even in crowded environments. The recent revolution in real-time object detection using deep learning has made a significant impact in the field of image recognition, with AI-based object detection algorithms demonstrating impressive robustness to various lighting conditions and background environments, and delivering precise results even in low-light conditions.

One particularly important application of object detection in the realm of public safety is the detection of weapons, such as guns and swords, in surveillance videos. By utilizing frame-level data, these systems are able to detect and localize such weapons, thereby alerting law enforcement authorities to potentially dangerous situations. Another significant use case for object detection is in the protection of infrastructure, including private properties, restricted zones, and railway passages. By implementing virtual fencing methods, these areas can be monitored to ensure that they are not being trespassed. This is achieved by marking prohibited regions in the optical image frame of surveillance cameras and comparing the localized pixel coordinates of detected objects against these marked regions to identify any violations.

The use of computer vision and AI in detecting vehicle number plates from traffic cameras is another crucial area where these technologies have demonstrated enormous potential. The fast movement of vehicles can cause significant motion blur in traffic videos, making it difficult to accurately identify vehicle number plates even for the human eye in some cases. However, AI-based vision algorithms have demonstrated exceptional performance in these applications, outperforming several state-of-the-art baseline methods.

1.5.2 Object tracking and behaviour analysis

Carrying out object detection in a video across multiple images allows for the modelling and evaluation of object behaviour. The trajectories of object coordinates in multiple video frames contain crucial information about the movements of objects and their interactions with other objects. For instance, the trajectories of automobiles in surveillance videos can be used to detect a range of events, such as speed limit violations and one-way traffic violations. By analyzing the trajectories of multiple vehicles, it is also possible to automatically detect traffic accidents.

In many applications, including smart cities, the movement analysis of people is of great significance for recognizing abnormal or violent actions, such as fights. By leveraging deep learning for behaviour analysis, it is possible to not only observe these actions but also predict the future movements of individuals. This enables the observation and classification of any deviations from normal behaviour, providing a valuable tool for

identifying potential security risks.

The use of object detection and behaviour analysis in video data has far-reaching implications, with applications not only in public safety, traffic management, and smart cities, but also in fields such as sports analysis, wildlife conservation, and even entertainment. The ability to track and analyze object behaviour in real-time opens up a range of possibilities for understanding and optimizing human and animal movements, as well as improving our understanding of the world around us.

1.5.3 Distributed deep learning

Developing effective deep-learning models requires a substantial amount of data collection and processing. To train practical deep learning models, millions of training samples must be incorporated, making the size and nature of the training data a critical factor. Specifically, videos contain data with multiple pixel values collected across various dimensions, including colour, spatial, and temporal. Depending on the number of frames, loading, and processing a video for training can be a computationally intensive task that requires a large amount of memory bandwidth.

To meet these memory requirements, the use of multiple high-end GPUs is a common approach. These GPUs are used to share the training data and to ensure that sufficient memory is available for processing. However, such high-end GPUs are not always readily available in consumer-level computers.

An alternative to using multiple GPUs is to use multiple computers as distributed nodes. This involves partitioning the data into smaller chunks and training independent models on each computer. This approach can be especially useful in scenarios where access to high-end GPUs is limited and allows for the efficient processing of large amounts of data.

1.6 Scientific challenges

Although there are multiple use cases for AI and computer vision in safety and security applications, following the state of current scientific research, there are several challenges that need to be addressed before these approaches are considered for practical implementation. These are discussed in the following:

- **Object detection and tracking** - In the core, most of the security applications of AI involve object detection in one or more forms. Current state-of-the-art methods for object detection do not always provide reliable results. Unlike other applications such as entertainment and gaming, in security-related applications, false

detections impose greater penalties in algorithm design. In videos, methods for object tracking are useful to filter out any false detections in individual frames using the detections from a series of previous frames. However, such filtering should be done heuristically for accurate tracking and behaviour analysis. These are not well discussed and experimented with in the literature.

- **Hardware and system complexity** - For improving the performance of object detection and tracking, advanced equipment such as depth cameras can be used. However, incorporating additional hardware increases overall system cost and may also introduce latency while processing this additional stream of data. Algorithms for action recognition extract visual features from videos which are used in training AI algorithms to achieve good classification accuracy. However, the cost of extracting well-representative visual features, using several state-of-the-art methods, is still prohibitive for devising approaches that work in real-time.

- **Datasets diversity** Datasets are very limited in security-related applications for use in deep learning. Successful deep learning architectures require millions of examples for meeting the accuracy needed in practical applications. However, due to laws such as GDPR such amounts of training data are not accessible in safety and security applications. Datasets with less training data lead to model overfitting.

- **Generalizability** For understanding if a trained deep learning model can perform well on samples not included in the training phase, cross-database validation studies are highly important. Exiting methods in the literature related to safety and security applications using AI do not conduct cross-database validations and the practical validity of the results is not well discussed.

- **Prototypes** For applying the research in practice, prototyping and stand-alone implements are significant. Much research in safety-related applications does not discuss the applicability of developed methods and the ways to adapt the approaches in practical use cases.

- **Remote processing** In applications where a central computing resource is involved, visual data is streamed to this remote central server which involves data compression. In literature, model sensitivity to noise and artefacts resulting from data compression is not studied.

- **Distributed deep learning** Currently there are no approaches proposed in the literature to address problems related to sharing the training data using data encryption and using distributed computing resources for deep learning. These can empower

the development of fast and novel technologies in securing related applications using AI.

1.7 Aims and objectives

The focus of the experimental research was to study the latest advancements in military technology, particularly in the use of AI for safety purposes. The integration of computer vision and artificial intelligence technologies holds great potential for providing an efficient and measurable response to events and issues in military bases and educational institutions.

Through the use of AI algorithms, including Human Activity Recognition (HAR) activity recognition, the researchers aim to develop automated decision-making tools that can assist in maintaining safety in these environments. The goal of this work is to explore the possibility of establishing a Digital Guard that can detect potentially dangerous situations in real-time, by signalling so-called 'trigger' objects and events and marking objects and individuals based on a risk scale.

The use of AI for safety applications in military and educational settings holds immense promise, as it has the potential to enhance the decision-making process and improve the overall security of these environments. By automating the process of identifying and responding to potential threats, AI has the potential to greatly improve the efficiency and accuracy of safety measures.

In short, the research aims to investigate the possibilities of using AI and computer vision technologies to improve safety in military and educational environments. The goal is to develop a Digital Guard that can efficiently detect potential threats and help prevent dangerous situations from arising. Through the use of AI algorithms and decision-making tools, this research contributes to the advancement of safety measures in these critical settings.

1.8 Scope of work

The scope of the work involves exploring and refining existing deep learning models for applications in safety and security. The research will also introduce new datasets for experimenting with these models.

However, due to privacy laws such as GDPR, the availability of real-world footage for training deep learning models is limited. As a result, the usage of synthetic training data has become more prevalent in computer vision. The researchers have previously explored using training data containing pasted object patches on real images, which has

been effective in tasks such as 2D object detection and human pose estimation. However, this approach may not fully capture the diverse and complex action patterns of violent or deceitful behaviour. As a result, the preparation and use of synthetic training data is outside the scope of the current work.

The research will focus on developing successful models and proposing approaches for their practical implementation. The work also discusses example prototypes and system architectures, but the development of specialized hardware specifically tailored to the proposed deep learning models is beyond the scope of the current work.

1.9 Research hypothesis

Based on the research objectives and the scientific challenges outlined, the thesis is driven to apply the research questions by addressing the following hypotheses:

- **H1**: Deep learning-based methods can effectively detect spoof attacks from human video frames.
- **H2**: Deep learning-based methods for action recognition can be adapted for violence detection.
- **H3**: Distributed ledger technology can effectively run high resource-intensive AI/deep learning algorithms.

1.10 Research methodology

Meeting these objectives of the work in this dissertation involves a two-stage research approach. In the first stage, a literature review was undertaken, examining the subject-related fields and in the second stage, new novel methods are proposed that overcome the shortcomings of the existing methods and advance the results achieved in state-of-the-art.

Among the previously developed methods and most of the modern deep learning methods for video replay spoofing detection, most use public datasets for experimenting with and comparing different methods. There are several inconsistencies among the existing public datasets for video replay spoofing detection using facial image data in terms of the subject position, image resolution, and cameras used for dataset generation. Cross-dataset evaluations in the literature have shown limitations of both the developed methods and existing datasets. Results in the literature show that spoofing detection datasets for biometric or live face detection are not relevant to other applications such as HAR, since they generate data in static sessions with minimum illumination variability.

In contrast, in HAR applications, users have the flexibility to freely choose their location, and eventually, it is not always possible to control the user's behaviour.

In some cases such as virtual games, some of the body parts including the face are partially or completely occluded by objects for interaction such as a football. Methods developed in the literature for spoofing detection in bio-metric face authentication applications may not extend to such cases. On top of that, if the activity recognition involves remote server processing, the image data may be subjected to resizing and/or compression. The dependency of such operations on spoofing detection accuracy is not discussed in the literature. Finally, the state-of-the-art deep learning-based methods for spoofing detection using facial image data involve pre-processing steps such as face alignment after face detection which is impossible in HAR use cases since the head position of a user is not fixed. In addition, in the literature, no such meticulous studies are using deep learning methods targeted at mobile devices. To address all of these issues, in this work, I developed new tools including a novel database and robust data-driven approach that jointly examine various regions of the captured images to detect spoofing. The proposed system enhances and integrates state-of-the-art deep learning algorithms for detecting replay attack spoof cases for HAR applications.

The key aspects in developing successful algorithms for violence detection are that these methods should be computationally fast, achieve the best classification accuracy, and adapt to scenarios that are not present in the training. To the best of my knowledge, there are currently very few cross-database validation studies reported in the literature for violence detection. Although methods that learn normal patterns or predict future frames for detecting violence are computationally very light, they are not suitable for practical applications because of poor generalizability.

From the results presented in the literature, it is evident that methods that employ spatiotemporal features, extracted from labelled videos that account for both motion and appearance information, in training can achieve good classification accuracy. However, it is important to note that the cost of extracting some of those features is still prohibitive for practical applications. The work related to violence detection adapts and extends computationally light 3D deep learning architecture X3D-M using various collected datasets to develop efficient methods for violence detection.

1.11 Dissertation organization

The dissertation is organized as follows: Chapters 2, 3 & 4 are self-contained. Each of them presents a brief introduction, followed by a summary of state-of-the-art methods, their limitations, and the main contributions to spoof detection, violence detection,

and decentralized AI computing using distributed ledger technology respectively. New scientific results from the current work (described in Chapters 2, 3 & 4) in essence are presented in Chapter 5. Chapter 6 briefly discusses the overall applications of the work. Finally, conclusions and future work are presented in Chapter 7.

Chapter 2

Live Spoofing Detection

2.1 Introduction

Human Activity Recognition (HAR) has become a widely researched topic in recent years, thanks to the availability of ubiquitous computers and smart wearable sensors. The goal of HAR algorithms is to provide information on the physical activities of humans, from simple to complex. To achieve this, the algorithms take data from various sensors as input and use machine learning and computer vision techniques to extract information about human activities. HAR algorithms have a wide range of applications, including medical diagnosis, keeping track of elderly people [20], [21], [22], smart homes [23], automated driving [24], military training and surveillance [25], [26], crime control, and motion-driven virtual games [27]. For instance, in the medical field, these algorithms can be used to monitor the physical activity of patients with chronic conditions and to detect changes in their activity patterns that may indicate a worsening of their condition. In smart homes, HAR algorithms can be used to automate lighting and heating systems based on the presence and movements of residents. In automated driving, they can be used to monitor the driver's activities and detect any signs of drowsiness or distraction.

In military training and surveillance, HAR algorithms can be used to monitor soldiers' physical activities and training progress and to detect any signs of injury or exhaustion. In crime control, these algorithms can be used to monitor public spaces and detect suspicious activities. Finally, in motion-driven virtual games, HAR algorithms can be used to track player movements and translate them into in-game actions.

In the context of automatic HAR, a spoofing attack occurs when a person deliberately provides misleading or false visual data to the activity recognition algorithm. The algorithm then reports this fake data as a successfully performed action. This type of attack is particularly concerning in the context of intelligent video surveillance systems, which are becoming increasingly popular in places such as large waiting rooms, campuses, and

public spaces. Intelligent video surveillance systems are designed to recognize simple or complex motion patterns and derive high-level subjective descriptions of actions and interactions among subjects and objects. However, these systems are vulnerable to spoofing attacks, particularly when people are equipped with smartphones. For example, a person could play a pre-recorded video on their smartphone screen in front of a surveillance camera, misleading the system into recognizing a false action.

Given the increasing use of intelligent video surveillance systems and the potential for spoofing attacks, it is critical to develop methods for detecting such attacks before making a final decision on recognized actions. This is essential for ensuring the reliability and accuracy of these systems and maintaining public trust in their use.

Sensors used for HAR can be divided into two main categories: wearable and external sensors [28]. Wearable sensors are those that measure the required data for activity recognition while in physical contact with the user. Examples of wearable sensors include accelerometers, gyroscopes, and magnetometers, which are used to translate human motion into signal patterns [29] for activity recognition. These sensors are often integrated into wearable devices, such as fitness trackers, that are worn by the user. On the other hand, external sensors are set at fixed points and require interaction from the user. They are commonly used in applications such as public safety and surveillance, where cameras are set up at fixed locations to capture human activity. With the advancements in deep learning methods, such as convolutional neural networks and recurrent neural networks, it is possible to achieve state-of-the-art results in activity recognition. These deep learning methods can automatically learn features from raw sensor data, improving the accuracy and reliability of the recognition process.

A use case that utilizes deep learning methods for Human Activity Recognition using external sensors is illustrated in Figure 2.1. This use case involves a user interacting with a digital gaming application designed to analyze and rate their freestyle football/soccer performance. The user is equipped with a football, which serves as a tool for interacting with the application. In this scenario, deep learning algorithms are employed to detect the body parts of the user and the football in real-time. The real-time detection information is then used to recognize the current activity of the user, enabling the application to accurately assess their performance. This information is then passed on to the evaluation stage, where the user can compare their skills to others and work on improving their abilities.

The use of deep learning approaches for Human Activity Recognition has numerous potential applications, including in the digital gaming industry. By detecting human activities in real time, deep learning algorithms can provide valuable insights into the user's performance and enable them to improve their skills. The use of external sensors, such as cameras, in this scenario, provides a unique perspective on the user's activities

and allows for a more comprehensive analysis of their performance.

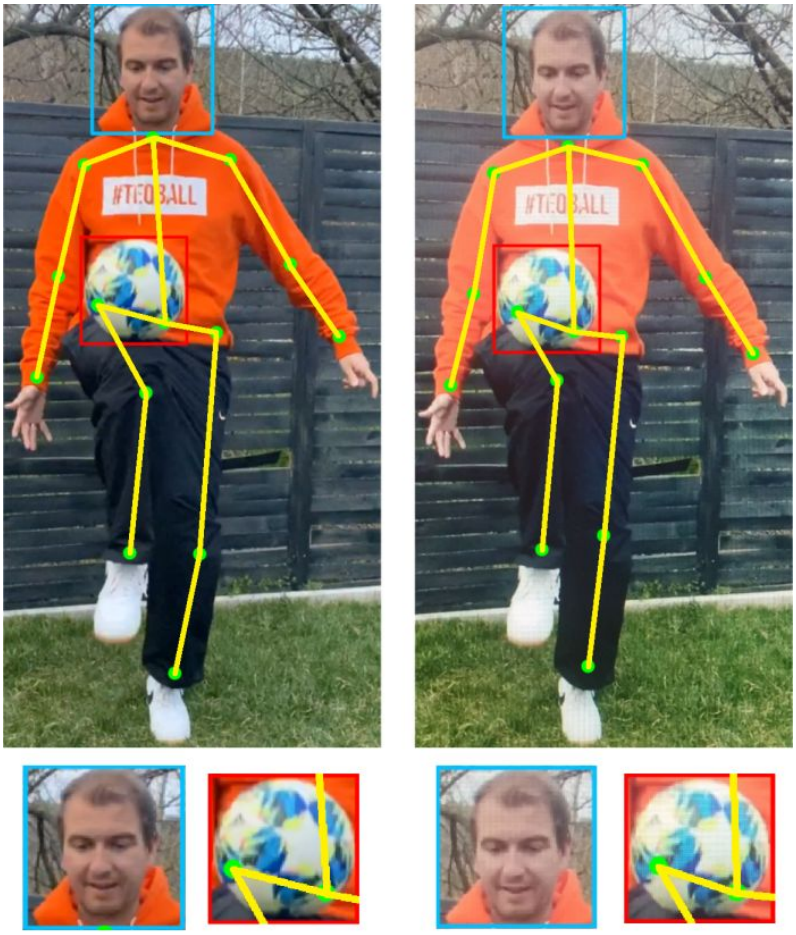

Figure 2.1: A sample event detection in a digital football game - user hits the ball with the right knee. Top row: left - detection on a single frame of a video that captured a real user, right - detection on the same frame of the same video captured on a computer monitor (fake user). Bottom row left to right: close-ups of the detected head and football of real and fake users respectively (original illustration by Viktor Huszár).

Despite the high recognition accuracy achieved by Human Activity Recognition (HAR) systems, they are still vulnerable to spoofing attacks where intruders can deceive the system by presenting fake visual data. One common type of spoofing attack is replaying a video of human activity on a digital screen. This is demonstrated in Figure 2.1, which shows how a deep learning algorithm employed in the system is unable to distinguish between real and fake human activities.

The presence of these spoofing attacks poses a significant security challenge for HAR systems. Fraudsters and intruders can come up with innovative ways to bypass the simplest security measures and penetrate computer vision systems. Therefore, it is crucial for future developments to effectively filter out such potential frauds and provide robust solutions for protecting against spoofing attacks.

The challenge of spoofing detection must also be addressed in accordance with existing infrastructural systems. This is because fraudsters and intruders can have creative ideas and solutions for circumventing computer vision systems and bypassing security measures. The high recognition accuracy of HAR systems makes it even more important to ensure the security of the system and prevent spoofing attacks.

Despite the advancements in Human Activity Recognition (HAR) systems, there are several challenges that need to be addressed. One of the challenges is related to the quality of the data that is collected by the sensors, particularly in the case of external sensors. For example, weather conditions like rain, fog, or snow may affect the image quality captured by cameras, leading to poor recognition results. Furthermore, the use of cameras for activity recognition may also be limited by visibility conditions, such as lighting and shadows.

Another challenge is the limited amount of data available for training deep learning algorithms, particularly Convolutional Neural Networks (CNNs). CNNs are data-driven algorithms, and thus require large amounts of data to learn and make accurate predictions. However, due to regulations such as the General Data Protection Regulation (GDPR), access to large datasets may be restricted, making it challenging to train the algorithms effectively.

In certain applications, such as military training and surveillance, the data is often classified and access to the data is limited by nature. This further complicates the process of training the algorithms and achieving accurate recognition results. The challenge lies in finding a way to train the algorithms effectively while adhering to data privacy regulations and ensuring the security of classified data. To overcome these challenges, it is important for researchers to explore alternative data sources and develop innovative techniques for data collection and processing.

The use of artificial intelligence in profiling competencies based on pre-defined risk criteria is an important aspect in the field of surveillance and security. This is because AI can help to process and analyze large amounts of image data from existing cameras to identify potential risks and security threats. The development of profiling competencies is based on various image analysis techniques, including object analysis and motion analysis. The objective of this process is to make use of the data captured by cameras to identify patterns and trends that can help to identify potential risks and security threats.

Moreover, the implementation of this technology requires the support of competent authorities such as the police, National Security Special Service, and other relevant organizations. The verification of the data obtained from AI-based systems can only be carried out with the support of these authorities, who can validate the results and take appropriate action if needed. Additionally, national public service universities, and military and police academies play a crucial role in defining the professional requirements and

methodology for further experimental development in the field of AI and security. These organizations have the expertise and resources needed to conduct research, analyze data, and develop best practices and protocols to ensure that the technology is implemented effectively and securely.

The task of spoof detection in the context of HAR is even more challenging due to the nature of the captured data. As users are always moving, this can lead to the generation of blurred images on captured video, making it difficult for spoof detection algorithms to accurately detect such attacks. Additionally, depending on the complexity of the action recognition, it may be necessary to stream visual data to a server for remote processing, which may involve video compression and resizing operations that further degrade the quality of the video and make the task of spoof detection even more challenging.

To overcome these challenges, I propose a deep learning-based approach for spoof detection in videos for HAR applications. My approach leverages the latest advances in deep learning and computer vision to accurately detect spoof attacks in real-time, even under challenging conditions. The proposed approach is specifically designed to be robust against the types of attacks that are commonly encountered in HAR applications and has the potential to significantly improve the overall security and accuracy of these systems.

In the current work, a lightweight deep learning-based algorithm that runs in parallel with HAR algorithms to detect and report cases of spoofing is proposed. Figure 2.2 shows the flowchart of my proposed system architecture. The incoming real-time visual data stream is fed in parallel to HAR and proposed spoofing detection algorithms. Irrespective of a successful or unsuccessful activity recognition using the HAR algorithm, on a positive detection of a spoof case, a spoofing attack is reported to the operator. If no spoofing attack is detected, further visual data is grabbed from the sensor for processing by HAR and spoofing detection algorithms. In particular, the contributions of this work are:

- A deep learning-based approach that can be used to detect spoof attacks from video frames capturing humans. The algorithm precisely detects the spoofing attacks and is also able to cope with video resizing and streaming artefacts, while it remains lightweight enough to run it on a mobile device alongside the main HAR algorithms.

- A strategy to combine detection of the proposed deep learning network, temporally on a captured video or on a live video stream for detecting video replay spoof attacks systematically while maintaining real-time performance.

- A new database comprising real and spoof videos captured from 30 different users

in different locations and under different lighting conditions. The database is diverse and precisely captures the required features for training my deep learning network.

- An additional evaluation of the performance of the proposed approach in the context of biometric recognition applications.
- An IOS mobile application that implements the proposed approach in real-time on a mobile device.

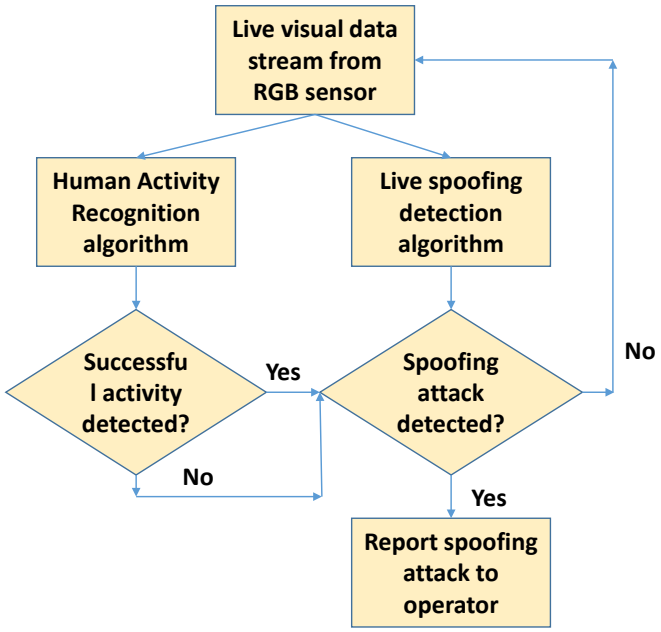

Figure 2.2: Proposed workflow for spoofing detection in HAR applications: A lightweight spoofing detection algorithm is run in parallel to HAR algorithm for live detection of spoofing attack cases (original illustration by Viktor Huszár).

2.2 Related work

There are vast varieties of techniques available for detecting replay attacks from visual data. Most of these are addressing spoofing detection in bio-metric recognition applications, predominantly using facial image analysis. Users are located close to the camera and looking towards the camera. To the best of my knowledge, my system is the first of its kind exploring spoof detection from videos in the context of HAR where users are located far away from the camera and not always looking at the camera. In the subsequent paragraphs, I present some of the more relevant works in the domain of spoof detection using facial image data.

In the field of replay attack detection, techniques proposed in the literature can be grouped into four major groups: user behaviour modelling, user cooperation, hardware-based, and data-driven characterization[30].

Behaviour modelling techniques are designed to track user actions, such as head movement and blinking. They are employed to determine whether the individual in front of the camera is a real user or a photograph of that user. Research in this field has also explored detecting subtle movements in a live human face using motion magnification, but this approach is not applicable to my current work as it only addresses the scenario where an attacker uses a photograph [31] [32]. In contrast, my focus is specifically on detecting video replay attacks, which require a different approach. Unlike traditional behaviour modelling techniques that are limited to identifying static photographs, my approach must be able to effectively detect and distinguish between real-time human activity and artificially generated video replays. To achieve this, I utilize advanced machine learning algorithms and techniques to analyze the video data and identify anomalies.

User cooperation-based techniques for detecting presentation attacks rely on secondary interaction between the user and the detection module [33]. These techniques typically require the user to perform a specific movement or action, which can serve as evidence that the user is indeed a real individual and not a video replay. However, in the context of automated video surveillance, it is often not practical to use user cooperation-based techniques as users may not be available or willing to engage in deliberate interaction with the system. In these cases, relying solely on user cooperation-based techniques would result in a less efficient and less reliable detection system. Therefore, it is important to consider other approaches that do not rely on user cooperation in order to effectively detect presentation attacks in automated video surveillance systems.

Hardware-based techniques make use of additional hardware such as infrared or depth sensors to understand the depth information of the scene [34] [35]. Within such techniques depth cues enable us to differentiate between a flat object such as a mobile telephone screen or real 3D objects such as humans. Thus, these methods provide better robustness against variations in illumination and pose of the user. However, in applications such as motion-driven games designed to work with smartphones, users may not always possess such additional hardware on their devices. Thus, the scope of the application for these methods is limited.

Finally, there are techniques that are data-driven and characterize the replay attacks by predicting/learning the artefacts/features of attempted attacks using the visual data that came from a standard acquisition sensor. This work focuses on such data-driven techniques using visual data obtained from fixed or mobile cameras. In the rest of this section, several techniques that are data-driven and relevant to the current work are discussed.

2.2.1 Depth analysis

There is a branch of techniques for spoofing detection based on scene depth analysis using optical-flow estimation. They intend to estimate the 3D structure of the face of a user to discriminate between a 3D live face and a 2D spoof face [36] [37]. Similar to depth camera-based spoofing detection, live faces are 3D objects and can be clearly differentiated from a 2D planar medium such as photographs. Such methods can be practically used to identify attacks from planar static mediums. But if the camera is not stationary or if the user is moving, obtaining reliable depth information for performing spoof detection can be very difficult and challenging. Moreover, depth or shape analysis examines several frames to make a single prediction which can make these approaches slow and resource intensive.

2.2.2 Texture analysis

The mediums used for spoofing such as paper or digital screens have different reflection properties than real and live faces. Texture analysis approaches make use of this observation for spoofing detection [38], [39], [40]. By modelling the interaction between the illuminating and the reflective surfaces, it is possible to extract albedo and normal maps [41], which can be used as features to discriminate between real and attack samples. These methods have faster response time than depth analysis-based approaches since they often examine one frame at a time. However, they have poor abilities to generalize to HAR applications, since there can also be reflective objects in real videos that can trick such approaches. Furthermore, in the case of HAR applications, users are dispersed and the lighting is mixed and uncontrollable. So, the basic assumptions do not hold in practice. Other texture-based approaches try to capture the high-frequency information using descriptors, such as a variation of local binary patterns (LBP) [42] [43].

2.2.3 Image quality

These methods estimate degradation in overall image quality which occurs during recapturing photographs or screens. Some factors that contribute to image quality degradation include blurriness and colour deformation. [44], [45], [46], [47]. To compute image quality, it is often needed to have a reference image that is not available. The usual approach is to simulate a degraded version of the source image and use it together with the source image for computing image quality. Here the hypothesis is that the objective quality degradation between the source and the simulated image is relatively less in cases of real images when compared to attack images. Thus, if acquisition conditions are not similar, this hypothesis does not hold, which makes these methods delicate to HAR scenarios.

2.2.4 Frequency domain analysis

Recapturing on digital spoofing mediums introduces high-frequency noise into the image data due to the discreteness of the presentation mediums. This frequency-specific information can be captured by Fourier analysis [48], [49]. Frequency domain-based methods explore these noise signals in the recaptured video to distinguish between live and spoof faces [50], [51]. These noise signals in the frequency domain are strong cues to detect attacks. However, it should be noted that using high-definition spoofing mediums can dampen this noise and the recognizable noise patterns are not always present which makes such approaches solely based on them unreliable.

2.2.5 Methods based on deep learning

In recent times, in the context of several image recognition applications, it is proven that models based on Convolutional Neural Networks (CNNs) achieved state-of-the-art performances. Such deep learning approaches learn to predict features or intermediate representations directly from the raw pixel data without the need for computing the features explicitly. Using CNNs for performing spoof detection in HAR applications is not represented in the literature. However, architectures are proposed for performing spoof detection, especially on videos in facial biometric authentication applications. Rodrigo B. et al., [30] proposed an approach using ResNet50 [52]. They train this network using several pre-computed maps such as depth, saliency, and illumination maps, which makes their approach context-dependent. Using such approaches on a mobile device can be cumbersome, since computing these maps can be highly resource intensive.

Yaojie L. et. al., [53] proposed an approach using a combination of CNNs and Recurrent Neural Networks (RNNs) for detecting spoofed faces. Particularly, they use RGB+HSV representation of images and use several video frames to make a single prediction. I argue that residual networks are usually slower since they work on multiple frames.

Atoum. Y. et. al., [54] introduced a two-stream CNN which computes a) local features from patches and b) depth maps and combines the output of these streams for replay attack detection. Using local features makes these methods robust. However, due to depth map computation, they may not be able to keep up with the real-time requirements, especially, when an activity recognition algorithm also has to run in parallel on a mobile device. In my current work, I also target the mobile device scenario where lightweight models are highly favoured.

2.2.6 Summary of state-of-the-art methods

Table 2.1 summarizes the performance of methods published in the literature for face spoofing detection on public datasets. Except for deep learning-based methods, intra-database results for the Idiap Replay-Attack and UVAD databases are results are reported using the Half Total Error Rate (HTER) (%) metric. Intra-database results for the CASIA FASD, MSU MFSD, and MSU USSA databases are reported using Equal Error Rate (EER) (%) metric. For the ZJU Eyeblink database classification accuracy is reported. Unless otherwise specified, all cross-database results are reported using the HTER metric.

Preliminary methods for spoofing detection using approaches such as blinking detection fail under video replay attacks. Other slightly advanced methods based on low-level texture descriptors and simple classifiers for spoofing detection are also shown to fail under challenging cross-dataset protocols [55]. It is important to note that among previously developed methods, and most of the modern deep learning methods, for video replay spoofing detection use public datasets for experimenting and comparing different methods. I argue that there are several inconsistencies among the existing public datasets for video replay spoofing detection using facial image data in terms of the subject position, image resolutions, and cameras used for dataset generation. Cross-dataset evaluations in literature have shown limitations of both the developed methods and existing datasets.

I also argue that spoofing detection datasets for biometric or live face detection are not relevant to HAR applications since they generate data in static sessions with minimum illumination variability. In contrast, in HAR applications, users can have the flexibility to freely choose their location, and eventually, it is not always possible to control user behaviour. Such uncontrolled behaviour can happen in school campuses, other institutions and military sites. Besides, in some cases such as virtual games, some of the body parts including the face are partially or completely occluded by objects for interaction such as a football. Methods developed in the literature for spoofing detection in bio-metric face authentication applications may not extend to such cases. On top of that, if the activity recognition involves remote server processing, the image data may be subjected to resizing and/or compression. The dependency of such operations on spoofing detection accuracy is not discussed in the literature. Finally, the state-of-the-art deep learning-based methods for spoofing detection using facial image data involve pre-processing steps such as face alignment after face detection which is impossible in HAR use cases since the head position of a user is not fixed. Also in the literature, there are no such meticulous studies using deep learning methods targeted at mobile devices.

To address all of these issues, in this work, I developed new tools including a novel database and robust data-driven approach that jointly examine various regions of the

Table 2.1: Performance of published methods on face spoofing detection. Please refer to the text concerning the metrics used for evaluation.

Method	State-of-the-art performance
Face motion analysis	**ZJU Eyeblink** [32] (Intra-DB, Cross-DB) [32]: (95.7, n/a) **Idiap Replay-attack** [38] (Intra-DB, Cross-DB) [31]: (1.25, n/a) **CASIA FASD** [56] (Intra-DB, Cross-DB) [57]: (21.75, n/a)
Depth analysis	**Idiap Replay-attack** (Intra-DB, Cross-DB) [58]: (12.5, n/a)
Image texture analysis	**Idiap Replay-attack** (Intra-DB, Cross-DB) [55]: (15.54, 47.1); [59]: (0.8, n/a); [60]:(2.9, 16.7); [61]:(1, n/a) **CASIA FASD** (Intra-DB, Cross-DB) [60]: (6.2,37.6) ; [61]: (7.2, 30.2)
Image quality analysis	**Idiap Replay-attack** (Intra-DB, Cross-DB) [62]: (7.41, 26.9); [44]: (15.2, n/a); **CASIA FASD** (Intra-DB, Cross-DB) [62]: (12.9, 43.7) **MSU MFSD** [62] (Intra-DB, Cross-DB) [62]: (5.82, 22.6)
Frequency domain analysis	**Idiap Replay-attack** (Intra-DB, Cross-DB) [51]: (2.8, 34.4) **CASIA FASD** (Intra-DB, Cross-DB) [51]: (14, 38.5) **UVAD** [51] (Intra-DB, Cross-DB) [51]: (29.9, 40.1)
Deep learning based methods	**PR FASD** [63] (Intra-DB, Cross-DB) [63]-ResNet: (n/a, 14.19); [63]-DenseNet: (n/a, 16.97) **MSU MFSD** (Intra-DB, Cross-DB) [63]-ResNet: (n/a, 20.53); [63]-DenseNet: (n/a, 21.78) **Idiap Replay-attack** [38] (Intra-DB, Cross-DB) [63]-ResNet: (n/a, 31.42); [63]-DenseNet: (n/a, 31.52)

captured images to detect spoofing. My system enhances and integrates state-of-the-art deep learning algorithms for detecting replay attack spoof cases for HAR applications.

2.3 Visual analysis for video replay spoofing detection

Here, I elaborate on my method based on training a CNN model that distinguishes between genuine users and spoof attacks. For experimenting as well as for describing my system, I consider the instance of motion-driven smartphone-based virtual games, in particular, SQILLER [27]. However, I highlight again that, my approach is not limited to this specific setting and can be used in higher security demanding applications such as military grounds, defence and border control. Also, for the sake of simplicity, I consider a single-user scenario. My approach is scalable and can be easily extended to multi-user scenarios.

2.3.1 Context selection for spoofing detection

Generally, depending on the user activity to be recognized and its complexity of it, activity recognition algorithms involve computing locations of one or several body parts/joints or in some cases, complete skeleton tracking of a user from an image [64]. In the case of virtual games using an interaction object such as a football, it is also needed to compute the location of the football in an image together with the user body parts, since here, the activity involves the interaction of the user with the football (see Figure 2.1). For this, I have created a diverse database consisting of almost 50500 full HD images of 38 users juggling a football in different backgrounds and lighting conditions. For collecting data, I selected both male and female users between the ages of 8 and 40. These images are extracted from several videos shot using various iPhones - iPhone 6, 6S, SE, 7, 8, X, and XS. Some representative images from my database are are shown in Figure 2.3. These images are manually labelled with bounding boxes that encapsulate the following data: human body parts - head, left shoulder, right shoulder, left elbow, right elbow, left hand, right hand, left hip, right hip, left knee, right knee, left foot, right foot and also the bounding box of the football. After gathering the database and labels, I trained YOLO [65] deep learning CNN architecture and later used it for detecting the required body parts together with the ball from the video frames in a single shot. I chose well-proven YOLO state-of-the-art CNN for this task because of its fast and robust performance even on mobile devices. Note that, in contrast to traditional face detection methods, I generate face bounding boxes that also contain a substantial background image to provide more contextual information.

Figure 2.3: Representative images from my video database. Videos consist of several users playing a digital football game in diverse locations with varying lighting and backgrounds (original illustration by Viktor Huszár).

In order not to overload the final application for activity recognition with the additional functionality of spoofing detection, I consider the already computed body parts location data for this purpose. Among the detected bounding boxes by my trained YOLO model, the head bounding box, comprising the human face is uncovered most of the time and is composed of different interesting structures and characteristics. Due to the wide usage in bio-metric authentication systems and given its potential in many applications such as surveillance, home, and institutional security and border control and military, to experiment further with spoofing detection, I chose to use the extracted head bounding box data from the trained YOLO model. Note that this also ensures the generalization of my approach to other HAR applications where there are not any similar interaction objects and also to cases where the lower human body is not visible.

2.3.2 Models

While working with videos, there can be several ways to combine the temporal information which makes it hard to use fixed-sized architecture. In my work, I handle videos as several short clips consisting of identical-sized frames. The idea is to use these several frames of a clip to learn spatiotemporal features. For the experiment, I consider a network inspired by one of the ImageNet challenge winning models, VGG16 net [66]. I chose this architecture in my experiments for various reasons: Firstly, the VGG blocks are smaller networks, support real-time performance, and are light enough to be embedded into mobile devices. Secondly, it is proven in the literature that VGG16net has the potential for representing complex visual relationships [67]. Finally, the VGG network offers flexibility to change input size and alignment which makes it easier to adopt and modify this network for experimenting with several input data dimensions. Particularly, I consider two VGG blocks, each comprising two 3X3 convolution layers followed by a 2X2 pooling layer for the experiment. The output of these VGG blocks is connected to two fully connected layers. The final layer is connected to a softmax classifier with dense connections (see Figure 2.4).

I carefully chose this simple model to achieve better run-time performance while still retaining the detection accuracy. I implemented my system in Keras [68]. I consider the CNN in 2.4 as baseline CNN and investigate approaches to combine information across the temporal domain (see Figure 2.5). Here I describe my trails for learning the spatiotemporal features.

Single frame model (SF)

For studying the potential of static frames in accurately classifying the genuine and spoof cases, I consider this model. SF takes every frame of a video and outputs the observations.

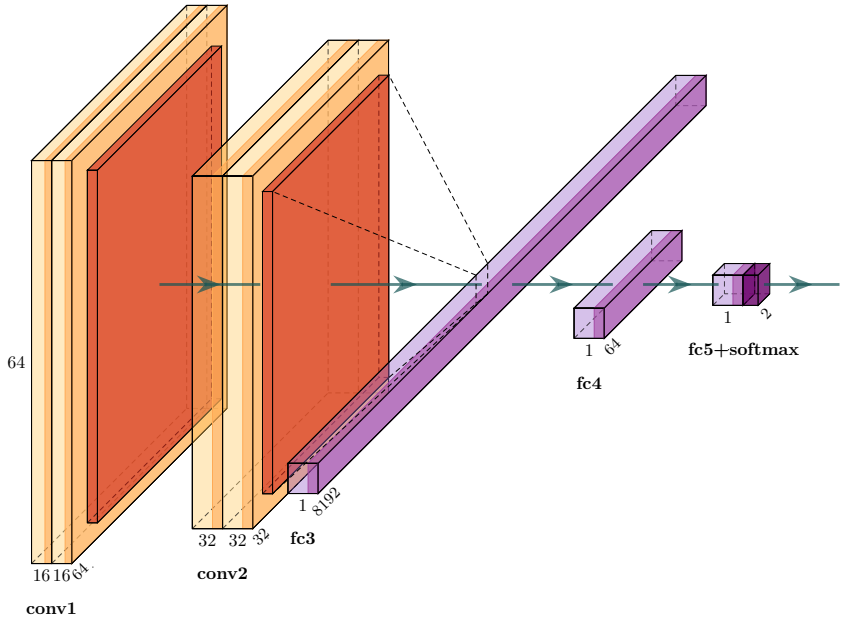

Figure 2.4: CNN architecture inspired from VGG16 net used for my experiments (original illustration by Viktor Huszár).

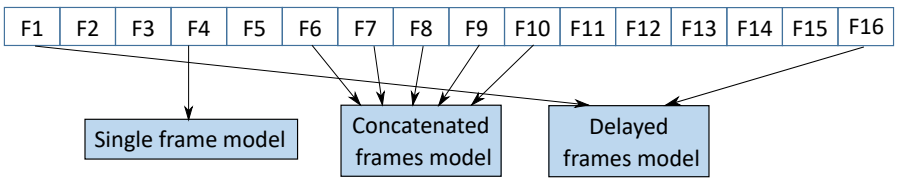

Figure 2.5: Some of the explored approaches for capturing spatiotemporal features: Single frame model processes one image at a time; Concatenated frames model processes consecutive 5 frames at a time; Delayed frames model processes two frames at a time that are 15 frames apart temporally (original illustration by Viktor Huszár).

Here, the extracted face bounding box from the current input Full HD frame is resized to 64X64X3 pixels and then fed to the aforementioned baseline CNN.

Concatenated frames model (CF)

The approach here is to combine the visual information across a time window containing contiguous frames in a video clip. This is achieved by adapting the filters on the convolutional layer of the first VGG block in the baseline model. Similar to the SF, the consecutive frames across the considered time window are resized to 64X64X3 pixels and the initial convolutional filters are extended to be of size 64X64X3XN pixels, where N is the size of the considered time window. I choose N=5 in my work for learning and detecting local patterns across the time window.

Delayed frames model (DF)

The delayed frames model uses two separate baseline models until the second VGG block and these two models is fed with two 64X64X3 resized frames that are P frames apart in the given video sequence. After the second VGG block, the outputs of these two models are merged together and connected to the rest of the fully connected layers on the baseline model. In this work, I use P=15 for learning and detecting global features.

Ensemble multi-stream model (EM)

The ensemble model investigates the resized head bounding box in three different streams of processing over three spatial resolutions. The approach is presented in Figure 2.6. Similar to the SF model, I consider one frame at a time. From the input HD video, single frames are extracted and fed to the trained YOLO model to extract the head region on the image. The detected head bounding box is resized to 64X64X3 pixels and processed in three different streams.

In the first stream, the resized head image is fed to the baseline model (same as SF). In the second stream, lower 64X32X3 pixels of the resized head image is cropped and supplied to the modified baseline model for processing. On the initial convolutional layer of the baseline model, filters are modified to be of the size of the cropped image in this stream. This stream mostly gets the visual data around the chin area and this way, bias coming from any users wearing eyeglasses or caps is minimized. The last stream is fed with the cropped central 32X32X3 pixel data from the resized image. This data is processed by another modified baseline model that has the required initial convolution filter size. The stream processes the central crop of the face and this way it has no bias from any disturbing patterns from the background. The detections from these three

streams are manually fused based on the detection probabilities and majority voting for arriving at an ensemble detection.

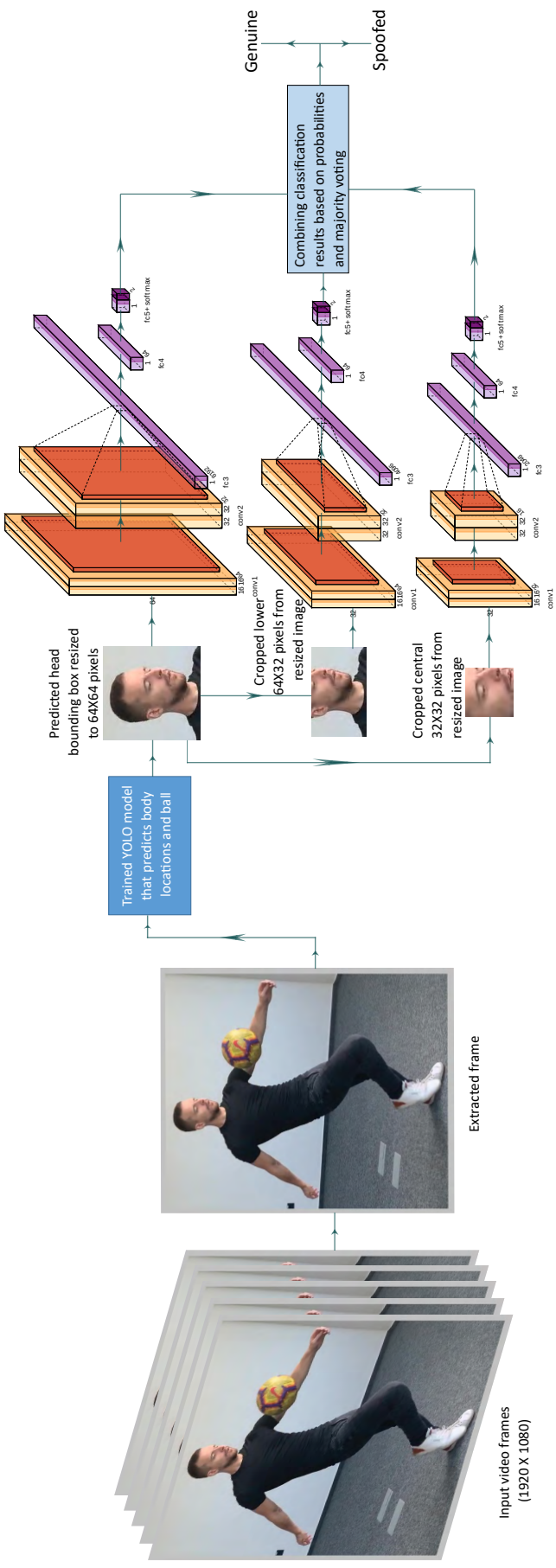

Figure 2.6: My ensemble multi-stream model: Frames are extracted from videos and one frame is processed at a time in three different streams. The output of the three streams is combined to derive the final classification results (original illustration by Viktor Huszár).

2.3.3 Learning

Data preparation and pre-processing

For learning the genuine and spoof cases, I used the same dataset used for training the YOLO body detection containing 50500 Full HD images obtained from 38 users in 38 videos playing Sqiller. For training my models, I generated an additional 50500 Full HD spoofed images using the original images by capturing the same videos on several monitors - a 27-inch Dell 4 Kilo (4K) monitor, a 15-inch Full HD Lenovo laptop monitor, a 13-inch MacBook Pro monitor with the resolution of 2560×1600. Before training the networks, I pre-process all video frames to extract the head region using the previously trained YOLO model and resize the head image to 64X64 pixels. I also use data augmentation techniques to reduce problems from overfitting - increase and decrease the pixel brightness by a value of 50, double and half the image contrast, and add Gaussian noise with variances of 50 and 100. These pre-processing steps are consistently added to all the video frames. Finally, before training, I shuffle the data and scale all the pixel intensities to the range [0, 1].

Optimization

I use Adam [69] available in Keras to optimize my models with the initial learning rate of $1e^{-4}$. The model is compiled with a binary cross-entropy loss function. For training, I use a batch size of 12 samples and 20% validation split. Models with improved validation accuracy are automatically saved every 100 epochs to retain the best models.

2.4 Results and discussion

2.4.1 Testing scheme

Since I are targeting to detect spoof cases on a continuous basis either on recorded or live video streams, I define chunks of overlapping video clips, each containing several frames (depending on the model under testing) with an overlap of 15 frames. For each clip, I provide a single prediction. To this end, within a considered clip, I run my models on three frames that are 15 frames apart and combine these predictions for obtaining a single prediction for this clip. Therefore, given a video stream, in the case of SF and EM models, the first prediction is made after 31 frames, and thereafter, predictions are made every 15 frames. In the case of the DF model, since it takes as input two frames that are 15 frames apart, the first prediction is made only after 45 frames and thereafter, predictions are made every 15 frames. In the case of the CF model, since it takes 5

consecutive frames as input, the first prediction is made after 35 frames and thereafter, predictions are made every 15 frames. Figure 2.7 demonstrates the testing schemes more distinctly.

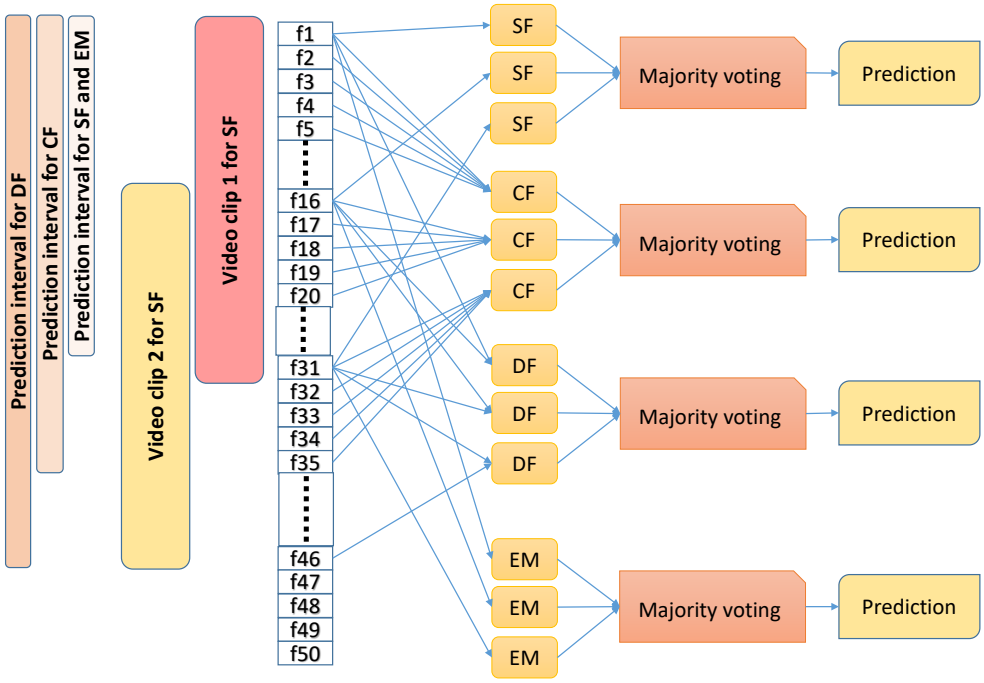

Figure 2.7: Schemes used for testing the proposed models: all models are tested on overlapping video clips. Within each video clip, three predictions are made and majority voting is applied to derive a prediction for this video clip (original illustration by Viktor Huszár).

2.4.2 Experiments on my dataset

As mentioned earlier, to the best of my knowledge, my system is the first of its kind to explore spoof detection from videos in the context of HAR. Since also as mentioned before, I prepared a dataset containing 101000 images, obtained from 38 different players. These images contain users located at varying distances from the camera so that all the body parts are visible. The face orientation of players widely varies across videos and sometimes the face is completely or partially occluded by the ball. I manually annotate these images with head bounding boxes containing facial and background information of the players which is then used for training a YOLO model that detects head regions in an image in real time. Figure 2.8 shows the histogram of InterPupillary Distances (IPD) on all the images used for training my models (64X64). With the highest number of samples at around 44 pixels, my database also contains samples where IPD is evidently located between 0-50 pixels which show the distribution of how far away users are located from

the camera. 0-pixel IPD may mean either the face is not completely visible or a player is facing in an orthogonal direction to the camera.

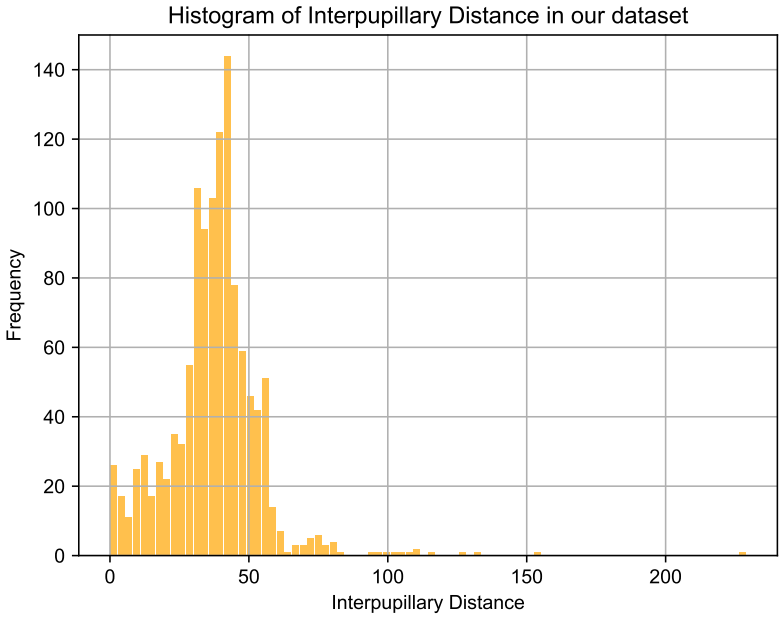

Figure 2.8: Histogram of interpupillary distances across various players in my database (original illustration by Viktor Huszár).

Training

I split the database by assigning 80% of the videos to the training set and 20% to the test set. To this end, I separate videos from 8 players in my database and use them for testing. All the models are trained for 1000 iterations and training each model took about 24 hours on average. Among the testing videos, predictions are made for each video clip containing 30 frames irrespective of the length of the video. To report results, I use results collectively from all video clips in the testing split.

Video-clip level predictions

I used videos from 8 users (20% of the whole database) for testing that are not included in the training. In these videos, users performed similar actions using football as in the training set in challenging indoor and outdoor lighting conditions. A fake database for the test set is also created using random monitors that are mentioned in section 2.3.3. Similar to the training set, the test set also contains videos from users who are both male and female in the age group 8-40. I report my results using a set of metrics commonly considered in bio-metric presentation attack evaluation - Attack Presentation Classifi-

cation Error Rate (APCER), Bona fide Presentation Classification Error Rate (BPCER) and Half-Total Error Rate (HTER) [70]. APCER and BPCER are analogous to False Acceptance Rate (FAR) and False Rejection Rate (FRR). Thus lower values of APCER, BPCER, and consequently, HTER indicate better performance of a model. Table 2.2 presents the results of the various described models on the testing split. For reporting these results, the outputs of the models are used as obtained and thresholding based on class prediction probabilities is not applied.

The ensemble model outperformed other methods by a considerable margin. This can also be validated by analyzing the trade-off between False Positive Rate (FPR) and True Positive Rate (TPR) on the testing set (see Figure 2.9. Higher value of Area Under Curve (AUC) represents the better performance of a model). Note that the single-frame method also provides reliable results. However, my experiments show that combining the frames across time did not yield better results. Particularly, in the case of the CF model, the APCER value is high indicating that several fraud users are accepted as valid users by this model. This suggests that combined frames may not always contain patterns that can be sufficiently characterized by my deep learning model and may be completely arbitrary. In the case of the DF model, the BPCER value is high indicating that several real users are categorized as fake users by this model. From this behaviour, I hypothesize that the DF model learned features accounting more to motion than to the patterns related to spoofing detection. It is important to note that my detected head region contains visual information about the background or the scene as well as the user costume. The cropped lower region contains lesser visual information about the background but contains information about the user's costume, and cropped central region contains mainly visual information about the user's face. I have observed that specular, glossy surfaces in the background, spectacles and face masks of users, and dense striped costumes can mislead spoofing detection. By utilizing three cropped regions extracted from the detected head bounding box, I believe that the EM makes smaller errors in spoofing detection than other methods used for comparison.

Table 2.2: Results (%) on my testing database

Method	**APCER**	**BPCER**	**HTER**
SF	9.9010	12.9496	11.4253
CF	36.4486	1.7007	19.0746
DF	3.1915	45.5556	24.3735
EM	8.9109	6.1151	**7.5130**

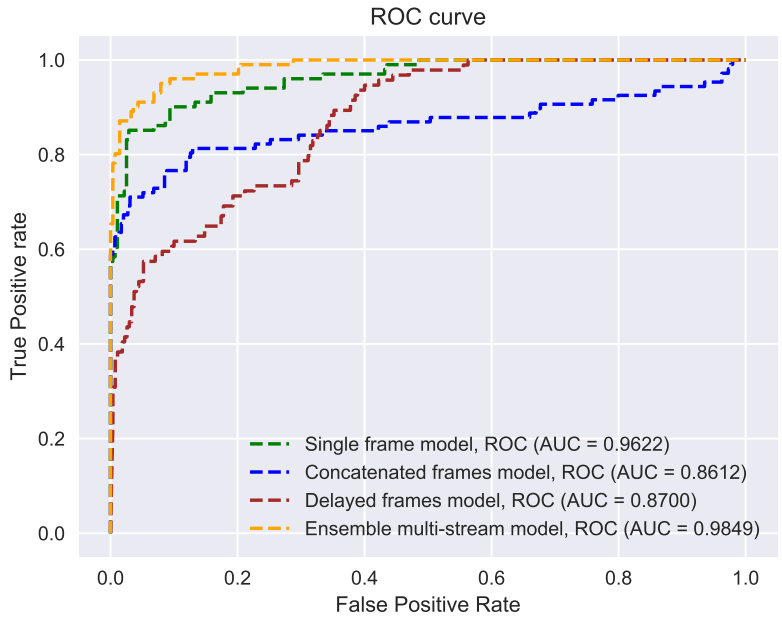

Figure 2.9: Receiver operating characteristics (ROC) of the methods evaluated on my testing database. Compared to other models, the Ensemble multi-stream model produced better classification results (original illustration by Viktor Huszár).

Comparison with existing baseline spoofing detection methods

For assessing the effectiveness of my ensemble method, I consider recent baseline spoofing detection methods for comparison. Please note that, as pointed out earlier, there are no relevant data-driven approaches in the literature for spoofing detection from videos when users are performing simple or complex activities and are free to choose their position. Particularly, I consider the following four descriptors for the experiment, which are reported as potential candidates in the literature for spoofing detection applications using facial image data: LBP histograms, Local Binary Pattern histograms from Three Orthogonal Planes (LBP-TOP) [71], Statistical Binary Pattern histograms (SBP) [72] and Statistical Binary Pattern histograms from Three Orthogonal Planes (SBP-TOP) [73]. Also, it is reported in the literature that the Support Vector Machines (SVM) classifiers achieved better accuracy than random forest classifiers in spoofing detection applications [73].

Consequently, for comparison, I have trained four different SVMs using the above-mentioned four descriptors. For training SVMs, I used the sci-kit-learn toolkit. While calculating these descriptors, I use radius, R = 1, and a number of points P = 8 (see [73] for more details on these parameters). For calculating SBP, I considered two orders of moments obtained with the mean and variance. For computing the orthogonal planes, I consider 3D frame stacks containing 20 consecutive video frames. For acquiring joint

histograms in the case of SBP and SBP-TOP, as described in [73], two binary images are computed additionally by thresholding the input or moment images with respect to their average value. For testing, I used the same testing dataset that is used for comparing the proposed methods. Table 2.3 presents the results from the baseline methods in the literature together with the results from the EM method using my testing dataset. Experimental results show that my EM method outperformed other methods by a considerable margin. My observations correspond to the results presented in [55] - that approaches using low or medium-level texture descriptors and training simple classifiers using these descriptors for spoofing detection are not effective under cross-dataset protocols.

Table 2.3: Results of existing baseline spoofing detection methods (%) on my testing database after training using my training database.

Method	APCER	BPCER	HTER
LBP	40.5780	48.1586	44.3683
LBP-TOP	27.2727	40.3226	33.7977
SBP	**3.1915**	45.5556	24.3735
SBP-TOP	15.4124	78.9660	47.1892
EM	8.9109	**6.1151**	**7.5130**

Experiments with video compression

As mentioned earlier, depending on the complexity of the action to be recognized, it is needed to stream the video to a remote server where the actual processing of the video stream happens. In cases like motion-driven virtual games, where user videos are streamed from a mobile telephone to a remote server, there is not always enough network bandwidth to stream the video in its native quality or resolution. Subsequently, video compression techniques are applied to reduce the video bitrate which introduces video artefacts. For evaluating the ability of my EM model to correctly classify the spoof cases, I generate compressed video streams with varying bitrates from 300 kbps to 1500 kbps.

For this experiment, a recorded video is considered and evaluation is done completely offline. Multiple videos are generated with different bitrates using FFmpeg [74]. Figure 2.10 shows the AUC values under various bitrates. Results show that my ensemble model performs robustly even under extreme compression (300 kbps).

2.4.3 Experiments with face recognition databases

To evaluate the ability of my spoofing detection model to generalize to other domains such as face recognition systems, I experiment with two widely known datasets in this context: Idiap REPLAY-MOBILE [75] and CASIA Face AntiSpoofing [56]. Particularly,

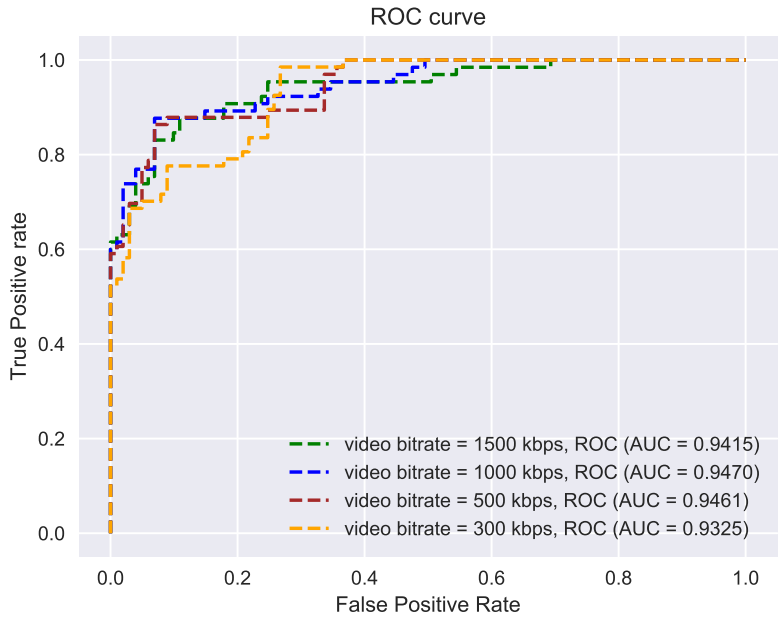

Figure 2.10: Evaluation of the sensitivity of my ensemble multi-stream model to video compression rates: the model produces robust results even at lower bitrates (higher compression) (original illustration by Viktor Huszár).

I identify the spoofing cases that occur in video replay cases and evaluate the performance of my algorithm on these attacks. The CASIA Face AntiSpoofing Database consists of 600 video clips of 50 subjects. Out of the 600 video clips, 150 clips represent video replay attacks. Compared to the Idiap database, the CASIA DB provides images captured using a variety of cameras (Sony NEX-5-HD, two low-quality USBs) to capture replay attacks displayed on an iPad. However, a significant deficiency of this database is that the video replay attacks are captured in very low resolution (640 X 480). The REPLAY-MOBILE dataset consists of 1190 video clips of photo and video presentation attacks (spoofing attacks) to 40 clients, under different lighting conditions. These videos were recorded with an iPad Mini2 (running iOS) and an LG-G4 smartphone (running Android) in full HD resolution.

Table 2.4 shows the evaluation results on testing sets of the considered face recognition databases. The facial images are cropped using the supplementary data obtained from the database and then resized to 64X64 pixels before feeding to my EM model. The test results are reported using my testing scheme (see Figure 2.7). My results show that my approach performs well on the REPLAY-MOBILE database irrespective of the fact that the users are located very close to the camera (higher IPD values than that of my database). My method also performs reasonably well on the CASIA database. The relative drop in the performance in the case of the CASIA database is due to the low-

resolution input images. Note that my method is trained on image data that is extracted from full HD input images and I hypothesize that my learned feature set does not correspond very well to images that are captured at lower resolution. Note that, this may not be a problem since the current generation sensors used in HAR applications, including surveillance systems, are capable to capture full HD images.

Table 2.4: Results (%) on considered face recognition databases

Method	APCER	BPCER	HTER
REPLAY-MOBILE	5.4678	13.7856	9.6267
CASIA	27.0916	16.0396	21.5656

2.4.4 Mobile implementation and performance

I implemented my Ensemble multi-stream model in swift for IOS [76]. The application is standalone and tested on iPhone 8 released in 2017 and has a 2.39 GHz Hexa-core 64-bit. My Keras models as well as the YOLO model for head bounding box detection are converted to CoreML API for running on the IOS device. The application takes as input incoming frames from the on-device-camera and runs my trained YOLO model to detect head bounding box information. The head bounding box was then resized to 64X64 for ensemble multi-stream processing. I follow the strategy shown in Figure 2.7 for processing the incoming frames for detection. The algorithm works in real-time.

The proposed method can run on iPhone 8 and processes a single frame on an average in 4 milliseconds. This translates to a frame rate of about 250 frames per second. Following my testing scheme shown in Figure 2.7, running EM every 15 frames further ensures that there are enough resources left on the device for running the activity recognition applications.

2.4.5 Performance Constraints

I compute head bounding box information using a pre-trained YOLO model with fixed anchor sizes for bounding boxes. The anchor sizes are chosen to assume that a user is located at a distance range of 0.5-2.5 meters away from the camera. Also, for training, I use those samples when a user is located within this range. If a user moves out of this range, bounding boxes may still be detected, but with less probability. However, due to fixed-size anchors, in such cases, the bounding box would contain relatively more information about the background. My algorithm fails to detect spoofing in such a situation.

my spoofed database contains samples captured from cameras held at angles close to zero (almost held vertically) along the camera Z axis (depth axis). My rigorous testing shows that if the recording camera (camera used for action recognition and not the

original camera that captured video for spoofing) rotates beyond the range of -15 to 15 degrees along the camera Z axis, my algorithm fails to detect spoofing attacks.

Also, my database consists of images of a user where at least one of the eyes is visible (the user can face sideward perpendicular to the direction pointed by the camera) or the face is covered with a football which is my considered object for interaction. If a user faces completely away from the camera (only hair or a bald head is visible), my algorithm also fails to detect spoofing in such cases.

Under natural lighting conditions, my EM method can precisely detect spoofing cases. To deal with artificial lighting conditions, I also captured enough samples in my database, where a scene is lit by a bulb having a flickering frequency as observable by cameras. However, thorough testing shows that if more than one light source is in the scene, my algorithm has issues detecting the spoofing cases.

2.5 Concluding remarks

The study investigates video replay spoofing detection in human activity recognition using RGB sensors. An ensemble multi-stream model was proposed to detect attacks by examining multiple regions of the face. The model's performance was evaluated using 38 subjects under diverse conditions and showed robust results even with partial face visibility. The algorithm was validated on a mobile device and demonstrated real-time performance with a minimal memory footprint. Future work may involve adapting the method to devices with advanced sensors and studying algorithms for detecting face and body coverings designed to confuse spoof detection systems.

Chapter 3

Violence Detection in Surveillance Videos

3.1 Introduction

The current landscape of video surveillance has shown the importance of having an efficient and effective way to monitor and detect anomalies in the massive amount of video data generated every day. The use of human labour to monitor this data is impractical and susceptible to human error, which reduces the efficiency of the process. In order to prevent crime and ensure public safety, detecting and recognizing incidents of violence from video data in real-time is a crucial task for law enforcement agencies. To tackle this challenge, there is a growing need for automatic and effective methods that can filter out normal activities from abnormal events in video data. These methods must be able to process large amounts of video data in a timely manner while being able to detect events such as violence that occur infrequently. The focus of this work is on detecting violence and other abnormal events between humans. The proposed solution utilizes advanced artificial intelligence algorithms to process and analyze the video data, identify anomalies, and highlight incidents of violence in real-time. This approach helps law enforcement agencies respond quickly to potential incidents and reduce the crime rate while improving public safety.

Video classification based on Human Activity Recognition (HAR) is a popular research topic in recent years and is analogous to the field of violence detection. HAR methods provide information on simple or complex physical activities of humans such as standing, talking, and cooking from incoming sensor data. Earlier methods for HAR followed the detection and tracking of human body parts in a set of consecutive video frames using image level descriptors such as Histogram of oriented gradients (HOG) or Histogram of oriented optical flow (HOF) [77]. Other advanced approaches involved

computing spatiotemporal descriptors for motion [78] [79]. One of the main drawbacks of such techniques is that they often need better lighting conditions and good visibility for successful operation. With the advent of depth cameras, several algorithms have emerged, that use depth measurements from sensors such as Microsoft Kinect [80] [81], ASUS Xtion2 [82] or Intel real sense [83] for HAR. One of the added advantages of such depth sensors is that they come with Software Development Kit (SDK) containing real-time skeleton detection algorithms [84]. Specifically, the skeleton joint coordinates can be obtained in 3-Dimensions (3D) in real-time, and a series of these coordinates, when tracked over time, can be used to detect and describe human actions. Following this notion, several algorithms have been proposed in the literature for HAR using depth sensors [85] [86] [87] [88] [89] or combining information from both colour and depth sensors [90]. However, the depth sensors, even the most modern ones come with substantial measurement noise and thus it's not possible for them to yield better results. Moreover, integrating depth sensors into use-cases such as surveillance increases the hardware costs and may not be always favourable.

Usage of Convolutional Neural Networks (CNNs) has become increasingly common in computer vision after their exceptional success in image recognition tasks [91] [92]. CNNs are evolving at a tremendous pace in many fields of research, and recent trends show that many solutions are expected in the future that will enhance the spread of such solutions. With the occurrence of big data and the exponential growth of computing power, these learning algorithms continue to have enormous development potential. Several successful methods have been recently proposed that extend the spatial CNNs that are used for image recognition tasks along the temporal domain for HAR in videos [93] [94] [95] [96] [97] [98] [99] [100]. One of the main advantages of using CNNs for HAR is that these methods are able to deal with defiances, such as changes in lighting conditions, background changes, camera movement, different dressing styles, and varying body shapes of people. They are also able to deal with videos involving partially or completely occluded human body parts.

It is important to note that the prevalent HAR algorithms that are based on CNNs follow supervised learning. In particular, during the training process, for every example in training data, CNNs are provided with the correct expected output, also known as labelled inputs and outputs. In supervised learning, using the labelled data, CNN models constantly estimate accuracy and learn over time. Given that the size of the training data is considerably large, such CNNs are able to learn to classify videos into action classes and also use this knowledge to generalize [101] to other unseen examples. However, if sufficient data is not available for training, generalization to new data is also likely to be poor[102].

Transfer learning or domain adaptation [103] refers to the situation where what has

been learned in one setting (i.e., distribution P1) is exploited to improve generalization in another setting (say distribution P2) [101]. Specifically, in transfer learning, a neural network model is first trained on another task that is similar to the current task that is being solved. Later, one or more layers of the trained model (on another task) are used to train the problem of interest. Transfer learning is especially useful when I have less number of samples for training CNNs on a problem of interest. I argue that the violence detection task falls into this category where I have relatively fewer real-world video samples available for training due to laws such as General Data Protection Regulation (GDPR) in Europe [104].

In this work, I address the problem of violence detection using transfer learning. Explicitly, my contributions from the current work are:

- A deep transfer learning-based method that can be used to filter violent and normal events from human actions in videos. The algorithm outperforms several state-of-the-art methods for violence detection and is also able to cope with video compression artefacts while remaining lightweight enough to run in real-time on a laptop or desktop computing device.

- A comprehensive database comprising violent and normal videos that combines and extends 7 different existing video databases, captured in different locations and under different lighting conditions.

- An extensive evaluation of the performance of the proposed approach on collected databases and cross-database validation to investigate model over-fitting.

- A stand-alone application that implements the proposed approach in real-time on a laptop computing device.

3.2 Related Work

In this section, I describe the several classes of algorithms proposed in the literature for violence or aggression detection using deep learning. Due to the lack of a substantial amount of labelled data containing a large number of real-world violence samples, many results reported in the literature used training data containing only a few videos. For some datasets, the exact time instances when the violence happens are not available. Algorithms trained on such data often strive to minimize unusual patterns among training samples to learn about rare violent activities [105] [106] [107] [108] [109]. These methods are described in the subsections 3.2.1 and 3.2.2.

3.2.1 Modelling normal patterns

This class of techniques learns patterns about normal behaviour from training videos containing no violence. Since only normal videos are used in the training phase, no specific labels are provided. During testing, these methods are expected to find samples that deviate from normal behaviour [110]. In [111], and [112], motion trajectories are used to learn about normal patterns. In [111], the authors suggest representing motion patterns using super-trajectories that describe the motion of local groups of similarly moving points and cluster these motion patterns hierarchically to derive prototype patterns for normal samples.

In [113] and [114], the authors use auto-encoders to learn regularities in video sequences. Auto-encoders are CNNs comprising of an encoder and decoder to efficiently learn about intrinsic and hidden patterns in the form of sparse feature representations from input data. A common assumption in the methods employing auto-encoders for violence detection is that the trained auto-encoder can reconstruct the motion patterns present in normal videos with low error but will not accurately reconstruct motion patterns in violent videos. As an input to the auto-encoder, they use state-of-the-art handcrafted motion features that consist of HOG and HOF with improved trajectory features [115]. In [116], authors used optical flows along with video sequences and construct multiple auto-encoders using reconstruction loss [117] to detect abnormal or violent events.

Authors in [118], [119] and [120] also incorporated auto-encoders to learn normal behaviours, but without explicitly computing local motion patterns. With this one-stage approach, not needing to deal with the object and feature detection, such methods are quicker in terms of computational speed. There were also approaches that augment memory modules [118] and/or optical flow images [121] [122] to the auto-encoders. In [118], authors augmented the output of an encoder, in a variation of auto-encoder CNNs, with a memory module that adaptively records prototypical patterns of normal data for more accurate detection of violent cases in a given database.

In [120] and [123], the authors employ a variation of auto-encoders to predict a future video frame to a given set of consecutive video frames. Then, they compute the per-pixel difference between predicted and ground truth frames to make a decision on whether the current video sequence is normal or not. Future frame prediction has gained increasing attention in light of its potential applications in unsupervised feature learning for video representation [124]. In [120], they also quantize the output of the encoder using a predefined code book that further narrows the explanation of normal events and aids in better future frame prediction in normal videos. Also in these methods, abnormal or violent events are detected based on the assumption that the trained models will find it difficult to generate future frames from given violent video sequences. In [125], to gen-

erate more realistic and accurate future frames, the authors imposed a loss in temporal space. In particular, they compute optical flows in video sequences using a pre-trained CNN [126] and formulate a loss function for an auto-encoder that ensures the optical flow of predicted frames is consistent with ground truth. All aforementioned works focus on predicting future frames. Apart from these, there were also efforts in literature to predict transformations needed for generating future frames [127] [128] [129] [130].

3.2.2 Multiple instance learning

These methods also aim to learn violent actions using video-level labels that are provided in the training phase. In contrast to the methods that are based on modelling normal patterns, these methods employ both normal and violent data with video-level annotations for training violent detection models [131] [132] [133]. With give video-level labels, a video belonging to a violent class can contain several segments that do not have violence. Multiple Instance Learning (MIL) [134] helps to model the patterns of normal and violent events in such challenging cases. In [131], Sultani et al. proposed to divide each video (in both normal and violent videos) into multiple temporal segments to form positive and negative bags. The positive bag contains instances of violent events and the negative bag contains instances of normal events. Then they extracted C3D [135] spatiotemporal features on each segment and used these features to train three fully connected layers to derive scores for the given positive and negative bags. Due to the absence of segment-level labels, they proposed a novel ranking loss function that encourages the computed score in the positive bag to be higher than the score in the negative bag. They also impose smoothness and sparsity constraints in their ranking loss to reduce false alarms. They showed that their approach works well on their database.

By extending the approach of Sultani et al., Zhu et al. in [132] introduced temporal context information into the MIL ranking loss to compute scores video-wise and not segment vise. They proposed a temporal augmented network that captures motion features using pre-computed optical flows. This network is similar to an auto-encoder that tries to reconstruct the input motion patterns specified in the form of stacked optical flows. They consider the encoded motion patterns (derived at the output of the encoder in their temporal augmented network) for training their MIL ranking model for accurately localizing anomalies.

[136] proposed a two-step approach where they first detect and track humans locally across a given segment of a video to form human tubes (spanning the entire segment) and then use multi-fold Multiple Instance Learning (MIL) with Support Vector Machine (SVM) [137] to learn about human tubes that contain the action described by the video-level labels. In [138], Yan et al. proposed a multi-task ranking model. In their approach,

they segmented videos into supervoxels, using a graph-based segmentation method, to generate action tubes and action–actor tubes. Action tubes are then used as proposals for actions, e.g., walking, adult running, and crawling. Features are extracted from each tube to train the ranking model to select the most characteristic action tubes.

Arnab et al. [139] proposed a probabilistic variant of MIL, where they estimate the uncertainty of an instance-level prediction. They used a pre-trained person detector trained on a large image dataset to detect persons over consecutive frames of a video to form person tubelets. A bag for MIL consists of all tubelets within a video, and it is annotated with the video-level label. During training, due to computational constraints, a whole bag cannot be processed simultaneously; therefore, they model the label noise through the uncertainty of sampling bags that do not contain any tubelets with the labelled action.

Mettes et al. in [140], strove to find the spatiotemporal locations of actions in videos using pseudo-annotations. They investigated spatiotemporal pseudo-annotations from different sources such as action proposals, object proposals, person detections, motion, and centre biases. Later, they combined the extracted pseudo-annotations using a correlation metric to train a classifier using MIL.

3.2.3 Supervised learning

Since suitably annotated databases are increasingly available in the past couple of years, there are also approaches proposed in the literature to classify violent videos using annotated databases and extract features using learning methods. Particularly, these methods are destined to work with well-calibrated video datasets containing precise visual information about the events for the relevant class. For example, there are little or no normal events in the videos belonging to the violence class. In [141], Long et al. proposed a method that employs the Motion SIFT (MoSIFT) algorithm to extract the low-level description of a query video. Later, they investigated Kernel Density Estimation (KDE) to filter any feature noise from the MoSIFT descriptor. On the reduced MoSIFT features, they applied sparse coding using a pre-defined dictionary to obtain a video-level feature vector. They trained an SVM using the computed video-level feature vectors to classify videos.

In 2012, Hassner et al in [142] suggested a method for real-time detection of breaking violence in crowded scenes. They used the Violent Flows (ViF) descriptor that captures optical flow information between consecutive frames of a given video. They use linear SVM to classify the videos using the computed ViF descriptors. They compared their approach to other methods existing at that time and showed that their method is effective in classifying videos containing crowd violence. The work of Zihan Meng[143] also

follows a similar approach. They proposed a combination of feature extraction and deep learning using CNNs for violence detection. In particular, they use Hough Forests spatiotemporal feature extractor in combination with a 2D CNN. Sudhakaran and Lanz [144] proposed to encode the difference between two successive frames using a combination of CNN and LSTM. They showed that their network trained on the frame difference has better representation and also performs better than a model trained on raw frames. Nouar AlDahoul et al. [145] also suggested a lightweight model with less number of parameters using CNN and LSTM. They aimed to capture spatial features using 2D CNN and applied LSTM block on captured spatial features for violent video classification. Fath U Min Ullah et al. [146] proposed a Violence Detection Network (VD-Net) where they first carry out object detection to detect humans and suspicious objects like guns to pre-filter the video sequences for violence detection. Then they use a combination of convolutional LSTM and gated recurrent units [147] for violence detection on filtered video sequences. Romas et al. [148] also proposed a CNN-LSTM-like architecture that is computationally light. In particular, they used MobileNet V2 for extracting spatial features which are used for training an LSTM network.

Chollet et al. [149] used a trained model which is inspired by XcpetionNet [150] to extract features and used a bi-directional LSTM, that analyses features in both temporal directions, for classification. Khan et al. [151] proposed a method that samples a video uniformly into several segments, and then they select a representative frame from each segment using computed levels of saliency. They fine-tune MobileNet [152] using the representative frames to classify the corresponding segment into violent and non-violent classes.

Li et al.[153] proposed an architecture based on DenseNet [154] 3D CNN that directly works on the video data without explicitly computing any features. They showed that they can achieve good accuracy on standard databases with a relatively lightweight deep learning model. Fernando et al.[155] also proposed an architecture based on a variant of DenseNet [156], however, they use DenseNet to extract feature maps. On the extracted feature maps, they link different positions of a single sequence using self-attention mechanisms [157] to generate a representation that focuses on the most relevant parts of the sequence. The output of the self-attention layers is connected to bi-directional LSTM blocks followed by fully connected layers for classification. They demonstrated that their method achieves good accuracy on four different databases. As an input to DenseNet, they used optical flow information computed from the videos. Additionally, they have also experimented with pseudo-optical flows as input to DenseNet obtained by subtracting two adjacent frames.

In [158], Dong et al. proposed multi-stream deep convolutional neural networks consisting of three streams - colour, optical flow, and acceleration for person-to-person vio-

lence detection in videos. Their main hypothesis is that violent events are more intense in terms of speed when compare to other normal events. The proposed acceleration stream aims to capture such important intense information. Three different LSTMs are trained using the features from these three streams and the outputs of the three streams are fused to classify a video.

Yukun Su et al. [159] followed a different approach to violence detection. They computed 3D skeleton point clouds from the human skeleton sequences extracted from videos and then performed interaction learning on these 3D skeleton point clouds. They introduced Skeleton Points Interaction Learning (SPIL) module that captures spatiotemporal features from extracted point clouds to model the interactions between skeleton points. They employed multiple SPIL modules together with fully connected layers to classify violent videos from normal videos.

Guankun Mu et al. in [160] proposed to use both visual and audio cues to detect violent scenes in videos. Their hypothesis is that visual information may not be always reliable for violence detection and using audio would increase the detection accuracy. For extracting audio features, they used a 40-dimensional Mel Filter-Bank (MFB) with a 25 ms analysis window and 10 ms shift. They used SVM in their implementation for classifying audio samples from input videos.

3.2.4 Limitations in the state-of-the-art methods

Although methods that learn only normal patterns in training show successful results on some databases, they are inefficient in generalizing to other databases. One of the common reasons is that some normal actions involving close physical interaction between humans can mimic violent actions and can mislead the deep learning algorithms to infer false negatives. Also, some violent human actions are visually complex, and using only normal visual data for deep learning may not be optimal. I suggest that it is important to incorporate both normal and violent behaviours in the training data.

Violence detection methods that follow future frame prediction also fail to generalize. While reconstructing future frames, if a current frame contains objects, for example, a bicycle, that are not available during training, the predicted frame differs significantly from the ground truth future frame and an anomaly is declared. Further, using memory modules, dictionaries, and codebooks in architectures like auto-encoders for future frame prediction substantially restricts the interpretation of normal actions by such models.

Some classes of techniques in literature use information such as optical flows or detected objects or human skeleton coordinates. Depending on the target hardware to deploy violence detection algorithms, computing such information is expensive. Even if sufficient hardware resources are available for implementation and execution, tracking

object coordinates frame-by-frame to compute robust trajectories is still challenging, especially, when long trajectories need to be extracted or difficult scenarios appear such as crowded scenes.

Table 3.1 compares different approaches for violence detection in videos proposed in the literature and lists their advantages and disadvantages. I suggest that the key aspects in developing successful algorithms for violence detection are that these methods should be computationally fast, achieve the best accuracy and adapt to scenarios that are not present in the training. It is evident from several efforts in the literature that methods that employ spatiotemporal feature detection from videos that account for both motion and appearance information can achieve high accuracy. However, it is important to note that the cost of extracting some of those features is still prohibitive for practical applications. In the current work, I adapt and extend X3D expandable architecture that has very few parameters and for fast and accurate violence detection. Although following MIL using spatiotemporal features is computationally light, these methods cannot yield better classification accuracy since they focus on predicting segment-level labels while neglecting modelling hidden temporal context information. In this work, I focus on using calibrated datasets to develop an efficient method for violence detection (refer section 3.3.1).

3.3 Fast and Accurate Violence Detection

ResNet [161] is a popular base architecture for image and video recognition tasks, known for its effectiveness and state-of-the-art results on benchmarks like ImageNet [162] and COCO [163] datasets. 3D ResNets [99] are an extension of the ResNet architecture, designed for learning spatiotemporal features from video data. They have achieved strong performance on various benchmarks and real-world applications, including the Kinetics-700 action recognition dataset [164] (where a variant called I3D [165] achieved state-of-the-art performance) and the Something-Something V2 action recognition dataset (where a 3D ResNet called R(2+1)D [166] achieved state-of-the-art performance).

3D ResNets have higher accuracy than counterparts like 3D-MobileNet [167] due to factors such as more layers for learning complex spatiotemporal features and skip connections between the input and output of each layer that allow input to bypass intermediate layers. However, they are generally more computationally intensive due to a large number of model parameters. To improve computational efficiency, model complexity can be reduced through techniques such as reducing the number of layers, using fewer filters in convolutional layers, and using smaller input data, though this may decrease accuracy on complex tasks. Christoph et al. [168] experimented with various parameters of the 3D ResNet architecture to understand the effect of reduced model complexity

Table 3.1: Comparison of approaches in the state-of-the-art for violence detection in videos.

Method	Strength(s)	Limitation(s)
Learning normal patterns	Computationally light	Poor generalizability and may not be suitable for practical applications.
Future frame prediction	Computationally light	Poor generalizability, difficulty interpreting normal actions, sensitivity to objects not present during training and memory modules in some architectures may restrict the interpretation of actions.
MIL using spatiotemporal features	Depending on features, can be computationally light	Trims videos into bags, and focuses on predicting bag-level labels, ignoring temporal context information between bags, may not achieve high classification accuracy.
3D deep learning architectures	Good classification accuracy, adaptable to new scenarios	Extraction of 3D volumes can be memory-intensive, computing optical flow in some architectures can be time-consuming, training can be slow, computing some features can be expensive, and may not be suitable for real-time applications.

on accuracy. They expanded the architecture along multiple axes to form spatiotemporal models and selected the axis that achieved the best trade-off between computational speed and accuracy, resulting in a series of models ranging from extra small (XS) to extra large (XL) in increasing complexity. Using the Kinetics-400 dataset [169], they showed that their expanded model, X3D-M, had the same accuracy as state-of-the-art video classification networks but with a 10X reduction in model parameters.

The X3D-M model is an appropriate choice for my violence detection task due to its high accuracy and reduced model complexity. As demonstrated by Christoph et al., the X3D-M model achieves similar accuracy to state-of-the-art video classification networks, but with a significantly lower parameter count. This reduction in model complexity makes the X3D-M model more efficient to train and deploy, particularly for resource-constrained systems. Also, the ResNet 3D backbone, which has a proven ability to learn complex spatiotemporal features, is particularly useful for my violence detection task, as it allows the model to capture the dynamic nature of the videos and learn robust representations of the data. The proposed system using X3D-M model architecture is detailed in section 3.3.2.

3.3.1 Datasets for experiments

As mentioned previously, due to laws such as GDPR, substantial real-world footage containing violence for training deep learning models is not available. In the recent past, the usage of synthetic training data is becoming more common in computer vision. Using training data containing pasted object patches on real images was shown to be effective in the case of 2D object detection tasks [170] [171] [172] and human pose estimation [173]. However, for violence detection, I postulate that such fabricated training data do not fully capture the complex and diverse action patterns of violent actions with various nuances. Therefore, preparing and using synthetic training data is considered to be out of the scope of the current work.

Table 3.2: Overview of the datasets used in my experiments. Apart from the existing databases in the literature (CV, HF, MF, RLVS, RWF-2K, UCFS and XD-V), I have annotated parts of two other datasets (UCFS and XD-V).

Dataset	Acronym	Total Samples	Avg. Frame Resolution(Pixels)	Avg. Video Duration(S) [min, max]	Avg. FPS [min, max]	Violent Samples	Normal Samples
Crowd Violence[142]	CV	246	320×240	3.49 [1.04, 6.44]	26.47 [25, 29.97]	123	123
Hockey Fights[174]	HF	1000	360×240	1.64 [1.6, 1.96]	25 [25, 25]	500	500
Movie Fights[174]	MF	200	320×240	1.74 [1.66, 2.04]	43.14 [25, 59.94]	100	100
Real Life Violence Situations[175]	RLVS	2000	variable	4.91 [1, 11.32]	27.21 [10.5, 37]	1000	1000
Real-World Fight-2000[176]	RWF-2K	2000	variable	4.99 [0.32, 5]	29.99 [25, 30]	1000	1000
UCF-Crime Selected	UCFS	1082	320×240	4.25 [0.3, 5.03]	29.99 [25, 30]	541	541
XD-Violence Selected	XD-V	1126	640×360	4.15 [0.3, 5.04]	24 [24, 24]	563	563

P. Sernani et al. in [177] proposed an AIRTLab dataset that contains videos with violence performed by non-professional actors. They studied 2D and 3D deep learning architectures for violence detection using their dataset. Their results indicate that the studied models adapt well to their settings where violence is mimicked by non-professional actors. However, they pointed out that their results cannot be considered general. Their architectures are not validated on more datasets and no cross-validation experiments were performed. Hence I also do not consider such datasets in my experiments.

In the current work, I have considered seven different datasets that are commonly used in the literature for experimentation with violence detection and to facilitate comparison of my results with other methods. I have also extended some of these datasets with annotations to assist in in-depth cross-validation experiments. These datasets are described in the following:

- **Crowd Violence** (CV) [142] dataset contains videos involving violence in crowds, collected from YouTube.

- **Hockey Fights** (HF) [174] dataset is a collection of fights between players in hockey games from the USA's National Hockey League (NHL).

- **Movie Fights** (MF) [174] dataset collects several scenes from action movies.

- **Real Life Violence Situations** (RLVS) [175] dataset gathered fighting videos from YouTube and also from real street cameras that contain many real street fights.

- **Real-World Fight-2000** (RWF-2K) [176] dataset is a collection of large-scale fighting videos from YouTube. The dataset contains trimmed video clips captured by surveillance cameras from real-world scenes.

- **UCF-Crime Selected** (UCFS) dataset is a subset of UCF-Crime [131] dataset. The UCF-Crime dataset contains long untrimmed surveillance videos that cover 13 real-world anomalies including Abuse, Arrest, Arson, Assault, Burglary, Explosion, Fighting, Road Accident, Robbery, Shooting, Stealing, Shoplifting, and Vandalism without annotations. Although this is a large-scale dataset, all videos in the violence class contain a mix of violent and normal actions which is undesirable. Among the anomalies, I selected the classes - Abuse, Explosion, Fighting, Road Accident, and Shooting and manually trimmed these videos to only contain violent parts for training and testing.

- **XD-Violence Selected** (XD-V) dataset contains a subset of videos from the XD-Violence [178] dataset. XD-Violence dataset contains several untrimmed videos

covering 6 anomalies including Abuse, Car Accidents, Explosions, Fighting, Riots, and Shooting gathered from action movies and YouTube. Similar to the UCF-Crime dataset, I selected a set of videos belonging to the classes - Abuse, Explosion, Fighting, Road Accident, and Shooting and manually trimmed these videos to only contain violent parts for training and testing.

All of the datasets also contain normal videos for training and testing that do not involve violence. In the case of the UCFS and XD-V datasets, I trimmed the normal videos to five-second video clips to match the average duration of normal clips in the other datasets. Additionally, in the case of the UCFS and XD-V datasets, I limited the maximum duration of a video clip containing violence to approximately five seconds. Table 3.2 provides more details about each of the datasets I used in my experiments.

3.3.2 Model Architecture

For accurate violence detection, it is important to have a properly calibrated dataset containing a large number of diverse examples for each class. Successful action recognition datasets such as Kinetics-400 [169] contain a minimum of 400 videos for each action class such as standing, sitting and talking, etc. All videos in Kinectics-400 dataset have a fixed time span of five seconds. They obtained clips for each class from YouTube and then used Amazon Mechanical Turkers (AMT) to decide if a given clip contains the desired action or not. Three or more confirmations (out of five) were required before a clip was accepted [169]. Further, the dataset was also de-duped to reduce redundancies in the environment.

In several cases, actions involving violence are more complex than actions like sitting and talking and the number of examples of violent videos collected in the existing datasets may not be enough for training a model that generalizes well and can lead to model over-fit. Also, as shown in Table 3.2, different datasets for violence detection contain clips having different time spans in seconds and they are not well-groomed to check for the validity of a specific action or for redundancies. To cope with these issues with existing datasets for violence detection, I follow training approaches that are inductive in nature. Specifically, I aim to make use of the knowledge learned using better-calibrated action recognition datasets to solve efficient violence detection problems. To this end, I propose two different learning configurations that are described in the following subsections.

Fine-Tuned X3D-M model

In Fine-Tuned X3D-M (hereinafter referred to as FT) model, I consider the X3D-M model architecture initialized with weights obtained by training on the Kinectics-400 dataset. Note that the original architecture used for training on the kinetics-400 dataset contains two fully connected layers with the output of the second fully connected layer representing the classification results for each class (the number of outputs of this layer is equal to the number of classes in the training dataset). Since, in my case, I aim to predict if a clip contains violence or not (binary), I modify the architecture into a regression model to generate a violence coefficient that indicates the probability of the existence of violence in a given video clip. Specifically, I trim the X3D-M model until the first fully connected layer and replace the second fully connected layer to output a floating point variable which is converted into range $[0, 1]$ using a sigmoid function to derive the violence coefficient. Plainly, during learning, for samples of video clips containing violence, I label the violence coefficient as 1 and for samples of video clips containing no violence, I label the violence coefficient as 0.

The architecture of the X3D-M model follows the Fast pathway design of SlowFast networks [179] with down-sampled temporal input. Therefore, I pre-process the input videos as per the requirements of the X3D-M model. In particular, for a given video clip, first, I extract 16 video frames by uniform sampling in the temporal domain. Then, I transform the pixel value range of the extracted frames to be in between $[0, 1]$ to obtain floating point images. Later, I normalize the video frames using mean and standard deviation and resize the frames so that the shortest side corresponds to 256 pixels. Finally, the resized frames are centre cropped to obtain 16 video frames with a spatial resolution of 256×256. Batches of pre-processed video frames are supplied to the FT model with corresponding labels for training. Note that here, the X3D-M model weights obtained by training on the Kinetics-400 dataset are only used for network initialization and these are further optimized during training datasets for violence detection. My FT architecture is shown in Fig. 3.1 and Table 3.3 presents the information on the corresponding model parameters.

Transfer-Learned X3D-M model

The Transfer-Learned X3D-M model, or TL for short, leverages the strengths of the X3D-M model, which has been trained on the large Kinetics-400 dataset, for feature extraction. To do this, the input video data is pre-processed and passed through the X3D-M model to extract a feature set. The feature set is then used to train three additional fully connected layers, which are shown in Fig. 3.2 and Fig. 3.3.

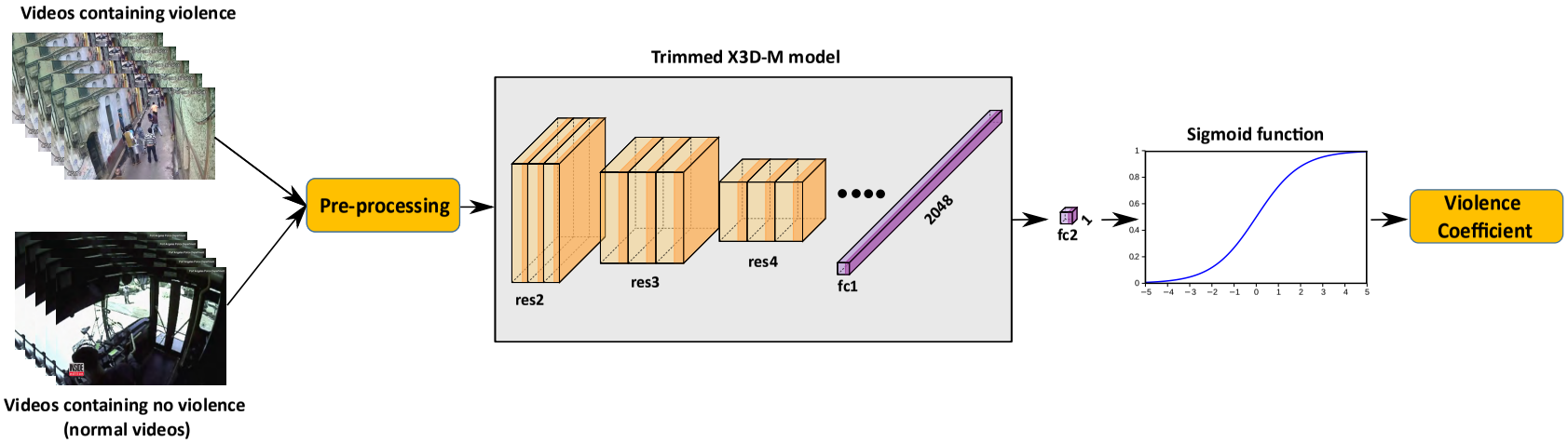

Figure 3.1: my FT model: Batches of videos containing violence and no violence are supplied for training. Each input video is pre-processed to obtain 16 uniformly sampled temporal frames for training a trimmed X3D-M model. The second fully connected layer of X3D-M model is replaced to output a floating point variable which is converted into range $[0, 1]$ using a sigmoid function to derive the violence coefficient (original illustration by Viktor Huszár).

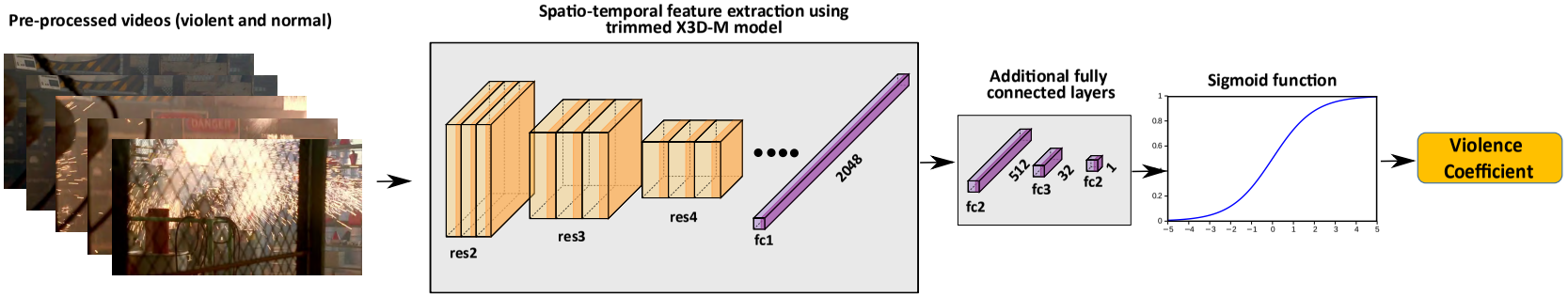

Figure 3.2: my TL model: Pre-processed videos containing both violence and no violence are inputted to a pre-trained X3D-M model for feature extraction. Three fully connected layers are trained using the extracted features to obtain the violence coefficient (original illustration by Viktor Huszár).

The resulting output is a floating point variable that is transformed into a violence coefficient in the range of $[0, 1]$ using a sigmoid function. The advantage of this approach is that it leverages the already trained X3D-M model to extract meaningful features from the input video data, reducing the amount of training required for the additional layers. This results in a more efficient and effective model for violence detection in videos. The information on the TL model parameters is presented in Table 3.4.

Table 3.3: Parameters involved in my FT model - combining the trimmed X3D-M model and the replaced second fully connected layer, I have 2976723 parameters in this model. Since the parameters of the trimmed X3D-M model are also optimized during training, all 2976723 parameters are trainable.

Component	Output Shape	Params #
Trimmed X3D-M model	(2048)	2974674
Dense (”fc2” in Fig. 3.1)	(1)	2049
Total Parameters	2976723	

Table 3.4: Parameters involved in my TL model - I have 4040211 parameters in this model. Since the parameters of the trimmed X3D-M model are not trained, 1065537 parameters are trainable and 2974674 parameters are nontrainable.

Component	Output Shape	Params #
Trimmed X3D-M model	(2048)	2974674
Dense (”fc2” in Fig. 3.2)	(512)	1049088
Dense (”fc3” in Fig. 3.2)	(32)	16416
Dense (”fc4” in Fig. 3.2)	(1)	33
Total Parameters	4040211	

3.3.3 Learning and optimization

I do not apply data augmentation techniques in the training of the proposed models. I use Adagrad [180] to optimize my models with an initial learning rate of $1e^{-3}$. Both models are compiled to minimize the Binary Cross Entropy (BCE) between the estimated and ground truth violence coefficients. For training the TL model, I use a batch size of 30 samples collected from shuffled pre-computed X3D-M feature vectors. Since the FT model takes videos as input, to account for higher memory usage during training, I consider a batch size of 4 samples collected from shuffled videos. For regularity, within a training batch, for both models, I concatenate a batch of violent video clips with a batch

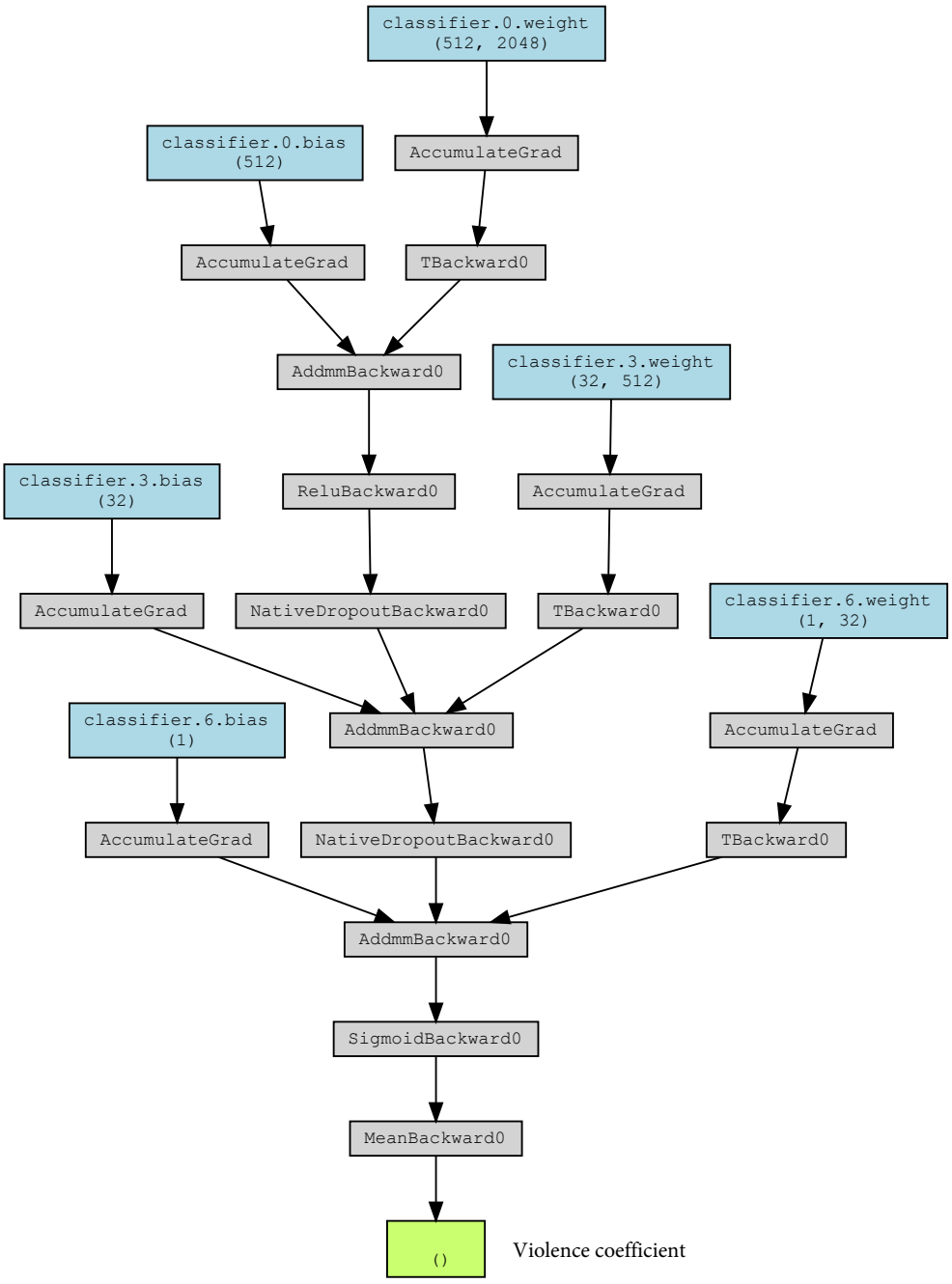

Figure 3.3: Network graph for the additional fully connected layers in the TL model (original illustration by Viktor Huszár).

of non-violent video clips. For ease of access, all my hyperparameters are listed in the table 3.5

Table 3.5: List of hyperparameters and their corresponding values used in my training.

Hyperparameter	Value
Learning Rate	0.001
Batch Size	TL-30 & FT-4
Number of Epochs	50
Optimizer	Adagrad
Loss Function	Binary Cross Entropy loss
Dropout Rate	0.1
L2 Regularization	0.001

3.4 Results and Discussion

In this section, I present the results from my various experiments using the proposed models and the various datasets described in section 3.3.1. Most of the datasets used in the study already have a training and testing data split with 80% of the data as the training set and 20% as the test set. For other datasets, for my experiments, I preserve this percentage and randomly select 20% of violent and non-violent samples to create a testing set for fair comparison across datasets. To facilitate fair comparison, all the models are trained for 50 epochs using a given training dataset. I use the PyTorch [181] deep learning library to train and test my models on a Nvidia GeForce GTX 1080 Ti GPU using the CUDA toolbox. I use the Ubuntu Linux operating system on an AMD Ryzen Threadripper 1950X 16-core processor. To evaluate the performance of various methods, I use the following metrics that are commonly used to evaluate the performance of classification algorithms using deep learning.

- **Accuracy** (ACC) [182] is the most popular metric for evaluating deep learning models for video classification. It is the ratio of the number of correct predictions (as violent or non-violent video clips) to the total number of predictions. To compute the accuracy, I used the provided ground truth binary labels - 0 (for video clips without violence) and 1 (for video clips with violence) - that are provided during training. Since I designed my networks to output floating-point violence coefficients, I round the predicted violence coefficients to the nearest integer before calculating the accuracy. In line with other methods in the literature, I report the accuracy score in percentages.

- **Area Under Curve** (AUC) [183] is a statistical measure to evaluate the performance of a classification model. It represents the area under the Receiver Operating Characteristic (ROC) curve, which graphically illustrates the effectiveness of a classifier in discriminating between the trained classes at various decision probability thresholds. Specifically, using the predicted violence coefficients, the ROC curve shows the relationship between True Positive Rate (TPR) (the number of times when violence cases are correctly identified as violence among the total cases when violence cases are correctly identified as violence and non-violence cases are correctly identified as non-violence) and False Positive Rate (FPR) (the number of times when non-violence cases are incorrectly identified as violence among the total cases when non-violence cases are incorrectly identified as violence and violence cases are incorrectly identified as non-violence). Higher values of the area under the ROC curve (that are close to 1) represent the ability of a model to effectively discern between violence and non-violence cases, while lower values represent the opposite.

I have conducted several experiments, including cross-dataset validation, to evaluate the performance of the proposed approaches using the considered datasets and metrics. The details and results of these experiments are presented in the following subsections.

3.4.1 Experiments on individual datasets

Most datasets already have pre-defined data splits for training and testing, with 80% and 20% of the data respectively. I used these splits without modification for unbiased comparison. For the remaining datasets, I maintained this proportion of training and testing data by randomly selecting 20% of violent and non-violent samples for testing. I trained my models on the training data split and evaluated their performance on the testing data split for each dataset separately. The testing results using the ACC and AUC metrics are presented in Tables 3.6 & 3.7 respectively. The tables also show the performance of state-of-the-art methods discussed in section 3.2 on the respective datasetes. As mentioned, I created the UCFS and XD-V datasets and I report the results on these datasets using only my methods.

It is worth noticing that only a few studies in the literature report evaluations using the AUC metric. I argue that in applications such as violence detection, false positives (incorrectly reporting non-violent events as violent) should be explicitly considered when evaluating the performance of a model and the ACC metric does not directly account for false alarms.

The experimental results on individual datasets show that both of my proposed methods perform well on individual datasets. Overall, my FT model outperformss most of

Table 3.6: The ACC(%) scores of my FT and TL models along with the state-of-the-art methods on individual datasets. Based on the ACC metric, my FT method outperforms most of the state-of-the-art methods on all datasets except HF, with relatively fewer model parameters.

Method	CV	HF	MV	RLVS	RWF-2K	UCFS	XD-V	Params #
ViF[142]	81.3	82.9	-	-	-	-	-	-
3-stream+LSTM[158]	-	93.9	-	-	-	-	-	-
MoSIFT[141]	89	94	-	-	-	-	-	-
Bilinski[79]	96.4	96.8	99	-	-	-	-	-
Sudhakaran[144]	94.5	97.1	100	-	-	-	-	9.6M
Zihan Meng[143]	-	94.6	99	-	-	-	-	-
Li et al.[153]	97.17	98.3	100	-	-	-	-	7.4M
Akti 5-Frames[149]	-	95	90	-	-	-	-	9M
Akti 10-Frames[149]	-	96	87.5	-	-	-	-	9M
Khan et al.[151]	-	87	99.5	-	-	-	-	-
Pseudo-OF[155]	94.8	97.5	100	94.1	-	-	-	4.5M
OF[155]	96.9	**99.2**	100	95.6	-	-	-	4.5M
CNN-LSTM-IOT[145]	-	-	-	73.35	73.35	-	-	**1.266M**
Romas et al.[148]	-	-	99.5	73.35	82.3	-	-	4.074M
SPIL [159]	94.5	96.8	98.5	-	89.3	-	-	-
VD-Net[146]	-	98.5	-	-	88.2	-	-	4.4M
Choqueluque-Roman et al.[184]	-	97.3	-	92.88	88.71	-	-	above 30M
ours (FT)	**99.5**	97.5	100	**96.7**	**91**	**90.1**	**93.59**	2.98M
ours (TL)	92	97.5	100	95.2	85	84.2	89.31	4.04M

the state-of-the-art methods, and my TL model also achieved decent performance on all datasets. I postulate that the FT model, which optimizes the parameters of the (trimmed) X3D-M model during learning, is more adaptable to a given dataset. On the MF dataset, the results for both TL and FT models suggest overfitting, which is consistent with the results from most methods in the literature. This suggests that the MF dataset may contain more regular examples with less diversity and may be less challenging for deep learning video classification models. The Tables 3.6 & 3.7 also show the model parameter count for various models under comparison and my models have fewer parameters than the state-of-the-art methods.

Bilinski et al.[79] achieved a higher accuracy than my TL model on the CV dataset. They used improved Fisher vectors for spatiotemporal feature extraction, which can be context-dependent. For example, the CV dataset only contains examples of violence involving a crowd, and their results show that their method performs better in such scenarios. It is important to note that statistical feature extraction methods like this can be sensitive to variations in the video capture environment and may result in false alarms. When evaluated using the AUC metric, my TL model performs better than the method of Bilinski et al.[79] on the CV dataset (see Table 3.7).

Sudhakaran et al. [144] used a pre-trained AlexNet model trained on ImageNet for

Table 3.7: The AUC scores of my FT and TL models compared to various state-of-the-art methods. According to the AUC metric, the FT model outperforms the state-of-the-art methods on most of the datasets and has fewer model parameters.

Method	CV	HF	MV	RLVS	RWF-2K	UCFS	XD-V	Params #
MoSIFT [141]	0.935	0.96	-	-	-	-	-	-
Bilinski[79]	0.87	-	-	-	-	-	-	-
CNN-LSTM-IOT[145]	-	-	-	0.82	0.82	-	-	**1.266M**
VD-Net[146]	-	0.994	-	-	0.91	-	-	4.4M
Choqueluque-Roman et al.[184]	-	0.993	-	0.913	0.914	-	-	above 30M
ours (FT)	**1.0**	**0.999**	**1.0**	**0.996**	**0.972**	**0.971**	**0.976**	2.98M
ours (TL)	0.99	0.994	**1.0**	0.993	0.944	0.936	0.959	4.04M

their method. They used the difference between consecutive video frames as input to capture temporal information. The results show that their method performs better on the CV dataset compared to my TL model. I should note that my TL model extracts features using a pre-trained X3D-M model trained on the Kinetics-400 dataset. This dataset contains a smaller number of examples with several people appearing in individual frames of the videos. In contrast, the ImageNet dataset contains a relatively higher number of examples with several people appearing in one frame. Therefore, we suggest that the extracted X3D-M features might be noisy and result in lower accuracy on datasets involving crowds such as the CV dataset.

Li et al.[153] used a DenseNet 3D-CNN to train and extract spatiotemporal features from videos. Their model was initialized with parameters from a pretrianed model trained on the Kinetics-400 dataset, similar to my FT model. However, their model had more CNN layers and higher model parameters, which contributed to its better accuracy on the CV and HF datasets compared to my TL model. It should be noted that DenseNet uses multi-layer feature concatenation for improved feature representation, but this approach requires more GPU memory and longer training times. Choqueluque-Roman et al.[184] followed an approach that used an I3D architecture in combination with a ResNet50 for feature extraction using human action tubes for training a deep learning model based on MIL. Their results showed that, according to the accuracy and AUC metrics, my models achieved better performance with relatively fewer model parameters, which confirms that training based on MIL may not achieve high classification accuracy.

Violence-Net [155] also used DenseNet for training and extracting feature maps. According to the ACC metric (see Table 3.6), their method using optical flow input achieved better scores than my FT model on the HF dataset. However, their architecture contains more model parameters and involves computing optical flow information, making it computationally more complex than ours. When pseudo-OF was used as input in their method, the accuracy decreased compared to my FT model. On the CV dataset, their

model with more number of parameters achieved higher accuracy than my TL model. As previously mentioned, the extracted X3D-M features from videos involving crowds can be noisy and lead to less accurate results.

The method proposed by Romas et al. [148] used MobileNet V2 architecture for spatial feature extraction and LSTM modules for learning about temporal associations. Despite having a similar number of model parameters as my TL model, my methods achieved higher accuracy. As demonstrated by my results, methods that capture 3D spatiotemporal features directly from the video data, such as my proposed models, represent temporal associations more accurately and are therefore more effective at detecting violence in videos. This is due to the ability of my proposed models to accurately capture the full context and dynamics of the events depicted in the video, leading to improved performance in violence detection tasks.

The SPIL method [159] achieved higher accuracy scores than my TL model on the CV and RWF-2K datasets. However, this method requires significant computational resources due to the need to estimate 3D skeleton point clouds for interaction learning, making it impractical for practical applications.

The Violence Detection Network (VD-Net) [146] achieved better accuracy on HF and RWF-2K datasets compared to my TL model and has slightly more model parameters. VD-Net first detects humans and suspicious objects such as guns, which requires more computational resources than my TL model. However, the AUC scores for the TL model are comparable to VD-Net.

Finally, the CNN-LSTM-IOT model [145] has fewer parameters than all of the models under comparison, including ours, and it has been demonstrated that it can run on a low-cost Internet of Things (IOT) device like a Raspberry Pi. However, the model relies on spatial features for learning and performs poorly on the RLVS and RWF-2K datasets.

In summary, my experiments on individual datasets demonstrated that my FT model outperformed most of the state-of-the-art methods on most datasets while having fewer model parameters. My TL model also achieved decent performance on all the datasets, despite having fewer trainable parameters than the FT model, as shown in Tables 3.3 & 3.4. This suggests that the TL model is relatively less adaptable to specific scenarios.

3.4.2 Experiments about generalizability

To study the adaptability of my proposed approaches to unseen videos, I conducted cross-dataset experiments where I trained a model on one dataset and evaluated its performance on another dataset. Table 3.8 shows the results from such one-on-one cross-validation tests in the top section (columns 5-8). It should be noted that, among the considered datasets, different datasets have different numbers of videos containing instances of vio-

lence and non-violence actions. In general, the number of samples available for training can greatly affect the learning capabilities of a deep learning model. Few and less diverse training samples can lead to model overfitting, where the model models some noise or random fluctuations in the training data is modelled very well, but it cannot generalize to new data. In my case, since I follow an inductive training approach using a pre-trained X3D-M model on the Kinetics-400 data, I suggest that my models are least influenced by the number of training samples, and my cross-validation results essentially show the ability of my models to learn the concept of violence.

Both ACC and AUC metrics show that there are several inconsistencies in the results across the considered datasets. To provide deeper insights into my cross validation results, I plot the ACC and AUC scores obtained by training on a specific dataset and averaging the testing scores on the rest of the datasets in Figures 3.4 & 3.5 respectively for both FT and TL models. Each plot also shows the standard deviation of the metric scores obtained from the testing datasets, indicated by the red colour lines. According to the metric scores, the trained FT and TL models on the CV dataset did not generalize well to other datasets (see bar plots in Figures 3.4(a) & 3.5(a)). This is anticipated since the CV dataset contains only examples of mass violence, and the other datasets do not contain many such examples. Also, the trained FT model on the HF dataset poorly generalized to other datasets, indicating that the HF dataset does not contain diverse examples of violence and contains monotonous fighting videos between hockey players. However, the TL model trained on this dataset showed better generalization than the FT model as indicated by the metric scores.

FT and TL models trained individually on datasets - MF, RLVS, RWF-2K, UCFS & XD-V performed satisfactorily in my cross-validation tests and generalized well to other datasets with average ACC scores close to or above 80% and average AUC scores close to or above 0.8. When considering both metrics, FT and TL models trained on UCFS and XD-V datasets exhibited the best generalization ability in my cross-validation studies. This suggests that these datasets, which I compiled, contain the most representative and diverse samples for violent and non-violent actions.

Table 3.8: Cross dataset experiment results - One-on-one cross-validation test results are shown in the top section, leave one out cross-validation test results are shown in the middle section, and the bottom section shows the performance of my models on the training/testing folds used in Violence-Net [155] . To compare, ACC scores for Violence-Net using both OF and Pseudo-OF inputs are also provided for relevant datasets.

Dataset Training	Dataset Testing	ACC-OF[155]	ACC-Pseudo-OF[155]	ACC-FT (ours)	ACC-TL (ours)	AUC-FT (ours)	AUC-TL (ours)
CV	HF	65.16	64.76	56.5	50.6	0.68	0.78
CV	MF	60.02	59.48	56	72	0.54	0.88
CV	RLVS	58.76	58.32	78.58	86.56	0.84	0.95
CV	RWF-2K	-	-	76.05	73.05	0.81	0.87
CV	UCFS	-	-	68.11	72.46	0.75	0.79
CV	XD-V	-	-	53.99	64.48	0.62	0.71
HF	CV	62.56	61.22	97.95	97.95	0.99	1
HF	MF	65.18	64.86	45.5	76	0.55	0.86
HF	RLVS	58.22	57.36	76.58	86.11	0.83	0.94
HF	RWF-2K	-	-	69.9	78.2	0.76	0.86
HF	UCFS	-	-	64.51	65.8	0.68	0.77
HF	XD-V	-	-	54.09	58.7	0.51	0.66
MF	CV	52.32	51.77	86.48	96.31	0.99	1
MF	HF	54.92	53.5	98.3	87.8	1	0.95
MF	RLVS	56.72	55.8	79.29	87.41	0.94	0.95
MF	RWF-2K	-	-	75.6	78.35	0.84	0.86
MF	UCFS	-	-	68.39	66.36	0.71	0.77
MF	XD-V	-	-	55.77	59.5	0.57	0.66
RLVS	CV	-	-	90.98	98.36	0.99	1
RLVS	HF	69.24	68.86	97	84	0.99	0.95
RLVS	MF	75.82	74.64	90	85.5	0.99	0.95
RLVS	RWF-2K	67.84	66.68	82.2	78.55	0.9	0.87
RLVS	UCFS	-	-	65.71	67.19	0.75	0.78
RLVS	XD-V	-	-	64.48	62.43	0.75	0.69
RWF-2K	CV	-	-	90.98	98.77	0.95	1
RWF-2K	HF	-	-	87.1	81.9	0.93	0.95
RWF-2K	MF	-	-	90.5	87.5	0.99	0.96
RWF-2K	RLVS	-	-	96.44	92.13	1	0.98
RWF-2K	UCFS	-	-	66.82	66.82	0.8	0.78
RWF-2K	XD-V	-	-	66.25	61.9	0.83	0.69
UCFS	CV	-	-	80.33	99.18	0.89	1
UCFS	HF	-	-	64	76.5	0.92	0.95
UCFS	MF	-	-	98	86.5	1	0.96
UCFS	RLVS	-	-	91.47	91.73	0.99	0.98
UCFS	RWF-2K	-	-	93.6	78.35	1	0.89
UCFS	XD-V	-	-	82.15	66.52	0.91	0.74
XD-V	CV	-	-	81.97	99.59	0.9	1
XD-V	HF	-	-	54.3	69.2	0.72	0.95
XD-V	MF	-	-	94	83	0.99	0.96
XD-V	RLVS	-	-	93.73	90.27	0.98	0.98
XD-V	RWF-2K	-	-	87.65	74.8	0.97	0.89
XD-V	UCFS	-	-	87.62	76.8	0.95	0.84
HF+MF+RLVS+RWF-2K+UCFS+XD-V	CV	-	-	88.93	72.13	0.95	0.92
CV+MF+RLVS+RWF-2K+UCFS+XD-V	HF	-	-	96.9	97.1	1	0.99
CV+HF+RLVS+RWF-2K+UCFS+XD-V	MF	-	-	97	99.5	1	1
CV+HF+MF+RWF-2K+UCFS+XD-V	RLVS	-	-	99.45	90.37	1	0.969
CV+HF+MF+RLVS+UCFS+XD-V	RWF-2K	-	-	97.25	91.45	1	0.98
CV+HF+MF+RLVS+RWF-2K+XD-V	UCFS	-	-	75.97	92.33	0.86	0.98
CV+HF+MF+RLVS+RWF-2K+UCFS	XD-V	-	-	70.25	93.07	0.91	0.98
HF+MF+CV	RLVS	70.08	69.84	99.45	98.95	1	1
HF+MF+RLVS	CV	76	75.68	77.87	80.74	0.95	0.95
HF+RLVS+CV	MF	81.51	80.49	99.5	99	1	1
RLVS+MF+CV	HF	79.87	78.63	61.4	97.1	0.94	0.99

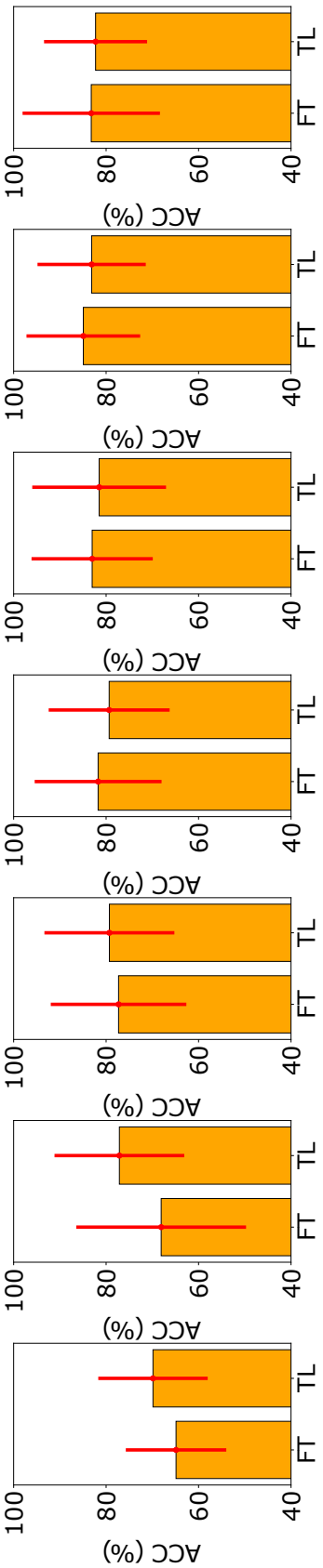

Figure 3.4: ACC scores for each of the datasets obtained by training on the specific dataset and averaging the testing accuracy scores on the rest of the datasets. Scores are indicated for both FT and TL models. Red lines indicate the standard deviation of the testing accuracy scores from the mean value (original illustration by Viktor Huszár).

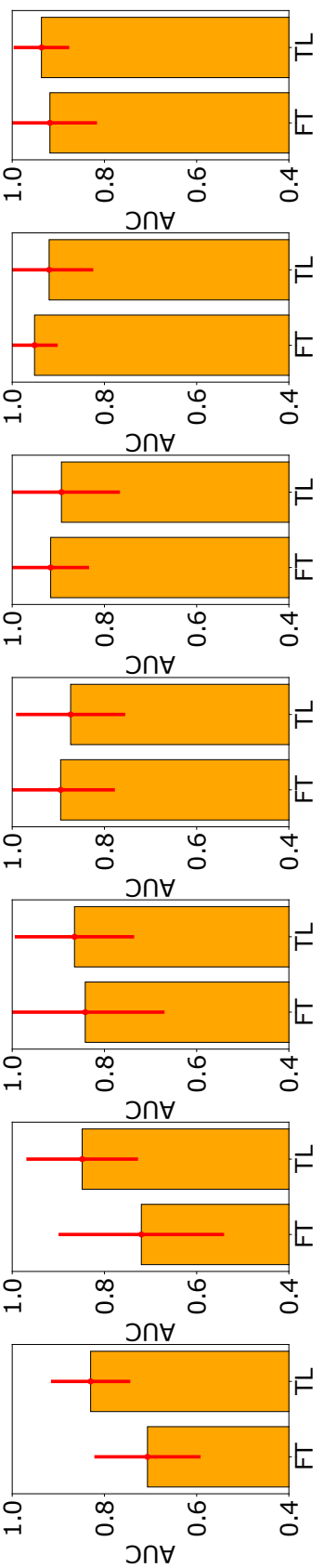

Figure 3.5: AUC scores for each of the datasets obtained by training on the specific dataset and averaging the testing accuracy scores on the rest of the datasets. Scores are indicated for both FT and TL models. Red lines indicate the standard deviation of the testing accuracy scores from the mean value (original illustration by Viktor Huszár).

For closer examination, I also conducted leave-one-out cross-validation tests where I trained my models on all datasets except one, which was reserved for testing. The results of these tests are presented in the middle section of Table 3.8. The tests suggest that when the CV dataset was left out of the training, the TL model did not achieve a good ACC score. This is expected because the TL model extracts features from training videos using a pre-trained X3D-M model that was trained on the Kinetics-400 dataset, which does not contain many examples involving crowd participation. However, the FT model achieved decent accuracy, indicating that the datasets other than CV contain a sufficient number of examples for learning about violence involving crowds. In line with the results obtained in the one-on-one cross-validation tests, leaving out the UCFS or XD-V datasets from training resulted in poor performance for the FT model. However, the performance of the TL model did not drop when these datasets were left out of the training, indicating that the TL model generalizes better than the FT model. To confirm this, I collected all instances of the one-on-one cross-validation tests when a specific dataset was being tested for further examination. In Figures 3.6 and 3.7, I plot the ACC and AUC scores obtained by averaging the testing accuracy scores on a specific dataset when all other datasets were used individually for training for both the FT and TL models. Each plot also shows the standard deviation of the metric scores obtained during testing, indicated by red lines. Based on these plots, it is evident that overall, the TL model showed the better capability to generalize and had lower standard deviation within the testing accuracy scores for individual datasets when compared to the FT model.

To the best of my knowledge, results from cross-validation studies are rarely presented in the literature for violence detection algorithms. For comparison, I have also included the cross-validation results from Violence-Net [155] using both OF and pseudo-OF inputs (columns 2-3) in the Table 3.8. Only ACC scores are provided since AUC scores are not presented in their original study. Also the authors of Violence-Net only used four datasets in their experiments, so results are presented only for these four datasets. The comparison results show that, on average, my TL and FT models consistently outperformed Violence-Net using both OF and pseudo-OF inputs. This suggests that my approaches are more accurate and better able to generalize to unseen scenarios for violence detection when compared to Violence-Net.

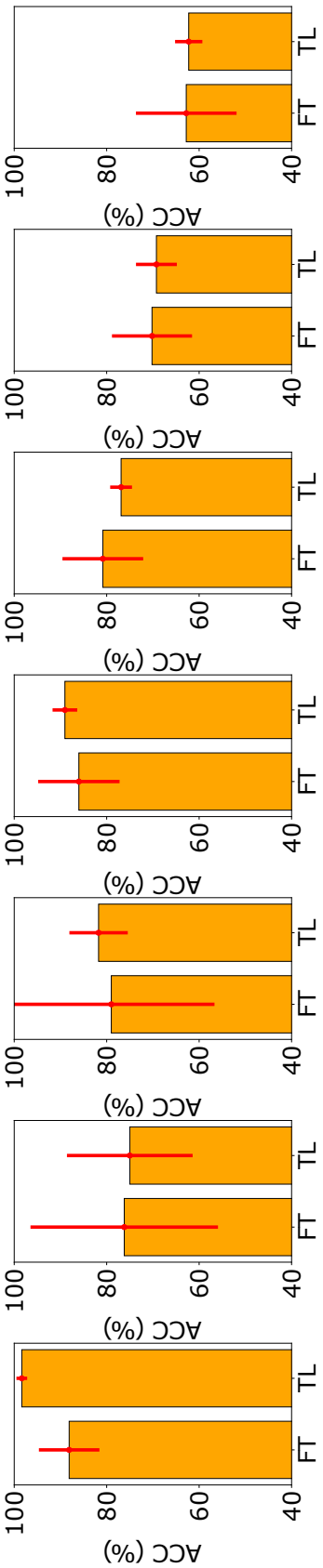

Figure 3.6: ACC scores for each dataset obtained by averaging the testing accuracy scores on the specific dataset when all other datasets are individually used for training. Average scores are indicated for both FT and TL models. Each plot also shows the standard deviation in ACC scores obtained during testing (original illustration by Viktor Huszár).

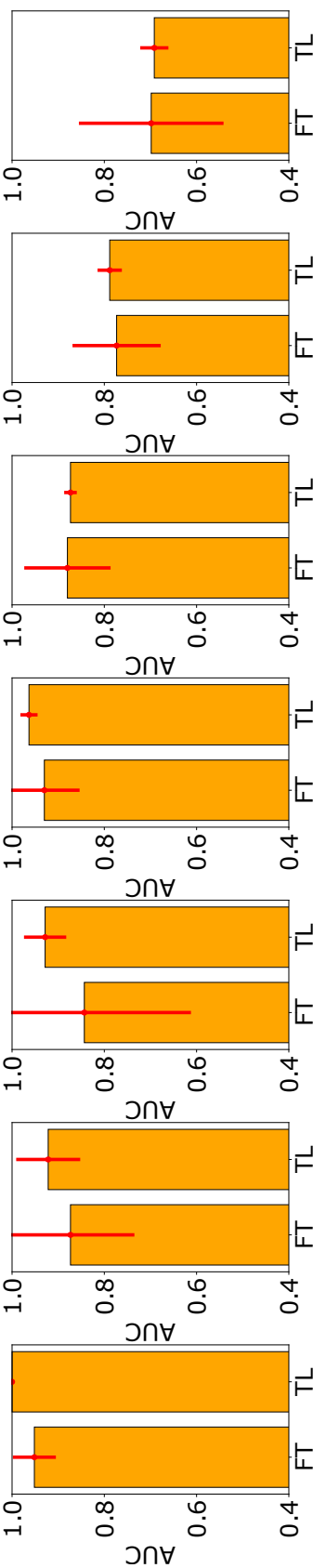

Figure 3.7: AUC scores for each dataset obtained by averaging the testing accuracy scores on the specific dataset when all other datasets are individually used for training. Average scores are indicated for both FT and TL models. Each plot also shows the standard deviation in ACC scores obtained during testing (original illustration by Viktor Huszár).

3.4.3 Experiments with all combined dataset

In this section, I describe my experiments using combined dataset and discuss the performance of the FT and TL models on this dataset. To ensure a fair distribution of training samples from each dataset, I selected and grouped the predefined 80% of the data from each dataset for training and the remaining 20% for testing. Figure 3.8 illustrates the proportion of samples from each dataset. The ROC curves, including the obtained ACC and AUC scores are presented in Figure 3.9. Results from both metrics suggest that my models performed satisfactorily on this dataset, with the FT model achieving slightly better performance. It is worth noting again that the TL model has fewer trainable parameters than the FT model.

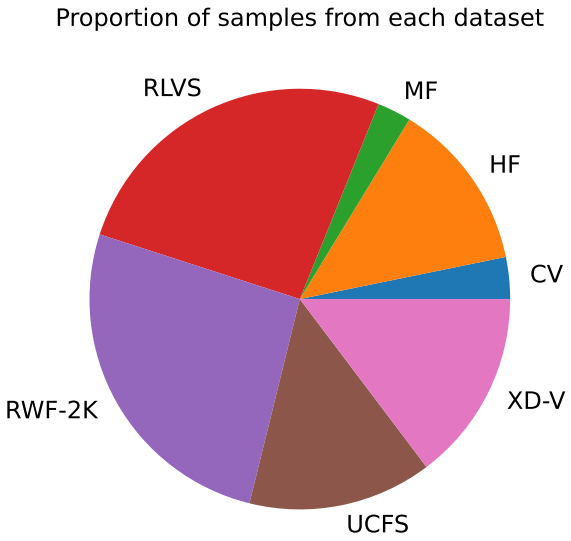

Figure 3.8: The pie chart illustrates the distribution of samples from different datasets used for training and testing in the combined dataset experiments (original illustration by Viktor Huszár).

For further analysis, I present the confusion matrices for both models in Figure 3.10. The rows of the confusion matrix represent the true labels, or the expected output, for the Violent (V) or Non-Violent (NV) classes, while the columns represent the predicted labels. In my case, the following are the four numbers presented in the confusion matrices:

- **True Positives (TP)** - the number of videos actually containing violence that were predicted as containing violence. TP are shown in the first row, first column of the confusion matrix.
- **False Negatives (FN)** - the number of videos actually containing violence that were predicted as not containing violence. FN are shown in the first row, second column of the confusion matrix.

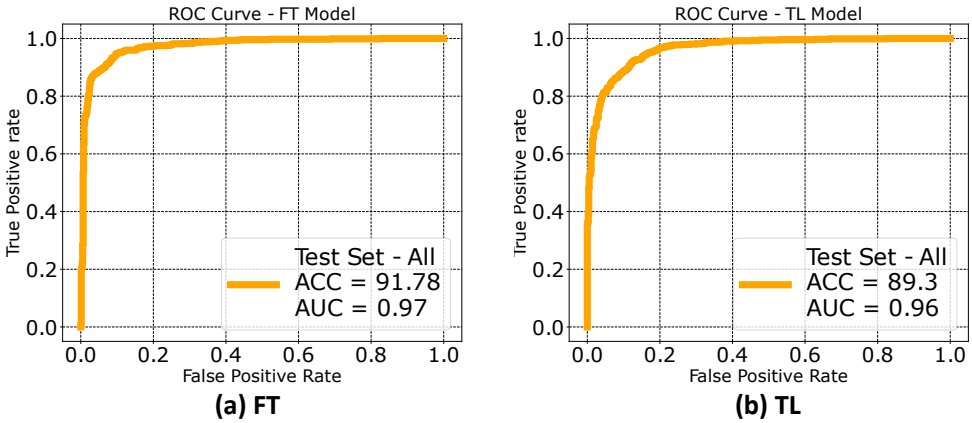

Figure 3.9: Performance of FT and TL models (left and right respectively) on all combined dataset shown using ROC curve. Both ACC and AUC scores show that the FT model performs better than the TL model on this dataset (original illustration by Viktor Huszár).

- **False Positives (FP)** - the number of videos actually not containing violence that were predicted as containing violence. FP are shown in the second row, first column of the confusion matrix.

- **True Negatives (TN)** - the number of videos actually not containing violence that were predicted as not containing violence. TN are shown in the second row, second column of the confusion matrix.

From the confusion matrices, it is evident that the TL model produced a greater number of combined FP & FN than the FT model. For detailed evaluation, I also studied and presented the metric scores and confusion matrices for individual datasets. Figures 3.11 & 3.13 show the results from the FT model, while Figures 3.12 & 3.14 show the results from the TL model. I note that overall, for both models, the number of FP & FN is balanced for all datasets, indicating that the training samples from both the violence and non-violence classes are balanced. Additionally, from the confusion matrices for individual datasets, it is clear that for most of the datasets, the TL model produced a greater number of combined FP & FN. My hypothesis is that the fixed nature of the extracted X3D-M features in the TL model does not provide sufficient flexibility to accurately recognize the attributes of violent actions.

3.4.4 Analysis of the datasets and Challenges

Even though the CV dataset has a smaller number of examples, both models trained on the combined dataset performed well on it. However, my leave-one-out cross-validation results indicate that when the CV dataset was excluded from training, the models did

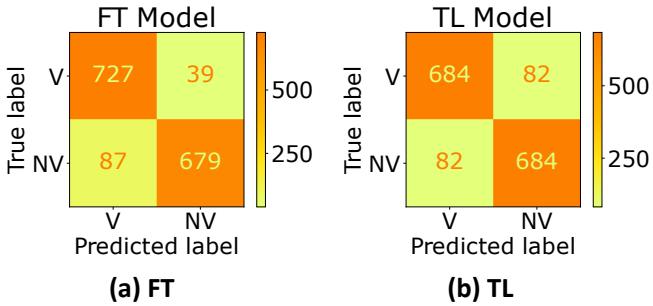

Figure 3.10: Confusion matrices for FT and TL models obtained after testing on all combined dataset. Results show that the TL model produced more combined false negatives & false positives than the FT model (original illustration by Viktor Huszár).

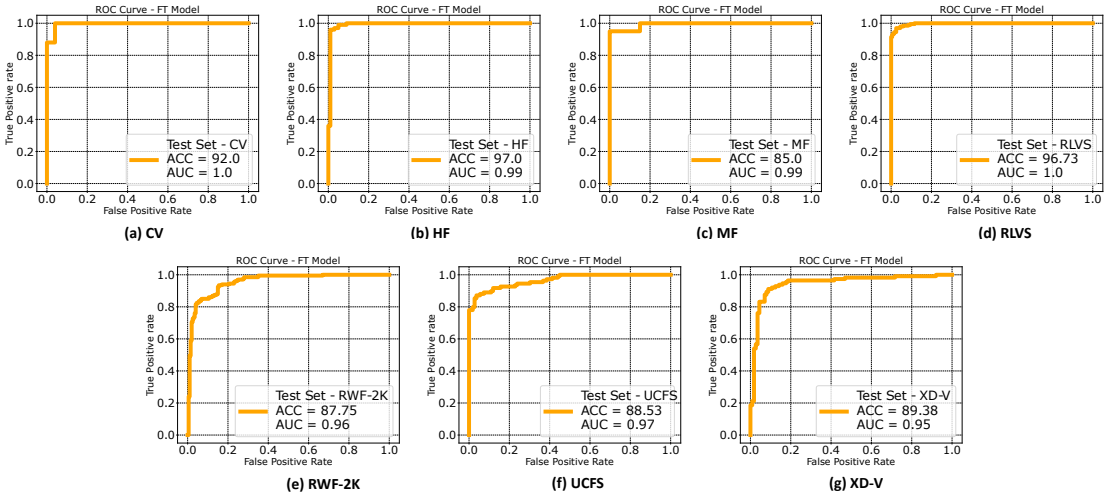

Figure 3.11: ROC curves including ACC and AUC scores obtained by testing on individual datasets using FT model trained on all combined dataset (original illustration by Viktor Huszár).

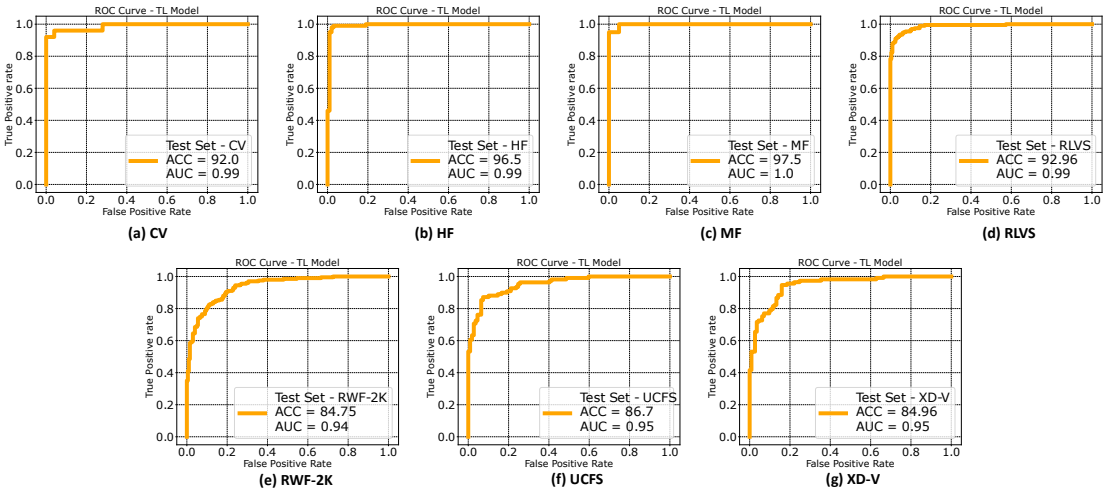

Figure 3.12: ROC curves including ACC and AUC scores obtained by testing on individual datasets using TL model trained on all combined dataset (original illustration by Viktor Huszár).

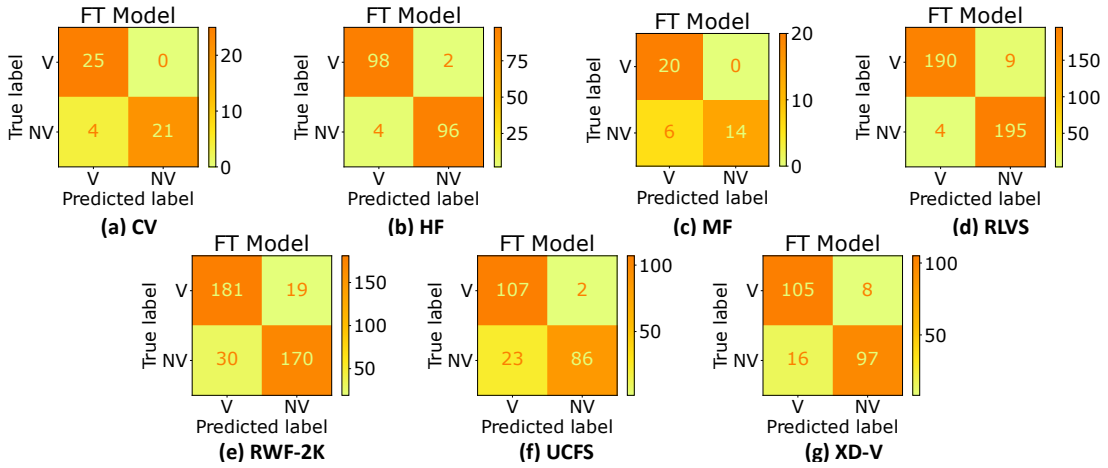

Figure 3.13: Confusion matrices obtained by testing on individual datasets using FT model trained on all combined dataset (original illustration by Viktor Huszár).

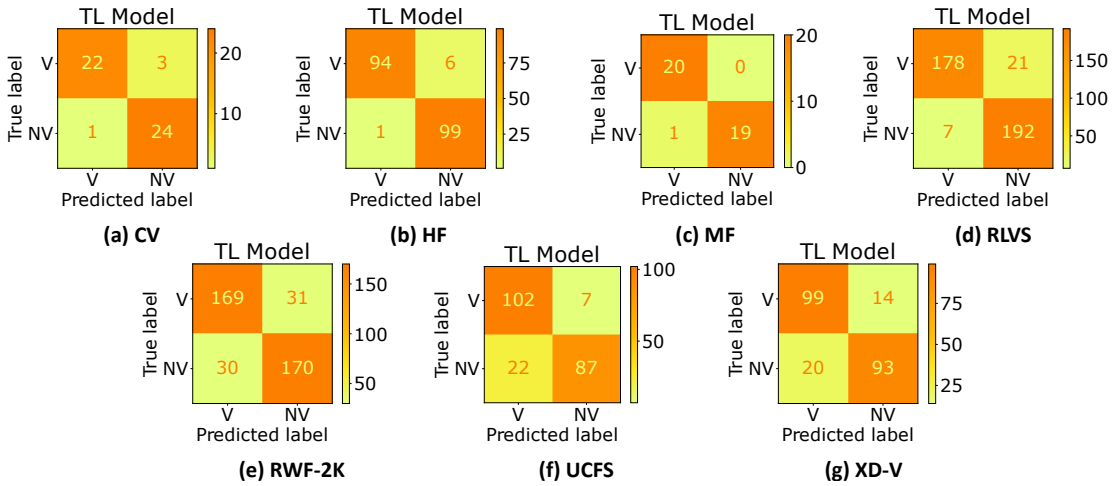

Figure 3.14: Confusion matrices obtained by testing on individual datasets using TL model trained on all combined dataset (original illustration by Viktor Huszár).

not perform well. This suggests that the CV dataset contains diverse and representative examples of crowd violence. However, it should be noted that the dataset only includes examples of violence involving crowds and the models trained on it did not generalize well to other types of datasets.

The HF dataset, on the other hand, contains a relatively larger number of training samples, primarily consisting of monotonous fighting videos between hockey players. Both my FT and TL models trained on the combined dataset performed well on this dataset as well. However, my leave-one-out cross-validation test revealed that excluding this dataset did not significantly decrease the accuracy of my models. Additionally, the model trained solely on the HF dataset did not generalize well to other datasets, as shown in figure 3.4. In line with my previous results on generalizability, highly monotonic datasets like the HF dataset are less useful for developing robust deep-learning models for violence detection.

Since my models use pre-processed input containing 16 uniformly sampled temporal frames, the duration of a video and the number of frames per second can affect the model's performance. The MF dataset has significant fluctuations in the FPS values of the training videos (as seen in table 3.2), which is not favourable for training my violence detection models. Additionally, this dataset has the least number of training samples compared to others and models trained solely on this dataset did not generalize well to other datasets. I hypothesize that these drawbacks of this dataset could be the reason for the decrease in the performance of the FT model (trained on the combined dataset) on this dataset. On the other hand, due to better generalizability, the TL model trained on the combined dataset performed well on this dataset.

In addition, my leave-one-out cross-validation test shows that the MF, RLVS, and RWF-2K datasets do not contribute significantly to model generalizability. The RLVS dataset mainly contains examples of two people fighting, which are also present in other datasets such as UCFS and XD-V. The RWF-2K dataset contains videos that are encoded at 30 frames per second, but I have observed that there are videos captured at very low fps, resulting in repeated frames to create 30 fps videos. Additionally, most examples in this dataset are repetitive in terms of environment and lighting conditions and lack diversity. However, it is important to note that the RLVS and RWF-2K datasets contain the highest number of examples, which can lead the model trained on the combined dataset to better represent scenarios in these datasets. I hypothesize that due to the aforementioned drawbacks specific to each of these two datasets, my models trained on the combined dataset did not perform very well on the RLVS and RWF-2K datasets.

Finally, my results show that models trained on my UCFS and XD-V datasets generalize better to other datasets (as seen in figure 3.4). Also, when these datasets were excluded from training, the performance of my models dropped significantly, indicating

that these datasets contain well-calibrated, diverse video footage, which is highly relevant for training practical deep learning algorithms for violence detection (as seen in table 3.8). However, these datasets contain a fewer number of training examples compared to RLVS and RWF-2K. Additionally, the UCFS and XD-V datasets contain forms of violence such as explosions and road accidents, which are not distinctly available in other datasets. Due to this, I hypothesize that my models trained on the combined dataset did not perform very well on the UCFS and XD-V datasets. Overall, with fewer false positives and false negatives, my FT model performed better on the combined dataset than the TL model.

3.4.5 Experiments with video compression

Depending on the available hardware resources, it may be necessary to stream the surveillance video to a remote server for actual classification and violence detection. Additionally, depending on the available network resources, there may not be sufficient bandwidth to stream the video in its native resolution and quality. In several fields where video streaming is involved, video compression techniques are commonly applied to reduce the video bit-rate, which can introduce artefacts in the video. To study the effect of such video artefacts on the performance of my TL and FT models, I generated compressed video streams with varying bit-rates - 300, 500, 1000 and 1500 Kbps.

For this experiment, I randomly selected two datasets, RWF-2K and CV and compressed the testing videos from these two datasets. Multiple videos Ire generated with the different bit-rates using ffmpeg [74]. I used the models trained on the combined dataset for this experiment and the testing results are presented in Table 3.9. My study shows that both TL and FT models did not show significant fluctuations in the performance and performed well even under extreme compression (300 Kbps). This suggests that my trained models did not model the noise in the training videos and focused on learning the concept of violence.

3.4.6 Standalone implementation and performance

We have implemented a standalone application for violence detection using the PyTorch deep learning library and using my FT and TL models that are trained on the combined dataset. The application design is outlined in Figure 3.15 and can be easily extended for usage in surveillance applications. The incoming video stream is divided into non-overlapping video segments of four seconds, from which 16 video frames are extracted per segment using uniform temporal sampling. These 16-frame blocks are pre-processed and then used as input for either the FT model or TL model to determine a violence

Table 3.9: Results from video compression experiments - The top section shows results for the CV dataset and the bottom section shows results for the RWF-2K dataset using ACC and AUC metrics.

Dataset (bitrate in Kbps)	ACC-FT	ACC-TL	AUC-FT	AUC-TL
CV (1500)	92.0	92.0	0.99	0.98
CV (1000)	92.0	92.0	1	0.98
CV (500)	92.0	92.0	0.99	0.98
CV (300)	92.0	92.0	0.99	0.98
RWF-2K (1500)	88.25	85.0	0.95	0.93
RWF-2K (1000)	89.0	85.0	0.95	0.93
RWF-2K (500)	88.5	84.25	0.95	0.93
RWF-2K (300)	88.5	83.0	0.96	0.92

coefficient for the current segment. The application was implemented on an Ubuntu Linux operating system using an AMD Ryzen Threadripper 1950X 16-core processor and a Nvidia GeForce GTX 1080 Ti GPU with the CUDA toolbox for running my trained PyTorch models.

my results indicate that, when combined with block extraction and pre-processing, both the FT and TL models require an average of 0.06 seconds on average to infer a violence coefficient for each four second-video segment. The pre-processing was implemented on the CPU, consuming an average of 0.04 seconds. Therefore, the average time required to run the FT or TL model is 0.02 seconds. It should be noted that, the dense or fully connected layers of the models consume minimal computational resources in practice. As a result, even though the TL model has more parameters than the FT model, the average time required to run both models is similar.

To give a thorough understanding of the performance of my standalone system, I have graphically represented the progression of violence coefficients over time using the FT model that showed the most optimal results on the combined dataset. In figures 3.16 to 3.20, I illustrate my classification outcomes on selected video samples that exhibit the capability of my system in identifying violence. Each figure comprises three sections: the top row displays the actual graph of violence coefficients over time, where the coefficients are set to one during the occurrence of violence. The middle row illustrates the predicted violence coefficients by my FT model on a series of non-overlapping video segments with a duration of four seconds. The bottom row shows key frames extracted from the videos.

Figures 3.16 to 3.19 demonstrate the performance of my standalone system on video clips from the testing set of the original UCF-Crime dataset. These video clips include

different scenarios such as instances of violence amidst normal events (as illustrated in figure 3.16), multiple occurrences of violence (as illustrated in figure 3.17), a crowd engaging in violence at a metro station (complex and long video sequence as illustrated in figure 3.18), and a single, short instance of violence in the form of shooting (as illustrated in figure 3.19). The predicted violence coefficients align closely with the ground truth, indicating the algorithm's capability to accurately identify and predict instances of violence in video segments of various lengths and complexities.

To further evaluate my system, I also created a video sequence by combining random video clips from the Smart-City CCTV Violence Detection Dataset [185], which was not used in my study. As shown in figure 3.20, my results exhibit outstanding performance on this compiled sequence as well. This illustrates the adaptability of my algorithm and its capability to perform well on new and unseen data.

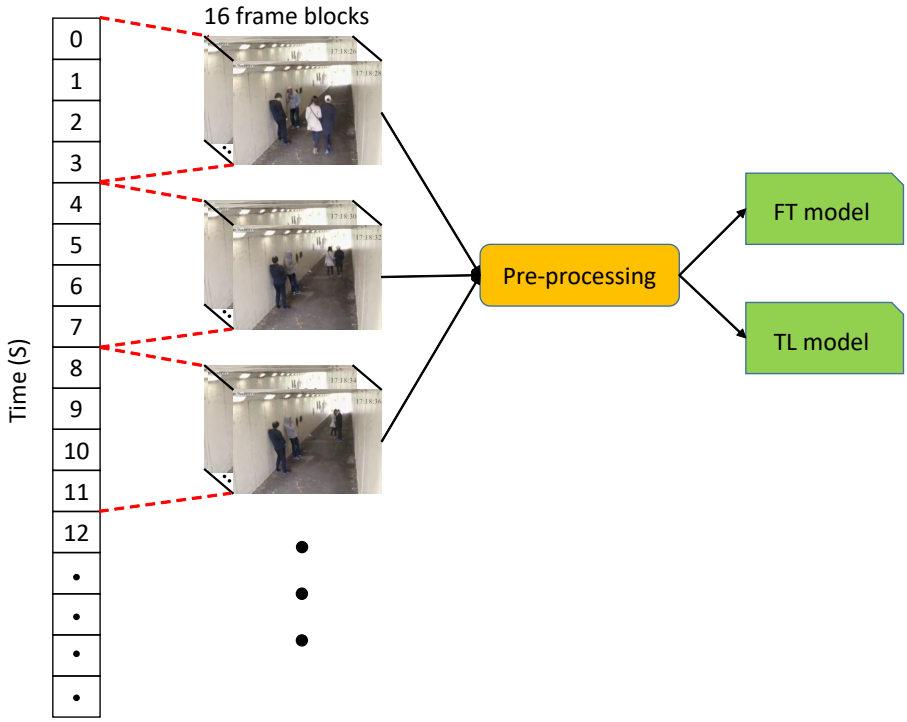

Figure 3.15: Schematic diagram of my standalone application. 16 frame-blocks are extracted from each four-second video clip which are pre-processed and input to FT or TL model (original illustration by Viktor Huszár).

Figures 3.16 to 3.20 also demonstrate the areas where my standalone system falls short, which require further improvement in future research. I have noticed certain situations where my tested FT model triggers false alarms in the standalone implementation. For instance, when a person suddenly starts running or crawling (as shown in the keyframes from the 28th second in figure 3.16), it is detected as violence, but with a lower level of violence coefficient. In the original UCF-crime dataset, activities such

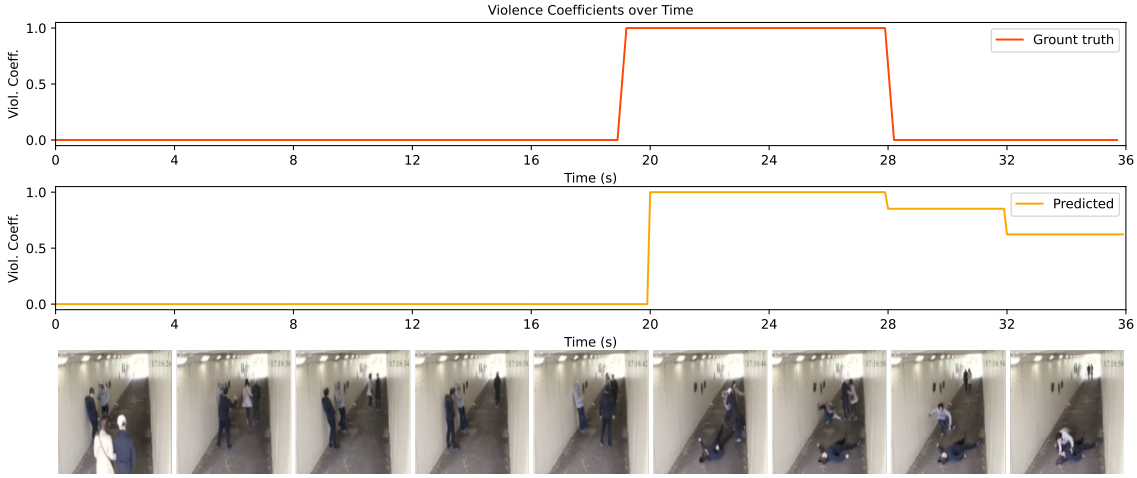

Figure 3.16: Results of my standalone system using FT model on a video clip from the testing set of the original UCF-Crime dataset. The video clip includes an instance of violence between two normal events. Row 1: Ground truth of violence coefficients over time. Row 2: Predicted violence coefficients on 4-second video segments. Row 3: Keyframes extracted from the video (original illustration by Viktor Huszár).

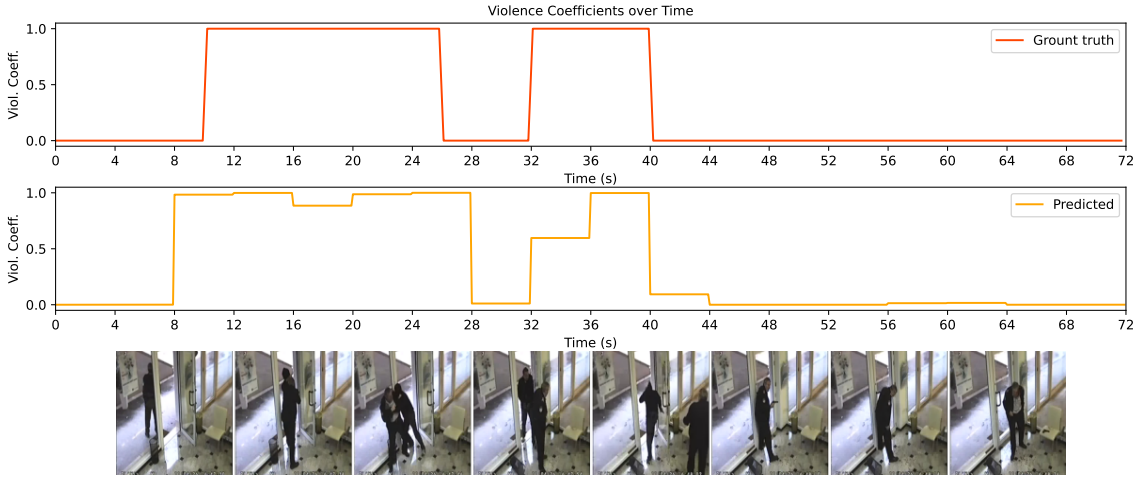

Figure 3.17: Results of my standalone system using FT model on a longer video clip from the testing set of the UCF-Crime dataset with two instances of violence. Row 1: Ground truth of violence coefficients over time. Row 2: Predicted violence coefficients on 4-second video segments. Row 3: Keyframes extracted from the video (original illustration by Viktor Huszár).

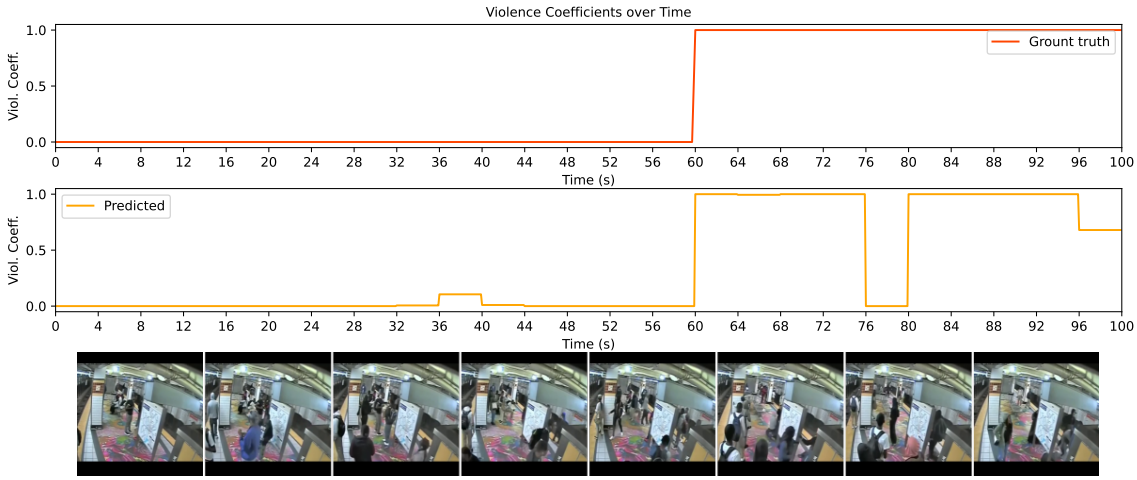

Figure 3.18: Results of my standalone system using FT model on a complex and longer video sequence from the testing set of the UCF-Crime dataset, showing a crowd involved in violence at a metro station. Row 1: Ground truth of violence coefficients over time. Row 2: Predicted violence coefficients on 4-second video segments. Row 3: Keyframes extracted from the video (original illustration by Viktor Huszár).

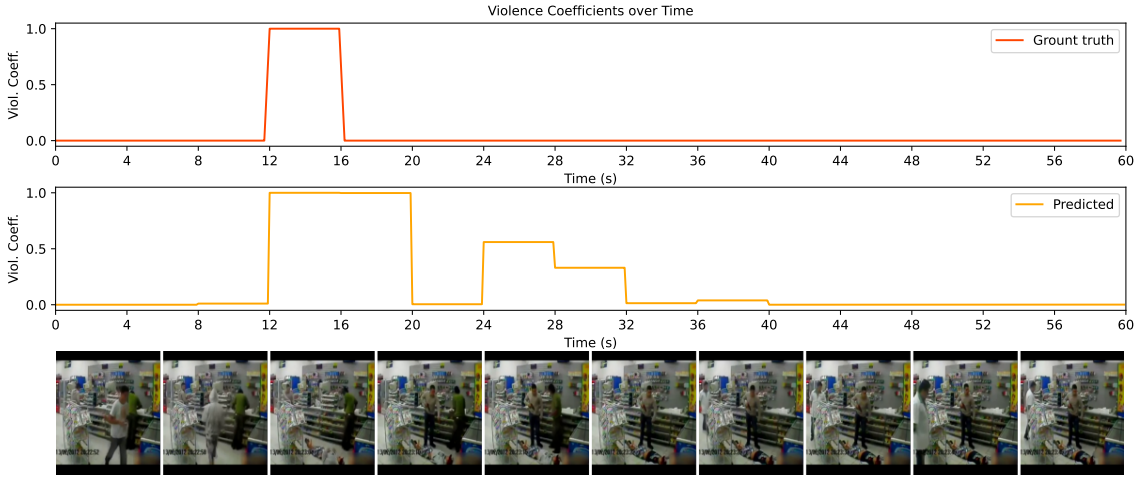

Figure 3.19: Results of my standalone system using FT model on a video clip from the testing set of the UCF-Crime dataset that contains a single and short instance of violence in the form of shooting. Row 1: Ground truth of violence coefficients over time. Row 2: Predicted violence coefficients on 4-second video segments. Row 3: Keyframes extracted from the video (original illustration by Viktor Huszár).

Figure 3.20: Results of my standalone system using FT model on a video compiled from 3 random videos of Smart-City CCTV Violence Detection Dataset, one violent and two non-violent videos concatenated in such a way that the violent video is placed in between two non-violent videos. Row 1: Ground truth of violence coefficients over time. Row 2: Predicted violence coefficients on 4-second video segments. Row 3: Keyframes extracted from the video (original illustration by Viktor Huszár).

as crawling or sudden fleeing are not considered violence. Nevertheless, in real-world surveillance scenarios, such actions may appear suspicious and require more investigation.

In situations where there is occlusion and the individuals or objects engaged in violence are only partially visible, the model may have difficulty identifying the violence. This can be observed in the predicted violence coefficients between the 32nd and 36th second in figure 3.17, where a person is holding a gun in his hand which is partly visible and hidden by his body.

As previously noted, videos that include people in crowds situated closely together can lead to inaccuracies in my system. Figure 3.18 between seconds 76 and 80 illustrates this scenario, where the predicted violence coefficient suddenly falls to zero even though violence is happening during this time. As mentioned earlier, my training dataset has limited examples of crowds, and including more such examples in future work is suggested.

3.5 Concluding remarks

I proposed two violence detection architectures, FT and TL, for automated surveillance applications using the computationally light X3D-M deep learning architecture. I evaluated the models using seven annotated video datasets, including Kinetics-400 and UCF/XD-Violence. My experiments showed that both models performed well, with FT outper-

forming most state-of-the-art methods on popular datasets. Cross-dataset validation showed that TL generalized better to unseen scenarios but had a higher number of false positives/negatives compared to FT on the combined dataset. My results indicate a need for diverse and representative large-scale violence datasets for practical applications. I also presented a computationally light system architecture for implementing the proposed models in practical surveillance, with a focus on optimizing computational speed and accuracy. In the future, I plan to construct a large-scale dataset and improve the system by handling scenarios where violence spans multiple video segments.

Chapter 4

Blockchain in AI

4.1 Introduction

Military defence areas for blockchain-based applications and solutions raise a wide range of scientific questions. In terms of military use, governmental development puts emphasis on hybrid warfare and globally substantially more funding is available for research and development in the field. Blockchain technology allows voluntary, distributed networks created for military purposes to cooperate with a cryptographic process without central and state control. Blockchain technology consequently results in a digital paradigm shift, characterised by decentralization, blockchain technology, machine learning, and artificial intelligence.

The potential applications of the blockchain raise numerous military technical challenges. Blockchain is a computer science term for a distributed data storage technology, a distributed ledger database that stores a list of entries in ever-increasing blocks. Each block also contains a link to the preceding block on each node that stores a structured chain. An essential feature of systems using blockchain is the storage of blockchain nodes in all structured entries using a consensus algorithm. Despite the fact, that bitcoin was an early adapter of the technology, today there are numerous systems under development that follow the same principle but differ fundamentally from bitcoin in their purpose and key technical elements. However, these systems together are often referred to as blockchain technology, not too precisely. The technology allows voluntary, distributed networks to be cryptographically deployed in a robust manner, without governmental and central state control. In addition to banking systems, virtual money and the development of Smart Contract [11], real estate sales, exchange of assets and movable property are also emerging. However, military applications are even more interesting, as data security must be a priority in defence administration and in the daily communication of authorities.

For blockchain-based military use, there are several scientific challenges depending on the used protocol. The need for centralized data storage and an uncontrolled security management system should be explored for the efficient use of resources. The problem extends to the artificial isolation of such systems and the military risks of „awaking" machine learning or programmed artificial intelligence. Science should investigate how data security, data integration, and isolation of the artificial intelligence decision-making environment and the framework for authorization levels can be achieved in such an automated distributed network-based military environment.

Currently, due to centralized data storage and central control, breakable data communication between military and police departments is a major problem, especially in less developed countries. The most important aspect in terms of the degree of vulnerability is the activity of the user or the organization and - closely related to this - the value of their data. Particularly popular targets of attackers are financial institutions and organizations dealing with state or classified documents.

Blockchain technology also raises the issue of user profiling. The problem is that the long-term use of the blockchain may allow monitoring of user behaviour and the use of profiling. The Government regulations require data protection measures based on the complete knowledge of a specific system, the personal data it processes and the related data processing operations [186].

Police forces aim at reducing illegal activities through appropriate regulations as there has been a transition from traditional means of payment to blockchain-operated cryptocurrencies. Central authorities, including the central bank of a country, may find it easier to filter the purpose of the cash flow, so money laundering, illegal substances or weapons will not be completed with cryptocurrency payments. The main argument supporting blockchain and cryptocurrencies like bitcoin is based on the transparency of the transactions, which are all public. On the other hand, the analysis of public blocks lacks any KYC process (which is increasingly propagated by regulators) to allow the identification and exclusion of prohibited operators [12].

4.2 Cyberspace

Cyberspace is characterized by networked systems and their physical attributes, where data is stored, exchanged, and modified. New technologies have emerged on the network that can revolutionize many industries. Disruptive innovations are Artificial Intelligence and computer vision. The potential uses of deep neural network learning capabilities represent military scientific challenges, consequently, they create possible new platforms for hybrid warfare. Very few scientific articles have explored the combination of Artificial

Intelligence and computer vision applications on blockchain networks. The result of the research is a fault-tolerant, real-time, cost-efficient military use of existing governmental computing power. Further research is needed to develop organic expertise in blockchain technologies within the Central Defense Management Authorities.

4.3 Hybrid warfare in cyberspace

In Hungary, the Zrínyi 2026 Defense and Armed Forces Development Program has recognized that new types of challenges require special attention in order to build, maintain and develop the Hungarian cyber defence capabilities [187]. In June 2019, the Cyber Training Center of the Hungarian Armed Forces was established, which serves as one of the most important pillars of our hybrid force development strategy. The cyber defence force's approach relies on the appropriate infrastructure and equipment, but the establishment of a Hungarian cyber academy created the ability needed for hybrid warfare: digitally trained soldiers. The Cyber Training Center can serve a dual purpose, because, in addition to supporting cyber defence capability development and harmonization, it can also directly become an institutionalized military technology research and development centre, which fundamentally determines the defence capability of a country [188]. Previous studies suggest that Hungary should specialize in IT areas such as electronics and software development. Given the fact, that blockchain-based military applications globally have not been explored, a key research target area should be distributed ledger systems. Countries like Hungary can gain a globally competitive military advantage with computer scientific superior knowledge, which results in government-civil-military interoperability, and educational, economic and social added value. The development of hybrid warfare is therefore beneficial for key national objectives.

Distributed General Ledger (DLT) technology is able to make optimal use of new innovations such as artificial intelligence and machine vision [189]. A distributed general ledger is a database of digital data distributed (decentralized), distributed, shared, and synchronized across multiple geographic locations, countries, or institutions. The potential use of deep neural network learning capabilities that can be run on such a general ledger poses a number of military scientific challenges and consequently creates new platforms for cyber operations.

The everyday work of military and police forces has been supported by security camera feeds. Computer vision and automated image analysis could save military resources. An image analysed means the actual image data structure changes: images become image descriptions, which can be classified into image groups. The purpose of computer vision is to create three-dimensional models based on images or videos [190]. This requires

data processing, analysis, and image recognition, both of which require high computing power, so current image analysis methodologies are often slow and do not operate in real-time. Blockchain-based networks achieved outstanding computing power capabilities. Shockingly, in 2013 all mining computers had a combined computing capacity of 250 times the capacity of the 500 largest supercomputers [191], and the mining community's aggregate consumption in 2017 was higher than the average annual electricity demand of 159 countries [192]. It is legitimate to use blockchain-based technologies to help machine vision, thus reducing expensive hardware and resource requirements. Without a resource-efficient IT backend, computer vision research expenses can get out of control. The price of the first Hungarian 5G automotive test track, proves the massive cost related to the implementation of innovative machine vision-based R&D results [193].

Blockchain technology might be a revolutionary research field for military engineering, but there is still space for understanding better the advantages and disadvantages of decentralization, by studying the international military applications that have already been implemented or are under current development.

Hungary's National Security Strategy 2020 states in its introduction that "Supporting the domestic defence industry, including R&D and innovation, is in the national security interest, as it can reduce import dependence, increase the security of supply and modernise defence equipment with domestically produced products." (Machine vision and related artificial intelligence research are a priority for the defence sector and are widely used in countries with the most advanced technology, but in the Hungarian law enforcement and defence environment the use of these solutions is rudimentary. The applications of machine vision-based artificial intelligence raise a multitude of military engineering science challenges. However, the last few decades have shown what we already knew - information technology is not only evolving but is disrupting many industries by offering unique opportunities for growth and development. Next-generation technologies, machine learning and artificial intelligence are giving the military and law enforcement the opportunity to transform and continue to retain their respective industries.

The United States of America was already using digital technology in the 1990s, such as network-centric operations, which combined "tactics, technologies and procedures that a networked force can employ to gain a decisive battlefield advantage" [194]. In 1995, Bower and Christensen shone the spotlight on "disruptive technologies" that can help competitors stay ahead of the competition. Their research showed that the majority of well-established companies are determined to stay ahead of their industry in the "development and market application of new technologies", and that this can range from step-by-step developments to progressive, industry-wide approaches - with the main aim of such developments being "targeted at the next generation of their customers' needs"

[8]. Around the same time, a pervasive information network based on blockchain technology emerged - a technology capable of storing data securely without central control and without modification [9]. Jump forward a few decades and you can see that blockchain DLT has had a profound impact on many sectors, greatly influencing economic systems, legal frameworks and information technologies [10]. It is worth noting that while commercial industries have done their utmost to gain an advantage in blockchain technology and have subjected the shortcomings of regulation to some degree of scrutiny, much less attention has been paid to the potential of blockchain technology and its vulnerabilities in the areas of military intelligence and law enforcement. The following subsections define blockchain technology in broad terms to anticipate the possible links between the technologies in terms of their potential applications in the military intelligence and law enforcement sectors.

The arena of hybrid warfare is intangible, ”cyberspace is magical”, a living, pulsating organism kept in constant motion by the innovation and digital revolution. The military definition of cyberspace is broader than the usual, civilian understanding of the term. Cyberspace is characterised by network systems and their physical properties where data is stored, exchanged and modified. The 2013 National Cybersecurity Strategy defines cyberspace as a collection of globally interconnected, decentralised information systems. The Force Operational Domains define four possible physical domains (land, air, sea, and space), but the intent behind data and information activity in cyberspace remains a key question and an important topic of military engineering science research. New technologies have emerged on networks that could transform military industries. The blockchain is putting cyberspace in a new context: distributed ledger technology (DLT) can harness new innovations such as the combination of artificial intelligence and machine vision in a resource-optimised way. A distributed ledger is a database of digital data distributed (decentralised), replicated, shared and synchronised by consensus across multiple geographic locations, countries or institutions. The potential use of deep neural network learning capabilities that can run on such a ledger poses a number of military science challenges and consequently creates new platforms for cyber operations.

4.4 Blockchain in cyberspace

The potential applications of blockchain raise a multitude of military engineering science problems. Indeed, the technology enables voluntary distributed networks to act together robustly using cryptographic techniques without state control. In addition to banking systems and virtual money, the development of smart contracts [11] is opening up new opportunities for the sale of real estate, the exchange of assets and movable property.

However, military applications are even more interesting, as data security is a key issue in defence management and in the day-to-day communication of public authorities.

For blockchain-based military applications, there are several scientific problems depending on the protocol. For efficient use of the resource, the justification of centralised data storage and an uncontrolled defence management system needs to be explored. The scope of the problem includes the artificial isolation of such a possible system, the military risks in case of a possible machine learning and programmed artificial intelligence 'waking up'. The science needs to investigate how such an automated, distributed network-based military environment of use can be made data secure, data integrated, and isolated from the AI decision environment, with a framework of privilege levels.

Currently, central data storage and central control means that hackable data communication between departments is a major problem. The most important aspect of the level of vulnerability is the activity of the user or organisation and, closely related to this, the value of their data. Financial institutions and organisations handling state or official secrets are particularly popular targets for attackers.

Another issue related to blockchain technology is user profiling. The problem is that the long-term use of blockchain may allow monitoring of user behaviour and so-called profiling. In the legislator's view, this issue can only be assessed in the context of a full understanding of a specific system, the personal data it processes and the related processing operations.

An interesting issue from a national defence perspective could be the reduction of illegal activities in the case of a shift from conventional to cryptocurrency through appropriate regulation. It would be easier to filter the purpose of the money flow so that cryptocurrency payments would not be used to obtain illegal drugs or weapons. One side says that blockchain and cryptocurrencies like bitcoin should of course be fully public, but anonymous. On the other side, the argument is that public blockchain analysis should be coupled with banks and KYC (know your customer) processes (a widespread principle of customer identification that regulators increasingly expect) to allow the flagging and exclusion of prohibited actors from the market [195].

In image analysis, the data structure changes: images become descriptions. Shape recognition operates on image descriptions and creates object classes. Finally, computer vision aims to build three-dimensional models from images or videos [196]. This requires processing, analysis and recognition, which also require large computational resources, so current image analysis methodologies are often slow and do not operate in real-time. The combined computational capacity of all mining computers already exceeded 250 times the capacity of the 500 largest supercomputers in 2013 [191], and the aggregate consumption of the mining community in 2017 was greater than the average annual electricity demand of 159 countries [192]. It is reasonable to consider the use of

blockchain-based technologies to aid machine vision, thus reducing expensive hardware and resource requirements. The cost of the Zalaegerszeg automotive test track shows how expensive it is to implement innovative machine vision-based R&D results [197].

4.5 Scientific problem

The integration of blockchain technology with artificial intelligence (AI) and computer vision brings about a range of complex military, technical, scientific, and legal issues. The collection of data, which can come from various sources such as air, land, sea, and space, is crucial to the operation of cognitive processes and the storage, modification, or exchange of data. To secure the data, capturing it is essential, as AI systems need access to data for their learning processes. However, obtaining military data can be a challenge as it is often classified and secret, making it both legally and financially difficult to access. There is also the risk of human error in the data obtained, even when encrypted, and the need for further validation.

The use of AI and computer vision in military applications poses additional challenges such as the source of data, the encryption algorithms used, and the AI modelling structure. Real-time processing of data can also be difficult, especially with limited resources. The centralised data storage and lack of control over security management are also areas of concern. Science must explore the possibility of achieving data security, data integration, and separating decision-making environments for AI in a distributed network-based military environment.

Communication between departments can be a challenge due to centralised data storage, making it vulnerable to attackers, particularly financial institutions and organisations dealing with state secrets. The use of manually obtained data can also conflict with regulations, such as the GDPR, which limits the use of personal data.

In image analysis, the data structure changes, with images becoming data descriptions for shape recognition. Computer vision aims to create three-dimensional models from images or videos, but this requires high computational power, making current methodologies slow and inefficient. Energy resources can also be a challenge, making it expensive to implement new machine vision-based R&D outcomes. Distributed network systems can offer solutions, but also bring about new cyber warfare dimensions such as network warfare, electronic warfare, and computer network operations.

The following subsections broadly define blockchain technology to project possible relationships between AI and computer vision technology concerning implementation and utilization within the military intelligence and law enforcement sectors. Computer vision-driven AI developments demand high computing power resources. Most govern-

ments have enormous computing power resources currently not capitalized: nodes of computers with less sensitive data, connected to a network, used on a daily or even continuous bases. AI systems rely on deep neural network computations, and faster results can mean more efficiency. Combining AI, computer vision, and data-encrypted computation with DLTs and blockchain seem an obvious choice. The potential applications of AI and computer vision obtained data raise numerous military technical scientific and legal issues. Data has always been sourced from cyberspace, independent of the source of the data coming from the domains of air, land, sea, and space. The captured data connects these physical domains with the cognitive processes that use the data for storage, modification, or exchange purposes. Cyber superiority can be obtained by capturing the data, as emergent AI systems need data feeds for faster learning curves.

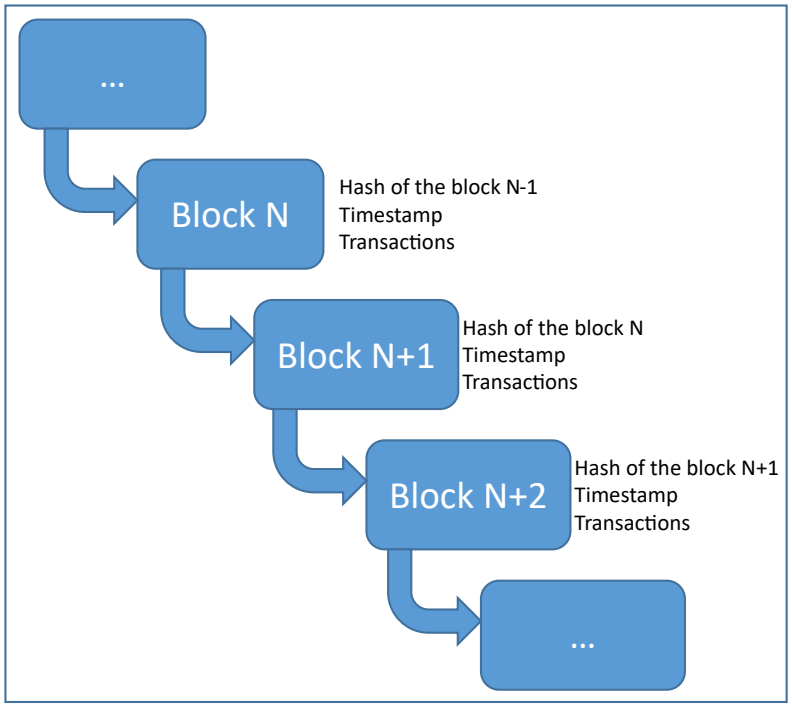

Figure 4.1: Structure of Blockchain (Based on the work of Pinna and Ruttenberg [198]).

4.6 Blockchain technology defined

Blockchain is a distributed data storage approach, where “blockchain” describe the list of data clusters (blocks) that are organized into a limitless chain creating a distributed database. Each block contains a link to the preceding block on each node (participant) that stores the blockchain. A basic feature of blockchain systems is the storage of blockchain nodes in sorted entries with unified agreement on the current state using a consensus approach. Though this approach to distributed storage has been popular-

ized through Bitcoin's distributed "cryptocurrency" framework, many alternative systems currently exist or are under development that follows the same principle but differ fundamentally in their purpose and key technology. As it stands, these systems are collectively and improperly referred to as "blockchain technology".

Blockchain can decentralize transactions while increasing security that has tech entrepreneurs investing in what is most likely the next technological revolution that will greatly impact and transform society, economies, and the internet (Tapscott and Tapscott 2016 [199]). The importance of blockchain's distributed fault tolerance and the seamless transaction has already been recognized by the military sector and research is underway to determine whether existing systems can be successfully migrated to the blockchain, in whole or in part.

In the context of blockchain technologies, the ledger is an entry repository where entries can be stored without the possibility of modification once they become part of the database - the ledger may also demonstrate narrow "ledger" semantics depending on the blockchain technologies applied but this is an irregularity. Blockchain technologies implement and expand a distributed ledger through continuous synchronization with nodes in the distributed network. The network can be geographically distant or owned by various companies, so each node has a copy of the ledger. Any additions to the ledger require agreement by other nodes, followed by the verified block's appearance within minutes or seconds on other nodes, depending on the solutions used. Any trusted central monitoring body can access the information stored in the entries, without involving said body's internal processes and rules [198].

The ledger is maintained by distributed network nodes based on several consensus algorithms employing cryptography to store and verify transactions. This allows the network to remain functional even with a large number of defective nodes, provided that the number of defective nodes is below the maximum allowed. It applies the distributed consensus algorithm or, more generally, the distributed consensus protocol a great deal. In each application context, the selection of the consensus protocol is influenced by factors such as hypothesized failure modes, maximum system size, consensus response time, and synchronization requirements. Accordingly, it is not surprising that different blockchain technologies utilize several different consensus protocols. What is uniform in blockchain technologies is that the problem of distributed consensus is addressed by some protocol.

Blockchain has a common structure and can be viewed as a transaction log (journal) whose data clusters are stored in blocks in a strictly chronological order. As shown in Figure 1, these blocks are time-stamped and identified by a selected cryptographic hash. Each block contains a reference to the block preceding it, and in this way, the blocks are organized into a backwards-chained list which, at its worst, can be processed from the first block to determine the current state of the distributed database. Of course, this

implies consensus between nodes on the blockchain. If inconsistent copies of a chain begin to spread within the network, the discrepancy is typically resolved through one of the following consensus protocols applied via mining nodes: proof-of-work (POW), proof-of-stake (POS), or round-robin (mining diversity)."

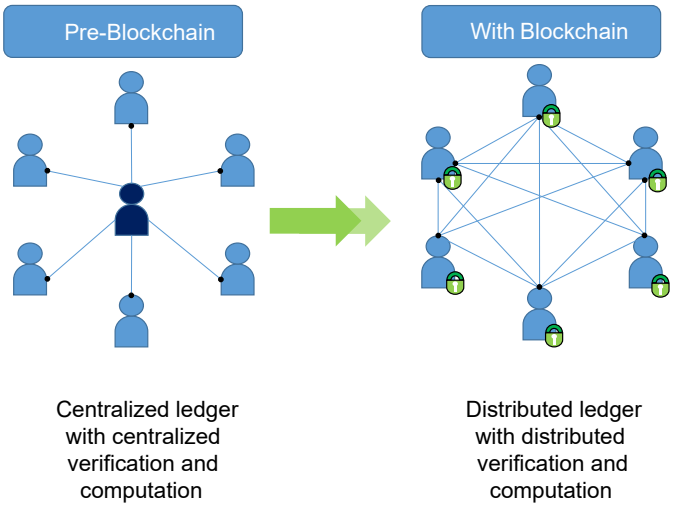

Figure 4.2: Centralized vs Decentralized Data Distribution (Based on the work of Perera et al. [200]).

The decentralized nature of blockchain technology (figure 4.2) removes the need for a central authority or checkpoint, which ultimately creates a fairer, more secure system. The way data is recorded on the blockchain reflects the value of decentralization (Dwyer 2014, [201]). Instead of relying on a central authority to secure transactions with other users, blockchain uses innovative consensus protocols on the node network to authenticate transactions and record data in an unbiased manner. Thus, the blockchain is not stored by a central data controller but by numerous computers.

The unbiased manner of recording and distributing secure, immutable data makes blockchain an asset capable of limitless potential with regard to cyber security and other military applications. To understand the range of blockchain technologies for tactical sustainment challenges, László Kovács suggests the military examines the potential of blockchain solutions to challenges associated with in-transit visibility, data integrity, reporting, operational contracting, and logistic estimation (Kovács 2019 [202]).

Authors McAbee, Tummala, and McEachen [203] undertook the survey of several examples of "military intelligence-specific guidance" frameworks considering the adoption of blockchain technology. The authors identified a key quality they deemed mandatory, that being the collaborative process in which several authors shared control. Peck's model illustrated flexibility "in potential employment, which suggests that even in the presence of other disagreeable factors blockchain technology may be worth considering in cases where the database is likely to be attacked." Military intelligence would benefit

				Problem			*Transaction*					*Participant*					*Solution*				*External Factors*	
Author	Sector	Type (C: Criteria, F: Flow Chart)	Year	Solvable without Blockchain	Intermediaries	Centralization	Rate	Size/Data Volume	Digital Nature	Privacy Expectations	Strict Immutability Required/Possible	Interdependency	Quantity	Shared Control	Integrity Motivations	Trust Level	System/Software Control	Stand-Alone System	Current Capability	Permission Type	Likelihood of Attack	Regulation Compliance
Greenspan	Blockchain Development	C	15	X	X	X						X	X	X		X				X		
Birch, Brown, Parulava	Finance	F	16										X	X	X					X		
Meunier	Blockchain Development	C	16		X		X	X		X	X		X	X		X	X	X				
Lewis	Blockchain Development	C	17		X	X						X		X								
Peck	Electrical Engineering	F	17	X			X			X			X			X	X			X	X	
Wüst-Gervais	Computer Science	F, C	17		X	X	X						X			X	X			X		
Mulligan	Finance	F	18		X		X	X	X	X	X			X	X	X	X		X	X		X

Figure 4.3: Trends in specific consideration points regarding blockchain applicability, highlighted columns indicate those of particular interest in military intelligence applications (based on the work of Ashley McAbee et al. [203]). Examples in the survey include Greenspan [204], Birch, Brown, Parulava [205], Meunier [206], Lewis [207], Peck [208], Wüst-Gervais [209] and Mulligan [210] (Based on the work of Perera et al. [200]).

keenly during periods of worst-case scenarios "when cyber, electromagnetic and physical attacks attempt to disrupt system operations when they will be needed most".

4.7 Computer security: Data integrity

Cyber defence is the closest low-cost, but the high-payoff application of blockchain technology. Blockchain technology is independent of secrets and trusts, unlike the systems that have been based on it. The blockchain preserves its authenticity in two ways. First, it ensures that digital events are widely distributed, forwarding them to other nodes in the network. Then, by using consensus, these events are stored in databases that can never be altered by an external party. In addition, blockchain enhances the perimetric security strategy of cyber defence, not by maintaining walls, but by constantly monitoring the walls and all the information inside. The increasing complexity of modern systems, including weapon systems, makes vulnerabilities more likely and less detectable.

A typical US warship like an Arleigh Burke class destroyer combines more than ninety missile launching cells with its radar systems, two independent Phalanx defence

systems and six torpedo launchers, not to mention a number of other weapon systems [211]. The challenge is to get all these combat systems to work together. The secret to the US Navy's success is systems integration, which is currently being accomplished by the Aegis Combat System. It is a centralised command and control system (CCS), making the right connections between sensors and weapons, like a boxer's brain connecting eyes and fists. But it's the centralisation that's the weak point - if the brain goes off, the whole system fails. That is why the possibility of using blockchain arises.

The Navy can structure its next-generation combat systems around decentralized decision nodes using a blockchain database architecture. This will speed up fire control, thereby (greatly) improving survivability. Artificial intelligence processors loaded into different weapon systems can coordinate their activities and check that they are working from the same data. In the 20th century, processing power was expensive, but data was cheap. So in 1969, it made sense to centralise on-board decision-making in a single Aegis brain. Today, processing power is cheap and data is more expensive. It is therefore likely that the Navy's twenty-first-century combat systems will use blockchain technology [212].

4.8 Supply chain management

Many industry organisations are working to use blockchain technologies in supply chain logistics and management. A growing concern is the supply chain management of defence systems which increasingly uses commercial off-the-shelf (COTS) components for embedded software systems. The problem is that these components may contain intentional vulnerabilities that an adversary can exploit at a time of his choosing. This threat was sensationalized by the recent Ghost Fleet incident, in which China disabled its entire fleet of F-35 aircraft with a deliberately embedded flaw in a commodity circuit card [213].

Blockchains offer a solution that traces the life of every circuit board, processor and software component from manufacturing to the user. A card design company can use blockchains to log the design iteration of each circuit. Manufacturers can log every model and the serial number of every card produced. Finally, distributors can report the sale of circuits to system integrators, who can log the distribution of circuits to a specific aircraft assembly, etc. In this context, blockchains create a permanent record of the transfer of assets between owners, thus creating a derivation. Many weapons systems are designed with a lifetime of 30 years or more. However, the computing technologies used in these systems are rarely produced for more than a decade. As a result, it becomes more difficult to replace obsolete components over time. Furthermore, in many countries, laws

prohibit the use by public authorities of components whose origin cannot be established. Disruption of ownership renders some parts unusable, even if they are functional and in high demand. Thus, resellers would also have an economic incentive to track their identified commercial off-the-shelf components in a block to preserve their provenance, which in turn increases their value.

In the Hungarian Defence Forces, decentralised technologies are not yet specifically addressed, but international research and development on the subject have already started. However, NATO C4ISR and the US Department of Defense (DoD) have already started their own blockchain programmes [214], and are developing a secure decentralised messaging application for the military called SBIR 2016.2.

4.9 Flexible communication

Bitcoin uses a peer-to-peer messaging model, which sends each message to every active node in the world within seconds. All nodes in the Bitcoin network contribute to this service, including smartphones. If a node's terrestrial, wireless or satellite Internet service is interrupted, a bitcoin message can be sent via alternative channels such as high-frequency radio, fax, or even barcoded and handwritten. Upon receipt, the service node checks the message and forwards it to each connected participant. Nodes can independently aggregate messages into new blocks [215]. Finally, the consensus mechanism ensures that invalid messages and blocks generated by dishonest actors are ignored. Together, these protocols ensure that authenticated message traffic can be reliably transmitted around the world, despite attacks on communication paths, individual nodes, or the blockchain itself.

4.10 Identification of vulnerabilities within military intelligence systems

Sam Mire, Market Research Analyst at Disruptor Daily, stated that there is a belief the US military's supply chains, cyber security and internal communications could benefit from implementing aspects of blockchain technology.

"With the world seemingly on edge and America's military manpower seemingly on the decline, exploring how blockchain can be used for defence purposes is a worthwhile pursuit" [216]. With the United States military's inability to create a "perfect" communication system thus far and ambitious communication programs like the Joint Tactical Radio System (JTRS) failing to live up to its potential in more ways than one [217], further exploration into blockchain technology's potential military applications include the improved visibility and traceability of expenditure, shipments, and contracts – through

blockchain's usage of transparent Distributed Ledger Technology, the military can eliminate fraud, waste, and reduction of losses. The United States Pentagon is worth an estimated $2.7 trillion dollars and failed its first official audit in 2018. Ernst & Young, among other private firms hired to perform the audit, could not complete the job due to the "DoD's financial records were riddled with so many bookkeeping deficiencies, irregularities, and errors that a reliable audit was simply impossible" [218]. Another application possibility of the technology is securing battlefield messaging – as late as 2017, United States Senator Ron Wyden expressed his concerns over the Defense Information Systems Agency's (DISA) lack of implementing encryption technology in daily communications, further noting that tech giants such as Google and Facebook employ standard STARTTLS encryption technology [219]. Using Bitcoin as an example of a peer-to-peer messaging model that delivers every message to every active node in the world in seconds, all nodes in the Bitcoin network contribute to this service, including smartphones. If a node's terrestrial, wireless, or satellite internet service is interrupted, a bitcoin message can be sent through alternative channels such as high-frequency radio, fax, or even barcode-based and manually. Upon receipt, the service node checks the message and forwards it to each associated participant. Nodes can independently aggregate messages into new blocks [220]. Finally, the consensus protocol ensures that invalid messages and blocks generated by rogue operators are ignored. Together, these protocols ensure that authenticated traffic can be reliably relayed anywhere in the world, even if communication paths, individual nodes, or the blockchain itself are attacked. Cyber superiority is not individually maintained by the nodes, but the network system can be kept controlled with current and expected data [221].

The technology can mean increased protection and preparedness against Cyber Warfare - the Defense Advanced Research Projects Agency (DARPA) is currently looking into blockchain's distributed consensus protocols to "evolve Cybersecurity for an Agile and Resilient Defense Posture" [222]. United States President Donald Trump signed a bill in December 2017 that includes a mandate for a blockchain-based cyber security assessment "of efforts by foreign powers, extremist organizations, and criminal networks to utilize such technologies;. . . [and] an assessment of the use or planned use of such technologies by the Federal Government and critical infrastructure networks" [223].

It is also expected, that DLTs will result in Improvements to the military's manufacturing processes – the Naval Additive Manufacturing department was a perfect use case for blockchain technology to illustrate its "ability to secure and securely share data throughout the manufacturing process (from design, prototyping, testing, production, and ultimately disposal)" [224]. Each key phase in Additive Manufacturing revolves around the use of data or a "Digital Thread: a single, seamless strand of data that stretches from the initial design concept to the finished part, constituting the information that enables

the design, modelling, production, use, and monitoring of an individual manufactured part" [225].

Blockchain technology's ability to highlight and detect hacking and network penetration attempts has led to an international arms race between China, the United States, and Russia all vying to solve vulnerabilities within supply chains and data integrity. Protecting and promoting data integrity of military supply chains can be therefore another application of DLTs. Ex-DISA Director Army Lt. Gen. Alan R. Lynn stated, "A few years ago, getting a 1-gigabyte or 2-gigabyte attack at the internet access point was a big deal. Now, we get 600-gig attacks on the internet access points and unique, different ways of attacking that we had not thought of before. There's now, we would call it the 'terabyte of death' – there is a terabyte of death that is looming outside the door" [226] Many weapon systems are designed with a life span of 2-3 decades or even more. However, the computing technologies used by these systems have a short shelf life, rarely lasting more than a decade. As a result, replacing obsolete parts becomes more difficult over time. Furthermore, in several countries, it is prohibited by law to use a component whose origin cannot be ascertained. Loss of ownership renders parts unusable, even if they are functional and in high demand. This would give the resellers an economic incentive to track their identified off-the-shelf commercial components in a block to retain their origin, which in turn adds value.

Decentralized technologies are not dealt with separately in the Hungarian Defense Forces, but international research and development are already underway. NATO's C4ISR and the US Department of Defense (DARPA - DoD) have already launched their own blockchain programs, developing a secure, decentralized messaging application for the military under the name SBIR 2016.2 [215].

All armed forces thrive to protect the weapon systems. The US Navy's' Aegis Weapon System (AWS) is a "centralized automated, command-and-control (C2) and weapons control system" that is vulnerable to cyber hacks and other threats. The challenge in controlling such a powerful weapon arises when you consider exactly what the system is meant to handle simultaneously: A typical destroyer like the "Arleigh Burke Class includes an upgraded SPY-1 multi-function, phased-array radar, Mk 41 Vertical Launching System, an advanced anti-submarine warfare system, advanced anti-air warfare missiles, and Tomahawk land-attack cruise missiles" [212]. Much like the British proved superior weapons integration outguns greater firepower in Jutland, 1916 during World War 1, blockchain technology can seamlessly integrate and operate multiple weapons' systems through its decentralized and distributed architecture [213] DARPA awarded Galois and Guardtime Federal a $1.8 million contract in 2016 to "verify the correctness of Guardtime Federal's Keyless Signature Infrastructure (KSI), a formal verification tools and all blockchain-based integrity monitoring systems" [227]

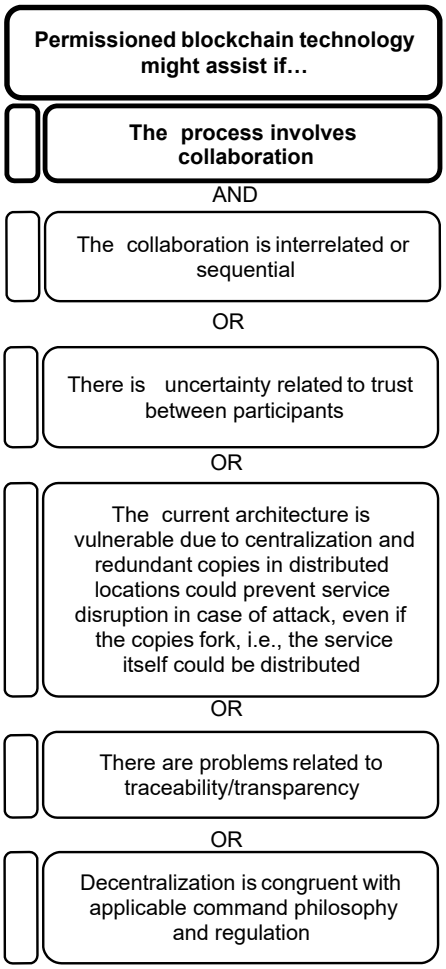

Figure 4.4: Critical factors in determining when blockchain technology might apply to military intelligence processes (based on the work of Ashley McAbee et al. [203]).

According to McAbee, Tummala and McEachen [203], figure 4.4 can act like a checklist if the military is interested in adopting blockchain. The authors propose the following: "If a system meets the first, mandatory tenet identified in bold and at least one of the others, it may be a reasonable candidate for a permissioned blockchain technology model". The authors also state that such a model will evolve during the study.

Kovács brings to light several challenges in his paper, National Cyber Security as the Cornerstone of National Security, including that of "the rapid modernization of infrastructure" and whether this exacerbates the "vulnerability of critical infrastructures" as well as the joint roles played by both private and public sectors (Kovács 2018). To solve these challenges, Kovács quotes from the national cyber security strategy of the United Kingdom, "Government has a clear leadership role, but we will also foster a wider commercial ecosystem, recognizing where the industry can innovate faster than us. This includes a drive to get the best young minds into cyber security".

4.11 Machine learning and Artificial Intelligence (AI)

Project Maven, also known as the Algorithmic Warfare Cross-Function Team, was launched in April 2017 and was overseen by Air Force Lt. Gen. Jack Shanahan. The main goal of Project Maven is to "integrate artificial intelligence and machine learning" with DoD operations and, in particular, to "turn the enormous volume of data available to the DoD into actionable intelligence and insights at speed" [228], "Project Maven and Some Reasons to Think About Where We Get Our Funding.") Maven was designed to "interpret video imagery, which could, in turn, be used to improve the targeting capability of drone strikes" and employs deep learning, and neural networks to constantly improve. Technologies developed through Maven have already been successfully deployed. Though the Project has received critical acclaim, "enormous organizational, ethical, and strategic challenges" remain where Maven soon became shrouded in controversy when over 3,000 Google employees signed a petition in protest against the company's involvement with a U.S. Department of Defense Artificial Intelligence study [229]. The project has since been taken over by Palantir.

One of the main challenges faced by governments across the world amounts to traditional legacy management and operational frameworks integrating with newer technologies. An example of this is military tanks developed in the first quarter of the 20th century, ultimately leading to the development of "stealth and precision-guided weapons technology in the 1970s. Such technologies created the "foundation for a monopoly, nearly four decades long, on technologies that essentially guaranteed victory in any non-nuclear war [230]. The amount of footage drones carry is so vast that human analysts can no longer cope with the sheer volume. Therefore, artificial intelligence is being harnessed, and thanks to machine learning, AI will constantly improve at recognizing and classifying objects.

Today, at least 90 countries have drones, including many non-state groups. While most of them will not be classified as sophisticated within the field of robotics, many are remotely controlled with operators hundreds, if not thousands of miles away. Autonomy is becoming increasingly apparent in the management of different vehicles, especially those used by the military. For example, the Guardium, developed by G-NIUS, is an Israeli unmanned ground vehicle (UGV) that "carries more than 660-pounds of cameras, electronic sensors, and weapons" and is used for combat and defence along the Gaza border. Though the vehicle is self-propelled, soldiers continue to be responsible for the weapons on board [231].

US Security Expert, Paul Sharre believes that Artificial Intelligence applications do not require major modifications to military tasks and can be integrated into weapon systems just as easily as civilian solutions [232]. According to General Shanahan, the United

States intends to stand apart from Russia and China. Both countries are currently testing their uses of Artificial Intelligence technology for military purposes, but raise ”serious concerns about human rights, ethics, and international norms” [233].

4.12 Artificial intelligence and law enforcement

In July 2018, Daniel Faggella, Head of Research and CEO of Emerj AI Research Institute, spoke at the Interpol/UNICRI Global Meeting on the potential and risks of AI and robotics for law enforcement. This marked the start of a dialogue on the use of AI in law enforcement, security and policing. The following Figure 4.5 is taken from a UNICRI report on the integration of AI into the law enforcement sector.

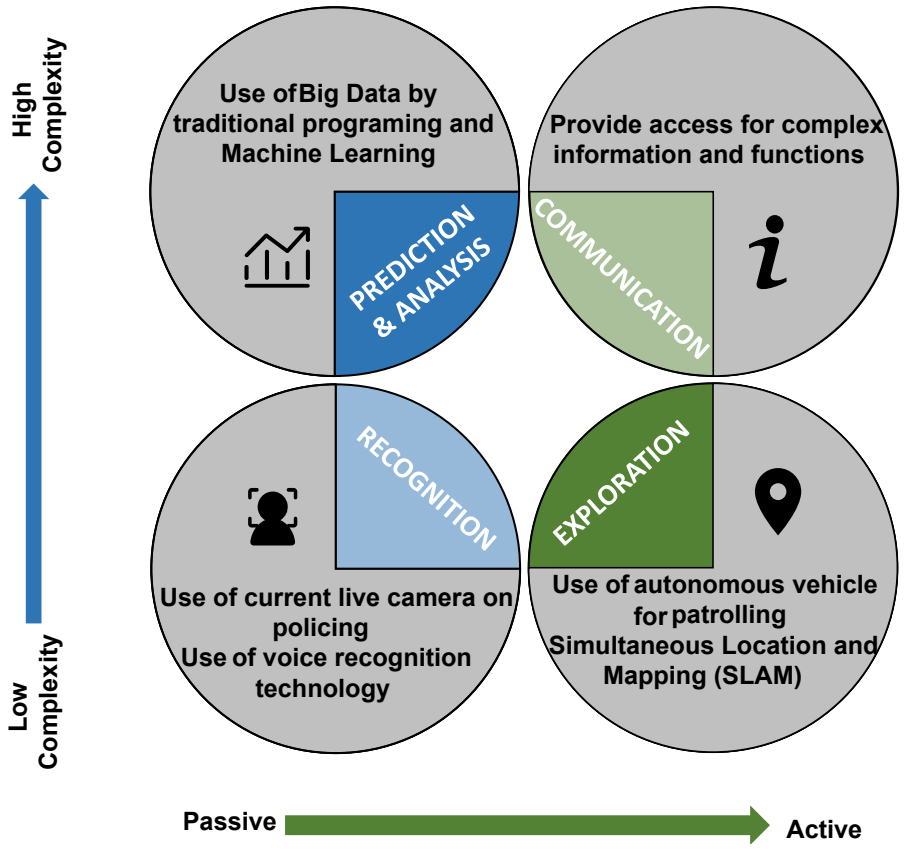

Figure 4.5: Artificial Intelligence and Robotics for Law Enforcement - UNICRI Global Meeting on the Potential and Risks of Artificial Intelligence and Robotics for Law Enforcement (Based on the work of [234]).

Kevin McCaney writes in Law Enforcement Using Analytical Tools to Predict Crime that law enforcement agencies are increasingly relying on predictive analytical software to prevent crime. Until recently, these technologies have typically been used by large companies in the competitive sector. An example is IBM’s Blue Crush (Criminal Reduction Utilizing Statistical History) software, which is used by the Memphis, Tennessee

police department to ”analyze crime and arrest data, compare it to weather reports, economic factors and information related to events such as paydays or concerts to create a predictive model” [16]. Police, courts and criminal justice institutions are working together to shape the criminal justice system. To achieve optimal levels of performance, these organisations need to have experts who can analyse crime data and simulate scenarios that can increase the accuracy of programmes using artificial intelligence. An example of this is the National Intelligence Model (NIM) in the UK, which is designed to improve and support police forces that rely on intelligence.

The UNICRI report concludes that AI and robotics can be used for weaponisation just as they can be used for the public good in policing and crime prevention, stating that ”a recent report by 26 authors from 14 institutions (academia, NGOs, industry) has looked at the issue in depth and concluded that many of the features that would make AI and robotics attractive to law enforcement (such as scale, speed, power and distance) make AI and robotics just as attractive to criminals and terrorist groups [18].

The report meticulously analyzed the various domains of attack and culminated in the identification of three crucial attack vectors that are of paramount importance. The first vector of attack is what we refer to as digital attacks. This category encompasses a wide array of malicious activities that leverages technological means to perpetrate cybercrime. This can be in the form of automated spear phishing techniques aimed at extracting sensitive information or exploiting vulnerabilities present in cyber systems.

The second vector of attack is political attacks. These attacks often have a more strategic objective in mind and can be used to sow the seeds of discord and sow chaos. The dissemination of false news and misleading media material, and the manipulation of videos through deep fakes and other fake means are examples of political attacks. Such tactics can be used to undermine the trust of political actors and even call into question the credibility of evidence presented in court.

The final vector of attack is physical attacks. This type of attack exploits the vulnerabilities in the physical domain. For instance, the usage of facial recognition capabilities by armed drones or smuggling drones is an example of physical attacks. The report also mentions that Artificial Intelligence can be used as a weapon in itself, as it can carry out harmful actions directly or can be used to corrupt other AI-based systems with malicious data sets.

It is imperative to take a multi-disciplinary approach to combat the threats posed by these three vectors of attack and protect our digital, political, and physical domains from malicious actors.

4.13 Concluding remarks

Distributed ledger technology (DLT) and blockchain technology will be priority areas for force development. International examples show that decentralised military solutions are responsible for the cybersecurity paradigm shift. New cyber operational dimensions and military-technical concepts are emerging as a consequence of decentralised digital interactions. The distributed ledger is trusted, transparent, encrypted and, depending on its type, the consensus is reached between users on the network. DLT is transparent and therefore secure, cryptographic solutions ensure that no network manipulation is possible. The best-known type of DLT is the blockchain. Blockchain networks are fault-tolerant, with trusted nodes being aligned and untrusted nodes being discarded. The Zrínyi 2026 force development research and development programmes should focus on developing special capabilities because Hungary can be internationally competitive in this area. R&D support for new types of cyber defence challenges can be an essential and one of the best investments in force development. The institutional framework described in this thesis can provide the infrastructure and tools to enable the development, maintenance and training of cyber operational intellectual capability at a high level. On the other hand, there is little chance that Hungary can become and remain a world leader in weapons or military vehicle production: this would require a lot of money and expertise, as well as access to closed data. It would place an unfair burden on force development reform and would not result in a particular specialisation opportunity globally. The military development potential of decentralisation and blockchain requires further serious and deeper research. The full range of applications of the technology is constantly changing, and it is therefore recommended to develop organic expertise, and monitor and process international results. Artificial intelligence and machine vision can link civil-military partnerships, resulting in professional collaborations for the development of DLT and blockchain-based technologies. Hungary is already one of the leading countries in the application of machine learning, machine vision and artificial intelligence. A success story for Zrínyi 2026 could be the specialisation of military capability in cybersecurity.

The effective military application of blockchain will be the subject of further research. The current utilisation and connectivity of state IT apparatus to a distributed ledger infrastructure raises technical and cyber security issues. International practical examples show that for data analysis tasks that challenge human capacity, machine learning and the network-based application of artificial intelligence are the designated research directions. In hybrid warfare, Hungary can gain a regional competitive advantage, develop cyber defence and offensive capabilities, and create internationally significant knowledge assets and scientific, institutionalised IT research workshops. Government-free IT computing capabilities can be pooled into decentralised systems in a resource-efficient manner. Such

a decentralized system, defined in its objectives, can do in a secure environment what humans cannot: make decisions accurately, quickly, without error, and produce informed, objective military leadership decision points.

Chapter 5

Summary of new scientific results

The results of this dissertation are organized into the following three main categories:

- Live spoof detection for Human Activity Recognition (HAR) applications.
- Automated fast violence detection for surveillance applications.
- The applicability of Distributed Ledger Technologies (DLT) and blockchain technologies in the military.

The respective contributions are outlined in the following sections of the thesis.

5.1 Thesis Group I - Live spoof detection for Human Activity Recognition (HAR) applications.

- **I created a deep learning architecture for detecting spoofing attacks in videos using replay techniques.**
- **I discovered the effectiveness and generalizability of the proposed architecture through comprehensive evaluation and testing.**

Relevant publications: [1] [2]

5.1.1 Novel deep learning architecture for spoofing detection

I created an ensemble multi-stream model that detects spoofing cases arising from video replay attacks. The model inspects distinct regions of human faces and combines these observations to provide robust classification between spoof and genuine cases.

- Frames are extracted and fed to the pre-trained YOLO model to extract windows corresponding to the human head region on the image. Each detected head bounding box on a video frame is resized to 64 × 64 × 3 pixels and processed in three different streams that use CNN architecture inspired by VGG16 net as a baseline model. The output of the three streams is combined using the obtained classification probabilities following majority voting to classify spoof and genuine cases. The three streams process:
  - Predicted and resized head bounding box of size 64×64 pixels.
  - Cropped lower 64×32 pixels from the resized image.
  - Cropped central 32×32 pixels from the resized image.
- The effectiveness of the proposed ensemble method is assessed using state-of-the-art methods in the literature for spoofing detection applications using facial image data: Local Binary Pattern (LBP) histograms, Local Binary Pattern histograms from Three Orthogonal Planes (LBP-TOP), Statistical Binary Pattern histograms (SBP) and Statistical Binary Pattern histograms from Three Orthogonal Planes (SBP-TOP). Experimental results show that the proposed EM method outperformed other methods considerably.

5.1.2 Temporal analysis for real-time spoofing detection

I formulated a strategy to combine multiple detections of the proposed deep learning network temporally on a captured video or on a live video stream to systematically detect video replay spoof attacks while maintaining real-time performance.

- Chunks of overlapping video clips, each containing several frames (depending on the model under testing) with an overlap of 15 frames are defined.
- Within a considered clip, the proposed deep learning models are run on three frames that are 15 frames apart and the obtained predictions are combined to derive a single prediction for this clip. Following this method, the first prediction is made after 31 frames and thereafter, predictions are made every 15 frames.

5.1.3 New database for spoofing detection

I created a new database comprising real and spoof videos captured from 38 different users in different locations and under different lighting conditions. The database is diverse and precisely captures the required features for training the proposed deep learning network.

- A diverse database consisting of almost 50,500 full HD images of 38 users (both male and female) between the ages of 8 and 40 juggling a football in different backgrounds and lighting conditions are collected. These are extracted from several videos shot using various iPhones — iPhone 6, 6S, SE, 7, 8, X, and XS.

- These images are manually labelled with bounding boxes that encapsulate the following data: human body parts — head, left shoulder, right shoulder, left elbow, right elbow, left hand, right hand, left hip, right hip, left knee, right knee, left foot, right foot and also the bounding box of the football.

- YOLO deep learning CNN architecture is trained using the captured and labelled database and used for detecting the required body parts together with the ball from the video frames in a single shot.

- This database is also used to learn the genuine and spoof cases. To this end, an additional 50,500 Full HD spoofed images were generated using the original images by capturing the same videos on several monitors: a 27-inch Dell 4K monitor, a 15-inch Full HD Lenovo laptop monitor, and a 13- inch MacBook Pro monitor with a resolution of 2560 × 1600.

- Before training the networks, all video frames are pre-processed to extract the head region using the previously trained YOLO model and resize the head image to 64 × 64 pixels.

- Data augmentation techniques are also used to reduce problems from overfitting including increasing and decreasing the pixel brightness by a value of 50, doubling and halving the image contrast, and adding Gaussian noise with variances of 50 and 100.

- The pre-processing steps were consistently added to all the video frames. Finally, before training, data was also shuffled. Further, all the pixel intensities are scaled to the range [0, 1] for training.

5.1.4 Analysis about model generalizability

I created empirical evidence through additional evaluation of the performance of the proposed approach in the context of biometric recognition applications, demonstrating its applicability in other domains.

- To evaluate the ability of the proposed spoofing detection model to generalize to other domains such as face recognition systems, I have experimented with two

widely known datasets in this context: Idiap REPLAY-MOBILE and CASIA Face AntiSpoofing.

  - The REPLAY-MOBILE dataset consists of 1190 video clips of photo and video presentation attacks (spoofing attacks) to 40 clients, under different lighting conditions. These videos were recorded with an iPad Mini2 (running iOS) and an LG-G4 smartphone (running Android) in full HD resolution.
  - The CASIA Face AntiSpoofing Database consists of 600 video clips of 50 subjects. Out of the 600 video clips, 150 clips represent video replay attacks. Compared to the Idiap database, the CASIA DB provides images captured using a variety of cameras (Sony NEX-5-HD, two low-quality USBs) to capture replay attacks displayed on an iPad. However, a significant deficiency of this database is that the video replay attacks are captured in very low resolution (640×480).

- Results show that the proposed ensembled approach performs very well on the REPLAY-MOBILE database irrespective of the fact that the users are located very close to the camera (higher IPD values than that of my database).

- Even though, the CASIA database consists of low-resolution images, the proposed method performed reasonably well also on this database.

5.1.5 Prototyping and stand-alone implementation

I created a solution by designing an IOS mobile application that implements the proposed approach in real-time, bringing the technology to the convenience of mobile devices.

- I have implemented the proposed Ensemble multi-stream model in swift for IOS. The application is standalone and tested on iPhone 8 released in 2017 and has a 2.39 GHz hexacore 64-bit.

- The developed models as well as the YOLO model for head bounding box detection are converted to CoreML API for running on the IOS device. The application takes as input incoming frames from the on-device-camera and runs my trained YOLO model to detect head bounding box information.

- The algorithm works in real-time. The proposed method can run on an iPhone 8 and processes a single frame on an average of 4 ms. This translates to a frame rate of about 250 frames per second.

- Following the proposed testing scheme, running the proposed model every 15 frames further ensures that there are enough resources left on the device to run the activity recognition applications.

5.1.6 Experiments with video compression

I created empirical evidence by simulating and evaluating the performance of the proposed ensemble model under extreme video compression (300 kbps) and discovered its robustness.

- Video compression techniques are applied to reduce the video bitrate to facilitate efficient streaming which introduces video artefacts. To evaluate the ability of the proposed ensemble model to correctly classify the spoof cases, I have generated compressed video streams with varying bitrates from 300 kbps to 1500 kbps. Multiple videos were generated with different bitrates using FFmpeg to experiment with video compression.

5.2 Thesis Group II - Violence Detection for automated video surveillance applications

- **I created a solution for automatic violence detection in video surveillance by exploring smart networks.**
- **I discovered a novel deep learning-based approach for violence detection that outperforms existing methods with fewer model parameters and demonstrated robust performance under compression artefacts commonly encountered in remote server processing.**

Relevant publications:

5.2.1 Efficient deep learning architecture for violence detection

I created a deep learning-based method that can be used to filter violent and normal patterns videos. Considering the popular video classification metrics for evaluation, the method outperforms several state-of-the-art methods for violence detection and is also able to cope with video compression artefacts, while remaining computationally lightweight.

- I have explored X3D-M deep learning architecture that is computationally lightweight to learn and detect violence patterns from videos. I proposed two architectures - fine-tuned and transfer-learned models for classifying video clips containing violence, which leverage action recognition features learned from the Kinetics-400 dataset. For both models, I have modified the architecture into a regression model to generate a violence coefficient that indicates the probability of the existence of violence in a given video clip.
  - The fine-tuned model trains all the parameters of the adapted X3D-M model on the datasets for violence detection. The second fully connected layer of the X3D-M model is replaced to output a floating point variable which is converted into range $[0, 1]$ using a sigmoid function to derive the violence coefficient.
  - The transfer learned model uses X3D-M for inferring video features and does not retrain the parameters of the original X3D-M model. Pre-processed videos containing both violence and no violence are inputted into a pre-trained X3D-M model for feature extraction. Three additional fully connected layers are trained using the extracted features to obtain the violence coefficient.
- The experimental results on individual datasets show that the fine-tuned model performed better than the state-of-the-art methods on most datasets with relatively fewer model parameters.
- Transfer learned model also achieved decent performance on all the datasets given that, it has less trainable parameters than fine-tuned model and thus relatively less adaptable to specific scenarios.
- When testing on all combined datasets, the fine-tuned model achieved better performance and the transfer learned model produced more combined false positives and false negatives.
- Further tests on individual datasets show that models trained on all combined datasets did not perform well in several cases when compared to the performance of models trained on individual datasets. This shows that there are multiple inconsistencies in the publicly available datasets for violence detection.

5.2.2 Comprehensive database for violence detection

I created a comprehensive database of violent and normal videos by combining and extending seven existing video databases, providing a robust resource for violence detection

research across various contexts.

- For experimenting with violence detection and to facilitate comparing the results with other methods, I have considered seven different datasets that are commonly used in literature. I have also extended some of the datasets with annotations to assist in-depth cross-validation experiments. These are described in the following:

  - **Crowd Violence** (CV) dataset contains videos involving violence in crowds, collected from YouTube.
  - **Hockey Fights** (HF) dataset is a collection of fights between players in hockey games from the USA's National Hockey League (NHL).
  - **Movie Fights** (MF) dataset collects several scenes from action movies.
  - **Real Life Violence Situations** (RLVS) dataset gathered fighting videos from YouTube and also from real street cameras that contain many real street fights.
  - **Real-World Fight-2000** (RWF-2K) dataset is a collection of large-scale fighting videos from YouTube. The dataset contains trimmed video clips captured by surveillance cameras from real-world scenes.
  - **UCF-Crime Selected** (UCFS) dataset is a subset of the UCF-Crime dataset. The UCF-Crime dataset contains long untrimmed surveillance videos that cover 13 real-world anomalies including Abuse, Arrest, Arson, Assault, Burglary, Explosion, Fighting, Road Accident, Robbery, Shooting, Stealing, Shoplifting, and Vandalism without annotations. Although this is a large-scale dataset, all videos in the violence class contain a mix of violent and normal actions which is undesirable. Among the anomalies, I selected the classes - Abuse, Explosion, Fighting, Road Accident, and Shooting and manually trimmed these videos to only contain violent parts for training and testing.
  - **XD-Violence Selected** (XD-V) dataset contains a subset of videos from the XD-Violence dataset. XD-Violence dataset contains several untrimmed videos covering 6 anomalies including Abuse, Car Accidents, Explosions, Fighting, Riots, and Shooting gathered from action movies and YouTube. Similar to the UCF-Crime dataset, I selected a set of videos belonging to the classes - Abuse, Explosion, Fighting, Road Accident, and Shooting and manually trimmed these videos to only contain violent parts for training and testing.

- All datasets also contain normal videos for training and testing that have no violence involved. In the case of UCFS and XD-V datasets, normal videos are trimmed to five-second video clips to match the average duration of normal clips

of other datasets. Also in the case of UCFS and XD-V, the maximum duration of a video clip containing violence is limited to approximately 5 seconds.

5.2.3 Analysis on the generalizability of proposed models

I discovered high-performance results through a thorough evaluation of the proposed approach on a comprehensive database of violent and normal videos, including cross-database validation to assess the generalizability of the methods.

- According to the metric scores trained fine-tuned and transfer-learned models on the CV dataset did not generalize well to other datasets. This is anticipated since CV contains only examples of mass violence and the other datasets do not contain plenty of such examples. Also, the trained FT model on the HF dataset has poorly generalized to other datasets indicating that the HF dataset does not contain diverse examples of violence and contains monotonous fighting videos between hockey players. However, the TL model trained on this dataset showed generalized better than the FT model as indicated by the metric scores.

- Both models trained individually on datasets - MF, RLVS, RWF-2K, UCFS & XD-V performed satisfactorily in the cross-validation tests and generalized decently to other datasets with average accuracy scores close to or above 80% and average AUC scores close to or above 0.8. Considering both metrics, models trained on UCFS and XD-V datasets exhibited the best generalization ability in the cross-validation studies. This indicates that the datasets, that are gathered in this study, have the most representative and heterogeneous samples for actions involving violence and non-violence.

- Overall, the transfer learned model showed a better capability to generalize and has less standard deviation within testing accuracy scores for individual datasets when compared to the fine-tuned model.

5.2.4 System implementation

I discovered the high performance of the developed approach through extensive evaluation on collected databases, including cross-database validation to assess the generalizability of the methods. I created a standalone functional system for automated violence detection, implementing the proposed methods using the PyTorch deep learning library. The resulting application can easily be adapted for use in surveillance applications.

- From the incoming video stream, non-overlapping video segments having a duration of four seconds are extracted. From each segment 16 video frames are extracted into a block following uniform temporal sampling. These blocks are pre-processed and then used as input to the trained models to get a violence coefficient for the current segment.

- This application is implemented on the Ubuntu Linux operating system using AMD Ryzen Threadripper 1950X 16-core processor. I have used Nvidia GeForce GTX 1080 Ti GPU using CUDA toolbox for running my trained PyTorch models.

- Implementation results show that together with block extraction and pre-processing, for each four-second video segment, both models require 0.06 seconds on average to infer a violence coefficient. The pre-processing is implemented on the CPU and this consumes 0.04 seconds on average. Therefore, the average time of running the proposed model is 0.02 seconds.

5.2.5 Video compression experiments

I uncovered the impact of video compression artefacts commonly encountered in video streaming fields on the performance of proposed models through comprehensive experiments and presented the results.

- For experiments, I have generated compressed video streams with varying bit-rates - 300, 500, 1000 & 1500 Kbps. Two datasets - RWF-2K and CV are used for the experiment and corresponding testing videos from these two datasets are compressed. Multiple videos are generated with the considered bit rates using FFmpeg.

- Results from the study pointed out that the proposed models did not show greater fluctuations in the performance and performed decently even under extreme compression (300 Kbps). This shows that the proposed models did not model the noise in the training videos and were precisely directed to learn the concept of violence.

5.3 Thesis Group III - Applicability of DLT and blockchain technologies in military

I created a system that leverages DLT technologies with a focus on blockchain as a computational data resource for training machine learning and deep learning algorithms while prioritizing data privacy through the use of specialized neural networks. Relevant publications: [3] [4] [5] [6] [7]

5.3.1 Data privacy, sharing, and tracking

I designed a concept where blockchain can be used to secure and encrypt data used for deep learning, making it more difficult for unauthorized parties to access or tamper with the data.

- Data privacy is ensured by using blockchain to encrypt and secure sensitive data, such as intelligence or surveillance data, that is collected by military assets. The data can be stored on the blockchain and is only accessible to authorized parties with the appropriate encryption keys. This can help to prevent unauthorized access or breaches of sensitive data.
- Secured data sharing in real-time between multiple parties, such as different military units or coalition partners, can also be facilitated by the use of blockchain to improve situational awareness and decision-making capabilities.
- The blockchain can also be used to track the origin and handling of data, providing transparency and accountability. For example, the blockchain can be used to record the provenance of intelligence data, such as the sources, handling, and analysis. This allows for the verification of data authenticity and integrity, and can also aid in the investigation of data breaches.

5.3.2 Model training and decentralized decision making

I formulated a concept where blockchain can be used to train machine learning models in a decentralized way, using distributed computing power and data, providing a more secure and transparent way of training large and complex models. While running the trained deep learning models, blockchain-based smart contracts can enable decentralized decision-making based on consensus among multiple parties.

- Blockchain can be used to secure and track the data used to train machine learning models for military applications. The data collected by military sensors can be encrypted and stored on the blockchain, ensuring that only authorized parties have access to it. Additionally, distributed computing power and data can be used to train models in a decentralized way, providing a more secure and private way of training models.
- Decentralized decision-making using smart contracts can be used to ensure that actions taken by military assets, such as unmanned aerial vehicles (UAVs) or autonomous weapons systems, are based on a consensus among multiple parties. For example, a smart contract could be used to ensure that a UAV only fires a weapon

if a consensus is reached among multiple operators, or if certain predetermined conditions are met. This can help to prevent unauthorized or accidental use of weapons, and can also improve the overall situational awareness and decision-making capabilities of military assets.

Chapter 6

Application of the work

6.1 Practical applications of spoof detection [scientific result 5.1.1]

Video replay spoof attack detection in biometrics recognition and in gaming is a crucial aspect of ensuring the security and integrity of systems. Deep learning algorithms have been used to develop video replay spoof attack detection systems that are able to accurately identify and prevent fake video replays, prevent fraud or cheating, and protect personal information.

- Biometrics Recognition: In biometrics recognition systems, such as facial recognition, video replay spoof attacks can be used to gain unauthorized access or to steal personal information. Video replay spoof attack detection can prevent such unauthorized access by detecting fake video replays and denying access to the system.

- Gaming: In gaming, video replay spoof attacks can be used to cheat or manipulate the outcome of games. Video replay spoof attack detection can prevent cheating by detecting fake video replays and taking appropriate action, such as banning the player or resetting the game.

- Real-Time Monitoring: Video replay spoof attack detection can be used to monitor video replays in real-time, providing instant alerts if a fake video replay is detected. This can help to prevent fraud or cheating in real-time.

- Improved Accuracy: Traditional methods of video replay spoof attack detection, such as manual review, can be time-consuming and prone to human error. Deep learning algorithms, on the other hand, can analyze footage more accurately and quickly, reducing the likelihood of missed attacks.

6.2 Divergent applicability of violence detection [scientific result 5.2.1]

The primary objective of HAR is the analysis of activity patterns of humans from digital videos. At border crossings, in prisons, at military camps, there is no current system that can automatically identify the variables that determine the activity of a person and determine whether the given human activity is otherwise correct/incorrect, normal or abnormal, or how much it deviates from a given pattern and, based on the classification of the difference, how much the volatility of the activity is considered illegal or dangerous, risky.

The essence of the proposed deep learning application for violence detection is that it can inform the designated decision makers (military staff, guard/reception service, security service, law enforcement authority, armed forces) in real-time about possible risk events that can be classified on the basis of a risk scale.

- Law Enforcement: The use of deep learning algorithms can assist law enforcement in identifying and preventing acts of violence before they occur. This can be especially beneficial in high-crime areas where the volume of footage can be overwhelming for human review.

- Security: In public spaces such as schools, malls, and airports, automatic violence detection can help to quickly identify and respond to potential threats, allowing security personnel to take appropriate action and potentially prevent a dangerous situation from escalating.

- Public Safety: In crowded public spaces, automatic violence detection can also play a critical role in ensuring public safety by detecting potential dangers and alerting authorities in real-time. This can help to reduce the risk of harm to innocent bystanders.

- Real-Time Monitoring: By using deep learning algorithms to automatically analyze large amounts of surveillance footage in real-time, automatic violence detection systems can identify and alert authorities of potential threats before they escalate.

- Improved Accuracy: Traditional methods of violence detection from surveillance footage often involve the manual review, which can be time-consuming and prone to human error. Deep learning algorithms, on the other hand, can analyze footage more accurately and quickly, reducing the likelihood of missed threats.

Based on sensor fusion theory, events can be classified based on multiple inputs. Demonstrated through examples, in a military institution, at a border, aggressive patterns of behaviour, fights, other than average movements, misbehaviour, dressing, and inappropriate objects (e.g., knife, or even a weapon) are all considered risk events that security or military personnel may not / cannot physically realize. Computer image processing systems can see what humans may not see, and with algorithms running on camera systems, the analysis can have a frame 'depth' precision of detection.

Certainly, a trigger event can be either a fire or an explosion, or smoke as well. Data from digital smoke detectors can be processed at the same time as camera image analysis, allowing complex information to be communicated to the authority based on the most efficient result. This increased accuracy can determine the risk level of the event as well.

AI health monitoring of military crews can save lives. It is possible to analyze the activity of a person with machine vision in real-time based on the information collected by different devices. A school porter, or even a stationed soldier, a police officer, or even the university security service, is usually at a workstation monitored by a camera, so from a physiological and safety point of view, every second can count in the event of a malaise. Malaise, such as a heart attack, is unfortunately preceded in many cases by a pattern of movement, but their course is not necessarily the same. In addition, if a person is alone, there is an additional risk that the 'trigger' event, the form of movement to be assessed, is not treated in real-time. This can cause tragic events, even death. An attempt to extract and analyze the spatiotemporal features from videos capturing such events could be a route of future research.

Based on the results of international public research, examining the DARPA / Air Force Research Laboratory's Mobile and Stationary Target Acquisition and Detection (MSTAR) program, Automatic Target Recognition (ATR) research, experimental results based on the MSTAR data set showed that a proposed research method is expected to be more efficient and accurate than existing manual systems.

6.3 Definition of development and the areas of use [scientific results 5.1.1 & 5.2.1]

This research also demonstrates the development of machine vision and competence supported by artificial intelligence. In the military, such a development is based on machine vision and artificial intelligence, which I achieve by developing neural networks and a teaching-learning methodology that enables us to achieve a more serious event marking of the critical information infrastructure.

Consequently, areas of application can be any camera-equipped unit, such as a uni-

versity campus, or school area, or portable devices, such as smart glasses or robotic law enforcement technologies, as well as UAV/drone competence, as well as facility security, personal security, and personnel surveillance. The application may lead to further syntheses that may result in a public interest digital data reporting application, thereby multiplying knowledge transfer due to development.

The European Union and the EU Strategy for the Security Union state that the creation of security is a common European goal, and that artificial intelligence will value new opportunities and innovations that can address these Security Union issues more effectively. Machine vision is a modern technology that can also develop existing competencies in quality control.

In the EU, research and development and innovation have played a key role in guaranteeing security, which can be brought into line with the EU Artificial Intelligence Strategy.

Many member countries established National Innovation Associations and the Artificial Intelligence Coalition, so government and commercial and non-profit industry players can actively participate in the continuous monitoring and regulation of Artificial Intelligence. Software development also provides an opportunity to help develop tool and method-specific requirements for security management hardware developments. With further development, a nation may be able to achieve a breakthrough and pioneering results in automating discovery, object detection, and object tracking. It can provide automatic transmission of image information in addition to the future 5G technology.

A very relevant area of use of the new technology could be military guarding (border control, institutional security, etc). The ‘digital guard’ project is a suggestion for further in-depth research. There is still a shortage and many vacancies in the traditional military/security guard positions. The military guard program builds on a living force of people who have a government, military, or a public task with special rights with or without the right to use weapons, but with coercive means. A military/security guard ensures that the members of the designated institution, the border of a country, and the institutional military staff / civilian employees who directly assist the work, can perform their duties without interruption. A ‘digital guard’ can be a real-time holistic supervisor: it prevents and interrupts acts that violate or endanger the safety of soldiers stationed at the military site, alerts decision-makers about endangered employees of the institution, and prevents illegal acts committed to their detriment. There are thousands of military sites, and guarded institutions in the EU where military guards could have been placed, but there is a limited number of guards, as many are not filled in roles, even though in most EU member states, this position is constantly being searched for, and according to the Crime Department of the member states Police Criminal Directorate, there are always vacant police, military, law enforcement guard positions open.

However, there are plenty of existing conflicts where the guardianship office questions the job competence of the School Guard, most of the time due to under-performance or just over-competence. The Digital School Guard provides an objective decision-making opportunity, so it is not a development aimed at replacing the School Guard, but a development aimed at filtering out violations, guardianship, and supporting law-abiding School Guards, which can significantly reduce the number of objectionable measures.

These statistics support the need for a "digital school guard" program, which will of course be able to be extended to other areas later than the "digital guard" program (universities, institutions, municipalities, etc.). The technology would not replace the existing stock but would facilitate efficient, real-time decision-making, and in places where there are no human resources, a 'digital' alert would be given to a central decision-making unit to signal a possible illegal activity.

During the development, a tool- and methodology-specific system of requirements will be developed with the involvement of person and object-tracking NKE and designated departments, from which I can determine more precisely on the basis of the feasibility study which area can most effectively achieve development results during the development cycle. Thanks to visual analysis, it may become possible to establish an intelligent university institutional system in Hungary that can alert competent managers responsible for university security in advance and in a preventive manner with motion and object detection and real-time background analysis supported by artificial intelligence. This increases the institutional security of the university, and in the event of an incident, the reaction time can be significantly reduced. In addition to proper training, the system can even be used in other areas, so it can increase public safety in order to increase criminal effectiveness, as depth checks can be performed even on images from surveillance cameras.

One of the most important security challenges is that so-called fraud detection must be built into the system, as anyone planning an illegal activity may have solutions to bypass machine vision, which could basically bypass the simplest security gates and possibly break through computer vision. Therefore, it is very important for development that you can effectively filter out such potential frauds and even be able to predict them using the integrated system. The challenge is that the capabilities of existing cameras may not be able to provide high-quality data at all times due to weather and visibility conditions. It is a challenge that it is naturally difficult to involve large amounts of data for teaching and artificial intelligence support due to GDPR and other regulations. It is expected that the image data obtained in a natural way will be supplemented with a hybrid methodology and synthetic data.

Therefore, by using artificial intelligence, profiling competencies based on pre-defined risk criteria can be developed by processing and analyzing image information from ex-

isting installed cameras. Image analysis is based on object analysis and motion analysis. The verification of the data is carried out by the competent authorities e.g. police, NAV, and NKH, which is also defined in the national strategies. Defining the professional requirements and methodology of the development during the project and processing the results of the feasibility phases is the task of the NKE as a consortium partner. The brand name “Digital School Guard” is also a conceptual element related to NKE.

6.4 Human activity recognition for public safety [scientific results 5.1.1 & 5.2.1]

With the availability of ubiquitous computers and smart portable sensors, computer vision-based technologies create a next-level opportunity for armed forces to be further researched and to be exploited. Artificial intelligence acts as a main driver for innovation in several industries and computer vision-based military applications can provide efficient decision-making support and reduce human error and bias. Computer image processing usually focuses on object detection, however, Human Activity Recognition (HAR) has become one of the most popular research topics in recent years. The purpose of HAR algorithms is to present information about simple or complex physical activity in humans. More generally, these algorithms take data from a variety of sensors and use machine learning and computer-based, machine vision techniques to extract information about human activities. Consequently, HAR can be widely used in many applications, such as medical diagnoses, elderly tracking, smart homes, and automated driving, but it’s military and defence applications remain to be explored, and additional civilian applications can be included in the research.

Motion-driven virtual games have helped market research accelerate R&D projects on the usability of HAR, however, the civilian application of computer vision in military/educational institutions as well as its school usability is special because access to data is limited. Additionally, military data access is further limited due to its nature (classified, limited access data). In educational institutions, only data with certain permissions are available due to the GDPR regulations and students ’privacy rights; it is very difficult to implement effective computer vision and object detection and HAR research development projects worldwide due to strict legal restrictions on the processing of personal data and the related social sensitivity.

However, military usability can be researched by creating artificial laboratory environments such as a border crossing zone, or controlled military campuses/institutions. More and more commercial, global technology companies have a set of competencies that can respond to the challenges posed by the current digitalization process in the de-

fence forces with practical research. Technology-based security issues affecting competence development capabilities should be supported by research and development that ensures the practical and independent applicability of independent military developments. Self-defence competence is not only a national defence but also a general national interest for every sovereign nation, but development that can be combined with the international interest of NATO membership and EU membership requires innovative thinking and pragmatic, goal- and solution-oriented research and development projects by each Central – European country.

With the arrival of 5G and other new innovations, divergent military applications are facing a technological revolution and an explosion, so countries like Poland and Hungary need to develop their own capacity to automate decision-making with research that can either make a decision on its own or a system to support human decision making in real-time in assessing objects, human activity patterns, trigger events (such as explosion, smoke, etc).

The Visegrádi 4 countries handle developments related to their own security with due care due to classified data management. Therefore, it is essential that R&D projects will be launched in the region as well, which can provide an independent capability for the defence forces and law enforcement and any other regional educational institution, the National Armed Forces, the Ministry of the Interior, and the National Service Agencies, regardless of exterior targeting authority and foreign developments. To support the objectives of the Police Headquarters, the expected results of my research can be utilized from a social, economic, and environmental point of view and are in line with the policy directions of NATO and the EU.

6.5 Practical use of AI and DLTs [scientific result 5.3]

There had been enough development which had occurred in DLTs and blockchain technology use cases beyond cryptocurrency. It is evident, that these technologies can be combined with other emerging ones, and an examination is required to understand, how the new technologies might be applied in the profession of arms. International research found that the armed forces of the most powerful countries, such as the US, China, and Russia have all publicly discussed the military applications of blockchain [235].

Hungary, being the gate for the European Union and Schengen zone, there are numerous challenges in monitoring the border. These challenges could be tackled by using distributed network-based computer vision detection tools, such as UAVs (drones) for the lengthy border and fence monitoring. The captured images of the UAVs could be automatically classified and analyzed using AI algorithms, which would improve the

decision-makers to send human resources (military) for further investigation of the designated border zone. Illegal migration and border trespassing could be also monitored in an automated way.

Another smart application of the technology would be to monitor human activity patterns at the borders. Computer vision-based Human Activity Recognition (HAR) is therefore a research area for the future, as it could provide relevant information to the authorities about abnormal activities and events.

A crucial feature of such recognition systems in these VUCA (volatile, uncertain, complex, and ambiguous) environments [236] is object detection. For law enforcement, manual video review can take enormous manpower and time, while human error is at a higher possible rate at these manual reviews. Using a distributed network for secure data storage, and artificial intelligence algorithms to detect different types of objects (with customization options, such as type of object, the colour of the object, etc) the armed forces, police, and other units could save significant time because of the improved accuracy of object detection.

6.6 Relevance of the project [scientific results 5.1.1, 5.2.1 &5.3]

The primary goal of artificial intelligence-assisted background analysis is to be able to predict and alert competent authorities in real-time in situations other than normal for any reason. NKE can provide a laboratory environment in both artificial and natural ways. As an institution itself, the university, and within the university, an artificial laboratory environment can be created that, by learning from empirical examples, facilitates experimental development. The analysis of the image information of the installed cameras with the help of artificial intelligence can perform both quality and internal control tasks. This risk analysis can be linked to jobs. HAR can even function by detecting gestures and forms of movement, and other movements can alert the competent control body in real-time. In order to ensure a safe university institution, it may be important to detect persons and clothing and to screen out unauthorized persons in a place that can only be accessed by certain persons or persons. Unauthorized access can be detected much more effectively. The sensor fusion system can also be used here, as the use of an unauthorized access system (eg unauthorized card use) can also be supported by the image information of the cameras. If someone has obtained an unauthorized card, the portrait of the unauthorized entrant currently using the card can be compared to the portrait assigned to the card.

The NKE and its departments provide effective support in defining professional re-

quirements, because the university is an extensive institution with many locations and many locations, and is the best simulation area for the school environment. In the vicinity of the NKE, the security camera system installed for the laboratory test, as well as additional laboratory tests can be performed in the NKE partner institutions and partner branches. The cooperation between the NCU and the ORFK ensures the possible national integration of the finished system.

The traditional modal analysis includes physically connected wired or wireless sensors for the analysis of structures and objects. However, this method has certain drawbacks due to the accuracy of the sensors, possibly the weight and low spatial resolution of the sensors, which limits the accuracy of the analysis, but also the high cost of optical sensors. In addition, installing and calibrating sensors is a time-consuming and labour-intensive process in itself.

Machine vision-based technologies can address these shortcomings. In my research, I want to develop the Convolutional Neural Network using residual modules as a small part of applied research and largely as an experimental development, which can serve as the backbone of the technology for machine vision-based decision support systems.

The challenge of the technology is that it is not enough to analyze individual pixels as a frame in the image of a camera, for the most accurate analysis, it is necessary to extract and process the time-spatial information in parallel with the data of the sensors. Therefore, time-space information can actually be recorded as a result of a separate sensor (e.g., in the case of a vibrating structure, its modal frequencies can be recorded). As a summary of sensors and data extracted from machine vision, more accurate image analysis systems can be built with proper training and support of artificial intelligence.

Demonstrating through the example of a smartphone, in addition to the image of the high-resolution camera built into the phone, in addition to the fixed-frame spatial-temporal information, much more accurate results can be obtained by extracting additional sensor data from the hardware device (in my example, the phone). Such sensors may be, without claiming to be exhaustive, a hardware device such as a smartphone:

- ISO: current iso value of the camera (sensor sensitivity), the ”unit” of light sensitivity. For film machines, the film determined the ISO value. In today’s digital cameras, we can set the sensitivity we want to work with without having to swap a film for it. In the case of a film or digital camera, the sensitivity of a light sensor circuit (CCR) (light) determines the lighting conditions under which you can shoot.

- MaxIso: maximum iso value that can be reached (this already indicates that the camera is trying to overcompensate for the dark)

- Expo: shutter time
- Expo_bias: this sensor helps where the sensor wants to be corrected compared to the metering to make the image bright or dark enough (if it can no longer compensate because we have reached the hardware limit)
- BackFacingCamera: we use a front or back camera, as the hardware and optics between the cameras are also defined, measurable
- FPS: frame per second, how many frames per second a given movie contains.
- FOV: field of view
- GravX-YZ: direction of gravity vector on three axes
- Heading: a compass that faces the hardware device (in my case the smartphone, but it could also be a well-equipped drone) to the north.

A system specializing in computer vision-based deep learning can only be developed in the course of research using reliable empirical results, which are more efficient, accurate, and autonomous than previous image analysis systems. The robustness of the deep learning model requires a wealth of measurement, and teaching, using different 'patterns', with varying sizes and outcomes. This is the only way to achieve a high level of sensing accuracy at the end of experimental development for more complex systems, e.g. for the analysis of security systems.

In such security systems, the analysis of more complex events may require more automated and diverse solutions than the analysis of individual images in a separate unit. Analyzing each frame separately does not clearly allow us to detect events where the object of the analysis is a shape or sequence of movements that cannot be identified from each frame.

In artificial intelligence neural network teaching, numerical data are generated using the artificial intelligence algorithm and other simulation algorithms. These are based on individual frames that exist as separate units and extract the coordinates of the objects on them. The chronological sequence of coordinates is given to the learning system.

6.7 Market and social innovation relevance of the project [scientific results 5.1.1, 5.2.1 &5.3]

The research and development activities resulting from this work hold the potential to greatly enhance the safety of educational institutions. The ultimate goal of this technological advancement is to provide Hungary with a competitive advantage on a global

scale, allowing the country to more effectively ensure its security through the implementation of cutting-edge, innovative solutions in line with its National Security Strategy. Despite the presence of limited security research initiatives in the EU, there is a lack of involvement from Hungarian partners and international market participants in these projects. This is a trend that is not expected to change, given the current trend in EU strategy papers and calls for proposals. This highlights the crucial importance of this research, which seeks to fill the gap in security research initiatives and provide Hungary with the necessary tools to address the challenges of modern security threats.

6.8 Hungarian applications of automated recognition technologies [scientific results 5.1.1 & 5.2.1]

In Hungary, automated recognition technologies are almost exclusively purchased from abroad, with limited in-house development in recent years. Although the Zrínyi 2026 Defence and Military Development Programme recognised that new types of challenges require special attention for Hungary to build, maintain and develop its cyber defence capabilities [237], the focus has not been on automation. In June 2019, the Cyber Training Centre of the Hungarian Defence Forces was inaugurated [238], which serves as a key pillar of my hybrid force development strategy, but this centre does not deal with machine vision-based detection, i.e. the categorisation of machine vision-based defence itself is questionable: Which force would be the category of a revolutionary new technology that can be used not only for the military but also for domestic purposes, and that could be useful not only for the police, border guards and secret services but also for the Hungarian defence forces? [239]

Hungary should specialise in IT fields such as electronics and software development because this is where the MH has a chance to play a special role, which could be internationally meaningful even as a NATO member [240]. Such a breakout point could be R&D on new types of challenges, automated machine vision-based recognition. In terms of value, this could also be a technological investment in which Hungary can be competitive in the military dimension, analysing the global situation, and which will result in government-civil-military interoperability, educational, economic and social added value. Supporting R&D of revolutionary technologies and the safe harmonisation of hybrid warfare is therefore in Hungary's national interest [241]

The military operational domains define the four possible physical domains (land, air, sea, and space), [242] but in general, machine vision is more appropriately classified in the fifth force domain, cyberspace, but the technology can be applied universally. The main strategic question is whether Hungarian innovations can be used in a resource-

optimised way by the Defence Forces, whether the combination of artificial intelligence and machine vision can be developed cost-effectively in Hungary, or whether it is better to buy foreign solutions, including their software risks. The potential use of deep neural network learning capabilities poses several military science challenges and consequently creates new platforms for the armed forces.

Chapter 7

Conclusion

Detection and prevention can now almost always be linked to some kind of camera system. The use of machine vision-based artificial intelligence to analyze the behaviour of individuals and groups can provide a new opportunity to prevent and respond more quickly to conflict situations. This requires R&D to leverage advances in machine vision using image recognition and image analysis, both of which require high computational power, current image analysis methodologies are often slow and do not work in real-time. With PAR, the causes of suspicious events or activities can be easily grouped (mass brawl, preparation of a terrorist act, smoke, weapons, etc.), thus making the prevention of border violations, terrorist acts, or other crimes and other national security tasks more efficient. Identification of criminals and wanted persons would not require as much time and resources, but the tracing of lost persons could be done more efficiently.

The security systems of the future will have to address new types of challenges that will drive automation, software solutions, and cybersecurity. The next generation of combat systems may be based on automated, decentralized decision-making mechanisms, which will in all cases be based almost entirely on machine vision and artificial intelligence, and will therefore be the technologies used in decision preparation. The human-added value may possibly be preserved in the final decision-making, but the depth of learning to make decisions will be another scientific question. An automated system can make a decision faster in a crisis situation, improving the chance of survival, but can a machine decide on people's lives without human approval? The potential for error can be filtered out if artificial intelligence processors loaded into various weapons systems can coordinate their actions and check the validity of the data. In the 20th century, military forces still used expensive computing power but cheap data. Centralized decision-making was justified. New challenges emerged because processing power became cost-effective, but the data itself became much more expensive. 21st-century militaries will use machine vision-based, artificial intelligence-based, and decentralized technologies.

7.1 HAR - Spoof detection

In this work, I studied video replay spoofing detection in applications where human activity recognition using RGB sensors plays a crucial role. Outdoor weather and visibility conditions are not controllable and therefore spoofing detection under such circumstances is a tedious task. I believe that, to the best of my knowledge, the problem in this background is studied for the very first time in HAR domain. In particular, I considered the virtual football game - SQILLER application for experimenting with various designed models. I formulated an ensemble multi-stream model that detects spoofing cases arising from video replay attacks. My model inspects distinct regions of the human face and combines these observations to provide robust classification between spoof and genuine cases. I collected a database comprising 38 subjects in widely varying lighting conditions and backgrounds. I also generated corresponding spoof videos using several monitors for training and testing my models.

I showed that my model provides robust results even if the face of the subject is partially visible. I also evaluated the performance of my trained model on compressed videos at different bitrates and the ability to generalize to face recognition systems. However, face recognition may need fine-tuning of my model to better fit such cases where IPD at full HD resolution is higher than that of my training images.

I also implemented and validated my algorithm on a mobile device and showed that my approach can work in real-time with minimal memory footprint, leaving enough room for running the activity recognition algorithms alongside. In the future, my method can be easily adapted to mobile devices containing advanced sensors like depth sensors and this would further improve the performance of my spoof detection approach. In addition to replay attacks, in the future, I also intend to study and explore algorithms for detecting masks or hoodies covering the face and/or other clothing that cover body parts and are designed for confusing spoof detection systems.

7.2 Panoramic HAR

A more comprehensive understanding of the activities in a crowd requires the development of a Panoramic Human Activity Recognition (PAR) system, which aims to simultaneously detect individual actions, social group activities, and universally patterned human activities. Such a recognition system is challenging, but it can answer practical problems in real-world applications. Such problems could be a crowd scene in the case of demonstrations or rallies, a dangerous situation with migratory groups at the border, or a sudden crowd scene in a public place. The paper describes the concepts and analyses a novel hierarchical graph neural network approach to represent and model progressively

human activity and mutual social relations in a crowd, how they look like to a human participant, from a machine vision perspective.

The 21st century has brought a digital technology revolution in military engineering and information science research topics. The digital paradigm shift is being completed by decentralization and automated processes, mainly defined by machine learning and artificial intelligence. Thanks to machine vision, human face and human activity recognition are now commonplace worldwide (Human Activity Recognition), but the accuracy of recognition of human activities and abnormal human activities is far from perfect. In addition to the recognition of individuals, another research challenge is the recognition of the behaviour of crowds, because in a crowd the individual often loses his 'face' and the collective consciousness overrides the individual behaviour pattern.

A more comprehensive understanding of crowd activities, therefore, requires the development of a Panoramic Human Activity Recognition (PAR) system, which aims to simultaneously detect individual actions, social group activities, and universally patterned human activities. Such a recognition system is challenging but can answer practical problems in real-world applications. Such problems could be an emerging crowd scene in the case of demonstrations or rallies, a dangerous situation with a migratory grouping at the border, or a sudden crowd scene in a public space in an observed environment. The research aims to develop a new hierarchical graph neural network approach to propose a progressive representation and modelling of human activities and mutual social relations in a crowd, what they look like to a human participant, from a machine vision perspective, and whether there are recognizable patterns of human behaviour in the crowd.

Machine vision is a "disruptive" technology that represents a revolutionary new solution, capable of taking an existing technological paradigm to the next level, the technology of automated cognition [8]. Internationally, several research and development programmes have already been launched in this field. NATO's C4ISR (Command, Control, Communications, Computers, Intelligence, Surveillance & Reconnaissance) and the US Department of Defense (DoD) have already launched their own automated, decentralized development programmes [214] because they recognize that they cannot analyze and process the vast amount of incoming data on a human scale.

The research aims to focus on a more comprehensive understanding of human activity in crowded crowd scenes. The practical benefits of such research are prevention, the development of automated video surveillance for crime prevention, and the enhancement of the effectiveness of warning systems because such a system can predict dangerous crowd scenes in time. Furthermore, the movements of individuals in a crowd can be interpreted, and individuals or small groups can be analyzed to see how their actions affect the crowd. Understanding video-based human activity is an important computer vision task, but also a major challenge in terms of accurate recognition performance.

7.3 Machine vision-based activity recognition

Machine vision technology has a crucial role in analyzing human activity in crowded environments. In order to understand the patterns of human behaviour, it is important to analyze individuals' actions, group activity, and the overall recognition of human activity.

HAR recognition focuses on the recognition of the global activity of a group of people. In other words, it aims to understand the interactions between individuals in a crowd. This type of recognition can be divided into three categories.

The first category is the recognition of the category of activity in a machine vision video. This involves recognizing the type of activity being performed by individuals in a crowd. The second category is the recognition of human interaction and activity in a crowd. This is a more comprehensive and multifaceted approach, known as panoramic human activity recognition, which aims to better understand human behaviour in crowded environments.

The third category is the recognition of social activity, which is part of a larger social activity analysis. This involves analyzing the behaviour of individuals in a crowd as they interact with one another, in order to gain insights into group dynamics and social patterns.

Overall, HAR theory is a crucial component of machine vision technology, as it allows for the simultaneous analysis of individual actions, group activity, and the overall recognition of human activity in crowded environments.

7.4 Violence Detection

In this work, I addressed the problem of efficient violence detection for automated surveillance applications. I adapted X3D-M deep learning architecture that is computationally lightweight to learn and detect violence patterns from videos. I proposed two architectures - FT and TL, for classifying video clips containing violence, which leverage action recognition features learned from the Kinetics-400 dataset.

For detailed analysis and performance evaluation of the proposed approaches, I collected and extended seven different datasets in my study. In the past, several deep learning-based methods for violence detection focused on datasets involving mostly fighting between two or more people for experiments. I remind that the spectrum of actions and visual patterns representing violence is far wider, for example, violence happening between a group of people in the form of a fight is visually very different from violence using objects such as guns or violence involving explosions. To also incorporate such cases, I annotated several videos from UCF and XD-Violence datasets for my experiments.

Using my collected videos, the FT model optimizes the X3D-M parameters learned from the Kinetics-400 dataset, while the TL model extracts spatiotemporal features first without modifying the X3D-M parameters (trained on the Kinetics-400 dataset) to train multiple fully connected layers. My experiments with individual datasets show that both models performed decently in terms of ACC and AUC scores on the collected datasets. However, the FT model performed better than most of the state-of-the-art methods on popular datasets with relatively fewer model parameters.

In the previous works for violence detection, cross-dataset evaluations are not thoroughly studied and I argue that these evaluations are very important for understanding the prominence of various datasets as well as deep learning models. In this work, I bridge this gap by providing a comprehensive evaluation including one on one cross dataset validation and leave one out cross-validation. My cross-dataset tests showed that the TL model generalizes better to unseen scenarios than the FT model. However, when testing on the combined dataset, the FT model achieved better performance and the TL model produced more combined false positives and false negatives. Further tests on individual datasets show that models trained on the combined dataset did not perform well in several cases when compared to the performance of models trained on individual datasets. This shows that there are multiple inconsistencies in the publicly available datasets for violence detection. Also, results from comparisons with several methods in the literature have shown limitations of both the developed methods and existing datasets.

I point out that the existing public datasets for violence detection are incoherent in terms of the video duration, FPS, number of videos available for training and testing as well as the forms of violence. Furthermore, existing datasets are not particularly representative of surveillance applications. My results show that, in the future, there is a great need for the development of diverse and meaningful large-scale datasets also involving footage from real-world surveillance. In the future, I plan to make steps toward constructing such a large-scale dataset. Once such datasets are available, I also plan to re-validate the models presented in the current work for more general results.

I also presented a computationally light and functional stand-alone system architecture for implementing the proposed models in practical surveillance applications. In this architecture, from the incoming video stream, I extracted and evaluated non-overlapping video segments having a duration of four seconds. Such a strategy fails to work in cases when an event of violence begins at the end of a segment and ends before the end of a consecutive segment. While processing such segments, my proposed models may not report accurate violence coefficients. In the future, I also plan to develop smart strategies to handle such scenarios, following techniques such as reducing the size of video segments adaptively and/or using overlapped segments. The main focus in developing such strategies will be on achieving the best computational speed and accuracy trade-off.

7.5 DLT + AI + BLOCKCHAIN

The integration of blockchain and AI technologies in the military sector is a rapidly evolving field. While blockchain-based applications in the military are not yet combat-ready, AI technologies are being widely used in various industries. This is largely due to the advancements in computer vision and machine learning, which have led to a revolution in biometric identification techniques such as face recognition.

The use of blockchain technology in defence logistics is already seeing significant benefits, as reported by Accenture in a research report. The report states that 86% of aerospace and defence companies plan to integrate blockchain within three years. Additionally, 93% of aerospace and defence executives believe that the next generation of intelligent solutions is moving into physical environments.

However, the integration of these technologies in the military sector also raises several challenges and dilemmas. When a nation decides to upgrade its cyber security strategies at a federal level, it must consider the recommendations made by international organizations and how these cyber challenges will be addressed. Kovács [243] emphasizes the importance of a country's national cyber security strategy being built on the basis of recommendations made by international organizations and the key regulatory issues that must be considered.

The rapidly evolving nature of cyberspace has prompted NATO and the European Union to develop individual policies and regulations regarding cyber security. NATO, in particular, views cyberspace as a domain of warfare and recognizes its far-reaching consequences for member states. This has led to a range of cyber security measures being put in place, including the delegation of cyber defence tasks to the army, the development of cyberattack capabilities, and the establishment of cyber commands.

Another key aspect of NATO's Cyber Pledge is to address the varying levels of cyber defence capabilities among its member states. These disparities highlight the importance of NATO's Cyber Operation Centre in not only the military sector but also the civil defence sector [244]. The Centre plays a crucial role in helping to bridge these gaps and strengthen the overall cyber defence capabilities of NATO member states.

NATO also recognizes the importance of international cooperation in the development and implementation of innovative technologies. To that end, the Alliance promotes collaboration between its member states and non-NATO countries in the field of cyber security. This cooperation can help to advance the state of the art in cyber security technologies and improve the overall defence against cyber threats.

Blockchain technology offers a number of unique advantages that make it a highly reliable and secure technology. One of the key features of blockchain technology is its reliability. Blockchain networks require the agreement of both internal and external users

in order to operate effectively. This consensus mechanism ensures that the network is secure and tamper-proof, as any attempted breaches are immediately apparent and can be dealt with promptly. Another key advantage of blockchain technology is its transparency. Unlike traditional computer security systems that rely on the functioning of individual nodes, blockchain networks are based on a cryptographic data structure that makes manipulation extremely complex. This structure ensures that the network is secure, even if some of the nodes within it are malfunctioning. Finally, blockchain networks are highly fault-tolerant. They coordinate trusted nodes to reject any untrusted entities, reducing the likelihood of failure and significantly increasing the cost of attempted breaches by foreign parties. This makes blockchain technology an ideal solution for organizations and institutions that require highly secure and reliable technology for their operations.

The development of in-house expertise in blockchain technologies within the Central Defense Management Authorities is deemed essential. This is because blockchain technologies have the potential to offer numerous benefits to the defence sector, and a strong understanding of these technologies will ensure that the defence sector can take full advantage of them. In order to further the development of blockchain-based technologies, it is recommended that the Central Defense Management Authorities look for strong partnerships with the industry. This will not only result in the development of cutting-edge technologies but will also result in mutual benefits for both parties.

The Hungarian Defense Forces could potentially adopt artificial intelligence-powered detections for applications such as border protection, institutional security, and public safety. This could be achieved by utilizing the existing infrastructure of CCTVs, thereby reducing the need for significant investments. Further research could be conducted into the use of unmanned aerial vehicles (UAVs) and portable optical infrastructures (such as self-driving cars used for Google Street View).

The rapidly evolving cyberspace requires more stringent regulations and accountability. While developed nations are struggling to keep up with the influx of cyber-related challenges, it is imperative that underdeveloped nations also prepare themselves to defend against these threats. Blockchain technology and Artificial Intelligence offer a way to develop smarter and more secure cyber defence systems. By developing a strong understanding of these technologies and leveraging their capabilities, countries can safeguard their citizens from cyber threats.

7.6 PAR - Overview of Human Activity Recognition and new types of challenges

Human interaction recognition by machines remains a complex challenge, with less attention paid to it compared to human action recognition. Previous research in the field of human interaction recognition has primarily focused on two-subject interactive activities, such as hugging, and has relied heavily on data samples collected from movies and TV shows [245]. While these sources provide an ample supply of data, they also present an artificial view of human interactions and may not accurately represent real-world interactions.

More recent studies have started to address human interaction recognition in multi-subject scenes [246], such as crowd scenes [247], to better understand the differences in spatiotemporal perception of human-human interactions. This is an important area of investigation as the perception of human interactions can vary greatly in crowded versus two-subject scenes. Additionally, there is a need to understand if traditional human interactions follow the same pattern across different cultures, for example, do Koreans hug in the same way as Brazilians? To fully address the challenges posed by human interaction recognition, it is important to continue conducting research in this area, exploring both two-subject and multi-subject interactions and considering cultural differences in human interactions.

The field of Panoramic Activity Recognition (PAR) aims to offer a comprehensive understanding of human recognition technology through the representation and modelling of human activities in crowd scenes. The focus is on capturing multiple individuals and their interrelated activities with a high level of detail and accuracy. While previous attempts have been made to recognize individuals in a crowd scene, they are often impractical and result in high error rates. To overcome this, it is suggested that research be conducted to investigate the feasibility of implementing a hierarchical graph network that integrates individual, pair, and crowd recognition technologies. This network would effectively combine existing technologies to achieve a more accurate recognition of human activities in a crowded scene.

The methods of analysis for Panoramic Activity Recognition (PAR) involve the implementation of complementary tasks aimed at accurately identifying individuals and their respective activities. These tasks can range from the recognition of individual activities to the recognition of social and societal interactions between individuals. For example, recognizing social activity in a conversation can help to determine when two individuals are facing each other, thereby providing valuable information about their interaction. Additionally, an understanding of individual actions and social group activities

in a crowded scene can enhance the recognition of general activity in a group. These methods of analysis work in tandem to provide a comprehensive overview of human recognition technology, effectively representing and modelling human activities at various levels of detail and interrelationships within crowded scenes.

7.7 Global Activity Recognition or Group Activity Recognition (GAR)

Group Activity Recognition (GAR) is a task in the field of human activity understanding that aims to recognize the collective activity performed by a group of individuals. Early research in this field focused on classifying recorded video frames into predefined activity categories. However, recent studies have demonstrated that considering individual actions and human-human interactions can greatly enhance the performance of GAR systems.

Several recent works have proposed incorporating individual action recognition labels as an additional supervision signal for GAR [248]. Other methods [249, 250] have focused on modelling the relationships between multiple individuals to better represent the group activity.

The PAR (Parallel Activity Recognition) framework aims to provide a comprehensive understanding of human activity by addressing the three tasks of individual action recognition, social group activity recognition, and global activity recognition simultaneously. Unlike previous works which deal with these subtasks separately or sequentially, PAR seeks to develop a unified framework that can handle all of them in parallel.

However, developing such a framework poses several technical challenges, as it requires the integration of multiple models and the consideration of various factors such as individual actions, human-human interactions, and global context. This requires the integration of multiple recognition models, as well as the consideration of various factors, such as individual actions, human-human interactions, and the global context. It is an open research problem and requires extensive experimentation and validation to ensure its effectiveness.

7.8 Using machine vision and artificial intelligence in PAR

The integration of Artificial Intelligence (AI) and Machine Vision into the Panoramic Activity Recognition (PAR) process introduces a multitude of technical, scientific, legal, and social implications that require examination. The collection of reconnaissance data from machine vision-based software, regardless of the platform - airborne, land, maritime, or

space - highlights the need for proper evaluation of the potential consequences. The deployment of AI-controlled cameras equipped with machine vision capabilities presents a new challenge in the decision-making process, as the camera would be responsible for selecting which street cluster to analyze by processing input data in real-time, without the need for central decision-making. Furthermore, the processed data would be transmitted through encrypted systems, with the AI determining which recognition data to transmit. The learning process of the deep neural network used by the AI-controlled cameras also presents a concern, as the processed data may not always infer the complete information of the input data, as seen in the example of a crowd fight. The ethical implications of AI-controlled cameras processing, learning from and transmitting data require further examination to ensure data privacy and ethical usage.

AI is based on learning, so these systems need continuously available data and data feeds to achieve faster learning curves. Obtaining data is expensive, as military information is classified and secret, so obtaining data for military applications of AI systems is increasingly difficult, a legal concern, and overall a very costly process. Intelligent solutions exist where encrypted data can be obtained from unverified or open source sources available in [251] form (e.g. the publicly available University of Central Florida offender video database), but the data can still generally be subject to human error and require secondary evaluation and confirmation. Data on military applications is even more interesting, as data security should be a priority in defence administration and in the daily communication of authorities. The use of manually obtained data may conflict with the use of processed personal data and related data processing operations. The GDPR and other regulatory policies, therefore, limit the source of data use [195]. In the near future, it is proposed to support research and development even with synthetic data, because data that can be applied in a given situation and cannot be obtained directly by other measurements will be needed to develop more efficient artificial intelligence algorithms for accurate machine vision-based activity recognition.

The use of machine vision and artificial intelligence in civil and military applications raises many questions [252]. While real-time analysis is not necessarily required for PAR, it would be an ideal preventive activity in terms of its usefulness. Computing real-time results require significant resources, as scarce capacity can make computing operations costly. However, hardware capabilities have improved rapidly in recent times, so the software solution for PAR will depend more on R&D results in the future. The scientific challenge extends to the artificial isolation of such a system and the military risks of machine learning or programmed AI awakening.

The practical international examples are also a guideline for Hungary, possibly also for the Zrínyi 2026 Force Development Programme. America is leading the way in automated development, such as the Maven project in the United States [253]. This pro-

gramme is a sub-programme of the Algorithmic Warfare Cross-Functional Team (AWCFT) programme, aimed at maintaining the competitive advantage of the US military through the use of automated machine learning and machine vision. The U.S. Department of Defense (DARPA) recognized years ago that they could not process the gigantic amount of data collected or acquired by the services and the military in a human capacity, and could only interpret it through machine learning. The process has been defined as Process - Exploit - Disseminate [254] and any digital photos and videos (Mid-Altitude Full Motion Video) taken by machine vision are being processed in this way. By applying artificial intelligence, data tagging, and algorithmic selection can be achieved, which speeds up the military decision mechanism. The HAR, PAR and all other systems that predict and recognise human activity can be applied to this data, and the efficiency of recognition will be improved thanks to machine learning, not only for individual ad hoc activities, and group activities but also for the recognition and classification (labelling) of objects and objects. Thanks to artificial intelligence, this technology has been more efficient than humans for years. In Hungary, such a recognition system could have helped in many cases: in the search for missing people, in the search for the perpetrator of the Béke Square bombing in the 13th district, or in the search for the perpetrator of the attack on police officers in Oktogon.

The Russian-Ukrainian conflict has also accelerated the development of UAV-UAS (Unmanned Aerial Vehicle and Unmanned Aircraft System) reconnaissance. It is currently estimated that more than 100 countries have military drones, of which 20 have used armed drones (not necessarily by state-affiliated organisations). Unmanned Aerial Systems (UAS) are often not yet very sophisticated in terms of robotics but are almost all remotely piloted. Autonomy is becoming increasingly important in the management of different vehicles. An example is the Guardium, an Israeli Unmanned Ground Vehicle (UGV) developed by G-NIUS, which is used for combat and defence along the Gaza border. The vehicle is self-driving, but the weapons on board are still operated remotely by humans. Machine vision-based artificial intelligence applications can be easily integrated into weapon systems for military missions [232] but also into civilian applications [255] (In the health sector I already see many examples of this, e.g. more accurate ECG results prediction and analysis [256]).

7.9 Summary

Detection and prevention can now almost always be linked to some kind of camera system. The use of machine vision-based artificial intelligence to analyse the behaviour of individuals and groups can provide a new opportunity to prevent and respond more

quickly to conflict situations. This requires R&D to leverage advances in machine vision using image recognition and image analysis, both of which require high computational power [191], current image analysis methodologies are often slow and do not work in real-time. With PAR, the causes of suspicious events or activities can be easily grouped (mass brawl, preparation of a terrorist act, smoke, weapons, etc.), thus making the prevention of border violations, terrorist acts or other crimes and other national security tasks more efficient. Identification of criminals and wanted persons would not require as much time and resources, but the tracing of lost persons could be done more efficiently.

With the Zrínyi programme, the research and development direction of Hungarian force development is partly following the traditional direction. The research and development programmes of the Zrínyi 2026 force development programme should also focus on the development of special capabilities because Hungary can be internationally competitive in this field. R&D support for new types of cyber defence challenges could be an indispensable and one of the best investments in force development. The Hungarian institutional background may be able to provide the infrastructure and tools to enable the development of automated systems, and the development, maintenance and training of Hungarian intellectual capability at a high level, but it will be more a question of money and a strategic issue of allocation of material resources. On the other hand, there is little chance that Hungary can become and remain a world leader in weapons or military vehicle production: this would require a lot of money and expertise, as well as access to closed data. In Hungary, conventional weaponry is the direction of development (drone and other purchases are those from outside contractors), because the existing stock of equipment is insufficient in quality and quantity for the current situation. Hungary is close to NATO requirements in terms of expenditure, but it is questionable whether the burden on the reform of the force development is proportionate to the results that can be achieved. Hungary's research direction could be the military applicability of specific specialisations, and one of the clearest areas of specialisation is machine vision and artificial intelligence.

This is underpinned by the fact that the military of the future will have to address new types of challenges that will drive automation, software solutions and cybersecurity. The next generation of combat systems may be based on automated, decentralised decision-making mechanisms, which will in all cases be based almost entirely on machine vision and artificial intelligence, and will therefore be the technologies used in decision preparation. The human-added value may be preserved in the final decision-making, but the depth of learning to make decisions will be another scientific question. An automated system can make a decision faster in a crisis, improving the chance of survival, but can a machine decide on people's lives without human approval? The potential for error can be filtered out if artificial intelligence processors loaded into various weapons systems can

coordinate their actions and check the validity of the data. In the 20th century, military forces still used expensive computing power but cheap data. Centralised decision-making was justified. New challenges emerged because processing power became cost-effective, but the data itself became much more expensive. 21st-century militaries will use machine vision-based, artificial intelligence-based, decentralised technologies [257].

List of Figures

List of Tables

Abbreviations

4K 4-Kilo.

5G 5th Generation mobile network.

ACC Accurcy.

AI Artificial Intelligence.

AIRT Artificial Intelligence and Real-Time system.

AMD Advanced Micro Devices.

AMT Amazon Mechanical Turkers.

API Application Programming Interface.

ATR Automatic Target Recognition.

AUC Area Under Curve.

AWCFT Algorithmic Warfare Cross-Functional Team.

AWS Aegis Weapon System.

C2 Command -and-Control.

C4ISR Command, Control, Communications, Computers, Intelligence, Surveillance Reconnaissance).

CASIA Institute of Automation, Chinese Academy of Sciences.

CCTV Closed-circuit Tele-Vision.

CD Compact Disc.

CEO Chief Executive Officer.

CF Concatenated Frames Model.

CNN Cellular Neural Network.

COCO Common Objects in Context.

CPU Central Processing Unit.

CUDA Compute Unified Device Architecture.

CV Crowd Violence.

DARPA Defense Advanced Research Projects Agency.

DB Data Base.

DF Delayed Frames Model.

DISA Defense Information Systems Agency.

DLT Distributed Ledger Technology.

DoD Department of Defense.

ECG ElectroCardioGram.

EER Equal Error Rate.

EM Ensemble Multi-Stream Model.

EU EUrope.

FAR False Acceptance Rate.

FASD Face Anti-Spoofing Database.

FFMPEG Fast Forward Moving Picture Experts Group.

FOV Field Of View.

FPR False Positive Rate.

FPS Frames Per Second.

FRR False Rejection Rate.

FT Fine-Tuned.

GAR Group Activity Recognition.

GDPR General Data Protection Regulation.

GHz GigaHertz.

GPU Graphics processing unit.

Grav Gravity vector.

GTX Giga Texel shader eXtreme.

HAR Human Activity Recognition.

HD High Definition.

HF Hockey Fights.

HOF Histogram of Oriented optical Flow.

HOG Histogram of Oriented Gradients.

HSV Hue Saturation Value.

HTER Half Total Error Rate.

IBM International Business Machines.

IOS iPhone Operating System.

IOT Internet Of Things.

IPD Inter-Pupillary Distance.

ISO International Standards Organization.

IT Information Technology.

JTRS Joint Tactical Radio System.

Kbps Kilobit per second.

KDE Kernal Density Estimation.

KSI Keyless Signature Infrastructure.

LBP Local Binary Patterns.

LBP-TOP Local Binary Pattern histograms from Three Orthogonal Planes.

LG Lucky Goldstar.

LSTM Long Short-Term Memory.

MF Movie Fights.

MFB Mel Filter-Bank.

MFSD Mobile Face Spoofing Database.

MIL Multiple Instance Learning.

MoSIFT Motion Scale-Invariant Feature Transform.

ms milli seconds.

MSTAR Mobile and Stationary Target Acquisition and Detection.

MSU Michigan State University.

NATO North Atlantic Treaty Organization.

NAV Nemzeti Adó- és Vámhivatal.

NCU National Coordination Unit.

NIM National Intelligence Model.

NKE Nemzeti Közszolgálati Egyetem.

NKH Nemzeti Közlekedési Hatóság.

ORFK Országos Rendőr-főkapitányság.

PAR Panoramic Human Activity Recognition.

POW Proof-Of-Work.

R&D Research & Development.

ResNet Residual Network.

RGB Red Green Blue.

RLVS Real Life Violence Situations.

RNN Recurrent Neural Networks.

ROC Receiver Operating Characteristic.

RWF-2K Real-World Fight-2000.

SBP Statistical Binary Pattern.

SBP-TOP Statistical Binary Pattern histograms from Three Orthogonal Planes.

SDK Software Development Kit.

SF Single Frame Model.

SIFT Scale-Invariant Feature Transform.

SPIL Skeleton Points Interaction Learning.

SVM Support Vector Machine.

Ti Titanium.

TL Transfer-Learned.

TPR True Positive Rate.

TV Tele-Vision.

UAS Unmanned Aerial Systems.

UAV Unmanned Aerial Vehicle.

UAV-UAS Unmanned Aerial Vehicle and Unmanned Aircraft System.

UCF University of Central Florida.

UCFS UCF-Crime Selected.

UGV Unmanned Ground Vehicle.

UK United Kingdom.

UNICRI United Nations Interregional Crime and Justice Research Institute.

US United States.

USB Universal Serial Bus.

USSA Unconstrained Smartphone Spoof Attack.

VD-Net Violence Detection Network.

VGG Visual Geometry Group.

ViF Violent Flows.

VUCA Volatile, Uncertain, Complex, and Ambiguous.

X3D-L EXpanded 3D architecture - Large.

X3D-M EXpanded 3D architecture - Medium.

X3D-S EXpanded 3D architecture - Small.

X3D-XL EXpanded 3D architecture - eXtra Large.

XD-V XD-Violence Selected.

YOLO You Only Look Once.

ZJU ZheJiang University.

Publications

[1] V. D. Huszár and V. K. Adhikarla, "Live spoofing detection for automatic human activity recognition applications," *Sensors*, vol. 21, no. 21, p. 7339, 2021.

[2] V. Huszár, "Hamisítás észlelési módszerek az automatizált emberi tevékenység felismerésben," *Hadtudomány*, vol. 32, no. 1, pp. 270–284, 2022.

[3] H. Viktor, "A blokklánc, a számítógépes látás és a mesterséges intelligencia alkalmazási lehetőségei a kiberhadviselésben," *Hausner, Gábor (szerk.) Szemelvények a katonai műszaki tudományok eredményeiből II., Ludovika Egyetemi Kiadó*, 2021.

[4] V. Huszár, "Kiberbiztonság mint a haderőfejlesztés kiemelt területe: a decentralizáció és a blokklánc-technológia lehetőségei a kibertérben," *Honvédségi Szemle: A Magyar Honvédség Központi Folyóirata*, vol. 148, no. 3, pp. 3–17, 2020.

[5] H. Viktor, "A decentralizáció és a blockchain-technológia felhasználási lehetőségei gépi látás és mesterséges intelligencia használatával a katonai szervezetekben," *Hadmérnök*, vol. 14, no. 4, pp. 179–189, 2020.

[6] V. Huszár, "Distributed intelligence: Cyber warfare application possibilities of computer vision and artificial intelligence," *Honvédségi Szemle–Hungarian Defence Review*, vol. 149, no. 1-2., pp. 4–19, 2021.

[7] V. Huszár, "Application possibilities of decentralization and blockchain technology using computer vision and artificial intelligence in defense management, military and police organizations," *Honvédségi Szemle: A Magyar Honvédség Központi Folyóirata*, vol. 148, no. 1.-SI, pp. 4–14, 2020.

Bibliography

[8] J. L. Bower and C. M. Christensen, "Disruptive technologies: catching the wave," 1995.

[9] S. Haber and W. S. Stornetta, "How to time-stamp a digital document," in *Conference on the Theory and Application of Cryptography*. Springer, 1990, pp. 437–455.

[10] The Economist, "Blockchains: The great chain of being sure about things, 2015," Available from https://www.economist.com/briefing/2015/10/31/the-great-chain-of-being-sure-about-things, visited on 29/10/2022.

[11] N. Szabo, "Formalizing and securing relationships on public networks," *First monday*, 1997.

[12] L. Cuen, "Most crypto exchanges still don't have clear kyc policies: Report," *coindesk. com*, vol. 27, 2019.

[13] J. Cohen, *Statistical power analysis for the behavioral sciences*. Routledge, 2013.

[14] O. Williams-Grut, "Banks are looking to use artificial intelligence in almost every part of their business: Here's how it can boost profits," *Business Insider*, 2017.

[15] V. A. Popescu, G. Popescu, and C. R. Popescu, "The amazing world of the internet-challenges of the internet age." *Manager (University of Bucharest, Faculty of Business & Administration)*, no. 12, 2010.

[16] K. McCaney, "Law enforcement using analytical tools to predict crime," 2010.

[17] S. Alzou'bi, H. Alshibl, and M. Al-Ma'aitah, "Artificial intelligence in law enforcement, a review," *International Journal of Advanced Information Technology*, vol. 4, no. 4, p. 1, 2014.

[18] Radhika Madhavan, "Artificial Intelligence in Policing – Use-Cases, Ethical Concerns, and Trends, 2019," Available from https://emerj.com/ai-sector-overviews/artificial-intelligence-in-policing/, visited on 29/10/2022.

[19] Brig. Gen. Mark T. Simerly and Daniel J. Keenaghan, "Blockchain for military logistics, 2019," Available from https://www.army.mil/article/227943/blockchain_for_military_logistics, visited on 29/10/2022.

[20] S. E. e. Jones, "Wearable smart sensor systems integrated on soft contact lenses for wireless ocular diagnostics," *Nat. Commun.*, 2017.

[21] S. E. Jones, V. T. v. Hees, D. R. Mazzotti, P. Marques-Vidal, S. Sabia, A. v. der Spek, H. S. Dashti, J. Engmann, D. Kocevska, J. Tyrrell *et al.*, "Genetic studies of accelerometer-based sleep measures in 85,670 individuals yield new insights into human sleep behaviour," *bioRxiv*, p. 303925, 2018.

[22] J. Kim, A. S. Campbell, B. E.-F. de Ávila, and J. Wang, "Wearable biosensors for healthcare monitoring," *Nature biotechnology*, vol. 37, no. 4, pp. 389–406, 2019.

[23] A. Jalal, M. Z. Uddin, and T. . Kim, "Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 3, pp. 863–871, 2012.

[24] Y. Xing, C. Lv, H. Wang, D. Cao, E. Velenis, and F. Wang, "Driver activity recognition for intelligent vehicles: A deep learning approach," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 6, pp. 5379–5390, 2019.

[25] A. R. Revathi and D. Kumar, "A survey of activity recognition and understanding the behavior in video surveliance," *Vis Comput.*, vol. 29, pp. 983–1009, 2013.

[26] A. Mukherjee, S. Misra, P. Mangrulkar, M. Rajarajan, and Y. Rahulamathavan, "Smartarm: A smartphone-based group activity recognition and monitoring scheme for military applications," in *2017 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*, 2017, pp. 1–6.

[27] SQILLER App, "The digital football game (2019)," Available at https://sqillerapp.com/, visited on 19/07/2021.

[28] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Communications Surveys Tutorials*, vol. 15, no. 3, pp. 1192–1209, 2013.

[29] H. F. Nweke, Y. W. Teh, M. A. Al-garadi, and U. R. Alo, "Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges," *Expert Systems with*

Applications, vol. 105, pp. 233–261, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417418302136

[30] R. Bresan, A. Pinto, A. Rocha, C. Beluzo, and T. Carvalho, "Facespoof buster: a presentation attack detector based on intrinsic image properties and deep learning," 2019.

[31] S. Bharadwaj, T. I. Dhamecha, M. Vatsa, and R. Singh, "Computationally efficient face spoofing detection with motion magnification," in *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 105–110.

[32] G. Pan, L. Sun, Z. Wu, and S. Lao, "Eyeblink-based anti-spoofing in face recognition from a generic webcamera," in *2007 IEEE 11th International Conference on Computer Vision*, 2007, pp. 1–8.

[33] E. Uzun, S. P. Chung, I. Essa, and W. Lee, "rtcaptcha: A real-time captcha based liveness detection system," in *NDSS*, 2018.

[34] Z. Zhang, D. Yi, Z. Lei, and S. Z. Li, "Face liveness detection by learning multispectral reflectance distributions," in *2011 IEEE International Conference on Automatic Face Gesture Recognition (FG)*, 2011, pp. 436–441.

[35] N. Erdogmus and S. Marcel, "Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect," in *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2013, pp. 1–6.

[36] W. Bao, H. Li, N. Li, and W. Jiang, "A liveness detection method for face recognition based on optical flow field," in *2009 International Conference on Image Analysis and Signal Processing*, 2009, pp. 233–236.

[37] M. De Marsico, M. Nappi, D. Riccio, and J.-L. Dugelay, "Moving face spoofing detection via 3d projective invariants," in *2012 5th IAPR International Conference on Biometrics (ICB)*, 2012, pp. 73–78.

[38] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *2012 BIOSIG - Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG)*, 2012, pp. 1–7.

[39] J. Määttä, A. Hadid, and M. Pietikäinen, "Face spoofing detection from single images using micro-texture analysis," in *2011 International Joint Conference on Biometrics (IJCB)*, 2011, pp. 1–7.

[40] J. Bai, T.-T. Ng, X. Gao, and Y.-Q. Shi, "Is physics-based liveness detection truly possible with a single image?" in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, 2010, pp. 3425–3428.

[41] X. Tan, Y. Li, J. Liu, and L. Jiang, "Face liveness detection from a single image with sparse low rank bilinear discriminative model," in *Computer Vision – ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 504–517.

[42] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face spoofing detection using colour texture analysis," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 8, pp. 1818–1830, 2016.

[43] J. Komulainen, A. Hadid, M. Pietikäinen, A. Anjos, and S. Marcel, "Complementary countermeasures for detecting scenic face spoofing attacks," in *2013 International Conference on Biometrics (ICB)*, 2013, pp. 1–7.

[44] J. Galbally, S. Marcel, and J. Fierrez, "Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 710–724, 2014.

[45] S. Unnikrishnan and A. Eshack, "Face spoof detection using image distortion analysis and image quality assessment," in *2016 International Conference on Emerging Technological Trends (ICETT)*, 2016, pp. 1–5.

[46] D. C. Garcia and R. L. de Queiroz, "Face-spoofing 2d-detection based on moiré-pattern analysis," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 778–786, 2015.

[47] S. R. Arashloo, J. Kittler, and W. Christmas, "An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol," *IEEE Access*, vol. 5, pp. 13 868–13 882, 2017.

[48] J. Li, Y. Wang, T. Tan, and A. K. Jain, "Live face detection based on the analysis of Fourier spectra," in *Biometric Technology for Human Identification*, A. K. Jain and N. K. Ratha, Eds., vol. 5404, International Society for Optics and Photonics. SPIE, 2004, pp. 296 – 303. [Online]. Available: https://doi.org/10.1117/12.541955

[49] G. Kim, S. Eum, J. K. Suhr, D. I. Kim, K. R. Park, and J. Kim, "Face liveness detection based on texture and frequency analyses," in *2012 5th IAPR International Conference on Biometrics (ICB)*, 2012, pp. 67–72.

[50] A. Pinto, W. R. Schwartz, H. Pedrini, and A. d. R. Rocha, "Using visual rhythms for detecting video-based facial spoof attacks," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 5, pp. 1025–1038, 2015.

[51] A. Pinto, H. Pedrini, W. R. Schwartz, and A. Rocha, "Face spoofing detection through visual codebooks of spectral temporal cubes," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 4726–4740, 2015.

[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[53] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," 2018.

[54] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu, "Face anti-spoofing using patch and depth-based cnns," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, 2017, pp. 319–328.

[55] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel, "Can face anti-spoofing countermeasures work in a real world scenario?" in *2013 International Conference on Biometrics (ICB)*, 2013, pp. 1–8.

[56] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face antispoofing database with diverse attacks," in *2012 5th IAPR International Conference on Biometrics (ICB)*, 2012, pp. 26–31.

[57] S. Tirunagari, N. Poh, D. Windridge, A. Iorliam, N. Suki, and A. T. S. Ho, "Detection of face spoofing using visual dynamics," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 762–777, 2015.

[58] S. S. W. Kim and J.-J. Han, "Face liveness detection from a single image via diffusion speed model," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2456–2465, 2015.

[59] D. Menotti, G. Chiachia, A. Pinto, W. Robson Schwartz, H. Pedrini, A. Xavier Falcao, and A. Rocha, "Deep representations for iris, face, and fingerprint spoofing detection," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, p. 864–879, Apr 2015. [Online]. Available: http://dx.doi.org/10.1109/TIFS.2015.2398817

[60] Z. Boulkenafet, J. Komulainen, and A. Hadid, "face anti-spoofing based on color texture analysis," 2015.

[61] S. R. Arashloo, J. Kittler, and W. Christmas, "Face spoofing detection based on multiple descriptor fusion using multiscale dynamic binarized statistical imagefeatures," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 11, pp. 2396–2407, 2015.

[62] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 746–761, 2015.

[63] J. Y. Bok, K. H. Suh, and E. C. Lee, "Verifying the effectiveness of new face spoofing db with capture angle and distance," *Electronics*, vol. 9, no. 4, 2020. [Online]. Available: https://www.mdpi.com/2079-9292/9/4/661

[64] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," 2019.

[65] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *CoRR*, vol. abs/1804.02767, 2018. [Online]. Available: http://arxiv.org/abs/1804.02767

[66] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[67] C. Feichtenhofer, "X3D: expanding architectures for efficient video recognition," *CoRR*, vol. abs/2004.04730, 2020. [Online]. Available: https://arxiv.org/abs/2004.04730

[68] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015.

[69] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[70] ISO, *ISO/IEC JTC 1 /SC 37 Biometrics Information technology – Biometric presentation attack detection – Part 3: Testing and reporting*. International Organization for Standardization 2017, sep 2017.

[71] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel, "Lbp top based countermeasure against face spoofing attacks," in *Computer Vision - ACCV 2012 Workshops*, J.-I. Park and J. Kim, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 121–132.

[72] T. P. Nguyen, N.-S. Vu, and A. Manzanera, "Statistical binary patterns for rotational invariant texture classification," *Neurocomputing*, 2016. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01245103

[73] Y. Zhang, R. K. Dubey, G. Hua, and V. L. Thing, "Face spoofing video detection using spatio-temporal statistical binary pattern," in *TENCON 2018 - 2018 IEEE Region 10 Conference*, 2018, pp. 0309–0314.

[74] FFmpeg Developers, "ffmpeg tool (Version be1d324), 2016," Available from http://ffmpeg.org/, visited on 19/07/2021.

[75] A. Costa-Pazo, S. Bhattacharjee, E. Vazquez-Fernandez, and S. Marcel, "The replay-mobile face presentation-attack database," in *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2016, pp. 1–7.

[76] Swift Developers, "Swift programming language," Available from https://developer.apple.com/swift/, visited on 19/07/2021.

[77] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[78] L. Liu, L. Shao, and P. I. Rockett, "Genetic programming-evolved spatio-temporal descriptor for human action recognition," in *BMVC*, 2012.

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

[80] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with microsoft kinect sensor: A review," *IEEE Transactions on Cybernetics*, vol. 43, no. 5, pp. 1318–1334, 2013.

[81] L. Cruz, D. Lucio, and L. Velho, "Kinect and rgbd images: Challenges and applications," in *2012 25th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials*, 2012, pp. 36–49.

[82] ASUS, "Xtion 2 depth sensor, 2017," Available from https://www.asus.com/ch-en/networking-iot-servers/smart-home/security-camera/xtion-2/, visited on 29/10/2022.

[83] Intel, "Real Sense Depth Camera D435, 2022," Available from https://www.intelrealsense.com/depth-camera-d435/, visited on 29/10/2022.

[84] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *CVPR 2011*, 2011, pp. 1297–1304.

[85] W. Ding, K. Liu, X. Fu, and F. Cheng, "Profile hmms for skeleton-based human action recognition," *Signal Processing: Image Communication*, vol. 42, pp. 109–119, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0923596516000138

[86] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 588–595.

[87] R. Chaudhry, F. Ofli, G. Kurillo, R. Bajcsy, and R. Vidal, "Bio-inspired dynamic 3d discriminative skeletal features for human action recognition," in *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 471–478.

[88] L. Xia, C.-C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2012, pp. 20–27.

[89] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1290–1297.

[90] J. Wang, Z. Liu, and Y. Wu, "Human action recognition with depth cameras," in *SpringerBriefs in Computer Science*, 2014.

[91] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[92] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[93] U. Buchler, B. Brattoli, and B. Ommer, "Improving spatiotemporal self-supervision by deep reinforcement learning," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 770–786.

[94] J. Wang, J. Jiao, L. Bao, S. He, Y. Liu, and W. Liu, "Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance

statistics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4006–4015.

[95] X. Long, C. Gan, G. De Melo, J. Wu, X. Liu, and S. Wen, "Attention clusters: Purely attention based local feature integration for video classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7834–7843.

[96] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 305–321.

[97] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6836–6846.

[98] L. Wang, W. Li, W. Li, and L. Van Gool, "Appearance-and-relation networks for video classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1430–1439.

[99] D. Tran, H. Wang, L. Torresani, and M. Feiszli, "Video classification with channel-separated convolutional networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5552–5561.

[100] K. Cao, J. Ji, Z. Cao, C.-Y. Chang, and J. C. Niebles, "Few-shot video classification via temporal alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 618–10 627.

[101] I. J. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016, http://www.deeplearningbook.org.

[102] R. D. Reed and R. J. Marks, *Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks*. MIT Press, 1999.

[103] S. Bozinovski, "Reminder of the first paper on transfer learning in neural networks, 1976," *Informatica*, vol. 44, no. 3, 2020.

[104] European Commission, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)," 2016. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj

[105] S. Wu, B. E. Moore, and M. Shah, "Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2054–2060.

[106] B. Antić and B. Ommer, "Video parsing for abnormality detection," in *2011 International Conference on Computer Vision*, 2011, pp. 2415–2422.

[107] B. Zhao, L. Fei-Fei, and E. P. Xing, "Online detection of unusual events in videos via dynamic sparse coding," in *CVPR 2011*. IEEE, 2011, pp. 3313–3320.

[108] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1446–1453.

[109] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 1, pp. 18–32, 2013.

[110] A. Del Giorno, J. A. Bagnell, and M. Hebert, "A discriminative framework for anomaly detection in large videos," in *European conference on computer vision*. Springer, 2016, pp. 334–349.

[111] R. Bensch, N. Scherf, J. Huisken, T. Brox, and O. Ronneberger, "Spatiotemporal deformable prototypes for motion anomaly detection," *International Journal of Computer Vision*, vol. 122, no. 3, pp. 502–523, 2017.

[112] S. Wu, B. E. Moore, and M. Shah, "Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes," in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 2054–2060.

[113] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 733–742.

[114] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked rnn framework," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 341–349.

[115] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3551–3558.

[116] D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe, "Learning deep representations of appearance and motion for anomalous event detection," *arXiv preprint arXiv:1510.01553*, 2015.

[117] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *CVPR 2011*. IEEE, 2011, pp. 3449–3456.

[118] H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 372–14 381.

[119] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," in *International symposium on neural networks*. Springer, 2017, pp. 189–196.

[120] S. Szymanowicz, J. Charles, and R. Cipolla, "Discrete neural representations for explainable anomaly detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 148–156.

[121] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe, "Abnormal event detection in videos using generative adversarial nets," in *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017, pp. 1577–1581.

[122] M. Ravanbakhsh, E. Sangineto, M. Nabi, and N. Sebe, "Training adversarial discriminators for cross-channel abnormal event detection in crowds," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1896–1904.

[123] L. Wang, F. Zhou, Z. Li, W. Zuo, and H. Tan, "Abnormal event detection in videos using hybrid spatio-temporal autoencoder," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 2276–2280.

[124] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," *arXiv preprint arXiv:1511.05440*, 2015.

[125] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection–a new baseline," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6536–6545.

[126] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convo-

lutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2758–2766.

[127] B. Chen, W. Wang, and J. Wang, "Video imagination from a single image with transformation generation," in *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, 2017, pp. 358–366.

[128] J. Van Amersfoort, A. Kannan, M. Ranzato, A. Szlam, D. Tran, and S. Chintala, "Transformation-based models of video sequences," *arXiv preprint arXiv:1701.08435*, 2017.

[129] C. Vondrick and A. Torralba, "Generating the future with adversarial transformers," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1020–1028.

[130] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee, "Decomposing motion and content for natural video sequence prediction," *arXiv preprint arXiv:1706.08033*, 2017.

[131] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6479–6488.

[132] Y. Zhu and S. Newsam, "Motion-aware feature for improved video anomaly detection," *arXiv preprint arXiv:1907.10211*, 2019.

[133] C. He, J. Shao, and J. Sun, "An anomaly-introduced learning method for abnormal event detection," *Multimedia Tools and Applications*, vol. 77, no. 22, pp. 29 573–29 588, 2018.

[134] F. Herrera, S. Ventura, R. Bello, C. Cornelis, A. Zafra, Dánel Sánchez-Tarragó, and S. Vluymans, *Multiple Instance Learning: Foundations and Algorithms*. Springer, 2016.

[135] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[136] P. Weinzaepfel, X. Martin, and C. Schmid, "Towards weakly-supervised action localization," *arXiv preprint arXiv:1605.05197*, vol. 2, no. 1, 2016.

[137] W. S. Noble, "What is a support vector machine?" *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[138] Y. Yan, C. Xu, D. Cai, and J. J. Corso, "Weakly supervised actor-action segmentation via robust multi-task ranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1298–1307.

[139] A. Arnab, C. Sun, A. Nagrani, and C. Schmid, "Uncertainty-aware weakly supervised action detection from untrimmed videos," in *European Conference on Computer Vision*. Springer, 2020, pp. 751–768.

[140] P. Mettes, C. G. Snoek, and S.-F. Chang, "Localizing actions from video labels and pseudo-annotations," *arXiv preprint arXiv:1707.09143*, 2017.

[141] L. Xu, C. Gong, J. Yang, Q. Wu, and L. Yao, "Violent video detection based on mosift feature and sparse coding," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 3538–3542.

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

[143] I. Serrano, O. Deniz, J. L. Espinosa-Aranda, and G. Bueno, "Fight recognition in video using hough forests and 2d convolutional neural network," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 4787–4797, 2018.

[144] S. Sudhakaran and O. Lanz, "Learning to detect violent videos using convolutional long short-term memory," in *2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS)*. IEEE, 2017, pp. 1–6.

[145] N. AlDahoul, H. A. Karim, R. Datta, S. Gupta, K. Agrawal, and A. Albunni, "Convolutional neural network-long short term memory based iot node for violence detection," in *2021 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*. IEEE, 2021, pp. 1–6.

[146] F. U. M. Ullah, K. Muhammad, I. U. Haq, N. Khan, A. A. Heidari, S. W. Baik, and V. H. C. de Albuquerque, "Ai-assisted edge vision for violence detection in iot-based industrial surveillance networks," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 8, pp. 5359–5370, 2021.

[147] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[148] R. Vijeikis, V. Raudonis, and G. Dervinis, "Efficient violence detection in surveillance," *Sensors*, vol. 22, no. 6, p. 2216, 2022.

[149] Ş. Aktı, G. A. Tataroğlu, and H. K. Ekenel, "Vision-based fight detection from surveillance cameras," in *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE, 2019, pp. 1–6.

[150] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[151] S. U. Khan, I. U. Haq, S. Rho, S. W. Baik, and M. Y. Lee, "Cover the violence: A novel deep-learning-based approach towards violence-detection in movies," *Applied Sciences*, vol. 9, no. 22, p. 4963, 2019.

[152] 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 preprint arXiv:1704.04861*, 2017.

[153] J. Li, X. Jiang, T. Sun, and K. Xu, "Efficient violence detection using 3d convolutional neural networks," in *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2019, pp. 1–8.

[154] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[155] F. J. Rendón-Segador, J. A. Álvarez-García, F. Enríquez, and O. Deniz, "Violencenet: Dense multi-head self-attention with bidirectional convolutional lstm for detecting violence," *Electronics*, vol. 10, no. 13, p. 1601, 2021.

[156] A. Diba, M. Fayyaz, V. Sharma, A. H. Karami, M. M. Arzani, R. Yousefzadeh, and L. Van Gool, "Temporal 3d convnets: New architecture and transfer learning for video classification," *arXiv preprint arXiv:1711.08200*, 2017.

[157] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[158] Z. Dong, J. Qin, and Y. Wang, "Multi-stream deep networks for person to person violence detection in videos," in *Chinese Conference on Pattern Recognition*. Springer, 2016, pp. 517–531.

[159] Y. Su, G. Lin, J. Zhu, and Q. Wu, "Human interaction learning on 3d skeleton point clouds for video violence recognition," in *European Conference on Computer Vision*. Springer, 2020, pp. 74–90.

[160] G. Mu, H. Cao, and Q. Jin, "Violent scene detection using convolutional neural networks and deep audio features," in *Chinese conference on pattern recognition*. Springer, 2016, pp. 451–463.

[161] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[162] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," http://www.image-net.org/challenges/LSVRC/2009/, 2009.

[163] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[164] J. Carreira and A. Zisserman, "A short introduction to the kinetics human action video dataset," *arXiv preprint arXiv:1906.05408*, 2019.

[165] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4724–4733.

[166] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546–6555.

[167] O. Kopuklu, N. Kose, A. Gunduz, and G. Rigoll, "Resource efficient 3d convolutional neural networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[168] C. Feichtenhofer, "X3d: Expanding architectures for efficient video recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 203–213.

[169] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[170] D. Dwibedi, I. Misra, and M. Hebert, "Cut, paste and learn: Surprisingly easy synthesis for instance detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1301–1310.

[171] G. Georgakis, A. Mousavian, A. C. Berg, and J. Kosecka, "Synthesizing training data for object detection in indoor scenes," *arXiv preprint arXiv:1702.07836*, 2017.

[172] J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Boochoon, and S. Birchfield, "Training deep networks with synthetic data: Bridging the reality gap by domain randomization," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 969–977.

[173] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, "Learning from synthetic humans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 109–117.

[174] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *International conference on Computer analysis of images and patterns*. Springer, 2011, pp. 332–339.

[175] M. M. Soliman, M. H. Kamal, M. A. E.-M. Nashed, Y. M. Mostafa, B. S. Chawky, and D. Khattab, "Violence recognition from videos using deep learning techniques," in *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)*. IEEE, 2019, pp. 80–85.

[176] M. Cheng, K. Cai, and M. Li, "Rwf-2000: an open large scale video database for violence detection," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 4183–4190.

[177] P. Sernani, N. Falcionelli, S. Tomassini, P. Contardo, and A. F. Dragoni, "Deep learning for automatic violence detection: Tests on the airtlab dataset," *IEEE Access*, vol. 9, pp. 160 580–160 595, 2021.

[178] P. Wu, j. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang, "Not only look, but also listen: Learning multimodal violence detection under weak supervision," in *European Conference on Computer Vision (ECCV)*, 2020.

[179] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.

[180] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization." *Journal of machine learning research*, vol. 12, no. 7, 2011.

[181] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[182] O. I. per a la Normalització, *Accuracy (trueness and Precision) of Measurement Methods and Results*. International Organization for Standardization, 1994.

[183] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[184] D. Choqueluque-Roman and G. Camara-Chavez, "Weakly supervised violence detection in surveillance video," *Sensors*, vol. 22, no. 12, p. 4502, 2022.

[185] T. Aremu, L. Zhiyuan, and R. Alameeri, "Any object is a potential weapon! weaponized violence detection using salient image," *arXiv preprint arXiv:2207.12850*, 2022.

[186] A. Péterfalvi, "A nemzeti adatvédelmi és információszabadság hatóság állásfoglalása a blokklánc („blockchain") technológia adatvédelmi összefüggéseivel kapcsolatban," 2017.

[187] kormany.hu, "Zrínyi 2026 Armed Forces Development Programme," Available from https://bit.ly/3X1OfK5. Accessed April 23, 2020.

[188] L. Kovács, *A kibertér védelme*. Dialóg Campus Kiadó, 2018.

[189] U. G. C. S. Adviser, "Distributed Ledger Technology: beyond block chain," Available from https://bit.ly/2lpNHhD. Accessed April 23, 2020.

[190] C. Dmitrij, "Digitális képelemzés alapveto algoritmusai," *Budapest: ELTE*, 2015.

[191] R. Cohen, "Global Bitcoin Computing Power Now 256 Times Faster Than Top 500 Supercomputers, Combined! Forbes, 2013." Available from https://bit.ly/3JsNmXw. Accessed April 23, 2020.

[192] O. Williams-Grut, "The electricity used to mine bitcoin this year is bigger than the annual usage of 159 countries," *Business Insider*, vol. 27, 2017.

[193] autoszektor.hu, "Mintegy 40 milliárd forintból épül járműipari tesztpálya Zalaegerszegen," Available from https://bit.ly/3RCyA2B. Accessed on 15 January 2019.

[194] J. J. Garstka, "Network-centric warfare offers warfighting advantage," *SIGNAL-FALLS CHURCH VIRGINIA THEN FAIRFAX–*, vol. 57, no. 9, pp. 58–60, 2003.

[195] L. Cuen, "Most Crypto Exchanges Still Don't Have Clear KYC Policies: report," CoinDesk, 2019. https://bit.ly/3XZOu9l.

[196] D. Csetverikov, "Basic algorithms for digital image analysis," *Course, Institute of Informatics, Eotvos Lorand University, visual. ipan. sztaki. hu*, 2003.

[197] A. Online, "Automotive test track to be built in Zalaegerszeg with HUF 40 billion," Available from https://bit.ly/3X8jFOI. Accessed on 15 January 2019.

[198] A. Pinna and W. Ruttenberg, "Distributed ledger technologies in securities post-trading revolution or evolution?" *ECB Occasional Paper*, no. 172, 2016.

[199] D. Tapscott and A. Tapscott, *Blockchain revolution: how the technology behind bitcoin is changing money, business, and the world*. Penguin, 2016.

[200] Perera and Wimal, ""Understanding Blockchain – How it works," Available from https://medium.com/the-capital/understanding-blockchain-how-it-works-5772e29421b8, visited on 28/04/2020.

[201] G. P. Dwyer, "The economics of bitcoin and other private digital currencies," *MPRA Paper*, vol. 57360, no. 8, 2014.

[202] L. Kovács, "National cybersecurity strategy framework," *Academic and Applied Research in Military and Public Management Science*, vol. 18, no. 2, pp. 65–76, 2019.

[203] A. McAbee, M. Tummala, and J. McEachen, "Military intelligence applications for blockchain technology," 2019.

[204] Greenspan, G., ""Avoiding the Pointless Blockchain Project," Available from https://www.multichain.com/blog/2015/11/avoiding-pointless-blockchain-project/, visited on 28/04/2020.

[205] D. Birch, R. G. Brown, and S. Parulava, "Towards ambient accountability in financial services: Shared ledgers, translucent transactions and the technological legacy of the great financial crisis," *Journal of Payments Strategy & Systems*, vol. 10, no. 2, pp. 118–131, 2016.

[206] Meunier, S., ""When do you need blockchain? Decision models," Available from https://medium.com/@sbmeunier/when-do-you-need-blockchain-decision-models-a5c40e7c9ba1, visited on 28/04/2020.

[207] A. Lewis, "Avoiding blockchain for blockchain's sake: Three real use case criteria," *Bits on Blocks*, 2017.

[208] M. E. Peck, "Blockchain world-do you need a blockchain? this chart will tell you if the technology can solve your problem," *IEEE Spectrum*, vol. 54, no. 10, pp. 38–60, 2017.

[209] K. Wüst and A. Gervais, "Do you need a blockchain?" in *2018 Crypto Valley Conference on Blockchain Technology (CVCBT)*. IEEE, 2018, pp. 45–54.

[210] K. Wüst and A. Gervais, "Blockchain beyond the hype: A practical framework for business leaders," in *2018 Crypto Valley Conference on Blockchain Technology (CVCBT)*. IEEE, 2018, pp. 45–54.

[211] MaidSafe, "Evolving Terminology with Evolved Technology: Decentralized versus Distributed," Available from https://bit.ly/2NuvScZ. Medium, 4 December 2015. Accessed on 15 January 2019.

[212] Naval Technology. n.d, "Arleigh Burke-Class (Aegis) Destroyer," Available from https://www.naval-technology.com/projects/burke/. Accessed April 20, 2020.

[213] S. Babones, "Smart 'Blockchain Battleships' Are Right Around the Corner," Available from https://nationalinterest.org/feature/smart-battleships-are-right-around-the-corner-25872. Accessed May 17, 2018.

[214] C. Thatcher, "Technology's dilemmas: Are we wired to respond," 2015.

[215] A. A. Malik, A. Mahboob, A. Khan, and J. Zubairi, "Application of cyber security in emerging c4isr systems," in *Cyber Security Standards, Practices and Industrial Applications: Systems and Methodologies*. IGI Global, 2012, pp. 223–258.

[216] S. Mire, "Blockchain for military defense: 7 possible use cases," Available from https://www.disruptordaily.com/blockchain-use-cases-military-defense/, posted on 09/11/2018.

[217] D. Axe, "Failure to Communicate: Inside the Army's Doomed Quest for the 'Perfect' Radio," Available from https://bit.ly/3JzuhTZ, posted on January 10, 2012.

[218] D. Lindorff, "Exclusive: The Pentagon's Massive Accounting Fraud Exposed," The Nation, 2019. Available from https://www.thenation.com/article/archive/pentagon-audit-budget-fraud/.

[219] R. Wyden, "DISA STARTTLS Letter," March 22 2017. Available from https://www.documentcloud.org/documents/3527403-Ron-Wyden-DISA-STARTTLS-Letter-March-22.html.

[220] M. Swan, *Blockchain: Blueprint for a new economy*. " O'Reilly Media, Inc.", 2015.

[221] Z. Haig, "Connections between Cyber Warfare and Information Operations," AARMS 8, no. 2: 329–337. Available from http://m.ludita.uni-nke.hu/repozitorium/bitstream/handle/11410/1900/13haig.pdf.

[222] U. S. D. of Defense Office of Prepublication and S. Review, "DoD Digital Modernization Strategy," Available from https://media.defense.gov/2019/Jul/12/2002156622/-1/-1/1/DOD-DIGITAL-MODERNIZATION-STRATEGY-2019.PDF. Accessed April 17, 2020.

[223] N. D. A. A. for Fiscal Year 2018, "H.R. 2810, 115th Cong (2017-2018)."

[224] J. L. McCarter, "DON Innovator Embraces a New Disruptive Technology: Blockchain," SECNAV, United States Navy. Available from https://www.secnav.navy.mil/innovation/Pages/2017/06/BlockChain.aspx. Accessed April 17, 2020.

[225] A. R. Nassar and E. Reutzel, "A proposed digital thread for additive manufacturing," in *2013 International Solid Freeform Fabrication Symposium*. University of Texas at Austin, 2013.

[226] L. Ferdinando, "'Terabyte of Death' Cyberattack Against DoD Looms, DISA Director Warns," Available from https://bit.ly/3WEPwGH. Accessed April 20, 2020.

[227] G. P. Release, "Galois and Guardtime Federal Awarded $1.8 Million DARPA Contract to Formally Verify Blockchain-Based Integrity Monitoring System," Available from https://galois.com/news/galois-guardtime-formal-verification/. Accessed April 20, 2020.

[228] T. Bretl, L. Righetti, and R. Madhavan, "Epstein, project maven, and some reasons to think about where we get our funding [ethical, legal, and societal issues]," *IEEE Robotics & Automation Magazine*, vol. 26, no. 4, pp. 8–13, 2019.

[229] P. Crofts and H. van Rijswijk, "Negotiating'evil': Google, project maven and the corporate form," *Law, Tech. & Hum.*, vol. 2, p. 75, 2020.

[230] G. C. Allen, "Project Maven Brings AI to the Fight Against ISIS," Available from https://thebulletin.org/2017/12/project-maven-brings-ai-to-the-fight-against-isis/. Dec, 21, 2017.

[231] J. Reed, "Israel's Killer Robot Carse," Foreign Policy, November 20, 2012. Available from https://foreignpolicy.com/2012/11/20/israels-killer-robot-cars/.

[232] P. Scharre, "Killer Robots and Autonomous Weapons With Paul Scharre," Podcast, June 1, 2018. Available from https://www.cfr.org/podcasts/killer-robots-and-autonomous-weapons-paul-scharre.

[233] O. Pawlyk, "If It's Not Ethical, They Won't Field It: Pentagon Release New A.I. Guidelines," Available from https://bit.ly/3XYeiT5. Accessed April 23, 2020.

[234] Interpol, "Artificial Intelligence and Robotics for Law Enforcement," Available from https://bit.ly/3wZU4NB. Accessed on 15 January 2019.

[235] B. Lilly and S. Lilly, "Weaponising blockchain: Military applications of blockchain technology in the us, china and russia," *The RUSI Journal*, vol. 166, no. 3, pp. 46–56, 2021.

[236] Z. Jobbágy, "Innovation methodologies for defence challenges: On design thinking and organic approaches," *Honvédségi Szemle–Hungarian Defence Review*, vol. 148, no. 2, pp. 50–64, 2020.

[237] O. Balogh, "The hungarian governmental defence and military development program "zrínyi 2026" as a response to the new threats in the region," *Studia Bezpieczeństwa Narodowego*, vol. 9, 2019.

[238] "Inauguration of the Cyber Training Centre of the Hungarian Defence Forces," Available from https://bit.ly/3HB2lNV. Accessed April 23, 2020.

[239] T. K. Csák, "The past, present, place, role and future challenges of military research and development in the development of the Hungarian armed forces in the light of the Zrínyi 2026 Defence and Military Development Programme," Honvédségi Szemle online, 2019/3. pp. 125-139., 2019.

[240] L. Sticz, "The role of the defence industry in the process of military capability development in the light of the privatisation of the defence companies," Military Engineer, 2009/3. pp. pp. 375 - 388.

[241] J. Pálinkás, "National interest in the era of global challenges," National Interest, New Stream, 2015/11-12, No. 93.

[242] S. Munk, "Some questions of the concept of cyberspace, Hadtudomány," Journal of the Hungarian Military Science Society 28: (1) pp. 113-131.

[243] L. Kovács, "National cyber security as the cornerstone of national security," *Land Forces Academy Review*, vol. 23, no. 2, pp. 113–120, 2018.

[244] N. A. T. O. (NATO), "Cyber Defence," Available from https://www.nato.int/cps/en/natohq/topics_78170.htm. Accessed March 17, 2020.

[245] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *2012 IEEE computer society conference on computer vision and pattern recognition workshops*. IEEE, 2012, pp. 28–35.

[246] A. Patron-Perez, M. Marszalek, I. Reid, and A. Zisserman, "Structured learning of human interactions in tv shows," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 12, pp. 2441–2453, 2012.

[247] J. Zhao, R. Han, Y. Gan, L. Wan, W. Feng, and S. Wang, "Human identification and interaction detection in cross-view multi-person videos with wearable cameras," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2608–2616.

[248] M. Ehsanpour, A. Abedin, F. Saleh, J. Shi, I. Reid, and H. Rezatofighi, "Joint learning of social groups, individuals action and sub-group activities in videos," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. Springer, 2020, pp. 177–195.

[249] M. S. Ibrahim and G. Mori, "Hierarchical relational networks for group activity recognition and retrieval," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 721–736.

[250] J. Wu, L. Wang, L. Wang, J. Guo, and G. Wu, "Learning actor relation graphs for group activity recognition," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2019, pp. 9964–9974.

[251] K. László and K. Csaba, "Digitális mohács 2.0: kibertámadások és kibervédelem a szakértők szerint," *Nemzet és Biztonság–Biztonságpolitikai Szemle*, vol. 10, no. 1, pp. 3–16, 2017.

[252] I. Porkoláb and I. Négyesi, "Research on the potential of Artificial Intelligence in the Armed Forces," Honvédségi Szemle, 2019/5, 6.

[253] S. Berta, "Maven project - Google is easily replaced," Sg.hu, 2018. https://bit.ly/3jdtixR.

[254] DEPUTY SECRETARY OF DEFENSE, Washington DC, "Maven project - Establishment of an Algorithmic Warfare Cross-Functional Team," Available from https://bit.ly/3DlTugx. Accessed April 23, 2020.

[255] N. Imre, "The vision of portable information technology equipment, Military Engineer," Military Engineer, Vol. 3, 2008/4, pp. 173-179.

[256] N. Scientist, "AI can predict if you'll die soon – but we've no idea how it works," Available from https://bit.ly/3H7jXQ2. Accessed April 23, 2020.

[257] N. Technology, "Arleigh Burke-Class (Aegis) Destroyer," www.naval-technology.com/projects/burke/.