Two-Stage Violence Detection Using ViTPose and Classification Models at Smart Airports

Iván López Torrecillas, İrem Üstek, Jay Desai, Jinjie Wang, Sofiane Abadou, Sandhya Rani Kasthuri, Quentin Fever, Yang Xing
Cranfield University, Bedford, United Kingdom

***Abstract*—This paper introduces an innovative violence detection framework tailored to the unique requirements of smart airports, where prompt responses to violent situations are crucial. The proposed framework harnesses the power of ViTPose for human pose estimation and employs a CNN-BiLSTM network to analyse spatial and temporal information within keypoints sequences, enabling the accurate classification of violent behaviour in real-time. Seamlessly integrated within the Saab's SAFE (Situational Awareness for Enhanced Security) framework, the solution underwent integrated testing to ensure robust performance in real-world scenarios. The AIRTLab dataset, characterized by its high video quality and relevance to surveillance scenarios, is utilized in this study to enhance the model's accuracy and mitigate false positives. As airports face increased foot traffic in the post-pandemic era, the implementation of AI-driven violence detection systems, such as the one proposed, is paramount for improving security, expediting response times, and promoting data-informed decision-making. The implementation of this framework not only diminishes the probability of violent events but also assists surveillance teams in effectively addressing potential threats, ultimately fostering a more secure and protected aviation sector. Codes are available at: https://github.com/Asami-1/GDP.**

Keywords—Violence detection, smart airport, integration, aviation sector, ViTPose, pose estimation, CNN-BiLSTM

I. Introduction

The aviation industry is highly sensitive to safety and timeliness, and airports face unique challenges in the violent events due to security and customs barriers that can complicate evacuations. As flight levels recover following the pandemic, airports are experiencing a surge in foot traffic [1] further heightening the aviation sector's sensitivity to violent events. While surveillance can assist in investigating violent events, the low proportion of abnormal frames in all video frames makes it particularly challenging to detect violent events manually. Smart airports are implementing advanced technologies to enhance operational efficiency and intelligence [2]. AI-based violence detection systems can identify threats in real-time, accelerate response times to emergencies, and facilitate data-driven decision-making regarding resource allocation, which contribute to reducing labor costs and preventing violent incidents in smart airport [3, 4].

The goal of violence detection is to automatically and accurately identify violent events [3]. However, the subjectivity of violence's definition poses a challenge in translating the concept into mathematical expressions or computer-understandable language. Past research has used feature descriptors based on low-level features such as gradients, optical flows, and intensities [4]. The accuracy of these methods depends on the selected feature descriptor and researchers' understanding of violence. In contrast, image-based deep learning methods can automatically learn the rules of violent behavior to adapt to different application scenarios and identify more complex violent behavior. Additionally, deep learning based techniques exhibit good accuracy, real-time performance, scalability, and transferability, which are advantageous for the smooth deployment and end-to-end optimization of violence detection systems in intelligent airports [5, 6].

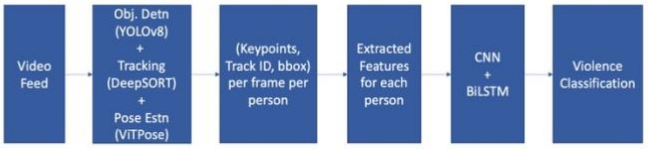

Video Feed
Obj. Detn (YOLOv8) + Tracking (DeepSORT) + Pose Estn (ViTPose)
(Keypoints, Track ID, bbox) per frame per person
Extracted Features for each person
CNN + BiLSTM
Violence Classification

Fig. 1: AI Pipeline

Therefore, to enhance airport security, this study aims to develop a two-stage violence detection framework to estimate human posture and detect violent behavior in real-time surveillance videos. This framework uses the pre-trained ViTPose model to detect human posture in each frame, preprocesses and extracts features from keypoints information, and then inputs these features into the CNN-BiLSTM network. This network further extracts features from keypoints information and learns the temporal information in keypoints sequences for violence detection. By identifying violence in real-time, this framework can reduce the likelihood of airport violence incidents and assist airport surveillance teams in making effective responses when such events occur.

The main contribution of this study is the use of a keypoint-based approach for violence classification, utilizing the state-of-the-art ViTPose model for pose estimation instead of directly extracting features from images (which may be high-dimensional and abstract). This approach simplifies feature extraction, reduces the sample size and computational resources required for training and deployment, and accelerates model training, inference, and the entire pipeline's execution. Additionally, the use of keypoints reduces data size, decreases storage burden, improves operation speed, and facilitates implementation and deployment optimization. By focusing solely on human posture and movement information, this method minimizes interference caused by changes in image backgrounds and human body shape differences, while also reducing the influence of differences between datasets [7]. Overall, this study not only designs a real-time violence detection system suitable for smart airports but also provides a more efficient and effective novel framework for real-time activity recognition tasks using different or multiple datasets. The remainder of this paper is organized as follows. Section II discusses related work. Section III explains the overall architecture of the system, the datasets used, the relevant models and algorithms, and the deployment at the real-time application end. Section IV evaluates and discusses the

{I.LopezTorrecillas.352, Irem.Ustek.197, Jay.Desai.336, Jinjie.Wang.647 Sofiane.Abadou.305, Sandhyarani.Kasthuri.167, Quentin.Fever.387, Yang.X}@cranfield.ac.uk

experimental results. Section V and Section VI respectively discuss future work and summarize the study.

II. RELATED WORK

A. *Pose Estimation*

Human pose estimation involves detecting the joints or keypoints of an individual from 2D or 3D images and videos [8]. There are classic methods for human pose estimation like HRNet, DCPose, and AlphaPose. HRNet can estimate high-resolution 2D or 3D poses but requires a lot of resources [9]. DCPose can detect and track human poses in real-time with high efficiency and robustness [10]. AlphaPose combines CNN and a bottom-up method to estimate the key points of multiple people in real-time [11]. Transformers have recently been applied to pose estimation in computer vision, with ViTPose+ being the current state-of-the-art method. It uses a simple visual transformer to extract features from input images and introduces expert blending in the backbone network to improve performance. ViTPose+ has achieved new records on four challenging benchmarks while maintaining the same inference speed as ViTPose, which uses single-instance detection and has lower computational costs [12].

B. *Violence Detection*

Violence detection is an application of Human Activity Recognition (HAR) in computer vision, which aims to accurately identify human actions from sensor data by analyzing their spatiotemporal features. This study focuses on using human body keypoints to train a classifier, which acts as an attention mechanism [6]. Deep learning based models offer promising solutions because they can automatically learn and extract spatiotemporal features from input. Two main approaches are commonly used for HAR: end-to-end 3D CNN and the two-stream architecture using RNNs [13].

The 3D CNN technique involves using 3D convolutional kernels to analyze a video's spatiotemporal information, enabling the model to learn features from it. Previous research has successfully used this method as an end-to-end architecture [14, 15]. However, 3D CNN requires a significant amount of computing resources and time to train and infer, making it unsuitable for some real-time applications. The CNN+RNN approach involves using separate neural networks for spatial and temporal information processing. RNNs can model temporal dynamics and dependencies between frames, but may suffer from the vanishing gradient problem, which can be addressed using LSTMs. Previous models achieved high accuracy using CNNs as spatial feature extractors and LSTMs for temporal information processing [16, 17].

III. METHODOLOGY

A. *Datasets*

Various datasets exist for pose estimation and violence detection individually. However, no common dataset is available, including ground truth for pose estimation and violence detection. Hence, we considered various separate datasets for each task and then combined them in a sequence to achieve the ultimate objective of violence detection using pose estimation.

1) *Dataset Selection*

a. *Datasets for Pose Estimation*

There are various datasets available for training pose estimation models. We are using a pre-trained model (ViTPose), demonstrating state-of-the-art results on datasets such as COCO, AIC (AI Challenger), MPII, CrowdPose, and WholeBody for human pose estimation.

b. *Datasets for Violence Detection*

We analysed widely used RWF-2000, UCF-Crime, and also AIRTLab dataset. The RWF-2000 dataset comprises 2,000 real-world video clips, evenly split between violent and non-violent content, sourced from movies, surveillance cameras, and social media [18]. The UCF-Crime dataset is a large-scale collection of real-world surveillance videos featuring 13 types of crime and regular activities, such as fighting, burglary, and vandalism [19]. Lastly, the AIRTLab dataset, created by the University of Rome III's Artificial Intelligence and Robotics Laboratory, consists of 350 MP4 video clips at 1920 x 1080 resolution and 30 fps, varying in duration from 2 to 14 seconds, designed specifically for automatic violence detection [20].

In AIRTLab Dataset, the clips are annotated as either "violent" or "non-violent" based on the presence or absence of physical aggression. The non-violent clips are specifically recorded to include behaviours (such as hugs, claps, and exulting) that can cause false positives due to the similarity of fast movements with violent behaviours for violence detection algorithms. The dataset is intended to train and test video violence detection algorithms and evaluate their robustness against false positives.

The AIRTLab dataset is characterized by its high video quality, an essential feature for accurate and reliable human pose estimation. Moreover, the dataset was collected using surveillance cameras, which makes it highly relevant to the real-world application scenarios of our project. This dataset has been selected to enhance the accuracy of our model by providing more precise keypoints data. Additionally, using this dataset helps mitigate false positives for fast movements, which is a frequent problem in human pose estimation. By utilizing the AIRTLab dataset, we aimed to improve the performance of our model and make it more effective in real-world scenarios. Due to these reasons, this dataset is selected.

2) *Dataset Preprocessing*

a. *Manual Labelling and Correction*

To achieve accurate violence recognition, a frame-level manual annotation was conducted on the dataset, categorizing individuals as neutral, aggressor, or victim. Noise-contributing boundary boxes, resulting from false detections by YOLO or discontinuous human behavior, were eliminated to improve training data quality. Misidentified boundary boxes by DeepSort were merged, and individuals' behavioral categories were labeled by evaluating their behavior in continuous frames. *Fig. 1* shows the output of our AI pipeline. *Fig. 3*, *Fig. 4* and *Fig. 5* show examples of manual labeling scenarios.

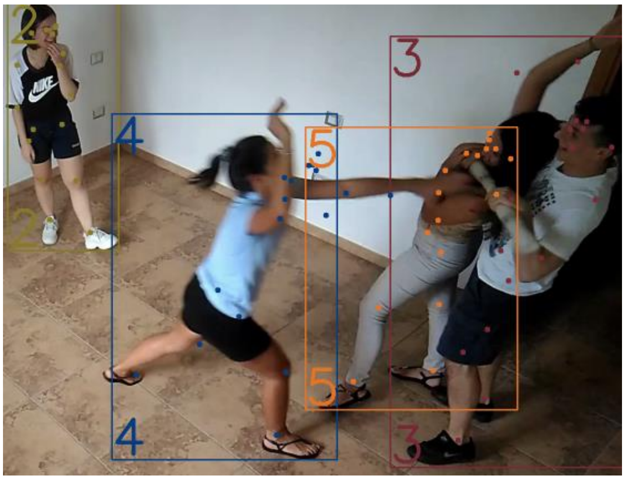


Fig. 2: An example frame after the three (YOLO, DeepSort and ViTPose) models have been applied

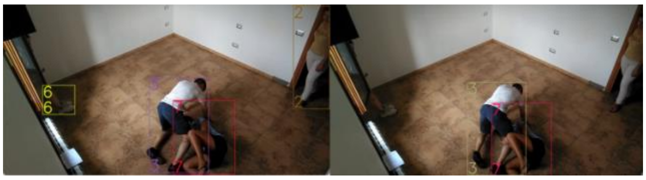


Fig. 3: An example of removing unnecessary track IDs with Python script (IDs 2 and 6 removed)

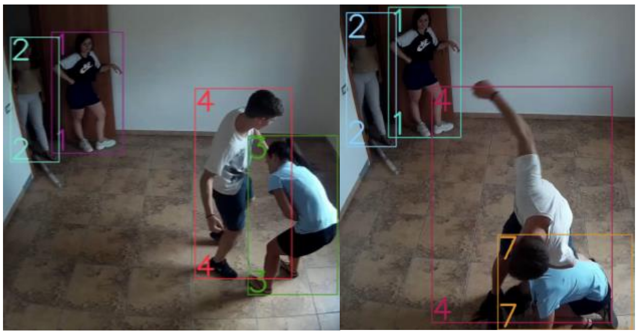


Fig. 4: An example frame having different IDs for the same person across frames

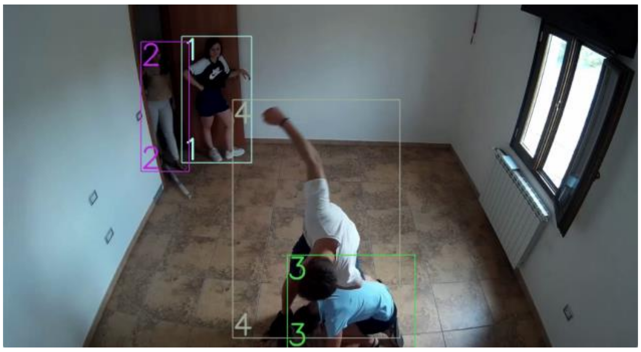


Fig. 5: An example of merging the IDs for the same person (IDs 3 and 7 are merged)

b. Feature Engineering

In this project, we used two feature engineering strategies from pose estimation for the classifier: distance calculation and angle calculation. Distance-based features detect violent behavior by assessing deviations from the body's normal state, such as leaning, tilting, or limb movements. Angle-based features utilize angles between body keypoints to provide information about body orientation and movement, aiding in violent and non-violent behavior classification. After comparing the two strategies, we found that distance-based approaches had better accuracy and selected models accordingly.

c. Extracting Each Individual

This project trains AI models using individual keypoints data to enhance robustness and generalization by capturing unique behavioral patterns while avoiding cross-interference. Despite requiring higher quality datasets and potentially affecting real-time operation speed, this approach is preferable given the project's focus on identifying aggressors and victims in violent events. Training with multiple individuals' keypoints data introduces additional challenges, such as increased model complexity, mutual interference, and potential information loss due to dimension reduction. Therefore, using individual body keypoints data for model training is adopted to meet project requirements and ensure higher recognition accuracy while maintaining a coherent methodology.

d. Data Augmentation

We compared AI model performance using two-time intervals: 1-second and 2-second intervals. The 2-second interval provided more information per data point but generated fewer data points, while the 1-second interval created more data points with less information. Using the Windows approach to analyze time series data, we broke the data into smaller windows to identify violent behavior patterns. Generating more data points with smaller window sizes enabled more frequent data sampling, providing a clearer view of the data and helping identify subtle behavioral changes. This approach improves the classifier's ability to accurately identify violent and non-violent behaviors, aiding in the development of effective interventions.

B. Tracking & Object Detection

In order to effectively carry out the pose estimation and classification, it is necessary to previously detect the bounding boxes of each person and track them through the video. While considering different methods and pre-trained models, it is important to obtain good accuracy while keeping inference time low, since the main objective is to deploy in a real-time application. In our work, two subsystems have been used separately: for the detection of people, it has been decided to use the smallest pretrained model of YOLOv8 (nano) that offers the fastest inference time with enough accuracy for this project; on the other hand, DeepSORT is used for tracking people across multiple frames. This method combines the classic Simple Online and Realtime Tracking (SORT) algorithm with a deep learning model which helps to reduce tracking errors due to missed detections or bounding box occlusion.

C. Pose Estimation by ViTPose

As in this project people body keypoints are used to identify and classify possible violent behaviors, one of the main systems in the AI engine pipeline is the pose estimator itself. ViTPose is an algorithm that combines the strengths of Vision Transformers (ViT) and pose estimation techniques to achieve high-precision human pose estimation in images or videos. Vision Transformers divide an input image into smaller patches and process them using a transformer architecture, while pose estimation focuses on identifying the spatial locations of key body parts in the images or videos. Once again, being inference time the deciding factor to choose a pretrained model, ViTPose+-S was implemented in the project.

D. Classification

As mentioned earlier, the output of the pose estimation along with the person bounding box track ID generated from previous subsystems, are preprocessed to extract the desired features (either distance or angle based) from each individual in a set duration sequence (one or two seconds long). These sequences are the input of the classification model.

In this project, four different model architectures were considered: on the one hand, Long Short-Term Memory (LSTM) networks which are Recurrent Neural Networks (RNN) specifically designed to address the vanishing gradient problem and effectively capture and store long-range dependencies within a data sequence; on the other hand, Bidirectional Long Short-Term Memory (BiLSTM) networks which are an extension of LSTMs that operate in two directions, allowing the model to learn from past and future contexts within a sequence; finally, a variation of the two previous networks mentioned having Convolutional Neural Network (CNN) layers as first layer of the architecture acting as feature extractors, capturing spatial information and local patterns within the input data. During training, CNN, LSTM and BiLSTM layers were followed by Dropout layers. Additionally, a custom callback early-stopped the training after 50 epochs with no improvement to avoid overfitting.

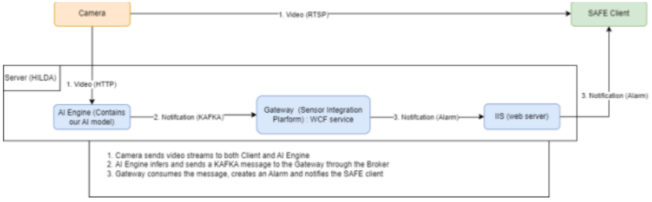
All four model architectures were trained in both distance-based and angle-based features and in 1-second and 2-second-long sequences. Furthermore, the models were hyperparameter tuned for different fixed values of Drop-out rates, Batch size, number of LSTM Layers in the architecture and Learning rate using SGD and Adam Optimizers.

E. Integration to SAFE

The work was conducted in collaboration with SAAB and DARTeC (Digital Aviation Research and Technology Centre) in order to integrate our AI model into one of their application. SAFE is an open-integration software platform for mission-critical operations. It is utilized in command-and-control rooms in various high-security areas such as airports, law enforcement operations centers, and maximum-security correctional facilities. The goal was integrating the developped AI model in the SAFE environment, allowing monitoring of its predictions in one place. Once the AI model has predicted conflict in the video feed, it sends a notification in the form of an alarm that would be received in the SAFE client, easing conflict detection for airport security

In order to do so, we first integrated our AI model into an AI Engine. Its role is to host the model and all the required dependencies, as well as to manage input acquirement and classification result forwarding. The engine was built using Flask, a python based lightweight and modular framework that allows to create applications interacting with AI technologies easily. We also built a docker container for the application to make it deployable regardless of the environment.

We then integrated this engine into the SAFE environment. To communicate with SAFE's application server, we used Kafka, a distributed streaming platform that offers high scalability, durability, and flexibility to build real-time streaming applications. The workflow is as follows as shown in *Fig. 6*.



1. Camera sends video streams to both Client and AI Engine
2. AI Engine infers and sends a KAFKA message to the Gateway through the Broker
3. Gateway consumes the message, creates an Alarm and notifies the SAFE client

Fig. 6: System Architecture

1) *Sending video feed to engine and client:* To perform Violence Detection, video feeds are captured from these cameras and given to the AI engine as well as the SAFE client terminal directly.

2) *AI engine inference:* AI engine receives the video feed from the camera system and performs Violence Detection on it. A block diagram for the same is given in *Fig. 1*.

3) *Notifying Gateway and SAFE client:* Gateway is used to consume the KAFKA message given by the AI engine and forward it to SAFE's client terminal for further display and alerts. All alerts for violence can be shown in real-time on this layout, and operator can then take appropriate action.

IV. EVALUATION

A. Classification

The study employed both angle-based and distance-based features derived from raw keypoint datasets with sequence lengths of 1-second and 2-seconds to train 2592 models (162 hyperparameter combinations x 4 different dataset features x 4 different model types). A custom call-back function was used to prevent overfitting and select the best performing models. Results were obtained in terms of average and maximum accuracy metrics and training time. It is observed that distance-based features dataset performed better compared to angle-based features dataset for both 1-sec and 2-secs of sequences, therefore, further analysis is only done through distance-based feature dataset.

B. Comparison of the Best Models

After conducting a comprehensive analysis of various models using multiple datasets (angle-based, distance-based, 1-second and 2-seconds of sequences) on 2592 models, several promising models were identified that meet specific evaluation criteria. The assessment included analyzing the models' train accuracy, test and validation accuracy, and examining the loss and accuracy plots to ensure that they are not overfit. In addition, the real-time performance of each model was evaluated using the AI Engine application to detect violent behaviour in camera footage. In general, it was observed that models trained on distance-based features performed better compared to angle-based features. It can be concluded that valuable information about violent behaviour can be extracted using distance features of body keypoints.

TABLE I. COMPARISON OF THE TWO CANDIDATE BEST MODELS FOR MULTICLASS CLASSIFICATION

Model	Dataset	Train Dataset Size	Train Accuracy	Test Accuracy
(A) CNN-BiLSTM	Distance-based & 1 second of seqs	(8048, 10, 24)	0.813	0.798
(B) CNN-BiLSTM	Distance-based & 2 seconds of seqs	(6558, 20, 24)	0.853	0.806

Model (B), trained on a distance-based dataset with 2-second sequences, achieved the highest train accuracy (0.853) and test accuracy (0.806) among several models. Model (A), trained on a standard 1-second dataset, performed closely to Model (B) in terms of test accuracy (TABLE I). However, in real-time classification, Model (A) was observed to outperform Model (B), which is important for the project's goal of real-time violence detection in live camera feeds (*Fig. 10*)

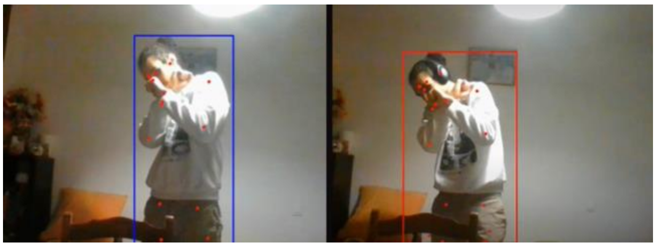


Fig. 7: Real-time performance comparison of Model (A) (on the left) and Model (B) (on the right) from the AI Engine (Blue bounding box: Victim, Red bounding box: Aggressor, Green bounding box: Neutral)

Model (B) struggles to consistently detect the person in *Fig. 7* as a Victim, unlike Model (A). This performance difference can be explained by dataset size and sequence length. Model (A) was trained on a larger dataset (8048 samples) and a shorter sequence length of 1 second, which may have provided more diverse and representative examples, leading to better generalization and sensitivity in capturing real-time violence cues. Therefore, Model (A) was better suited for real-time violence classification from live camera feed and was selected as the preferred choice for the final model.

C. *Final Model Results*

Based on the analysis of model results, the CNN_BiLSTM model utilizing a distance-based approach and a dataset with 1-second sequences has been identified as the optimal choice for multiclass classification. The model was trained using a training dataset size of (8048, 10, 24), a learning rate of 0.1, batch size of 64, dropout rate of 0.4, and the stochastic gradient descent (SGD) optimizer. The model achieved a train accuracy of 0.813 and a test accuracy of 0.798 at the best epoch (114).

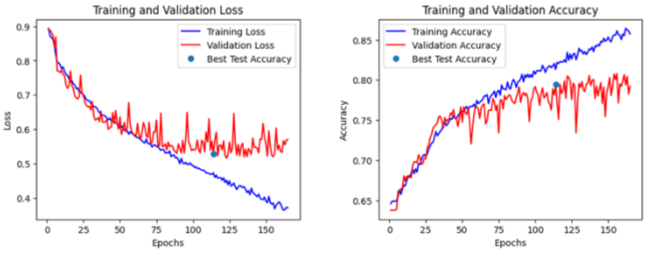


Fig. 8: Loss and accuracy plots of final model in train and validation datasets

TABLE II. CLASSIFICATION REPORT OF THE FINAL MODEL

	Precision	**Recall**	**F1-score**	**Support**
Neutral	0.86	0.91	0.89	1465
Aggressive	0.59	0.61	0.60	376
Victim	0.52	0.28	0.36	205

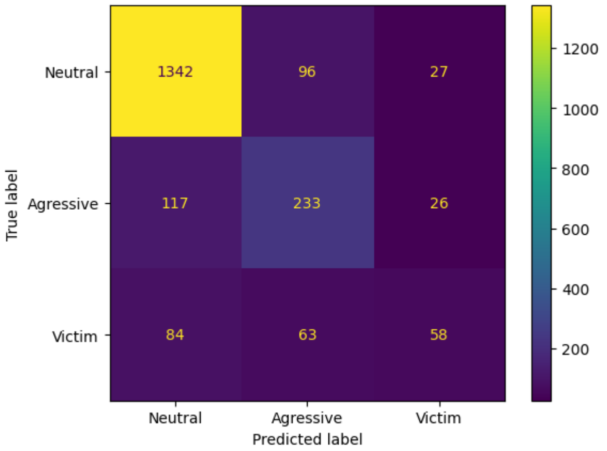


Fig. 9: Final model confusion matrix

The model's loss and accuracy plots indicate that it achieved its highest test accuracy around the 115th epoch while preventing overfitting using a custom callback (*Fig. 8*). The F1-score, which provides a balanced measure of both precision and recall, was used to evaluate the model's performance, and it showed that the 'Neutral' class had the highest F1-score of 0.89, while the 'Victim' class had the lowest F1-score of 0.36 (TABLE II). However, this could be attributed to the smaller number of instances in the 'Victim' class compared to other classes. Overall, this final model is utilized in the AI engine to detect violent behavior in real-time by classifying individuals as Aggressor, Victim, and Neutral.

D. *Integration to SAFE*

We successfully deployed our development model in HILDA, received frames from DARTeC's camera to AI engine for inference, and classified violent behavior in the video. *Fig. 10* shows the output of our model's classification on DARTeC's camera. Our model managed to classify agressors and victims adequately.

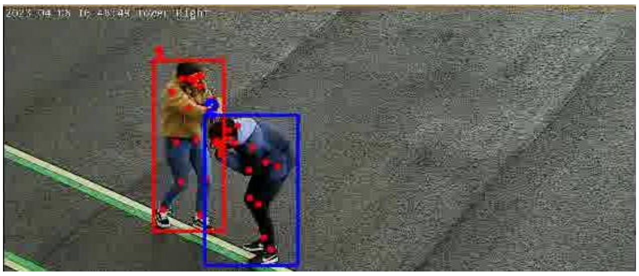
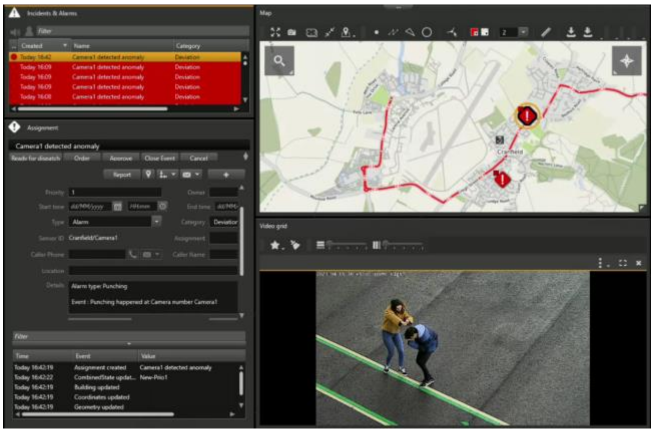


Fig. 10: Output of the model on DARTeC's camera feed

Fig. 11 shows SAFE's layout, we can see that our KAFKA messages were successfully received, alarms have been activated on the client's side.



Created	Name	Category
Today 16:42	Camera1 detected anomaly	Deviation
Today 16:09	Camera1 detected anomaly	Deviation
Today 16:09	Camera1 detected anomaly	Deviation
Today 16:09	Camera1 detected anomaly	Deviation
Today 16:08	Camera1 detected anomaly	Deviation

Assignement

Camera1 detected anomaly

Ready for dispatch Order Approve Close Event Cancel

Report

Priority: 1 Owner:

Start time: dd/MM/yyyy HH:mm End time: dd/MM/

Type: Alarm Category: Deviation

Sensor ID: Cranfield/Camera1 Assignment:

Caller Phone: Caller Name:

Location:

Details: Alarm type: Punching

Event : Punching happened at Camera number Camera1

Time	Event	Value
Today 16:42:19	Assignment created	Camera1 detected anomaly
Today 16:42:22	CombinedState updat...	New-Prio1
Today 16:42:19	Building updated	
Today 16:42:19	Coordinates updated	
Today 16:42:19	Geometry updated	

Fig. 11: SAFE layout during inference of the model

V. FUTURE WORK

The proposed activity recognition framework shows promising results, there are potential avenues for future work to improve the framework's performance and adaptability. One way to improve the classifier is to incorporate attention mechanisms. By integrating attention mechanisms into the classifier, the model can better capture temporal relationships between keypoints, potentially improving model performance. Additionally, attention mechanisms can help identify the most prominent parts of input data, reducing model reliance on less informative keypoints and mitigating the impact of noise. Another improvement would be to replace traditional AI models (CNN, LSTM, BiLSTM) networks with transformer models. Transformer models effectively handle long-range dependencies and parallel computation, making them suitable for tasks involving large amounts of input data. This could enhance the framework's ability to capture complex temporal patterns and relationships between keypoints, thus improving violence detection performance. Moreover, the parallelization capabilities of transformer models may speed up training and inference, making the proposed framework more scalable and suitable for real-time applications. Data diversity and augmentation are also essential to improve the framework's performance. The current study relies on a single dataset, limiting the diversity of violent situations and scene types. As the sample size is relatively small and the distribution of different types of behaviors is imbalanced, future work should explore using multiple datasets to balance behavior type distribution and apply data augmentation techniques to increase sample size and enrich train data. It is also important to consider camera calibration and normalization to ensure effective generalization across varying angles and scenes, making the framework more robust and capable of adapting to different camera angles and orientations, which is crucial for real-world applications.

VI. Conclusion

In conclusion, the implementation of a violence detection model employing a two-stage approach - pose estimation and violence classification - provides an effective solution for detecting violent behaviour in real-time. This system is particularly beneficial in high-security environments such as smart airports, where ensuring public safety is of utmost importance. The success of this approach is primarily attributed to the use of ViTPose for pose estimation, the CNN-BiLSTM model for violence classification, and the carefully curated AIRTLab dataset, which minimizes false positives.

To ensure the practicality and effectiveness of the system, end-to-end testing was conducted at the DARTeC building at Cranfield University, incorporating live feed from the camera and real-time alert notifications on the SAFE client. This comprehensive testing demonstrated the model's ability to detect violence accurately and efficiently for real-world environment.

Future research and development in this area should focus on refining the model further, trying out different model architectures and hyperparameters, and exploring the integration of additional modalities and datasets to enhance its capabilities and applicability.

References

[1] International Air Transport Association, " Air Passenger Numbers to Recover in 2024," IATA, March 1, 2022, https://www.iata.org/en/pressroom/2022-releases/2022-03-01-01/.

[2] K. Liu, L. Chen, and X. Liu, “Research on Application of Frontier Technologies at Smart Airport,” Communications in Computer and Information Science. Springer Singapore, pp. 319–330, 2021. doi: 10.1007/978-981-16-5943-0_26.

[3] B. Omarov, S. Narynov, Z. Zhumanov, A. Gumar, and M. Khassanova, “State-of-the-art violence detection techniques in video surveillance security systems: a systematic review,” PeerJ Computer Science, vol. 8. PeerJ, p. e920, Apr. 06, 2022. doi: 10.7717/peerj-cs.920.

[4] F. U. M. Ullah, M. S. Obaidat, A. Ullah, K. Muhammad, M. Hijji, and S. W. Baik, “A Comprehensive Review on Vision-Based Violence Detection in Surveillance Videos,” ACM Computing Surveys, vol. 55, no. 10. Association for Computing Machinery (ACM), pp. 1–44, Feb. 02, 2023. doi: 10.1145/3561971.

[5] 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. Institute of Electrical and Electronics Engineers (IEEE), pp. 160580–160595, 2021. doi: 10.1109/access.2021.3131315.

[6] J. Kim and D. Lee, “Activity Recognition with Combination of Deeply Learned Visual Attention and Pose Estimation,” Applied Sciences, vol. 11, no. 9. MDPI AG, p. 4153, May 01, 2021. doi: 10.3390/app11094153.

[7] L. Song, G. Yu, J. Yuan, and Z. Liu, “Human pose estimation and its application to action recognition: A survey,” Journal of Visual Communication and Image Representation, vol. 76. Elsevier BV, p. 103055, Apr. 2021. doi: 10.1016/j.jvcir.2021.103055.

[8] Viso.ai, "Pose Estimation: The Ultimate Overview," Viso.ai, accessed May 3, 2023, https://viso.ai/deep-learning/pose-estimation-ultimate-overview/.

[9] J. Wang et al., “Deep High-Resolution Representation Learning for Visual Recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 10. Institute of Electrical and Electronics Engineers (IEEE), pp. 3349–3364, Oct. 01, 2021. doi: 10.1109/tpami.2020.2983686.

[10] Z. Liu et al., “Deep Dual Consecutive Network for Human Pose Estimation.” arXiv, 2021. doi: 10.48550/ARXIV.2103.07254.

[11] H.-S. Fang et al., “AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time,” IEEE Transactions on Pattern Analysis and Machine Intelligence. Institute of Electrical and Electronics Engineers (IEEE), pp. 1–17, 2022. doi: 10.1109/tpami.2022.3222784.

[12] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, ‘ViTPose+: Vision Transformer Foundation Model for Generic Body Pose Estimation’. arXiv, 2022. doi: 10.48550/ARXIV.2212.04246.

[13] H. H. Pham, L. Khoudour, A. Crouzil, P. Zegers, and S. A. Velastin, ‘Video-based Human Action Recognition using Deep Learning: A Review’. arXiv, 2022. doi: 10.48550/ARXIV.2208.03775.

[14] S. Ji, W. Xu, M. Yang, and K. Yu, ‘3D Convolutional Neural Networks for Human Action Recognition’, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1. Institute of Electrical and Electronics Engineers (IEEE), pp. 221–231, Jan. 2013. doi: 10.1109/tpami.2012.59.

[15] S. N. Boualia and N. E. B. Amara, ‘3D CNN for Human Action Recognition’, 2021 18th International Multi-Conference on Systems, Signals &amp; Devices (SSD). IEEE, Mar. 22, 2021. doi: 10.1109/ssd52085.2021.9429429.

[16] A.-M. R. Abdali and R. F. Al-Tuma, ‘Robust Real-Time Violence Detection in Video Using CNN And LSTM’, 2019 2nd Scientific Conference of Computer Sciences (SCCS). IEEE, Mar. 2019. doi: 10.1109/sccs.2019.8852616.

[17] R. Nale, M. Sawarbandhe, N. Chegogoju, and V. Satpute, ‘Suspicious Human Activity Detection Using Pose Estimation and LSTM’, 2021 International Symposium of Asian Control Association on Intelligent Robotics and Industrial Automation (IRIA). IEEE, Sep. 20, 2021. doi: 10.1109/iria53009.2021.9588719.

[18] M. Cheng, K. Cai, and M. Li, ‘RWF-2000: An Open Large Scale Video Database for Violence Detection’, 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, Jan. 10, 2021. doi: 10.1109/icpr48806.2021.9412502.

[19] W. Sultani, C. Chen, and M. Shah, ‘Real-world Anomaly Detection in Surveillance Videos’. arXiv, 2018. doi: 10.48550/ARXIV.1801.04264.

[20] M. Bianculli et al., ‘A dataset for automatic violence detection in videos’, Data in Brief, vol. 33. Elsevier BV, p. 106587, Dec. 2020. doi: 10.1016/j.dib.2020.106587.