International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Video Based Violence Detection Using Deep Learning CNN-CHA-SPA Double Attention Mechanism with Mosaicking

V. Elakiya*[1], Dr. P. Aruna[2], Dr. N. Puviarasan[3], Dr. R. G. Suresh Kumar[4]

Submitted: 27/01/2024 **Revised**: 05/03/2024 **Accepted**: 13/03/2024

Abstract: Violence detection refers to the use of various technologies and methods to identify, keep track of, and react to instances of physical or verbal aggressiveness, threatening conduct, or violent acts. Security, public safety, and online content filtering are just a few areas where this use is vital. Due of the differences in the human body, it is challenging to capture more accurate and discriminative features for video-based violence detection. Automatically spotting aggressive behaviour in places with video surveillance, such train stations, gyms, and psychiatric facilities, is crucial. As a result, this research focuses on creating a violence prediction system with improved feature extraction and classification techniques while researching various and efficient feature extraction techniques. Constructing an improvised violence detection system has some difficulties. Deep neural self-attention and CNN feature extraction methods are used to determine if a video contains violent content or not in order to solve the aforementioned complexity, such as focusing on the types of attacks and improving the accuracy of violence detection. The Proposed Method CNN-CHA-SPA Double Attention Mechanism with CNN helps to extract the frames correctly and detect the video is Violent or not. Here, a cutting-edge deep learning approach using video mosaicking is suggested. The extracted images from the video are combined with these mosaic images in the preprocessing stage, which offer a more thorough perspective of the scene and which will aid in accurately extracting the feature and helps to obtain time consistent outcomes and on the other hand improve the performance of the algorithm. This proposed mechanism provides the accurate result compared to the other mechanisms available.

Keywords: *Feature extraction, CNN, Mosaicking, enhancement, attention technique, Violence detection***.**

1.Introduction:

In this adverse society there are a variety of hazards to be focused also the insufficient labor available there is a need to have a violence automated detection. As a part of society, facing a lot of Hazards is unavoidable and there exist an unavailability of persons to monitor such Problems. Therefore, there is a necessity for the automatic system for the detection of violence. Video surveillance phenomenon through the cameras is very cost- effective and it is more effective in ensuring the people safety. The crime rate has been significantly reduced to the deployment of surveillance cameras in public places. Hence this research focus on developing a violence detection with better feature extraction and classification approaches in studying them in elaborate. This current study emphasizes on feature extraction process combined with classification approaches to building improvised violence detection accuracy. There are certain challenges in establishing improvised violence detection. Due to the noisy images, huge data, and standards the conventional violence detection system execution is considered a trial. While identifying violence, image clarity and computational constraints are also other issues.

The video frames are first extracted with which the mosaicked frames that is obtained in the pre preprocessing phase are combined and the model is trained well. The detection is done with the test data set where when a video is given to test it states whether the given test video is violent or not. The method of stitching video frames together to formulate a complete understanding of the scene is called mosaicking. In addition to improve the mosaic image enhancement process before the stitching process is done. The resultant mosaic image is a compacted illustration of the video data is the resultant mosaic image. The mosaicking of the video block is frequently used in the video compression and analysis applications. This Mosaic images are also added to the extracted images which propose a more comprehensive view of the scene and which will help to extract the feature appropriately and do classification of the video with more accuracy.

2.Review of Literature

For the reason that of the reckless movement and the overlying elements from sealing, and chaotic backdrops, recognizing the violence in crowded scenarios is particularly difficult. With the use of clever methods, this work intends to put the upgraded model for detection of the violence into practise. The Motion Weber Local Descriptor (MoWLD), the Histogram of Oriented

[1] *Research Scholar, Department of Computer Science and Engineering, Annamalai University, Chidambaram, Tamil Nadu, India.*
[2] *Professor, Department of Computer science and Engineering, Annamalai University, Chidambaram, Tamil Nadu, India*
[3] *Professor and Head, Department of Computer and Information Science, Annamalai University, Chidambaram, Tamil Nadu, India.*
[4] *Professor, Department of Computer Science and Engineering, Rajiv Gandhi College of Engineering and Technology, Puducherry*
** Corresponding Author Email: elakiyaloganathan@gmail.com*

Gradients (HoG), and the Motion Boundary Scale Invariant Feature Transform (MoBSIFT) are employed in the extraction process of the feature. The best feature pick is further adopted. In order to find the best features for a multi-objective function, the Spider Monkey-Grasshopper Optimisation algorithm (SM-GOA) is used. After that, the Deep Neural Network (DNN), whose algorithm that is trained is strengthened by the same SM-GOA, classifies violent and nonviolent video frames[1]. The most important area of explore in the vision of the computer is behaviour recognition. The specific behaviour semantics may be automatically understood through the analysis of movement of the person and physique action inside scene frame. By shifting from multi-class behaviour identification to (VioBD) violence behaviour detection's two-class behavioural categorization, which identifies merely human behaviour [2]. The two main parts of the VioBD technique that have been mentioned are the classification of behaviour and the description of behaviour. In order to capture crucial behavioural, the focud is mainly done on the features of data from the video-based footage[3]

Different researchers had revealed their capacity for identifying abnormalities by advancing the algorithm in deep learning (DL) achievement rate in categorising of photos. In the meantime, extraction of the feature procedure is avoided and the straight data is handled, DL approaches do not require human intervention. However, because to the complicated background, blurriness, low resolution, scale fluctuations, illumination, and occlusion in video surveillance, the process of violent event recognition appears to be laborious. The use of texture-based feature descriptors based on the Local Optimal Pattern (LOOP) to identify any abnormal group of video frames was demonstrated in a study that carried out this task. Finally, to identify violent occurrences in the movie frames, the prominent features were applied and categorised using the support vector machine technique (SVM)[4]. But the most difficult thing to understand is how to classify violent episodes in surveillance recordings since it is subjective and involves a wider range of factors in the study and interpretation of the audial and graphic communication[5].

The custom STACOG features were investigated to be those that study the features of violent behaviours as of surveillance footage[6].This method involves of two main stages: first, the abstraction of gradient-based autocorrelation structures; second, the use of SVM model classifiers for discriminative knowledge of violence and non-violence behaviours. Comparing single evaluation and multimodal evaluation for programmed semantic detection of violence acts was another aim of the study[7].The use of violence detection is in the dropped violence finding approach taught over (MoBSIFT) motion frontier SIFT and filtering of the motion.[8]

To create a workable violence detection system, it is suggested using frame-grouping and spatiotemporal attention modules. MSM was created to extract salient regions from motion boundaries for spatial attention. Introducing the T-SE block for temporal attention, which allowed temporal features to be recalibrated with a limited number of extra parameters. Specifically, a technique called frame-grouping was presented, which involved average the channels and organizing three consecutively channel-averaged images for use as a 2D CNN input[9].It is not enough to just identify the background environment as being indoors or outside; additional scene classification might be applied during pre-processing.Temporal segments and semantics descriptions can greatly improve motion detection as well as tracking techniques[10]

When cross-correlation was used to the multiple channels image for image recognition, the Two-Dimensional CNN approach produced impressive results. However, because videos are made up of temporal frame patterns, the technique has limitations when it comes to applying it to analysing them. Nevertheless, this CNN 2D is capable of recording dynamic-motion information's[11][9].Recurrent neural networks (RNNs), which are simplified techniques, were used to produce spatiotemporal video representations for the CNN layer's outputs. However, because the techniques don't execute convolution across multiple pictures in previous network levels, they have revealed flaws in their execution rate. Given that traditional two-dimensional CNN frameworks don't include time information, video-based activity identification appears to be a difficult issue[12].

The HAJJ dataset features a variety of anomalous behaviours seen in large-scale crowd footage. Large-scale crowds may be put in danger by these strange actions. Second, a GAN-based optical flow methodology is presented that uses transfer learning to identify aberrant behaviour in settings with large numbers of people. The framework uses U-Net as well as Flow net to identify and classify people's normal and aberrant activities within the large crowd[13].The study compares the capacities of multi-modal identification annotated of ongoing video sequences versus single-mode semantics violent identification[14]Data regarding a video clip's historical and prospective trajectories can be used to more accurately forecast and pinpoint the location of violent events inside a frame[15].

3.Proposed Methodology

3.1 Preprocessing and Mosaicking

For hockey dataset, the research methodology explained a fight detection algorithm. The model is filled with hockey

play video clips from the same dataset that include fight frames and those without. The video detection stage will display that anticipated outcome. Videos from the dataset of hockey fights and non-fights are input to the model.

After basic preprocessing steps the video is split into frames and then the mosaicking process takes place as a part of preprocessing. To improve the fight detection in the input movies, mosaicking phenomena is used at the pre-processing stage. In addition, a method of Video mosaicking is proposed. It involves piecing together video frames to get a full understanding of the scene. The mosaic image that is produced is a compressed version of the video's data. Applications for surveillance and video compression frequently use the Video Mosaicking block. Once the Input Video is Processed the First step is Feature extraction which is the Frame Registration Part which Comprises of the Key Point Extraction and Key Point Matching which is the Most Important Task to be done First and after this Key Frames are identified the Screen-Cut is Done and later the Stitching of the video is done to a Mosaic image. This Mosaic images are also added to the extracted images which propose a more comprehensive view of the scene and which will help to extract the feature appropriately and do classification of the video with more accuracy. During the pre-processing stage, the mosaicking technique divides the image frames into separate subframes with a 40:20 enlargement ratio. CLAHE is an enhancement technique used to improve the contrast of an image while avoiding over-amplification of noise in areas with low contrast. The input frame sizes are 720x576 pixels, and apply an image enhancement

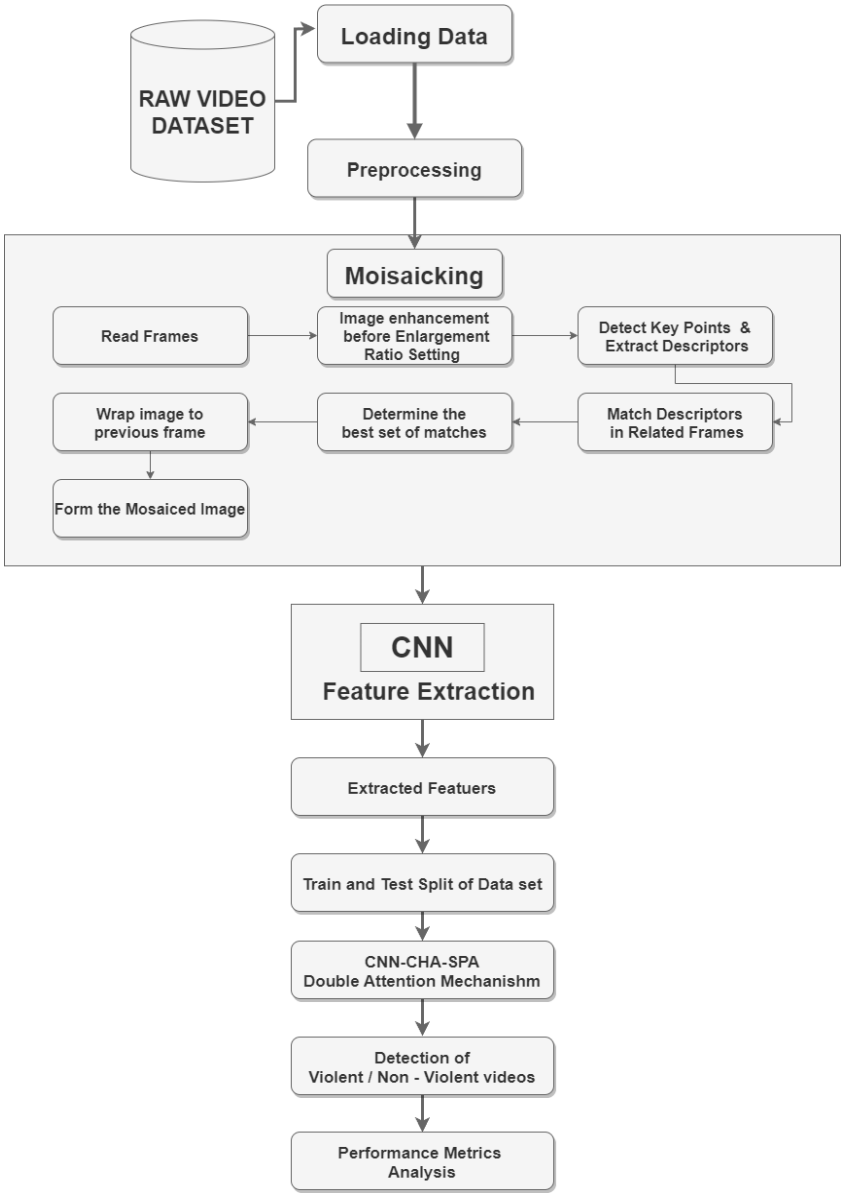

Fig 1: CNN-CHA-SPA Double Attention Mechanism Workflow

technique CLAHE, the output frame size would typically remain the same as the input frame size. Image enhancement techniques like CLAHE operate on the pixel

values of the input image without changing its spatial dimensions. Therefore, the output frame size would still be 720x576 pixels, maintaining the original resolution of the image.By reducing the search space, the mosaicking method makes use of this segmentation to create image mosaics that improve the extraction and classifying phase performance and help produce results consistently over time. The image mosaics shown in fig 2, the breaks between picture frames are gradually matched with picture frames from both fights and non-fights.. After reconstructing the image frame once more, the process of extracting the feature is carried out. The retrieved

features' feature maps are produced by combining the Attention layers, which are made up of numerous convolutional layers, trailed by a max-pooling layer and an activation function. By reducing the search space, the mosaicking method makes use of this segmentation to create image mosaics that improve extracting and classification phase performance and help produce results

consistently over time.Fig:1 shows the entire system workflow with mosaicking Process.

Step 1: Load the Dataset

Step 2: Preprocessing is done. Split the dataset into train and test data.

Step3: The individual frames are extracted from the train video dataset.

Step 4: Mosaicking Process in done.

Step 4.1: Image enhancement algorithm CLACHE is applied on the frames.

Step 4.2: Based on the Threshold value fixed. The keyframes are extracted.

Step 4.3: Matching points on the frames are taken.

Step 4.5: Stitching process is done to get the mosaicked images.

Step 4.6: These images are added to the normal frames that are already processed.

Step 5: The feature extraction is done with the deep learning neural network CNN.

Step 6: The features are extracted.

Step 7: Train the model.

Step 8: The CNN double CHA-SPA attention mechanism is applied to detect whether the given test dataset video is a Violent or non-Violent.

3.2 Feature Extraction and Image Enhancement:

In the realm of image processing, picture enhancement plays a significant role in increasing image quality. This is done by emphasizing pertinent information and reducing irrelevant information. The Histogram Frequency Weighting method is used to enhance the image. The histogram frequency weighted technique considers the relationship between picture grayscale frequency and histogram equalization. In particular, changing the image's initial frequency results in the desired enhancing effect. The use of traditional histogram frequency weighting-based image enhancing techniques has increased recently. A contrast enhancing technique that employs weighted histogram equalization (WHE), which divides the final result into a fixed proportion for the HE result and a fixed proportion for the current gray value setting ratio.

3.3 Attention Mechanism:

Through the acquisition of the convolutional data segment, features were retrieved. The mosaicked images are added to the frames extracted. The CNN is used in feature extraction and the model is trained. If the value of this ReLU activation function is positive, it increases a value; if not, it returns o value. A two-dimensional convolution with a filter-window size was used in the investigation with k = 3×3 is used for the convolution layer arrangement. With $d\hat{}=\max\ 〚(d)〛$ and the max-pooling operations is defined in the last two steps. Higher level vector representation was used to characterize the features-maps produced by these convolutional procedures. Thus, in order to reduce this representation In order to avoid the overfitting problem, the max pooling layer, which comes after the convolution layer, helps select important data features by removing weak activation information.

Although the CNN model is effective at processing image frame feature extraction, it may lose some image feature when processing a large number of frames of images from video clips. The CHA-SPA attention mechanism is applied for the detection process. In Channel Attention a feature map or a particular feature in the data is usually referred to as a "channel" in CNNs. The goal of channel attention mechanisms is to assign varying weights to distinct channels according to their significance. This makes it easier for the model to ignore less valuable features and concentrate on more pertinent ones. The mechanics of spatial attention concentrate on particular spatial areas within a picture. This can be especially helpful for applications when it's important to understand the spatial relationships between various image components. When channel attention and spatial attention are combined, it may indicate that the model is concurrently selectively attending to particular channels and spatial places. The model gains the ability to balance channel and physical locations, giving it a finer-grained emphasis on pertinent data.

The input to the Deep Neural attention technique mechanism is a input sequence of batch size, sequential length and input dimension of the image , after which the

query is computed, key and value matrices is computed to find the attention scores and apply the softmax function to obtain the weights of the attention and finally we compute the output sequence of Weighted Scores. Finally, classify the input videos as Fight and Non-Fight Videos. Channel and spatial attention combines the benefits of both types of attention. Both significant objects and regions are adaptively chosen. In order to learn more accurate spatial features, the spatial attention block (SPAB) combines multiscale knowledge from high-level and low-level stages, and the channel attention block (CHAB) allocates channel feature answers to strengthen the most important channel information while restricting the unimportant through the acquisition of the convolutional data segment, features were retrieved channels. First two layers are taken CHA-SPA attention mechanism is applied and again the next layers are then taken, and the attention mechanism is applied and hence the name the CHA -SPA double attention. The process is repeated unit the saturation point is attainted.

In the above pseudocode, image is the input image, a and b are the minimum and maximum intensity values to map to 0 and 255, respectively. The min() and max() functions compute the minimum and maximum pixel intensities of the image. The factor

Fig 2: Output Screenshots

variable is the scaling factor to map the pixel values between a and b to the range [0, 255]. The clip() function clips the pixel values to the range [0, 255], and the as type() function converts the pixel values to integers. Finally, enhanced_image is the output contrast-enhanced image.The enhanced image is then applied to the Moisacking Process and then the moisaicked images and the normal images which are extracted and applied to the self Attention Mechanism for the classification Process.

4.Results and Discussion

The mosaicked image frames in the hockey fight dataset were split 70% for the training set and 30% for testing.

Confusion Matrix:

The confusion matrix's has TP-true positive value, which totalled values for every kind of image frame, indicated the proportion of sample images that were accurately classified as battle or non-combat clips. The a forementioned figure 3 displays this confusion

matrix. The confusion matrix's bottom right corner, denoted by the number, is the true negative value (TN), which shows how many picture frames were accurately classified as not being under the combat category. The confusion matrix's FP-False Negative value, which also has a value of 0, indicates the number of samples that were mistakenly assigned to a different fight category, while the upper right corner's FN-False Positive value, which has a value of

0, indicates the number of frames of images that were assigned to this type of fight category.

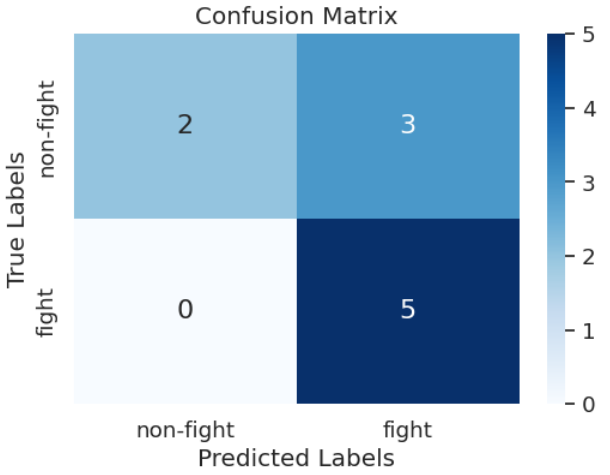

Fig 3: Confusion Matrix

4.1. Performance Assessment

ROC CURVE

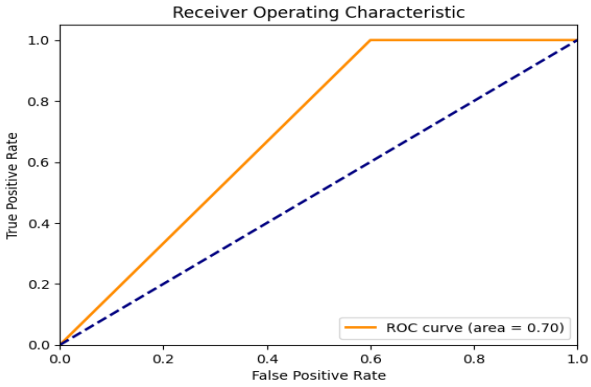

Fig 4: Performance Assessment with ROC Curve

The accuracy parameters evaluation of performance of the current technology and the suggested violence detection method is made clearer in Fig:4 with the ROC Curve.

Accuracy of datasets	
Approaches	**Violence**
AdaBoosts in addition Support Vector Machine	86.40%
Houghs Forest and 2Dim CNN	95.70%
ViF(Onoly)	83.80%
Enhanced Fisher Vectors	94.60%
Prevailing system	95.40%
Proposed system(CHA-SPA)	**98.00%**

Table 1. Proposed framework Comparative evaluation

Different methods of detecting violence, acquired from various datasets, display varying degrees of accuracy in identifying violent incidents. The numerous traditional methods, employing varied datasets, decide. Table 1 shows the results, which show that the

proposed CHA -SPA method classified Violence and non-violence scenes shows more accuracy of 98.00% compared to other traditional techniques. The below fig:5 shows the Training Loss and Training Accuracy for the Proposed method.

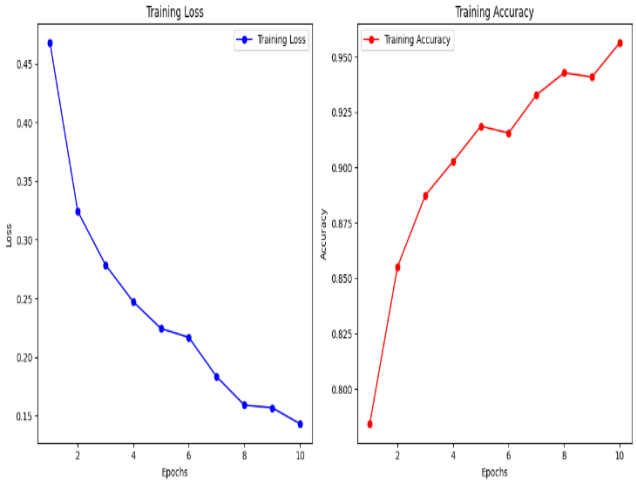

Fig: 5 Training Loss and Training Accuracy

5. Conclusion

The discipline of action detection has shown significant growth during the past fifteen years. A particular category of this component, fight recognition in videos, is especially important in unmanned surveillance and mass videos. The effort provided with well-known technique in evaluating violent action provided from the datasets by using this violent recognition structure and the mosaicking procedure of the image implemented at the pre-processing phase, which finds the most effective matches with priority and thus significantly improves the performance in detection when compared to the normal frames. The proposed model was constructed using several frames of the image from the clips of the video in the dataset. The mosaicking technique is used in this pre-processing stage together with the Contrast enhancement to improve the image. Next, the framework is used to CNN with attention mechanism provides the important features to be identified. The attention layer receives feature extracts for feature categorization under the presumption that all unique self-Attention Mechanisms exist. In the same way as the internal parameters are adjusted. By contrasting the suggested model's accuracy in identifying violence with other conventional approaches, the CHA-SPA model's validity was established that provided the accuracy in large value compared to other mechanism.

Author contributions

V.Elakiya: Conceptualization, Methodology, Software, Field study, Writing-Original draft preparation **Dr. P Aruna:** Data curation, Software, Validation., Field study **Dr. N Puviarasan :** Data Visualisation, Logical Reasoning, Optimization., Field study **Dr. R G Suresh Kumar :**Visualization, Investigation, Writing-Reviewing and Editing., Field study.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Naik AJ, Gopalakrishna MT. Automated Violence Detection in Video Crowd Using Spider Monkey-Grasshopper Optimization Oriented Optimal Feature Selection and Deep Neural Network. Journal of Control, Automation and Electrical Systems 2022; 33: 858–880.

[2] Colton D, Hofmann M. Sampling Techniques to Overcome Class Imbalance in a Cyberbullying Context. Journal of Computer-Assisted Linguistic Research 2019; 3: 21.

[3] Fan M, Zhang X, Hu J, Gu N, Tao D. Adaptive Data Structure Regularized Multiclass Discriminative Feature Selection. IEEE Trans Neural Netw Learn Syst 2022; 33: 5859–5872.

[4] Lohithashva BH, Aradhya VNM. Violent Video Event Detection: A Local Optimal Oriented Pattern Based Approach. Communications in Computer and Information Science, Springer Science and Business Media Deutschland GmbH 2021, 268–280.

[5] Peixoto B, Lavi B, Bestagini P, Dias Z, Rocha A. MULTIMODAL VIOLENCE DETECTION IN VIDEOS. .

[6] Deepak K, Srivathsan G, Roshan S, Chandrakala S. Deep Multi-view Representation Learning for Video Anomaly Detection Using Spatiotemporal Autoencoders. Circuits Syst Signal Process 2021; 40: 1333–1349.

[7] Mensa E, Colla D, Dalmasso M *et al.* Violence detection explanation via semantic roles embeddings. BMC Med Inform Decis Mak 2020; 20.

[8] Albadi N, Kurdi M, Mishra S. Investigating the effect of combining GRU neural networks with handcrafted features for religious hatred detection on Arabic Twitter space. Soc Netw Anal Min 2019; 9.

[9] Vijeikis R, Raudonis V, Dervinis G. Efficient Violence Detection in Surveillance. Sensors 2022; 22.

[10] Gowsikhaa D, Abirami S, Baskaran R. Automated human behavior analysis from surveillance videos: a survey. Artif Intell Rev 2014; 42: 747–765.

[11] Verma P, Charan C, Fernando X, Ganesan S. Lecture Notes on Data Engineering and Communications Technologies 106 Advances in Data Computing, Communication and Security Proceedings of I3CS2021. .

[12] Kang M-S, Park R-H, Park H-M. Efficient Spatio-Temporal Modeling Methods for Real-Time Violence Recognition. IEEE Access 2021; 9: 76270–76285.

[13] Alafif T, Alzahrani B, Cao Y, Alotaibi R, Barnawi A, Chen M. Generative adversarial network based abnormal behavior detection in massive crowd videos: a Hajj case study. J Ambient Intell Humaniz Comput 2022; 13: 4077–4088.

[14] Mahdi MS, Mohammed AJ, Jafer MM. Unusual Activity Detection in Surveillance Video Scene: Review. Journal of Al-Qadisiyah for Computer Science and Mathematics 2021; 13.

[15] Halder R, Chatterjee R. CNN-BiLSTM Model for Violence Detection in Smart Surveillance. SN Comput Sci 2020; 1: 201.