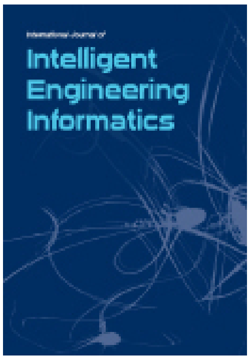
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of
Intelligent Engineering Informatics

International Journal of Intelligent Engineering Informatics

ISSN online: 1758-8723 - ISSN print: 1758-8715
https://www.inderscience.com/ijiei

Adaptive and effective spatio-temporal modelling for offensive video classification using deep neural network

Balika J. Chelliah, K. Harshitha, Saharsh Pandey

DOI: 10.1504/IJIEI.2023.10055332

Article History:

Received:	06 May 2022
Last revised:	20 January 2023
Accepted:	21 January 2023
Published online:	03 May 2023

Adaptive and effective spatio-temporal modelling for offensive video classification using deep neural network

Balika J. Chelliah*, K. Harshitha and Saharsh Pandey

Department of Computer Science and Engineering,
SRM Institute of Science and Technology,
Chennai, Tamil Nadu, India
Email: ballikaj@srmist.edu.in
Email: kh3388@srmist.edu.in
Email: sp5238@srmist.edu.in
*Corresponding author

Abstract: Security cameras have recently been widely implemented in public locations, and the overall crime rate has decreased dramatically due to these omnipresent gadgets. These cameras are typically employed to offer cues and evidence after crimes have occurred rather than to prevent or deter criminal activity in the first place. Manually monitoring a lot of video footage from surveillance cameras takes significant time and effort. As a result, it is critical to automatically recognise aggressive actions from video signals. It is also critical to detect violence in videos to protect children from inappropriate content. This paper tackles the difficult subject of detecting violence in videos. Unlike previous work focusing on merging multimodal features, we take it further by including and utilising visual subtypes connected to violence. The proposed approach is implemented using Jupyter notebook and Tensorflow, with better accuracy of 76.79% on the proposed database test set.

Keywords: deep learning; surveillance video systems; violence detection; transfer learning; motion vector; neural network.

Reference to this paper should be made as follows: Chelliah, B.J., Harshitha, K. and Pandey, S. (2023) ‘Adaptive and effective spatio-temporal modelling for offensive video classification using deep neural network’, *Int. J. Intelligent Engineering Informatics*, Vol. 11, No. 1, pp.19–34.

Biographical notes: Balika J. Chelliah is an Associate Professor in Computer Science and Engineering at SRM Institute of Science and Technology, Ramapuram, Chennai, India. He received his Master’s and PhD degrees in Computer Science and Engineering from SRM Institute of Technology. He has authored more than 50 papers in journals and conferences.

K. Harshitha earned her BTech in Computer Science from SMR Institute of Science and Technology in 2022. She is a proficient coder and programmer who appreciates learning and growing to contribute to new technology breakthroughs. She worked on machine learning.

Saharsh Pandey earned his BTech in Computer Science from SMR Institute of Science and Technology in 2022 and is pursuing a Master in IT Management from the University of Sydney. He is a proficient developer and programmer who appreciates learning and growing to contribute to fascinating technology breakthroughs. He worked on a machine learning and artificial intelligence project because he is always been curious about their applications.

1 Introduction

Problems with public security have sparked significant concern in recent years. Surveillance video systems are frequently utilised for monitoring public situations and play a significant role in public security (Heng et al., 2019). Hundreds of thousands of surveillance cameras are distributed throughout the city to ensure public safety (Mumtaz et al., 2018). An escalating range of security monitors are being put in open spaces, such as airports, railway stations, and highways, for extensive surveillance, and greater quality lenses, such as full HD and 4K webcams, are increasingly being employed in real scenarios (Heng et al., 2019). These cameras are valuable tools for monitoring people's habits, detecting their existence in a certain location, and possibly studying their actions, all of which can aid in the prevention, detection, and reduction of crime (Rashaideh et al., 2019). A crime is an illegal behaviour for which a person can be prosecuted under the law and can be divided into several categories. Crime research is a type of law enforcement work that systematically examines a crime pattern to identify and pinpoint the source of the crime (Prabakaran and Mitra, 2018). Since crimes are distinct every time, it's critical for the system to be educated on broad data that allows it to judge what defines a crime. One method is to create task-specific methods, such as those for violent acts identification and road crime identification; however, these frameworks are context limited and would not function as a generic crime recognition method in the actual situation. Another option is to train your model on various crimes to identify crime as a whole rather than just the one it was trained on (Puthige et al., 2021).

In terms of outstanding multimedia services such as multimedia processing, multimedia information understanding, and multimodal integration, among others, identifying inappropriate visuals is a recently emerging scientific issue (Ding et al., 2021). Substantial population segments are gaining easy access to multimedia content, yet data management is inadequate. Due to the large volume of multimedia content, physically creating annotations is a difficult operation (Giannakopoulos et al., 2010). A digital video is a form of visual capturing system which uses a digital signal instead of an analogue stream (Kang, 2007). Videos are innately multimodal and have extremely rich and complicated semantics (Sharma and Kumar, 2020). Methods for detecting high-level semantics in various unrestricted videos can be used in many scenarios, such as online video searches (Jiang et al., 2018). Although there has been a lot of research into still image facial identification, there has been less study into video identification. Unlike still photos, video often provides significantly more material for identification, including temporal and multiview features. However, we're focused on solving real-world video recognition challenges in which the environment is chaotic and subjects aren't actively interacting with the camera (Mau et al., 2013).

The categorisation of audio-visual content based on the presence of specific human actions is a topic that is becoming increasingly popular. Detection of violent events is receiving much attention in surveillance systems for people's protection in public spaces, especially after the multiple terrorist strikes in India over the last several years (El-Gamal et al., 2020) and to identify regular human actions from unusual activities. Violence detection is critical in developing computerised safety monitoring systems. Regular life-engaging activities, like taking a stroll, sprinting, running, and hand gesturing, are frequently classified as typical human behaviours. In contrast, unusual ferocious acts, such as fights involving two or more people, are subjected to violence (Mumtaz et al., 2018). Most violent incidents in video data are defined by specific audio events (e.g., explosions) (Giannakopoulos et al., 2010).

There are a few challenges in classifying violence in surveillance video systems. The data these cameras create is in petabytes (El-Hasnony et al., 2021). Yet, there is not enough human labour to evaluate it because most places have a low human-to-camera ratio. As a result, some occurrences and actions are skipped, and questionable conduct is not identified in time to avoid problems (Alkhonin et al., 2020). Due to several constraints, such as dynamic lighting shifts or lens motions in the backdrop image. Dimensions, shadowing, angle, and noise backdrop can generate intra-class differences in human movements, making it harder to discern aggression (Xu et al., 2014). The inability to recognise might threaten public safety (Liu et al., 2020). Another issue is that human involvement causes selection bias and indigeneity issues. This fact makes developing an automatic classification system for violent videos even more important (Shinichi and Terumasa, 2015). As a result, in an emergency, the governing officials will be notified and take immediate action against any observed violence (Mumtaz et al., 2018).

These issues can be solved using an automated intelligent anomaly detection system. Anomaly denotes happenings out of the ordinary, irregular, unexpected, and unpredictable, diverging from current designs (Tamang and Roy, 2021). As a result, detecting an anomaly entails recognising various conditions from a video clip, where the anomaly occurs for only a short time (Adnan et al., 2021). However, few studies seek to identify violent scenarios using visual cues in the field. Most techniques include security cameras and background removal methods to detect humans in the area. On the other hand, these methods are unsuitable for films where the camera travels quickly, with numerous shot changes (Giannakopoulos et al., 2010). Furthermore, no one has previously investigated 'visual codebooks' to fill the barrier between the fundamental and relevant fine feature patterns and the top-level ontology of interest.

Driven by the constraints of previous methods and the huge trend of employing deep neural networks for video classification (Jiang et al., 2018), in this paper, subclasses are added and exploited visually related to violence (Thepade et al., 2020). Motion boundary histogram, oriented gradient histogram, and optical flow histogram are three trajectory-based descriptors we utilise. Deep learning classifiers are trained for each subclass and the overall violence class (Elkabbash et al., 2021). By systematically comparing violence detection with and without subclasses, the effectiveness is re-justified for violence video detection, resulting in better accuracy on the test set. The dataset contains a total of 1,951 clips divided into two sets: violence and non-violence.

2 Related work

Many concepts and procedures were studied before proceeding with our paper on violence recognition. The basic structure for our work would require it to recognise faces. In facial recognition, there has been a lot of development. Many methodologies have been created for the same, for example, face recognition using convolutional neural networks with an ensemble on trunk-branch where TBE-CNN shares distinct CNN's low- and middle-level layers to obtain representations of the whole facial picture and facial components quickly and effectively. In this method, we counter the problem of blurred images, which is one of the major issues in recognition systems and causes the system's accuracy to drop low. Here the training dataset to be used is composed of both visuals that are crisp and still and also has images that are artificially blurred to allow and encourage the system to learn face representations with a lot of blurs (Huddar et al., 2020). To further enhance the quality of CNN features, the system extracts facial information and features of images containing pose variations and obscurity, such as patches and spots around facial regions (Ding and Tao, 2018).

Another method for the same could be using the application of cross Euclidean-to-Riemannian metric learning for face recognition from video to blend the average appearance and pattern variation of faces in a single movie. A unique metric learning structure is utilised for training a distance measure spanning a Riemannian manifold and a Euclidean space. The suggested metric learning framework can handle three common video-based face recognition tasks: video-to-still, still-to-video and video-to-video. The discriminant metric in the common space improves face recognition from videos by learning information on heterogeneous material with the shared label (Huang et al., 2018).

Generally, in surveillance videos, facial recognition of the interested individual is distributed over multiple feeds from a network of cameras (Tamang and Roy, 2021). The video-to-still conversion's performance because facial features acquired in the uncontrolled operational region with numerous video cameras have a separate underlying data distribution than faces recorded under controlled settings in the enrolment domain with a still camera, facial recognition algorithms can suffer (Ashraf et al., 2021). This set of problems has been tackled in the method presented in domain-specific face synthesis for video face recognition from a single sample per person. This domain-specific face synthesis (DSFS) technique takes advantage of the OD's representative intra-class variance information (Kumari and Bhatia, 2021). Before the system operation through affinity propagation clustering, a small set of faces from unknown persons is selected in captured condition space. A small group of artificial face images resembling people of interest is developed within capture parameters relevant to the OD (Kriti and Garg, 2021). The DSFS artificial face features create a cross-domain dictionary that compensates for ordered sparsity in a specified application based on sparse representation categorisation, with dictionary blocks combining every person's original and artificial faces. In such cases, using the presented DSFS method can result in greater precision (Mokhayeri et al., 2019). An important method to focus on would be motion detection and recognition using something similar to action recognition by dense trajectories (Wang et al., 2016). We sample closely packed points in every frame and monitor those using dispersion data off a closely-packed optical flow field (Bibi et al., 2022).

But along with facial recognition, other factors can be recognised and understood, such as emotion detection and object detection in videos. Object detection can be achieved using one of the support vector machine (SVM) techniques, i.e., video-based entity recognition using latent bi-constraint SVM. We present two large datasets for video-based object recognition in the following method. Second, we offer LBSVM, a maximum-margin framework for video-based object recognition (Tao et al., 2020). LBSVM is focused on structured-output SVM; however, it is extended to tackle noisy video input and assure output fairness throughout; this is done by latent variables and the two new constraints in the technique (Arabaci and Mohamed, 2020). The first constraint enlarges training videos and necessitates correctly classifying all subsequences (Kumari and Bhatia, 2021). This trains the classifier to recognise several testing videos from different angles. Second, the score function is monotonic regarding the inclusive connection among video substrings. Finally, the latent variable is for view selection and is used to cut out undesirable views of an object in the video sequence. These two constraints and the latent variable allow LBSVM to base the recognition decision on the entire video, preventing output inconsistency (Liu et al., 2018). Emotions are another component that can aid in recognising violence in videos (Arabaci and Mohamed, 2020). This can be done using a strategy for transmitting information from disparate external factors, such as textual information and image, to enable emotion recognition and summarisation (Xu et al., 2018).

3 Methodology

3.1 Transfer learning

Researchers have recently paid more attention to deep learning, which has already been effectively utilised in various practical situations. Deep learning techniques aim to explore wide features from huge volumes of data, putting it ahead of typical machine learning (Tan et al., 2018). Compared to hand-crafted feature descriptors, feature learning is desirable because it allows users to learn sophisticated underlying data representations, which is especially useful for difficult tasks like picture recognition. The learned features acquired from mastering a particular issue are re-used to rectify a different problem in a new activity, referred to as transfer learning. The idea of transfer learning may have originated in educational psychology (Zhuang et al., 2021). It's a fundamental technique for dealing with data scarcity (Tan et al., 2018). Figure 1 illustrates the overall transfer learning process with the source domain converted into the target domain, categorising two output classes of target datasets.

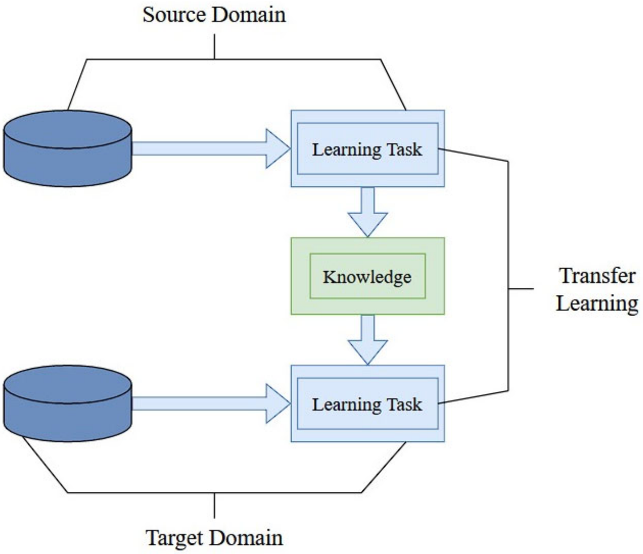
3.1.1 Definition

The term 'transfer learning' is expressed based on domains (D) and tasks (T). Considering a source domain D_s and learning task T_s, a target domain, D_t and learning task T_t, in which:

$$D_s \neq D_t \quad \text{or} \quad T_s \neq T_t \tag{1}$$

Transfer learning attempts to aid in the training of this target domain feature in D_t by leveraging the understanding from D_s and T_s.

Figure 1 Illustration of the transfer learning process (see online version for colours)



3.1.2 Notation

The domain D is divided into two halves: a parameter set X and an unconditional probability density $P(X)$, in which $X = \{x_1, \ldots, x_n\} \in X$. With the given domain $D = \{X, P(X)\}$, a task is divided into two halves: a label space Y and a prediction mapping that is objective $f(\cdot)$, which is depicted as $T = \{Y, f(\cdot)\}$, and this is deduced from the training sample sets $\{xi, yi\}$, wherein $xi \in X$ and $y \in Y$. Therefore, a task is depicted as $T = \{Y, P(Y|X)\}$ (Liu et al., 2020).

The previous study focuses on a multi-instance learning system based on representative prototype selection integrating deep instance classification methods. However, the said method is time demanding and inaccurate when dealing with diverse features. A potential drawback that hampers the success of this strategy is the inability to learn the relationship between components. The CNN deep model was designed to be data-driven, requiring a big labelled dataset to train. Preparing annotated datasets is a difficult and time-consuming operation. The notion of transfer learning is used to alleviate the issue of overfitting for short datasets when using recent deep-learning network topologies. This paper uses 800 violent and non-violent video segments and a Violent Scene Detection Dataset (VSD) 2014 (Ding et al., 2021) as a pre-trained reference system design containing attributes extracted from the internet. This is used to tackle the issue of categorising the proposed dataset as violent or non-violent and obtain the performance metric.

3.2 Working process

Step 1 Obtain the pre-trained model.

Step 2 TensorFlow trains a new dataset using this pre-trained model.

Step 3 The new dataset containing violent and non-violent videos is introduced in the code as a list of test sets is formed.

Step 4 The videos from the test are broken down into image frames (RGB Channel), and motion information is gathered (Optical Flow Channel).

Step 5 Sigmoid function and ReLU Activision are used at the end of the RGB channel and optical flow stream, respectively, to obtain outcomes subjected to max temporal pooling to receive output.

Step 6 This output is used to obtain performance metrics.

3.3 ReLU activation

ReLU activation function is an acronym for rectified linear activation function. If the input to this function is positive, it will output immediately; otherwise, it will output zero. Since models that utilise this function are quicker to train and generally produce higher performance, it has become the typical activation function for many neural networks. This function achieves better results because it can be used on many layers. Other functions cannot work due to the vanishing gradient problem, which the ReLU activation function is overcome.

The equation represents the activation function of ReLU:

$$f(x) = \{0 \ for \ x < 0, x \ for \ x \geq 0\} \tag{2}$$

3.4 Sigmoid function

The sigmoid function is a sort of activation mapping that works as a squelching function, ensuring that the outcome of a neuron in a neural network is within acceptable limits (usually 0 and 1 or –1 and 1). With Gaussian blur type training sets, the sigmoid function is frequently used. This function:

$$g(z) = 1/\left(1 + e^{-z}\right) \tag{3}$$

is advantageous since it is differentiable, and that’s crucial for weight-learning methods. The sigmoid function expands the number of possible hypotheses the system can reflect.

4 Experiments

The experimental aspects of the proposed transfer learning methodology are described in this section.

4.1 Dataset

The first step is gathering the dataset, which includes two clips: violence and non-violence. There were a total of 1,951 clips in the dataset. Figure 2 depicts the number of clips in each set, whereas Figures 3 and 4 depict some of the violent and non-violent dataset frames, respectively. Unusual ferocious acts, such as fights involving two or more people, are subjected to violence. In contrast, regular life-engaging activities, like taking a stroll, sprinting, running, and hand gesturing, are frequently classified as non-violence.

Figure 2 No. of clips in each set (see online version for colours)

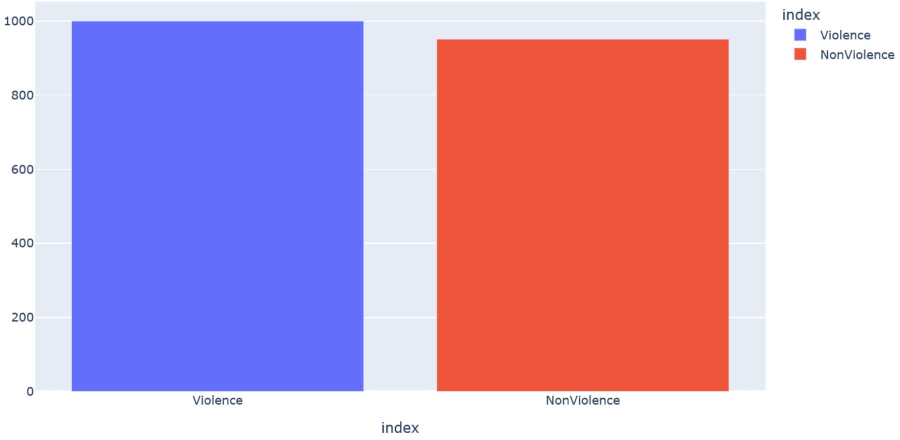


Figure 3 Violent dataset (see online version for colours)

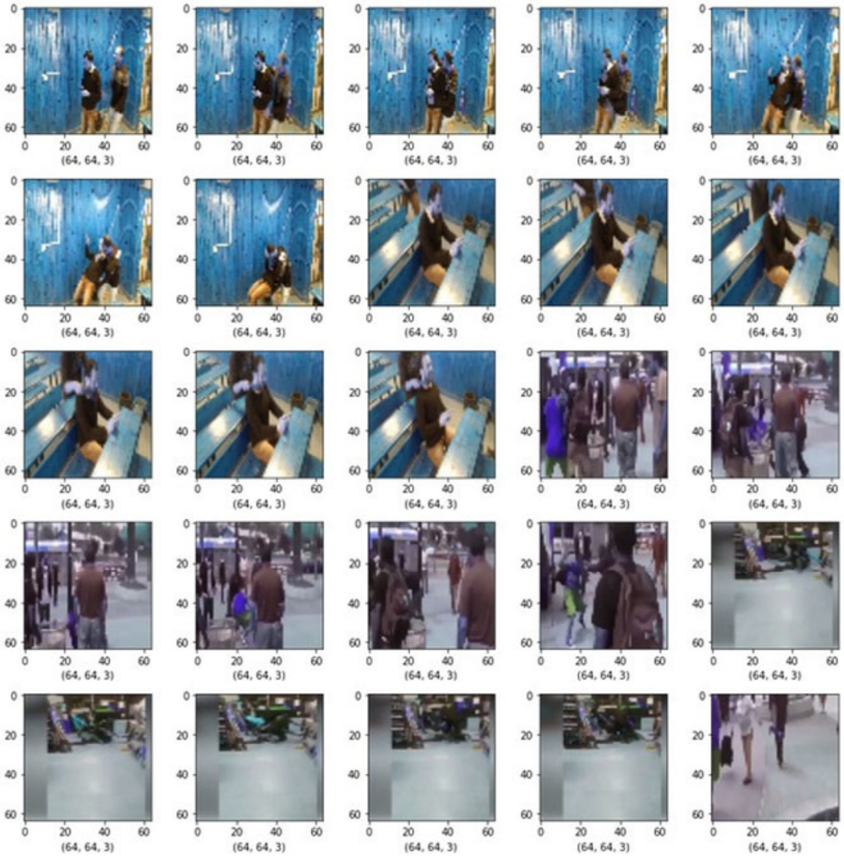
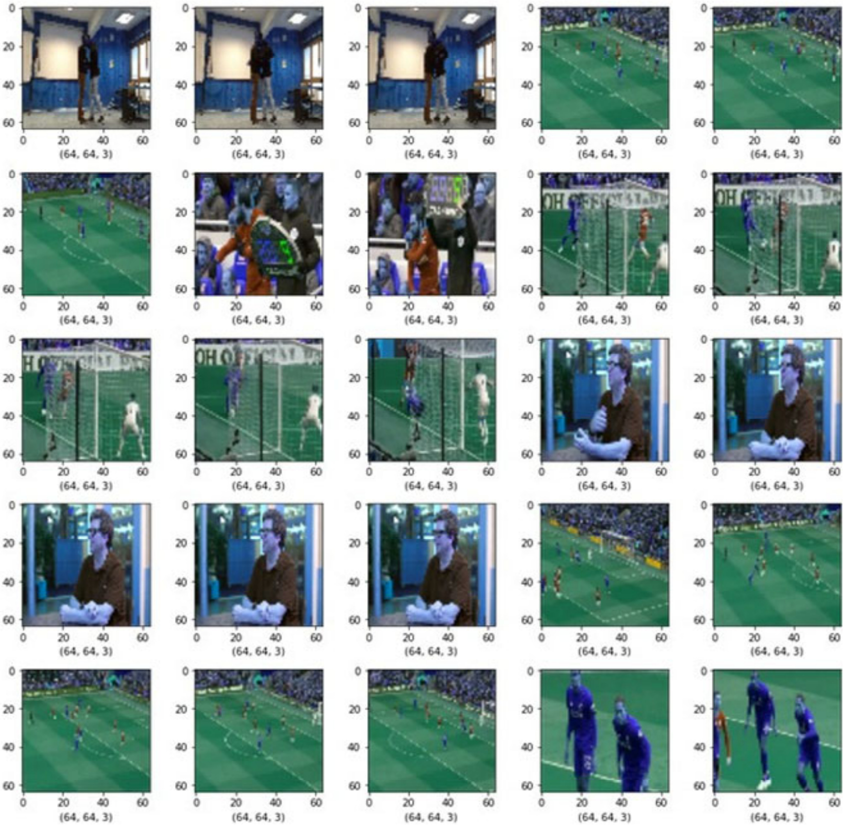


Figure 4 Non-violent dataset (see online version for colours)



4.2 Pre-processing

Pre-processing can be considered a set of processes performed on each frame. Each frame contains decoding, computation, and encoding activities. Decoding is converting a video frame from compressed to raw format. Computation is a specific operation that we must perform on the frame. Encoding is the process of returning the processed frame to its compressed condition. This is a sequential operation. Each frame contains spatial information, whereas the series includes temporal information. We employ a composite structure that involves computations (for perceptual processing) and recurring layers to simulate the two elements (for temporal processing). Pre-process the frames with Gaussian blur, thresholding, dilation, and erosion.

4.3 Training

The first step in the training process is to create a new model. The training data in train_list and train_label is fed to the model, after which it learns to associate frames and labels. The trained model makes predictions about the test set, the test_list array. The predictions are verified to check the accuracy, and if needed, the model is re-trained. Figure 7(a) depicts the steps required to obtain the trained model.

4.4 Feature extraction

The video codec is used to extract motion information. The motion vector, the magnitude of motion at each pixel per frame, is obtained from the codec. This motion vector is a two-dimensional vector with the same dimensions as a frame from a video sequence. A motion feature representing the amount of motion in the frame is generated from this motion vector. To create this motion feature, the motion vector is first divided into twelve equal-sized sub-regions by slicing it along the x and y axes into three and four regions, respectively. The amount of motion along the x and y axes at each pixel in these subregions is aggregated and used to generate a two-dimensional motion graph for each frame. Figures 5 and 6 display the frames plotted for feature extraction on the x and y-axis from both types of clips in the dataset.

Figure 5 Plotted image frames in a violent set (see online version for colours)

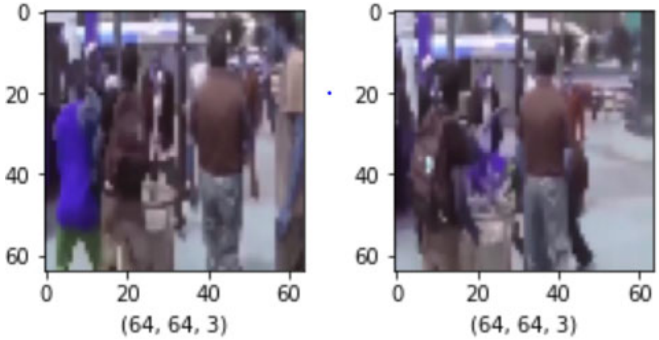
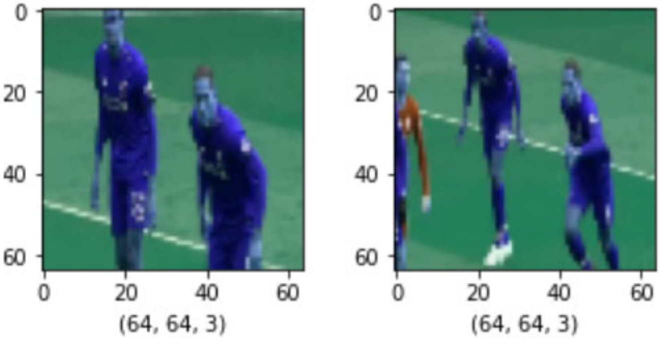


Figure 6 Plotted image frames in the non-violent set (see online version for colours)



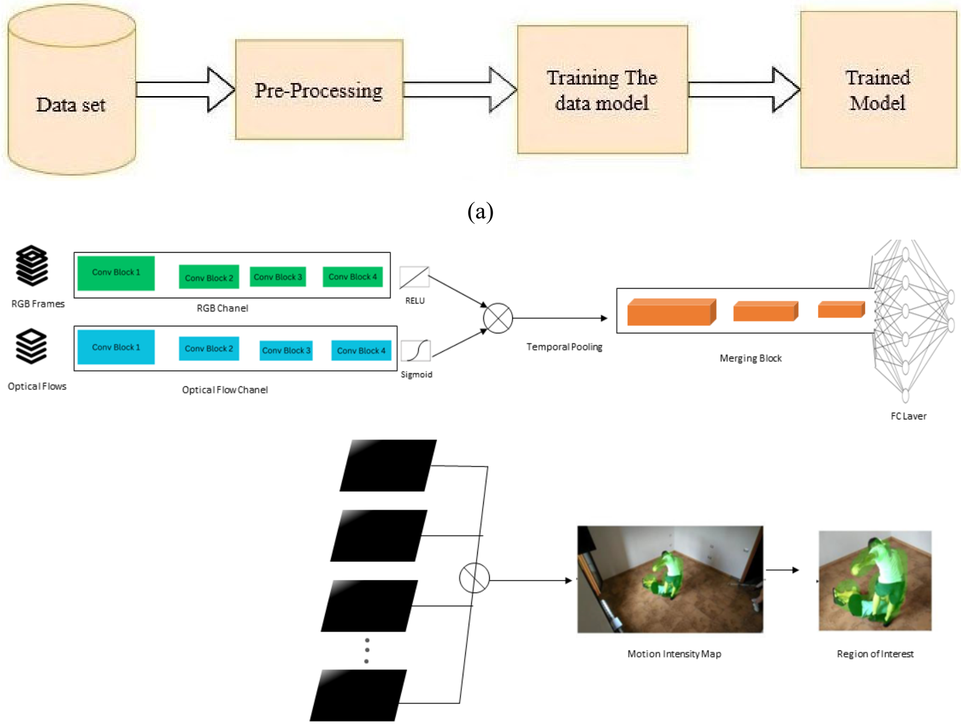
4.5 Classification and evaluation

The next step is to carry out the process of classification, which is a type of pattern recognition. This discovers a matching pattern on the test set based on training data. These predictions are made about some frames on the test set using tf.Keras models. Finally, the model is evaluated, and a list of two values is returned. The first value will be the model's loss on the data, while the second will be the model's accuracy on the dataset. The loss value is omitted because we are just concerned about accuracy.

5 Violence video detection

The suggested dataset is split into two parts: the training part, which is 90%, and the test set is the remaining 10% of the dataset. One part of the clips consists of violent behaviours and actions, whereas the other half is normal videos without any details or violence.

Figure 7 (a) Steps to obtain the trained model (b) The architecture diagram involved in obtaining the output from the trained model (see online version for colours)



(a)

(b)

To help design a pooling mechanism, the model uses an optical flow channel branch. The sigmoid function has been used at the end of the Optical Flow field, whereas Relu activation is used at the end of the RGB field, as depicted in Figures 7(a) and 7(b). The outcomes from the Optical and RGB channels are then amplified, after which they are subjected to temporal max-pooling processing. Because the sigmoid function's outcome is between 0 and 1, it can be used as a scalability component to change the RGB channel's outcome. Furthermore, because max-pooling can only save local maximums, the RGB channel's result amplified by one has a higher chance of being kept. At the same time, the magnitude proliferated by zero is likely to be lost. This process is a type of self-taught grouping method that uses a flow of an optical branch; as a result, it decides which data the framework should keep or discard.

6 Results and discussion

In summary, through our approach, we learnt that deep learning-based techniques frequently outperform traditional feature extraction-based models. Furthermore, most cutting-edge outcomes use multiple-channel inputs (e.g., raw RGB images and optical flows). Simultaneously, complicated models are not highly resistant to over-fitting. Optical flow uses a novel pooling mechanism that can use temporal feature pooling rather than using human-designed techniques. The research aims to apply the proposed hard-attention mechanism to action detection tasks. The larger datasets in the action recognition challenge provide a data-rich environment for the deep reinforcement learning agent. Yet, the greater diversity of actions creates a larger search space, resulting in additional difficulties. A promising next direction for this research is to apply the proposed hard attention mechanism to multi-attention scenarios in collaborative actions. The proposed method is evaluated against two datasets categorised as violent videos and non-violent videos. The test set comprises random videos from both these sets, and our proposed system was able to accurately recognise the video type from the set of input videos.

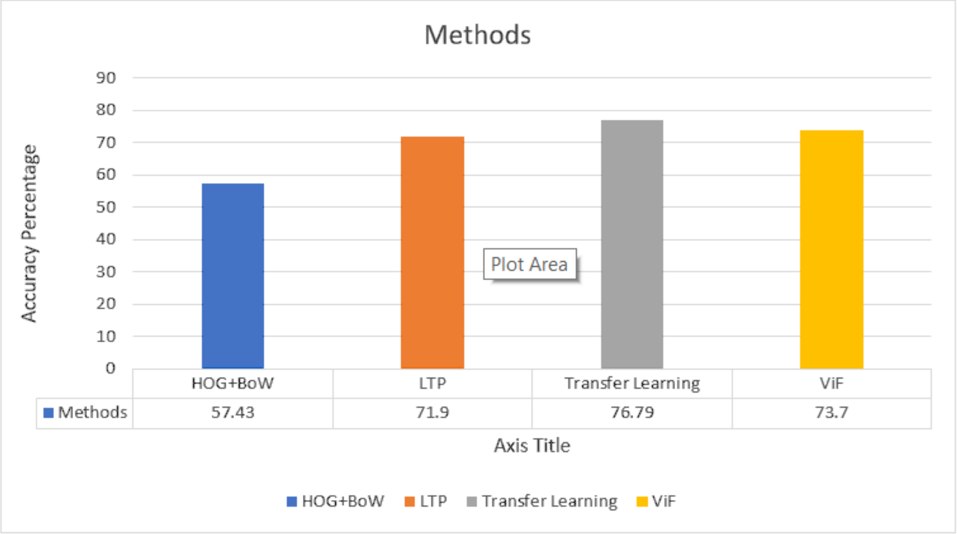
Results show an accuracy of 76.79% obtained on the dataset. Hence, this proposed transfer learning method achieved cutting-edge accuracy in the data. This model can distinguish between violent and non-violent actions from various clips. Above all, the proposed approach obtained desirable accuracy outcomes in 10 training epochs on the dataset, significantly lowering the steps needed to train the model for the target dataset. As a result, the suggested system has well-defined formal features and achieves a well-balanced equilibrium among multiple factors (Table 1). Transfer learning helps tackle complicated real-world issues with multiple restrictions by making deep learning more approachable. Experiments were conducted for offensive video classification on various methods, namely LTP, violent flow descriptor (ViF), and histogram of gradient (HOG) + bag-of-words (BoW) models on the proposed dataset. A comparison chart was plotted, which depicted all the algorithms' different accuracy results (Figure 8).

Table 1 Accuracy representation of distinct methods for the proposed dataset

Method	*Accuracy*
HOG + BoW (Mau et al., 2013)	57.43%
LTP (Puthige et al., 2021)	71.90%
ViF (Puthige et al., 2021)	73.70%
Transfer learning	76.79%

In Table 2, we discuss the different methods and their key advantages and disadvantages, which are highly responsible for their resultant performance metric. Starting with the method of histogram of gradient in combination with bag-of-words, the key advantage is that when provided with larger scales of data, it provides more global information. In contrast, when the scale of data is comparatively smaller, the data received is more fine-grained than global. The second method mentioned is the local ternary pattern, which has the primary advantage of being resistant to noises in images, giving its accuracy the required boost. However, the disadvantage of this method is that it is susceptible to variations in lighting, which lowers its accuracy return. The third method is the violent flow descriptor which has the significant advantage of showing the amplitude and direction of the motion. Still, unfortunately, the results of this method aren't transparent, which becomes a key limitation. Finally, the fourth method is our proposed algorithm, Transfer learning which saves resources due to the re-use of the pre-trained features from the source model to the target model. But the main limitation of this method is a negative transfer which occurs when it becomes difficult to solve a task using pre-trained features.

Figure 8 Graphical representation of accuracy results for different methods such as HOG+BoW, LTP, transfer learning and ViF (see online version for colours)



Note: Our proposed method, i.e., transfer learning, achieves the best results, depicting that it helps tackle complicated real-world issues by making deep learning more accessible.

Table 2 Advantages and disadvantages of the distinct methods

Method	*Advantages*	*Disadvantages*
HOG + BoW (Mau et al., 2013)	This has the advantage of providing more global information at larger scales while providing more fine-grained detail at smaller scales.	The end feature vector grows in size, which means it takes longer to retrieve and train with a particular classifier.
LTP (Puthige et al., 2021)	LTP are more noise resistant.	Not resistant to variations in lighting.
ViF (Puthige et al., 2021)	It shows both the amplitude and the direction of motion.	The results are not transparent.
Transfer learning	When developing novel models, resources are saved, and performance is increased.	Negative transfer is one of the most significant drawbacks of transfer learning.

7 Conclusions and future work

In conclusion, we have presented an adaptive and effective algorithm for transfer learning for offensive video classification. Deep learning-based techniques outperform traditional feature-extraction methods most of the time. Furthermore, multi-channel sources are used in most cutting-edge outcomes (e.g., raw RGB images, optical flows, acceleration maps). Optical flow employs a one-of-a-kind pooling mechanism that can replace human-designed tactics with periodic feature pooling; the research aims to implement the hypothesised difficult technique in motion detection methods. The transfer learning algorithm was used to recognise violent behaviours from videos automatically. This model provides constructive results and is easy to use. The accuracy obtained is 76.79% for this model by analysing the outputs obtained. In an attempt to get numerical findings and use our local dataset as a standard dataset for assessing abnormal activity, which can be applied in detection methods, additional studies will also look into evaluating this technique in other settings and will improve it. Another possible future approach for this work is applying the hard attentiveness technique to multi-attentional situations. In the approach, alternatives that do not specifically involve optical flow can be researched.

References

Adnan, R.M., Mostafa, R.R., Islam, A.R.M., Gorgij, A.D., Kuriqi, A. and Kisi, O. (2021) 'Improving drought modelling using hybrid random vector functional link methods', *Water*, Vol. 13, No. 23, p.3379, https://doi.org/10.3390/w13233379.

Alkhonin, A., Almutairi, A., Alburaidi, A. and Saudagar, A.K.J. (2020) 'Recognition of flowers using convolutional neural networks', *International Journal of Intelligent Engineering Informatics*, Vol. 8, No. 3, pp.186–197, https://doi.org/10.1504/ijiei.2020.111246.

Arabaci, H. and Mohamed, M.A. (2020) 'A knowledge-based diagnosis algorithm for broken rotor bar fault classification using FFT, principal component analysis and support vector machines', *International Journal of Intelligent Engineering Informatics*, Vol. 8, No. 1, pp.19–37, https://doi.org/10.1504/ijiei.2020.10027093.

Ashraf, N.M., Mostafa, R.R., Sakr, R.H. and Rashad, M.Z. (2021) 'Optimising hyperparameters of deep reinforcement learning for autonomous driving based on whale optimisation algorithm', *PloS ONE*, Vol. 16, No. 6, p.e0252754, https://doi.org/10.1371/journal.pone.0252754.

Bibi, A., Khan, M.A., Javed, M.Y., Tariq, U., Kang, B.G., Nam, Y. and Sakr, R.H. (2022) 'Skin lesion segmentation and classification using conventional and deep learning based framework', *Computers, Materials & Continua*, Vol. 71, No. 2, pp.2477–2495, https://doi.org/10.32604/cmc.2022.018917.

Ding, C. and Tao, D. (2018) 'Trunk-branch ensemble convolutional neural networks for video-based face recognition', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 40, No. 4, pp.1002–1014, https://doi.org/10.1109/TPAMI.2017.2700390.

Ding, X., Li, B., Li, Y., Guo, W., Liu, Y., Xiong, W. and Hu, W. (2021) 'Web objectionable video recognition based on deep multi-instance learning with representative prototypes selection', *IEEE Transactions on Circuits and Systems for Video Technology: A Publication of the Circuits and Systems Society*, Vol. 31, No. 3, pp.1222–1233, https://doi.org/10.1109/tcsvt.2020.2992276.

El-Gamal, A.H., Mostafa, R.R. and Hikal, N.A. (2020) 'Load balancing enhanced technique for static task scheduling in cloud computing environments', *Internet of Things – Applications and Future*, pp.411–430, Springer Singapore, Singapore, https://doi.org/10.1007/978-981-15-3075-3_28.

El-Hasnony, I.M., Mostafa, R.R., Elhoseny, M. and Barakat, S.I. (2021) 'Leveraging mist and fog for big data analytics in IoT environment', *Transactions on Emerging Telecommunications Technologies*, Vol. 32, No. 7, https://doi.org/10.1002/ett.4057.

Elkabbash, E.T., Mostafa, R.R. and Barakat, S.I. (2021) 'Android malware classification based on random vector functional link and artificial jellyfish search optimiser', *PloS ONE*, Vol. 16, No. 11, p.e0260232, https://doi.org/10.1371/journal.pone.0260232.

Giannakopoulos, T., Makris, A., Kosmopoulos, D., Perantonis, S. and Theodoridis, S. (2010) 'Audio-visual fusion for detecting violent scenes in videos,' *Artificial Intelligence: Theories, Models and Applications*, pp.91–100, Springer Berlin Heidelberg, Berlin, Heidelberg, https://doi.org/10.1007/978-3-642-12842-4_13.

Heng, W., Jiang, T. and Gao, W. (2019) 'How to assess the quality of compressed surveillance videos using face recognition', *IEEE Transactions on Circuits and Systems for Video Technology: A Publication of the Circuits and Systems Society*, Vol. 29, No. 8, pp.2229–2243, https://doi.org/10.1109/tcsvt.2018.2866701.

Huang, Z., Wang, R., Shan, S., van Gool, L. and Chen, X. (2018) 'Cross Euclidean-to-Riemannian metric learning with application to face recognition from video', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 40, No. 12, pp.2827–2840, https://doi.org/10.1109/TPAMI.2017.2776154.

Huddar, M.G., Sannakki, S.S. and Rajpurohit, V.S. (2020) 'Attention-based word-level contextual feature extraction and cross-modality fusion for sentiment analysis and emotion classification', *International Journal of Intelligent Engineering Informatics*, Vol. 8, No. 1, pp.1–18, https://doi.org/10.1504/ijiei.2020.105430.

Jiang, Y-G., Wu, Z., Wang, J., Xue, X. and Chang, S-F. (2018) 'Exploiting feature and class relationships in video categorisation with regularised deep neural networks', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 40, No. 2, pp.352–364, https://doi.org/10.1109/TPAMI.2017.2670560.

Kang, B. (2007) 'A review on image and video processing', *International Journal of Multimedia and Ubiquitous Engineering*, Vol. 2, No. 2, pp.49–62, http://dx.doi.org/10.14257/ijmue.2007.2.2.04

Kriti, N.A. and Garg, U. (2021) 'Modified silhouette based segmentation outperforming in the presence of intensity inhomogeneity in the hyperspectral images', *International journal of Intelligent Engineering Informatics*, Vol. 9, No. 3, p.260, https://doi.org/10.1504/ijiei.2021.10041954.

Kumari, N. and Bhatia, R. (2021) 'Systematic review of various feature extraction techniques for facial emotion recognition system', *International Journal of Intelligent Engineering Informatics*, Vol. 9, No. 1, pp.59–87, https://doi.org/10.1504/ijiei.2021.10039255.

Liu, D., Bellotto, N. and Yue, S. (2020) 'Deep spiking neural network for video-based disguise face recognition based on dynamic facial movements', *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 31, No. 6, pp.1843–1855, https://doi.org/10.1109/TNNLS.2019.2927274.

Liu, Y., Hoai, M., Shao, M. and Kim, T-K. (2018) 'Latent bi-constraint SVM for video based object recognition', *in IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 28, No. 10, pp.3044–3052, https://doi.org/10.1109/TCSVT.2017.2713409.

Mau, S., Chen, S., Sanderson, C. and Lovell, B.C. (2013) *Video Face Matching Using Subset Selection and Clustering of Probabilistic Multi-Region Histograms*, arXiv [cs.CV], http://arxiv.org/abs/1303.6361.

Mokhayeri, F., Granger, E. and Bilodeau, G-A. (2019) 'Domain-specific face synthesis for video face recognition from a single sample per person', *IEEE Transactions on Information Forensics and Security*, Vol. 14, No. 3, pp.757–772, https://doi.org/10.1109/tifs.2018.2866295.

Mumtaz, A., Sargano, A.B. and Habib, Z. (2018) 'Violence detection in surveillance videos with deep networks using transfer learning', *2nd European Conference on Electrical Engineering and Computer Science (EECS)*, Vol. 8, No. 18, pp.558–563, https://doi.org/10.1109/EECS.2018.00109.

Prabakaran, S. and Mitra, S. (2018) 'Survey of analysis of crime detection techniques using data mining and machine learning', *Journal of Physics. Conference Series*, Vol. 1000, p.012046, https://doi.org/10.1088/1742-6596/1000/1/012046.

Puthige, I., Bansal, K., Bindra, C., Kapur, M., Singh, D., Kumar Mishra, V., Aggarwal, A., Lee, J., Kang, B., Nam, Y. and R. Mostafa, R. (2021) 'Safest route detection via danger index calculation and K-means clustering', *Computers, Materials & Continua*, Vol. 69, No. 2, pp.2761–2777, https://doi.org/10.32604/cmc.2021.018128.

Rashaideh, H., Shaheen, A. and Najdawi, N.A. (2019) 'Real-time image encryption and decryption methods based on the Karhunen-Loeve transform', *International Journal of Intelligent Engineering Informatics*, Vol. 7, No. 5, pp.399–421, https://doi.org/10.1504/ijiei.2019.103623.

Sharma, S. and Kumar, V. (2020) 'Low-level features based 2D face recognition using machine learning', *International Journal of Intelligent Engineering Informatics*, Vol. 8, No. 4, pp.305–330, https://doi.org/10.1504/ijiei.2020.10034278.

Shinichi, G. and Terumasa, A. (2015) 'Violent scenes detection based on automatically-generated mid-level violent concepts', *Proceedings of 19th Computer Vision Winter Workshop*, Vol. 19, No. 5, pp.11–18 [online] https://cmp.felk.cvut.cz/cvww2014/papers/13/13.pdf (accessed 27 March 2022).

Tamang, S.K. and Roy, P. (2021) 'COVID-19 drugs invention using deep neural network models: an artificial intelligence approach', *International Journal of Intelligent Engineering Informatics*, Vol. 9, No. 2, pp.176–192, https://doi.org/10.1504/ijiei.2021.10040084.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C. and Liu, C. (2018) 'A survey on deep transfer learning', *Artificial Neural Networks and Machine Learning*, Vol. 11141, pp.270–279, Springer, Cham, https://doi.org/10.1007/978-3-030-01424-7_27

Tao, H., Al-Sulttani, A.O., Salih Ameen, A.M., Ali, Z.H., Al-Ansari, N., Salih, S.Q. and Mostafa, R.R. (2020) 'Training and testing data division influence on hybrid machine learning model process: application of river flow forecasting', *Complexity*, Vol. 2020, pp.1–22, https://doi.org/10.1155/2020/8844367.

Thepade, S.D., Abin, D., Das, R. and Sarode, T. (2020) 'Human face gender identification using Thepade's sorted N-ary block truncation coding and machine learning classifiers', *International Journal of Intelligent Engineering Informatics*, Vol. 8, No. 2, pp.77–94, https://doi.org/10.1504/ijiei.2020.109094.

Wang, X., Chen, D., Yang, T., Hu, B. and Zhang, J. (2016) 'Action recognition based on object tracking and dense trajectories', *2016 IEEE International Conference on Automatica (ICA-ACCA)*, pp.1–5, https://doi.org/10.1109/ICA-ACCA.2016.7778391.

Xu, B., Fu, Y., Jiang, Y-G., Li, B. and Sigal, L. (2018) 'Heterogeneous knowledge transfer in video emotion recognition, attribution and summarisation', *IEEE Transactions on Affective Computing*, Vol. 9, No. 2, pp.255–270, https://doi.org/10.1109/taffc.2016.2622690.

Xu, L., Gong, C., Yang, J., Wu, Q. and Yao, L. (2014) 'Violent video detection based on MoSIFT feature and sparse coding', *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.3538–3542, https://doi.org/10.1109/ICASSP.2014.6854259.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H. and He, Q. (2021) 'A comprehensive survey on transfer learning', *Proceedings of the IEEE. Institute of Electrical and Electronics Engineers*, Vol. 109, No. 1, pp.43–76, https://doi.org/10.1109/jproc.2020.3004555.