International Journal of Creative and Open Research in Engineering and Management
ISSN: 3108-1754 (Online)
Volume 02 Issue 11 November-2026 | Impact Factor: 3.5



AI-Based Crowd Violence Detection Using Vision Transformer and Temporal Transformer

Vinutha P H, Madhushree J, Mangala Biradar, Lavanya K Y, Sakshi L C
Department of Computer Science and Engineering Akshaya Institute of Technology, Tumakuru, Karnataka, India
vinupalegar100@gmail.com, madhushreej674@gmail.com, madhubiradarm@gmail.com yambalavanya45@gmail.com, chakalabbisaakshi@gmail.com

How to Cite this Article:
H, V. P., J, M., Biradar, M., Y, L. K. & C, S. L. (2026). AI-Based Crowd Violence Detection Using Vision Transformer and Temporal Transformer. International Journal of Creative and Open Research in Engineering and Management, 02(11), 1-9. https://doi.org/10.55041/ijcope.v2i11.162
License:
This article is published under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.
© The Author(s). Published by International Journal of Creative and Open Research in Engineering and Management.





https://doi.org/10.55041/ijcope.v2i11.162

***Abstract*—This work describes a crowd-violence classification pipeline that analyses short video sequences rather than making a decision from a single image. Eight frames are sampled from each clip and encoded with the pretrained google/vit-base-patch16-224-in21k Vision Transformer. The resulting 768-dimensional descriptors are reduced to 256 dimensions and processed by a two-layer Temporal Transformer to predict Fight or Non-Fight. The experiments use the RWF-2000 dataset and the recorded project split, which contains 1,594 training videos and 400 validation videos. The model obtained 92.75% accuracy on that validation set. A webcam prototype was also implemented to display predictions and demonstrate sound alerts, email notifications, and timestamped JPEG capture. The reported score is specific to this validation split; evaluation on an independent test set and additional surveillance conditions remains necessary.**

***Index Terms*—crowd violence detection, Vision Transformer, Temporal Transformer, RWF-2000, video classification, deep learning, surveillance.**

I. Introduction

Monitoring a large collection of surveillance recordings requires an operator to inspect many minutes of footage, even when an incident appears only briefly. The project addresses this review burden by assigning a preliminary label to each video clip. Clips predicted as Fight can then be prioritised for human examination; the system is designed to support an operator rather than make a final judgement about an incident. A short sequence may contain both routine movement and actions that appear aggressive. Looking at one image alone can therefore be misleading: physical proximity or a sudden move-ment may have a non-violent explanation. The implemented method retains the chronological order of selected frames so that the classifier can use changes across the clip, not just the appearance of an isolated image.

The video-processing stage is divided into two operations. A pretrained Vision Transformer (ViT) converts each selected frame into a feature vector, after which a second Transformer models the order of those vectors and produces one clip-level decision. A webcam prototype connects this classifier to a live camera and demonstrates prediction display, sound alerts, email messages, and the saving of image evidence.

The work combines ViT frame encoding with temporal classification and demonstrates webcam inference on the recorded RWF-2000 split.

II. Related Work

Prior studies have considered violence recognition as a video-analysis problem, but differ in the information they extract and how they combine it. Negre *et al.* organise the field through a review of datasets, input preparation, frame selection, and outstanding research issues [8]. Their survey is useful for positioning individual system designs within the wider area.

A related approach by Rendo´n-Segador *et al.* combines ViT representations with a temporal decision procedure based on an adaptive sliding window [9]. In a broader crowd-monitoring setting, Qaraqe *et al.* combine Video Swin Transformer features with crowd-count and optical-flow cues to analyse crowd state and violence level [10]. He *et al.* use two input streams, RGB and optical flow, with an Inflated 3D ConvNet and Transformer-based processing [11].

This project evaluates a more limited configuration than the multi-cue systems described above. For each RWF-2000 clip, eight frames are encoded individually by the pretrained ViT; their feature vectors are then supplied, in sequence, to an independent Temporal Transformer. The classifier has two outputs, Fight and Non-Fight. Accordingly, the reported experiment is an evaluation on the project's recorded validation partition, not a comparison across datasets.

III. Background

A. Vision Transformer

The selected image is represented by a set of patch tokens before it is processed by the Vision Transformer. Self-attention relates information from different patches, allowing the final image representation to incorporate more than one local region. This architecture was introduced for image-recognition tasks by Dosovitskiy *et al.* [1]. In this work, the pretrained model is used to obtain a descriptor for every sampled frame.

B. Temporal Transformer

The sequence of frame descriptors represents how the scene changes during a clip. A temporal model can learn relationships between earlier and later observations, rather than making a separate decision for every frame. Space-time attention has been studied for video understanding [2], while violence-recognition research has explored adaptive temporal



decisions and combined spatial-temporal processing [9], [11]. Here, a separate Transformer encoder performs this sequence-level processing after the ViT stage.

C. RWF-2000 Dataset

The RWF-2000 dataset provides surveillance-video examples for studying the recognition of violent activity [3]. It has also been used in subsequent work on automated violence analysis [9], [11]. The classifier in this project maps each selected clip to one of two labels, Fight or Non-Fight.

IV. Proposed System

Figure 1 shows the sequence of operations implemented in the project. The input clip is first represented by eight selected frames. The ViT encodes those frames, a projection layer adjusts the feature size, and the Temporal Transformer processes the ordered sequence. A classification head then returns one of the two labels. The webcam module reuses the trained classifier and activates the configured notification and evidence-saving functions.

A. Frame Sampling and Preprocessing

The loader obtains eight frames from each video and preserves their time order. Each image is resized and normalized according to the input requirements of the selected ViT checkpoint. Consequently, every clip is represented by the same number of frame inputs, while the chronological relationship among the selected images is maintained.

B. Spatial Feature Extraction

The project uses the pretrained model `google/vit-base-patch16-224-in21k`. Each sampled frame is passed through the ViT to obtain a 768-dimensional feature vector. For eight frames, the video is represented as

$$X = [x_1, x_2, \ldots, x_8], \quad x_t \in \mathrm{R}^{768}. \tag{1}$$

This representation has eight entries per clip, and each entry contains 768 values. The ordered set is used as the input to the temporal stage.

C. Temporal Feature Learning

Before entering the temporal encoder, each 768-value descriptor is transformed by a learned linear layer into a 256-value representation. The sequence is processed by two Transformer encoder layers with eight attention heads per layer. The feed-forward block uses 512 units, and the dropout parameter is 0.2. After sequence processing, mean pooling aggregates the frame representations into a single clip descriptor.

D. Classification

The pooled representation is passed to a classification head consisting of layer normalization, a dense layer, GELU activation, dropout, and the final two-class output layer. The two output scores correspond to Fight and Non-Fight; the larger score determines the predicted label.

TABLE I
Recorded Dataset Split

Subset	Videos	Purpose
Training	1,594	Model training
Validation	400	Performance evaluation
Total used	1,994	Recorded split

V. Implementation

A. Dataset Preparation

The RWF-2000 dataset was used for the experiments. The recorded project split contains 1,594 videos for training and 400 videos for validation. The validation set contains 200 Fight videos and 200 Non-Fight videos. The results in this paper refer to this recorded split.

B. Model Training

The training input consists of the frame-feature sequences prepared by the ViT stage. These sequences are used to fit the temporal encoder and its two-class prediction head. Performance is reported on the 400 clips reserved for validation. Since the available experiment does not include a separately evaluated independent test set, the results below describe this validation partition only.

C. Webcam-Based Detection

A webcam application was developed using OpenCV to capture live frames and run the trained model. The application displays the predicted class and confidence. Prediction smoothing is used to reduce rapid changes in the displayed result.

When the configured Fight detection condition is met, the prototype can generate an audio alert using a background thread. A cooldown is used to limit repeated alerts. The application also supports email notification through SMTP and saves timestamped JPEG evidence locally. Email credentials are configured outside the source code. These functions demonstrate the notification workflow; they do not by themselves measure real-world response time or guarantee that every incident will be detected.

VI. Results and Discussion

A. Evaluation Measures

The validation output is summarised with accuracy, precision, recall, and F1-score. In the calculations below, a Fight prediction is treated as positive. TP represents a Fight clip correctly identified, whereas TN represents a Non-Fight clip correctly rejected. FP refers to a Non-Fight clip assigned the Fight label, and FN refers to a Fight clip missed by the classifier:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{2}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN}, \tag{3}$$

$$\text{F1} = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{4}$$

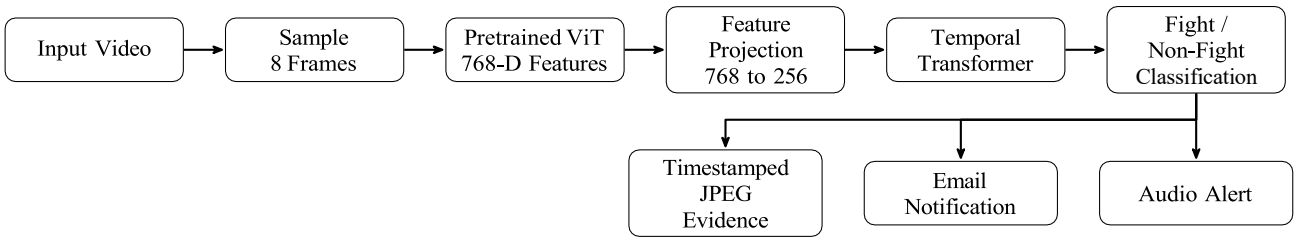


Fig. 1. Main stages of the proposed crowd violence detection system.

TABLE II
RECORDED VALIDATION CLASSIFICATION RESULTS

Class	Prec.	Recall	F1	N
Non-Fight	91.71%	94.00%	92.84%	200
Fight	93.85%	91.50%	92.66%	200
Macro avg.	92.78%	92.75%	92.75%	400
Weighted avg.	92.78%	92.75%	92.75%	400

TABLE III
VALIDATION CONFUSION MATRIX

Actual Class	Pred. Non-Fight	Pred. Fight
Non-Fight	188	12
Fight	17	183

B. Validation Results

The validation partition is balanced, containing 200 examples from each category. According to the confusion matrix, 371 of the 400 predictions agree with the reference labels, while 29 do not. This corresponds to a validation accuracy of 92.75%.

For the Non-Fight category, the classifier returned the correct label for 188 clips and marked 12 as Fight. Among Fight examples, 183 were identified correctly and 17 were missed. Those 17 false negatives matter in a screening setting because the corresponding clips would not be prioritised for human review by this model.

C. Discussion

The 92.75% figure applies to the validation clips used in this experiment. It does not establish the same level of performance for cameras or locations that were not represented in that partition. Differences in viewpoint, lighting, scene congestion, occlusion, and movements that resemble fighting could change the model's output. The webcam demonstration confirms that prediction and notification components are connected, but it does not evaluate field reliability or response latency. Those properties require separate tests using independent recordings and timed runs.

VII. CONCLUSION

This project implemented a clip classifier that combines ViT features from eight sampled frames with a separate Temporal Transformer. On the recorded set of 400 validation videos, it achieved 92.75% accuracy. A webcam prototype also connects the classifier to prediction display, smoothing, audio and email alerts, and timestamped image storage. Evaluation on independent footage and measurement of runtime behaviour remain necessary before drawing conclusions about deployment.

REFERENCES

[1] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.
[2] G. Bertasius, H. Wang, and L. Torresani, "Is Space-Time Attention All You Need for Video Understanding?" in *Proc. International Conference on Machine Learning*, vol. 139, pp. 813–824, 2021.
[3] M. Cheng, K. Cai, and M. Li, "RWF-2000: An Open Large Scale Video Database for Violence Detection," in *Proc. 25th International Conference on Pattern Recognition (ICPR)*, pp. 4183–4190, 2021, doi: 10.1109/ICPR48806.2021.9412502.
[4] A. Vaswani *et al.*, "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
[5] PyTorch Foundation, "PyTorch Documentation." [Online]. Available: https://docs.pytorch.org/docs/
[6] Hugging Face, "Transformers Documentation." [Online]. Available: https://huggingface.co/docs/transformers/
[7] OpenCV, "OpenCV Documentation." [Online]. Available: https://docs.opencv.org/
[8] P. Negre, R. S. Alonso, A. Gonza´lez-Briones, J. Prieto, and S. Rodr´ıguez-Gonza´lez, "Literature Review of Deep-Learning-Based Detection of Violence in Video," *Sensors*, vol. 24, no. 12, Art. no. 4016, Jun. 2024, doi: 10.3390/s24124016.
[9] F. J. Rendo´n-Segador, J. A. A´ lvarez-Garc´ıa, and L. M. Soria-Morillo, "Transformer and Adaptive Threshold Sliding Window for Improving Violence Detection in Videos," *Sensors*, vol. 24, no. 16, Art. no. 5429, Aug. 2024, doi: 10.3390/s24165429.
[10] M. Qaraqe *et al.*, "Crowd Behavior Detection: Leveraging Video Swin Transformer for Crowd Size and Violence Level Analysis," *Applied Intelligence*, vol. 54, pp. 10709–10730, 2024, doi: 10.1007/s10489-024-05775-6.
[11] D. He, Y. Gao, F. Jin, H. Yan, and W. Liu, "A Two-Stream Gated Network for Violence Detection Using Inflated 3D ConvNet and Transformer," in *Design Studies and Intelligence Engineering*, Frontiers in Artificial Intelligence and Applications, 2025, doi: 10.3233/FAIA250257.