DOI: https://doi.org/10.15276/aait.09.2026.03
UDC 004:83

A fragment-based spatial-temporal video analysis method for detecting anomalous violent events

Tetiana V. Normatova[1)]
ORCID: https://orcid.org/0009-0004-3503-6350; tetiana.normatova@nure.ua
Sergii V. Mashtalir[1)]
ORCID: https://orcid.org/0000-0002-0917-6622; sergii.mashtalir@nure.ua. Scopus Author ID: 36183980100
[1)] Kharkiv National University of Radio Electronic, 14, Nauky Ave. Kharkiv, 61166, Ukraine

ABSTRACT

In this work, we address the problem of violent event detection and classification in video streams under realistic computational constraints. Many safety-critical events, such as violent interactions or abnormal behavior, are characterized by short-term and spatially localized motion patterns, while the majority of video content remains static or irrelevant. Conventional deep learning approaches typically process full video frames or dense spatio-temporal representations, which leads to high computational cost and inefficient use of computational resources. We propose a fragment-based spatio-temporal video analysis method inspired by principles of video coding. Each video frame is divided into fragments, and motion activity is estimated using dense optical flow between consecutive frames. Only fragments exhibiting significant temporal changes are selected for further processing, while static regions are suppressed at an early stage. The fragment size is adaptively adjusted according to local motion intensity, allowing finer spatial resolution in dynamic regions and coarser representation in static areas. The selected fragments form a compact representation that is subsequently used for event classification via lightweight temporal aggregation. By reducing spatio-temporal redundancy prior to feature extraction, the proposed method significantly lowers computational complexity while preserving discriminative motion cues. The proposed method is evaluated on the UBI-Fights dataset, with additional training data augmentation using the Video Fight Detection (VFD2000) dataset. Experimental results demonstrate that the method achieves competitive performance with area under the receiver operating characteristic curve up to 0.72, area under the precision-recall curve up to 0.63, and binary F1-score up to 0.60, while maintaining efficient inference speed. These results indicate a favorable trade-off between accuracy and efficiency compared to dense frame-based baselines, making the method suitable for real-time and resource-constrained video analysis systems.

Keywords: video analysis; spatio-temporal video processing; event detection; anomalous behavior detection; violent event detection; fragment-based representation; adaptive fragmentation; motion-based analysis; video anomaly detection.

For citation: Normatova T. V., Mashtalir S. V. “A fragment-based spatial-temporal video analysis method for detecting anomalous violent events ”. *Applied Aspects of Information Technology.* 2026; Vol.9 No.1: 37–50. DOI: https://doi.org/10.15276/aait.09.2026.03

INTRODUCTION

Video-based recognition of violent and aggressive interactions has become an increasingly important problem in modern computer vision and intelligent surveillance systems. Physical altercations, fights, and other forms of violent behavior often emerge suddenly and evolve rapidly, posing serious risks to public safety. Early and reliable detection of such events enables timely intervention by security personnel, law enforcement, or emergency services, potentially preventing escalation and reducing harm.

Automatic detection of violent events is particularly relevant in public and semi-public environments, such as transportation hubs, stadiums, shopping centers, schools, hospitals, and urban surveillance networks. In these settings, continuous human monitoring of multiple camera feeds is impractical, while delayed response to aggressive incidents can lead to severe consequences.

Violent interactions differ from many conventional actions in that they are typically short, highly localized in space and time, and characterized by abrupt, irregular motion patterns. In real-world surveillance footage, violent events often occupy only a small portion of the frame, while the surrounding background remains static or irrelevant.

In practical surveillance systems, the majority of frames depict normal behavior, and violent interactions occur rarely and unpredictably. This class imbalance further complicates model training and evaluation, increasing the importance of efficient data representation and robust motion modeling. Therefore, there is a strong need for methods that can suppress redundant spatial information early, while preserving sensitivity to short and localized violent events.

In this work, we propose a fragment-based spatial-temporal video analysis method for detecting anomalous violent events. Inspired by principles of video coding, the method performs early motion-guided fragment selection, allocating finer spatial resolution to regions with intensive activity and coarser representation to static areas. By reducing spatial–temporal redundancy prior to feature extraction, the proposed framework enables efficient and scalable fight detection while maintaining high sensitivity to abrupt motion patterns characteristic of violent interactions.

1. RELATED WORKS

There are several approaches for violence recognition in video streams, which can be grouped as follows:

– action recognition-based approaches: early approaches to violent event and fight detection often formulate the problem as a supervised action recognition task. Classical spatio-temporal convolutional architectures such as C3D (Convolutional 3D Network) [1], [2], [3]**,** I3D (Inflated 3D Convolutional Network) [4], [5]**,** and R(2+1)D (Factorized Spatio-Temporal Convolutional Network) [6], [7] learn motion-aware features by applying 3D convolutions over short video clips. These models have demonstrated strong performance on multiple benchmarks for action and violence recognition. More recently, transformer-based models for video understanding have been proposed to improve long-range temporal modeling. Architectures such as ViViT (Video Vision Transformer) [8] process videos as sequences of patch tokens and apply self-attention across spatial and temporal dimensions. While these methods provide flexible representations, they rely on dense frame-level tokenization, which leads to high computational cost when applied to long surveillance videos;

– motion-based representations and optical flow: motion cues play a central role in recognizing violent interactions, as such events are typically characterized by abrupt and irregular movement patterns. Optical flow estimation has therefore been widely adopted in video analysis pipelines. Classical two-stream approaches combine RGB appearance with motion features extracted using optical flow networks such as PWC-Net (Pyramid, Warping, and Cost Volume Network) [9] and RAFT (Recurrent All-Pairs Field Transforms) [10]. More recent violence detection frameworks integrate dense motion representations extracted using GMFlow (Global Matching Flow) [11] together with 3D convolutional backbones to enhance sensitivity to aggressive motion patterns;

– attention mechanisms and adaptive focus: to alleviate the cost of dense spatio-temporal processing, several works incorporate attention mechanisms and adaptive focus strategies. CBAM (Convolutional Block Attention Module) [12] enhances feature representations by applying channel-wise and spatial attention to emphasize informative regions related to violent interactions. Recent two-stage violence detection frameworks combine dense motion representations with CBAM-enhanced 3D convolutional networks [13] to improve classification accuracy under weak supervision. Such attention mechanisms improve performance but operate on already extracted dense feature maps. Adaptive spatial resolution methods aim to further reduce unnecessary computation. AdaFocus (Adaptive Focusing for Efficient Video Recognition) [14], [15], [16] dynamically adjusts the spatial resolution of video regions, but still relies on an initial dense analysis stage;

– video anomaly detection and weakly supervised frameworks: An alternative formulation treats fight detection as a video anomaly detection problem, where violent interactions are considered rare abnormal events within predominantly normal video streams. Weakly supervised frameworks methods such as RTFM (Robust Temporal Feature Magnitude) [17], [18] treat violence detection as anomaly detection and improve temporal localization without frame-level labels.

Positioning of the proposed method: in dense feature extraction followed by attention-based reweighting or adaptive focusing, the proposed method performs motion-guided fragment selection prior to feature extraction. By dividing frames into fragments and selecting only motion-salient regions with adaptive spatial resolution, spatial-temporal redundancy is reduced at the input level.

2. PROBLEM STATEMENT

Detection of rare and safety-critical events in long video streams remains a challenging problem due to the high dimensionality of video data and the localized nature of target events. In many real-world scenarios, such events are characterized by short-term and spatially confined motion patterns, while the majority of video content depicts normal behavior and static background regions.

Although traffic incidents and violent interactions belong to different application domains, they share common properties from the perspective of video analysis. Both types of events occur infrequently, exhibit abrupt motion changes, and occupy only a small fraction of the spatial-temporal video volume. This makes them representative examples of motion-driven anomalies in video streams.

Existing video analysis methods typically rely on dense frame-level processing and extract spatio-temporal features from entire video frames. Such approaches lead to high computational cost and limit scalability, particularly in long-duration surveillance scenarios. While attention mechanisms and anomaly detection frameworks partially address temporal redundancy, spatial redundancy is often handled only after feature extraction.

A fragment-based spatial-temporal video analysis method was previously developed to address these challenges in the context of traffic event classification. The method employs motion-guided adaptive fragmentation to reduce spatial-temporal redundancy at an early stage and has demonstrated effectiveness in detecting traffic incidents under computational constraints.

The aim of this work is to investigate the applicability of the same fragment-based spatial-temporal analysis framework originally built for traffic event analysis [19] to the problem of violent event detection that prioritizes motion-salient regions while maintaining competitive classification quality under real-world resource constraints.

The main tasks of the study are as follows.

1. Use the developed architecture for training and evaluating the proposed method on a violence datasets (UBI-Fights [20] and VFD2000 [21]) and assess the impact of additional training augmentation data, using a fixed validation/test protocol to prevent leakage.

2. Compare the proposed method against other methods in terms of both effectiveness (classification metrics) and efficiency/complexity.

3. Provide qualitative validation of the proposed method by analyzing temporal stability of window-level predictions and the alignment between motion activity and adaptive patch allocation, including threshold-sensitivity analysis.

The proposed method is expected to achieve reliable violent event classification while maintaining efficient inference speed, providing an explicit accuracy-efficiency trade-off relative to dense baselines. Qualitative analysis should confirm stable temporal confidence and motion-aligned adaptive patch allocation, supporting interpretability of the model's decisions.

3. PROPOSED METHOD

This work builds upon our previously introduced adaptive sparse video representation framework, originally developed for traffic accident analysis, and extends it to the task of violence detection in video sequences. While the core idea of motion-driven fragment selection is preserved, the proposed method is adapted to handle the distinct spatio-temporal characteristics of human violent interactions.

Overview. Given an input video clip, the method processes a short temporal window of frames and represents each frame as a sparse set of adaptive patches selected according to local motion activity. The extracted patches are embedded and aggregated using transformer-based attention mechanisms to model both spatial relationships within frames and temporal dynamics across frames. The resulting video-level representation is used for binary classification into violent and non-violent categories.

Motion-Guided Adaptive Patch Selection. Unlike traffic scenes, where motion patterns are often structured and constrained by road geometry, violent interactions in videos exhibit highly irregular and localized motion patterns. To capture such dynamics efficiently, we employ dense optical flow estimation to measure frame-to-frame motion intensity.

Each frame is partitioned into candidate regions, from which patches of varying spatial resolution are selected:

- smaller patches for regions with high motion activity,
- larger patches for low-motion or static areas.

This adaptive strategy allows the method to focus computational resources on regions most relevant to violent behavior, while suppressing background motion and irrelevant scene elements.

Spatial Patch Encoding. For each selected patch, a visual embedding is computed and augmented with positional information. A spatial transformer encoder processes the set of patch embeddings within a frame, enabling the model to capture contextual relationships between different motion regions. A frame-level classification token aggregates spatial information into a compact representation. This design allows the model to reason about interactions between multiple moving actors without requiring full-frame processing.

Temporal Modeling of Video Dynamics. To capture temporal dependencies characteristic of violent actions, frame-level representations are processed by a temporal aggregation module. The model analyzes the evolution of motion patterns across time windows, allowing it to distinguish sustained violent interactions from short, non-violent movements. Multiple temporal windows are evaluated, and their predictions are combined using a robust aggregation strategy to obtain a final video-level decision.

Classification Head. The aggregated representation is passed to a lightweight classification head, producing a probability estimate for the presence of violent activity in the input video.

4. POST-PROCESSING

Due to the temporal variability of violent events, a single video clip may contain both neutral and aggressive segments. As a result, individual frame-level or window-level predictions may be unstable and insufficient for reliable video-level classification. To address this issue, a post-processing stage is employed to aggregate intermediate predictions into a single, robust video-level decision. This post-processing step plays a crucial role in violence detection scenarios, where the positive class is typically underrepresented and the choice of decision threshold has a significant impact on the trade-off between false alarms and missed events. By refining and consolidating the outputs of the temporal transformer, the proposed post-processing strategy improves the overall stability and reliability of the final classification results.

The input video is divided into overlapping temporal windows of fixed length. For each temporal window, the method processes a uniformly sampled subset of frames and outputs raw prediction scores that converted to probabilities via softmax. The value $p_i \in [0,1]$ denotes the predicted probability of the violent class for the i-th window. Each p_i is computed independently, resulting in a sequence of window-level probabilities $\{p_1,\ p_2, \dots, p_N\}$.

To obtain a robust video-level prediction, an aggregation strategy is applied to combine the window-level probabilities produced by the temporal transformer. Due to the temporal variability of violent events, a single video may contain both neutral and aggressive segments, which makes direct averaging of window scores sensitive to noise, camera motion, or partial occlusions. Instead of using a simple mean (1), a trimmed mean aggregation (2) is employed. Unlike a simple average, which is sensitive to outliers and short-term noise, the trimmed mean reduces the influence of abnormal windows caused by camera motion, partial occlusions, or brief non-representative segments. Specifically, the highest and lowest α % of window-level probabilities are discarded, and the remaining values are averaged. By suppressing extreme outlier scores, the trimmed mean reduces the influence of spurious or noisy windows while preserving consistent evidence of violent activity across the clip. This strategy improves the stability and reliability of the final video-level decision, particularly in scenarios with rare positive events and uneven temporal distribution of violence.

Formally, the video-level probability is computed as

$$p_{video} = \frac{1}{|S|} \sum_{i \in S} p_i, \qquad (1)$$

where S denotes the set of retained window indices after trimming.

Equivalently, let p_i denote the sorted window-level probabilities, N be the total number of windows, and k = $\lfloor \alpha N \rfloor$. The trimmed mean can then be expressed as

$$\hat{p} = \frac{1}{N-2k} \sum_{i=k+1}^{N-k} p_i\ , \qquad (2)$$

This formulation explicitly removes the highest-confidence false positives and lowest-confidence neutral segments, while preserving the dominant temporal patterns of violent behavior.

Finally, a threshold-based decision rule (3) is applied:

$$\hat{y} = \begin{cases} 1, if\ p_{video}\ \geq\ \tau\ , \\ \ 0, otherwise \end{cases} \qquad (3)$$

where threshold τ is selected on the validation set to maximize the F1-score.

The proposed post-processing scheme improves the stability of predictions and enables flexible trade-offs between precision and recall depending on the application requirements.

5. TRAINING

The proposed method is trained on the UBI-Fights training split with additional training augmentation from VFD2000, while all evaluation is performed exclusively on the fixed UBI-Fights validation split.

The use of multiple datasets during training allows the method to learn a broader spectrum of motion patterns and interaction dynamics, thereby improving generalization across different scenes, camera viewpoints, and recording conditions.

UBI-Fights Dataset. The UBI-Fights dataset is a widely used benchmark for video-based fight detection. It consists of short video clips depicting violent (fight) and non-violent (normal) scenes captured in real-world surveillance-like conditions. The dataset includes variations in illumination, crowd density, camera angles, and background complexity, making it suitable for evaluating the robustness of violence detection methods.

Each video is annotated at the clip level with a binary label indicating the presence or absence of a fight. Due to the inherent nature of surveillance data, the dataset exhibits a moderate class imbalance, with normal scenes being more frequent than fight scenes.

VFD2000 Dataset. To increase the diversity and volume of training data, we additionally incorporate the VFD2000 (Violence Fight Dataset 2000). This dataset contains approximately 2000 video clips collected from different sources, including surveillance footage and staged scenarios, and covers a wide range of violent interactions such as punches, kicks, pushes, and group fights, as well as non-violent activities.

Compared to UBI-Fights, VFD2000 introduces greater variability in:

- camera motion and resolution;
- scene context (indoor vs. outdoor);
- number of participants;
- duration and temporal structure of violent events.

The inclusion of VFD2000 enables the proposed method to better capture diverse motion dynamics and reduces overfitting to dataset-specific visual patterns.

Dataset Merging Strategy. The training set is formed by merging the training split of UBI-Fights with the full VFD2000 dataset, while the validation set is kept fixed and identical to the original UBI-Fights validation split. This strategy ensures a fair and controlled evaluation protocol while allowing the model to benefit from additional training data.

Formally:

- Training set:
- UBI-Fights (training split);
- VFD2000 (all available clips).
- Validation set: UBI-Fights (validation split only).

By keeping the validation data unchanged, all reported evaluation metrics remain directly comparable to existing works that rely solely on the UBI-Fights benchmark.

Preprocessing and Frame Sampling. Each video clip is uniformly sampled to extract a fixed number of frames T (in our experiments, T_{sample}=12). In our experiments, we set the number of sampled frames to T_{sample} = 12, which provides a balance between capturing short-term motion bursts typical for violent interactions and maintaining computational efficiency. Empirically, using fewer frames (e.g., T = 8) same as for the accidents detection in traffic videos led to unstable recall, while larger values (T ≥ 16) did not yield consistent improvements. Uniform sampling ensures that both short and long clips are represented consistently, capturing key temporal moments without introducing redundancy.

All frames are resized to a spatial resolution of *128×128* pixels which was found sufficient to preserve motion cues relevant for violent activity detection while keeping the computational cost low. RGB frames are used for patch extraction and feature embedding, while grayscale versions are generated for optical flow computation.

Dense optical flow is estimated between consecutive frames using the Farnebäck algorithm [22]. The resulting flow magnitude maps are normalized and used to guide adaptive patch selection.

Motion-Guided Patch Generation. For each pair of consecutive frames, dense optical flow is computed using the Farnebäck algorithm. The magnitude of the optical flow serves as a motion saliency map that guides adaptive patch selection. For each frame, a base grid with step size b = 8 is applied. Motion intensity within each grid cell is estimated using the average magnitude of optical flow. Based on quantile thresholds of motion magnitude, each frame is partitioned into non-overlapping patches of different sizes:

- small patches (8×8 pixels) for regions with high motion intensity,
- medium patches (16×16 pixels) for moderate motion,
- large patches (32×32 pixels) for low-motion or static regions.

To preserve global context, a small number of fixed grid-based patches is additionally included in each frame. The selected patches are non-overlapping, resized to a fixed embedding size, and converted into feature vectors. This strategy reduces effective spatial sampling density in low-motion regions by approximately 30–60% (relative to uniform dense patching), depending on scene dynamics.

Model Optimization. The model is trained from scratch using the AdamW optimizer [23], which provides stable convergence for transformer-based architectures with an initial learning rate of 1.5×10^{-4} and a cosine learning rate scheduler with warmup is employed to gradually adjust the learning rate during training. . Due to the strong class imbalance in violence detection datasets, class-weighted binary cross-entropy loss is used. The positive class (fight) is assigned a higher weight proportional to the inverse class frequency. Additionally, weighted random sampling is applied during training to ensure balanced mini-batches. The batch size is selected based on available hardware resources. Training is performed for a fixed number of epochs with early stopping based on the best F1-score on the validation set to prevent overfitting.

Regularization and Training Stability. To improve generalization, dropout is applied in both the patch embedding network and the transformer layers. Gradient clipping is used to stabilize training, particularly when processing clips with strong motion bursts. All experiments are conducted with fixed random seeds to ensure reproducibility. Model checkpoints corresponding to the best validation F1-score are saved and used for final evaluation.

Implementation Details. All experiments are conducted using the PyTorch framework. Training and inference are performed on both CPU and GPU environments to evaluate computational efficiency. The adaptive patch selection and optical flow computation are optimized to minimize overhead, allowing the model to operate under realistic resource constraints.

Post-training Aggregation. During evaluation, each video is processed using multiple overlapping temporal windows. The final video-level prediction is obtained by aggregating window-level probabilities using a trimmed mean, which suppresses outliers and stabilizes predictions for noisy clips.

Summary. By combining UBI-Fights and VFD2000 during training while preserving a fixed validation protocol, the proposed method benefits from increased data diversity without compromising evaluation fairness. This training strategy, together with motion-guided adaptive patching and factorized spatial-temporal transformer architecture, enables the model to achieve a strong balance between accuracy, robustness, and computational efficiency for video-based violence detection.

6. EXPERIMENTS

The experimental evaluation is conducted to assess the effectiveness, robustness, and computational efficiency of the proposed method when transferred from traffic accident analysis to the task of violence detection in video sequences. All experiments are performed on publicly available datasets and follow a consistent evaluation protocol.

The proposed method is trained primarily on the UBI-Fights dataset, which consists of short video clips depicting violent (fight) and non-violent (normal) scenes captured in real-world surveillance conditions. To improve the diversity of motion patterns and visual contexts, the training set is extended with additional samples from the VFD2000 dataset, which contains videos of violent and non-violent human interactions recorded in uncontrolled environments.

All evaluation is performed exclusively on the fixed validation split of the UBI-Fights dataset, which is kept unchanged across all experiments to avoid data leakage and to ensure comparability of results. All frames are resized to a spatial resolution of 128 × 128 pixels. Each video is divided into overlapping temporal windows of length $T = 12$ frames. Optical flow between consecutive frames is computed using the Farnebäck algorithm and serves as a motion saliency map for adaptive patch selection. For each frame, at most $K = 64$ patches are extracted, combining motion-driven adaptive patches with a small set of fixed grid patches to preserve global context.

The model is trained using the AdamW optimizer with a cosine learning rate schedule. Class imbalance is handled via balanced sampling. The best model is selected based on the highest F1-score on the validation split. All experiments are conducted using fixed random seeds to improve reproducibility.

For baseline comparison, we reproduced several representative methods including Two-Stream CNN+LSTM, Violence Flow CNN, and 3D CNN using publicly available implementations. All models were evaluated on the same datasets. Due to differences in original implementations and training pipelines, exact equivalence of training conditions cannot be fully guaranteed.

To comprehensively assess the performance of the proposed method, multiple evaluation metrics are employed, reflecting both threshold-dependent and threshold-independent aspects of the violence detection task. Since violent events constitute a minority class in real-world video streams, relying solely on accuracy is insufficient and may lead to overly optimistic conclusions. Since the dataset is imbalanced we evaluate the proposed method using Accuracy, Binary F1-score [24], Macro-F1 [25], ROC-AUC [26], and PR-AUC [27], which are commonly used in violence detection tasks. Since datasets are typically imbalanced, threshold-independent metrics (ROC-AUC, PR-AUC) and class-balanced metrics (Macro-F1) are particularly important.

Binary F1-score is computed as the harmonic mean of precision and recall. Macro-F1 represents the unweighted average of class-wise F1 scores.

ROC-AUC measures the discriminative ability of the classifier across all decision thresholds, while PR-AUC better reflects performance on the positive (violent) class under class imbalance.

In addition to accuracy-related metrics, computational efficiency is analyzed in terms of model size, number of processed patches per frame, and inference speed measured in frames per second (FPS). These metrics are critical for practical violence detection systems, which are often deployed under limited computational resources and strict latency constraints. Unlike full-frame video models that process dense spatial representations, the proposed method operates on a sparse set of motion-guided patches, reducing redundant background detail and enabling content-adaptive spatial sampling under a fixed token budget. The number of processed patches per frame directly reflects the computational load of the spatial transformer, while the total number of model parameters indicates memory consumption and scalability. Inference speed is measured on both CPU and GPU to assess the feasibility of real-time or near real-time deployment in surveillance scenarios. This comprehensive evaluation allows analyzing the trade-off between detection performance and computational efficiency across different methods.

The proposed method is compared against representative state-of-the-art approaches for video-based violence detection (Table 1, and Table 2). For a fair comparison, all methods are evaluated on the same fixed UBI-Fights validation split; VFD2000 is used only as an additional source of training augmentation where explicitly stated.

VFD2000 datasets. The selected baselines cover the main methodological directions commonly used in this domain, including two-stream convolutional neural networks that process RGB frames and optical flow separately, optical-flow-based CNN architectures that explicitly model motion patterns, 3D convolutional networks that jointly capture spatial and temporal information, as well as pretrained spatio-temporal models designed for action recognition. In particular, the comparison includes two-stream CNN models with temporal aggregation via recurrent layers, optical-flow-driven convolutional approaches focusing on motion saliency, 3D CNN architectures operating on short video clips, and pretrained spatio-temporal networks such as I3D and related variants. These methods represent widely adopted baselines in the violence detection literature and provide a meaningful reference for evaluating both classification performance and computational efficiency.

To provide a transparent efficiency assessment, inference speed is measured as end-to-end throughput including all processing stages: optical flow computation, adaptive patch selection, backbone forward pass, and temporal aggregation. FPS is defined as the number of raw input frames processed per second by the full pipeline.

Experiments are conducted on an Intel Xeon @ 2.20 GHz (2 vCPUs, 12 GB RAM) and an NVIDIA Tesla T4 GPU (16 GB). Under this configuration, the proposed method achieves approximately 12 FPS on CPU and up to 180 FPS on GPU, depending on video length and window configuration. This enables near real-time processing for surveillance-oriented applications.

During inference, each video is processed using sliding temporal windows of length T_{window} = 48 frames with a stride of 8 frames, which corresponds to approximately 83 % temporal overlap between consecutive windows. Such dense overlap improves robustness to temporal misalignment, since violent events may occur at arbitrary positions within a video.

The backbone model is trained to operate on sequences of fixed length T_{sample}= 12, each 48-frame window is uniformly sampled into 12 frames before adaptive patch extraction and tokenization.

Each temporal window produces an independent violence probability. To obtain a stable video-level prediction, we apply trimmed mean aggregation over window-level probabilities. In the validation protocol, $N \approx 9$ windows are sampled per video, and the highest and lowest predictions are discarded (trim = 1), which corresponds to removing approximately 22 % of extreme values ($\alpha \approx 0.22$). The remaining probabilities are then averaged to obtain the final video-level score. This aggregation strategy reduces sensitivity to occasional noisy windows caused by camera motion, abrupt cuts, or background dynamics, and provides more stable predictions compared to simple mean or max pooling.

Despite its lightweight design and reduced effective spatial sampling density, the proposed model demonstrates strong discriminative capability. On the UBI-Fights validation split, the proposed method achieves 0.856 accuracy, 0.757 Macro-F1, and 0.626 PR-AUC, showing strong class-balanced performance (Macro-F1) and a favorable efficiency–accuracy trade-off compared to representative CNN-based and two-stream baselines.

Table 1. **Methods classification metrics (part 1)**

Method	Accuracy	Precision	Recall
Adaptive Spars ViT (our method)	0.856	0.759	0.5
Two-Stream CNN + LSTM	0.89	0.72	0.68
Violence Flow CNN	0.91	0.75	0.71
3D CNN	0.92	0.78	0.74

Source: **compiled by the authors**

Table 2. **Methods classification metrics (part 2)**

Method	Binary F1	Macro – F1	ROC-AUC	PR-AUC
Adaptive Spars ViT (our method)	0.603	0.757	0.722	0.626
Two-Stream CNN + LSTM	0.69	0.7	0.88	0.68
Violence Flow CNN	0.73	0.73	0.9	0.71
3D CNN	0.76	0.76	0.92	0.74

Source: **compiled by the authors**

These results, summarized in Table 3, confirm that adaptive patch selection combined with temporal aggregation provides a favorable trade-off between accuracy and efficiency for violence detection in video streams.

Table 3. **Methods complexity and efficiency metrics**

Method	Patches per frame (b=8, input 128×128)	Number of parametes, millions	Frames per second (CPU/GPU)
Adaptive Spars ViT (our method)	up to 64 patches per frame (8x8/16x16/32x 32)	~4.2	~12/~180
Two-Stream CNN + LSTM	Full frame	~45	~8/~15
Violence Flow CNN	Full frame	~32	~10/~20
3D CNN	Full frame	~60	~4/~8

Source: **compiled by the authors**

Several limitations should be acknowledged. First, the method relies on classical optical flow estimation, which introduces additional preprocessing cost and may limit efficiency on low-power devices. Second, performance depends on window length and aggregation parameters, which are selected on the validation set. Third, reproducing all baselines under fully identical conditions is difficult due to differences in published implementations and training recipes; therefore, some variation in fairness of comparison may remain. These factors are reported transparently to ensure correct interpretation of the experimental results.

The proposed method includes a motion estimation stage prior to feature extraction, which introduces additional preprocessing overhead compared to standard feed-forward architectures. Based on empirical runtime profiling, optical flow computation accounts for approximately 20-30 % of total inference time, while adaptive patch selection contributes about 10-20 %, with the backbone forward pass remaining the single most time-consuming stage. The method operates efficiently on standard CPUs and GPUs; however, performance may decrease on low-power or embedded platforms, such as smart cameras or edge-based CCTV devices that process video directly on-device without dedicated acceleration for motion estimation, where this stage can become the primary bottleneck.

7. RESULTS AND DISCUSSIONS

This section analyzes the qualitative behavior of the proposed Adaptive-Sparse-ViT model in the context of video-based violence detection, emphasizing the role of motion-driven representation learning. Visual examples are provided to demonstrate how motion information, estimated through optical flow, influences the spatial sampling strategy of the model. In particular, regions characterized by pronounced human motion are represented using a denser set of small patches, while areas with limited or no motion are covered by fewer, larger patches. Such a mechanism enables the model to focus on informative spatial regions while avoiding unnecessary computation on static background content.

An example of the optical flow magnitude visualization is shown in Fig. 1, where motion intensity is superimposed on the original video frame.

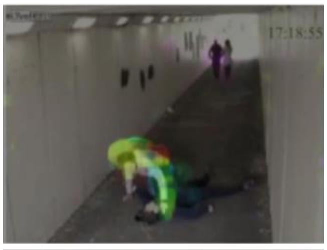

Fig. 1. **Optical flow visualization for a violence scene**
Source: **compiled by the authors**

Fig. 2 illustrates the result of the adaptive motion-guided patch selection applied to the same video frame. The colored rectangular regions correspond to patches of different spatial resolutions selected by the model. Smaller patches are densely concentrated around areas with active human motion, such as the interacting subjects in the center of the scene, allowing the model to capture fine-grained spatial details of potentially violent actions. In contrast, larger patches are assigned to static background regions, including walls and corridors, where detailed spatial modeling is less critical. This heterogeneous patch allocation demonstrates how the proposed method dynamically balances spatial precision and computational efficiency by adjusting patch granularity according to local motion intensity.

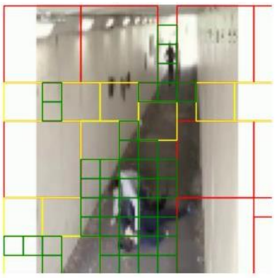

Fig. 2. **Adaptive patch grid for a violence scene based on motion intensity**
Source: **compiled by the authors**

To better understand the decision-making process of the proposed Adaptive-Sparse-ViT model and to provide qualitative evidence supporting the quantitative evaluation results, we analyze the temporal behavior of window-level predictions together with the spatial distribution of adaptive patches. Such an analysis allows us to assess not only the final classification outcome but also the stability of model confidence over time and the correspondence between motion dynamics and patch allocation.

The Fig. 3 illustrates the temporal evolution of window-level violence probabilities produced by the proposed method for a video containing a fight. Each point corresponds to the predicted probability p (fight) for an individual temporal window, plotted as a function of the window start time. As shown in the figure, the model consistently assigns high probability values in the range of 0.60-0.67 across most of the video duration, indicating stable and confident detection of violent activity.

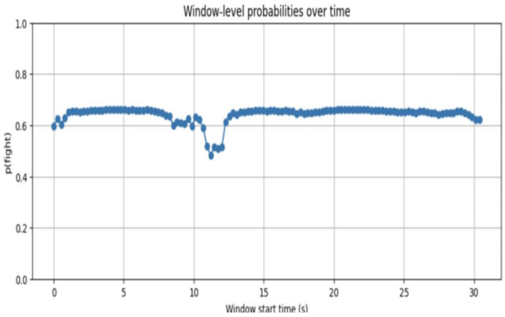

Fig. 3. **Window-level violence probabilities p(fight) over time, showing sustained high confidence across most temporal windows.**
Source: **compiled by the authors**

A brief drop in probability is observed around the middle of the sequence (approximately 10-12 seconds), where the score temporarily decreases to about 0.50. This local minimum corresponds to a short interval in which aggressive motion is partially occluded or momentarily less pronounced (e.g., changes in camera viewpoint, body overlap, or pauses between actions). Importantly, the model quickly recovers after this interval and resumes high-confidence predictions, demonstrating robustness to short-term ambiguities and temporal noise.

The Fig. 4 visualizes the effect of different decision thresholds on the same window-level predictions. Horizontal dashed lines indicate commonly used thresholds (τ = 0.25, 0.40, and 0.50). For this violent video, the vast majority of windows remain above all three thresholds, including the most conservative one (τ = 0.50). As a result, nearly all windows are classified as positive, and the final video-level decision remains unchanged across a wide range of threshold values.

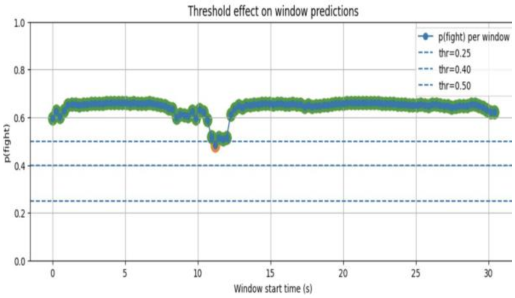

Fig. 4. **Effect of different decision thresholds on window-level predictions, demonstrating stability of the model output across a wide threshold range**
Source: **compiled by the authors**

This behavior highlights two important properties of the proposed method. First, the window-level predictions are temporally consistent, with limited variance despite local fluctuations in motion intensity. Second, the model exhibits low sensitivity to threshold selection in the presence of sustained violent activity, which is crucial for real-world surveillance scenarios where fixed thresholds must operate reliably under varying conditions.

Overall, these results confirm that the proposed adaptive sparse representation, combined with temporal aggregation, enables stable violence detection over time while remaining robust to short-term motion irregularities and local uncertainty.

Below we illustrate a qualitative example of the proposed Adaptive-Sparse-ViT model applied to a video fragment containing a violent interaction. The visualization corresponds to an explicitly selected frame (frame index 933), which lies within a temporal window spanning frames 912–960 and was used for model inference.

The Fig. 5 shows the original input frame, where a violent episode is clearly visible as a person is lying on the ground following an aggressive action.

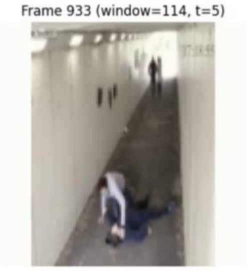

Fig. 5. **Original video frame extracted from the most confident temporal window**
***Source:* compiled by the authors**

The Fig. 6 presents the corresponding token density heatmap constructed from the spatial distribution of selected patches. High-density regions (highlighted in red and yellow) concentrate around the interacting individuals, indicating that the model allocates a larger number of tokens to areas exhibiting strong motion and semantic relevance. In contrast, static background regions are assigned fewer tokens, resulting in low-density areas (blue).

Fig. 7 overlays the adaptive patch grid on top of the input frame together with the token density map. Motion-adaptive patches of smaller spatial size are redominantly placed over the actors and regions with rapid motion, while larger grid-based context patches cover the surrounding background. This behavior demonstrates that the proposed patching strategy successfully focuses computational resources on dynamically salient regions while preserving global scene context.

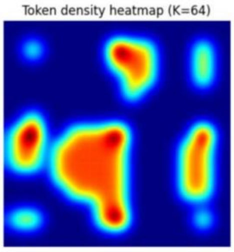

Fig. 6. **Token density heatmap corresponding to the selected frame, illustrating the spatial distribution of selected patches**
***Source:* compiled by the authors**

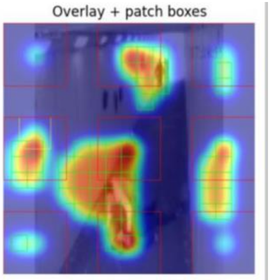

Fig. 7. **Final visualization with adaptive patch boxes, where smaller patches are concentrated in regions of intense motion related to the violent interaction, while larger patches cover static background regions**
***Source:* compiled by the authors**

Overall, this example confirms that the model's adaptive patch selection mechanism aligns well with the spatial structure of violent events, providing an interpretable link between motion cues, token allocation, and the resulting high window-level violence probability (p = 0.621) for the selected temporal segment.

To analyze the behavior of the proposed model in the absence of violent activity, we consider a video sequence containing only normal interactions. Fig. 8 presents the window-level fight probabilities p(fight), where each point corresponds to the model confidence for a temporal window and values lie in the range [0,1]. Throughout the entire video, the predicted probabilities remain low, typically around 0.12-0.20, indicating weak evidence of violent behavior.

Fig. 9 illustrates the effect of different decision thresholds applied to the window-level probabilities. The dashed horizontal lines correspond to threshold values of 0.25, 0.40, and 0.50. Only a small number of windows exceed the lowest threshold, while no windows surpass higher thresholds, demonstrating that the model does not produce confident false positives in non-violent scenes.

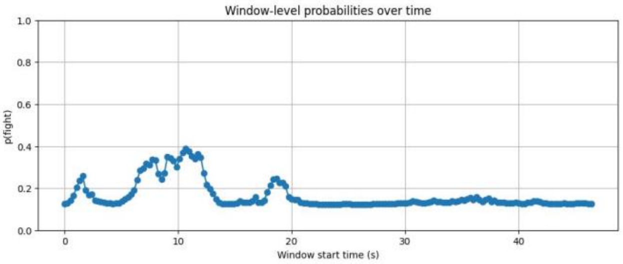

Fig. 8. **Window-level fight probability over time, showing consistently low confidence scores across the entire video**
***Source:* compiled by the authors**

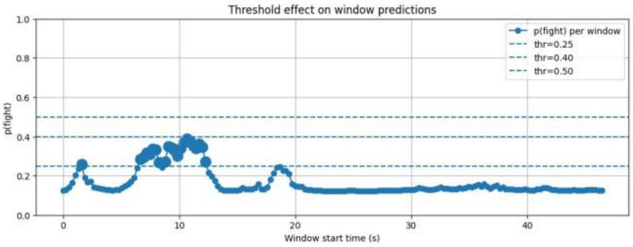

Fig. 9. **Effect of different decision thresholds on window-level predictions, illustrating the suppression of false positives in normal scenes**
***Source:* compiled by the authors**

The selected input frame Fig. 10 corresponds to a non-violence action. The token density heatmap on Fig. 11 shows that motion-adaptive tokens are distributed over regions with moderate motion, such as walking or hand gestures, while static background areas receive little attention.

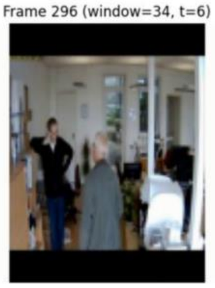

Fig. 10. **Example input frame extracted from a representative window without violent activity**
***Source:* compiled by the authors**

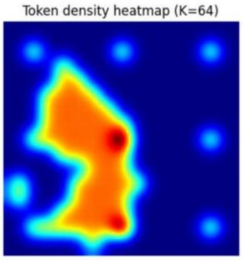

Fig. 11. **Token density heatmap highlighting the spatial distribution of motion-adaptive tokens, with attention focused on moderate human movement regions**
***Source:* compiled by the authors**

The overlay visualization on Fig. 12 confirms that fine-grained patches are sparsely allocated and do not form concentrated clusters, reflecting the absence of violent dynamics.

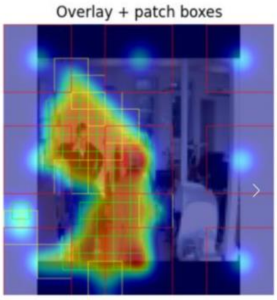

Fig. 12. **Overlay of adaptive patch boxes on the input frame, where fine-grained patches are sparsely assigned to moving subjects and coarse patches cover static background areas**
***Source:* compiled by the authors**

Thus, these results demonstrate that the proposed adaptive patching strategy not only enhances sensitivity to violent interactions but also preserves stability and precision in non-violent scenarios, which is crucial for real-world deployment.

CONCLUSIONS AND PROSPECTS OF FURTHER RESEARCH

This work extends the study of adaptive fragment-based video analysis previously introduced for traffic accident detection to a new application domain - violence detection in video sequences. The experiments confirm that the proposed motion-guided sparse tokenization strategy is not limited to a specific task, but represents a more general approach to efficient video understanding in scenarios where critical events are temporally sparse and spatially localized.

The proposed method combines motion-driven adaptive patch selection with a lightweight spatio-temporal transformer architecture, enabling the model to focus computational resources on semantically meaningful regions while avoiding redundant processing of static background areas. The use of optical flow-based motion estimation allows dynamic adjustment of patch sizes, preserving fine details in regions of intense activity and applying coarser representations in low-motion regions. Experimental evaluation on the UBI-Fights dataset and the merged UBI-Fights + VFD2000 dataset demonstrates that the method achieves a favorable balance between recognition quality and computational efficiency, with stable ROC-AUC and PR-AUC scores and near real-time inference performance.

An important outcome of this research is the demonstration of the method's generalization capability across domains. While the original study focused on traffic video analysis, the current work confirms that the same methodological principles remain effective for detecting violent behavior in unconstrained surveillance videos. This supports the interpretation of the proposed method as a domain-agnostic framework for sparse-event video understanding.

Despite these advantages, several limitations remain. The current implementation relies on heuristic motion estimation based on classical optical flow algorithms and manually defined quantile thresholds for patch selection. Although effective, this design introduces additional preprocessing overhead and limits the adaptability of the model in scenes with subtle or ambiguous motion patterns. Furthermore, the patch selection mechanism is not learned jointly with the classifier, which restricts the potential performance improvements that could be achieved through end-to-end optimization.

Future research will therefore focus primarily on advancing the method itself rather than further extending it to additional application domains. A promising direction is the development of fully learnable adaptive tokenization mechanisms, where the model automatically learns to allocate computational resources through attention-based or differentiable patch selection modules. Another important direction is the integration of faster and more robust motion representations, including modern neural optical flow models or learned motion encoders trained jointly with the classification network. Additional improvements may involve exploring reinforcement learning-based token budget control, incorporating multi-scale temporal attention mechanisms, and performing a more formal analysis of the trade-off between computational efficiency and recognition accuracy.

Overall, the conducted studies suggest that adaptive motion-guided tokenization represents a promising foundation for building efficient and generalizable video understanding systems, and the presented work forms a solid basis for further development toward fully learnable, resource-aware video transformers.

ACKNOWLEDGEMENTS

The results of this research were obtained under the international research project INITIATE under the grant No.101136775-HORIZON-WIDERA-2023-ACCESS-03.

REFERENCES

1. Mahmoodi, J., Nezamabadi-pour, H. & Abbasi-Moghadam, D. "Violence detection in videos using interest frame extraction and 3D convolutional neural network". *Multimedia Tools and Applications*. 2022, https://www.scopus.com/authid/detail.uri?authorId=27367449500. DOI: https://doi.org/10.1007/s11042-022-12532-9.

2. Su, J., et al. "Violence detection using 3D convolutional neural networks". *18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. Madrid, Spain. 2022. DOI: https://doi.org/10.1109/avss56176.2022.9959393.

3. Tran, D., et al. "Learning spatiotemporal features with 3D convolutional networks". *IEEE International Conference on Computer Vision (ICCV*). Santiago, Chile. 2015. DOI: https://doi.org/10.1109/iccv.2015.510.

4. Freire-Obregón, D., et al. "Inflated 3D ConvNet context analysis for violence detection". *Machine Vision and Applications*. 2021; 33 (1), https://www.scopus.com/authid/detail.uri?authorId=23396618800. DOI: https://doi.org/10.1007/s00138-021-01264-9.

5. Pan, J., et al. "An improved two-stream inflated 3D ConvNet for abnormal behavior detection". *Intelligent Automation & Soft Computing*. 2021; 29 (3): 673–688. DOI: https://doi.org/10.32604/iasc.2021.020240.

6. Wang, Y., et al. "Action recognition in videos with spatio-temporal fusion 3D convolutional neural networks". *Pattern Recognition and Image Analysis*. 2021; 31 (3): 580–587. DOI: https://doi.org/10.1134/s105466182103024x.

7. Murugan, N. & Sathasivam, S. "Real-time human action recognition by using R(2+1)D convolutional neural network". *3rd International Conference on Artificial Intelligence for Internet of Things (AIIoT).* Vellore, India. 2024. DOI: https://doi.org/10.1109/aiiot58432.2024.10574741.

8. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M. & Schmid, C. "ViViT: A video vision transformer". *ICCV*. 2021, https://www.scopus.com/authid/detail.uri?authorId=15765012100. DOI: https://doi.org/10.48550/arXiv.2103.15691.

9. Sun, D., et al. “PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume”. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* Salt Lake City, UT, USA. 2018. DOI: https://doi.org/10.1109/cvpr.2018.00931.

10. Teed, Z. & Deng, J. “RAFT: Recurrent all-pairs field transforms for optical flow”. *Computer Vision – ECCV*, 2020. p. 402–419, https://www.scopus.com/authid/detail.uri?authorId=57219544640. DOI: https://doi.org/10.1007/978-3-030-58536-5_24.

11. Xu, H., et al. “GMFlow: Learning optical flow via global matching”. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* New Orleans, LA, USA. 2022. DOI: https://doi.org/10.1109/cvpr52688.2022.00795.

12. Woo, S., et al. “CBAM: Convolutional block attention module”. *Computer Vision – ECCV*. 2018. p. 3–19. DOI: https://doi.org/10.1007/978-3-030-01234-2_1.

13. Mahmoud, M., et al. “Two-stage video violence detection framework using GMFlow and CBAM-enhanced ResNet3D”. *Mathematics*. 2025; 13 (8): 1226. DOI: https://doi.org/10.3390/math13081226.

14. Wan, Y., et al “Adaptive focus for efficient video recognition”. *IEEE/CVF International Conference on Computer Vision (ICCV).* Montreal, QC, Canada. 2021. DOI: https://doi.org/10.1109/iccv48922.2021.01594.

15. Wang, Y., et al. “AdaFocus V2: End-to-End Training of spatial dynamic networks for video recognition”. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* New Orleans, LA, USA. 2022. DOI: https://doi.org/10.1109/cvpr52688.2022.01943.

16. Wang, Y., et al. “Uni-AdaFocus: Spatial-temporal dynamic computation for video recognition”. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2024. p. 1–18. DOI: https://doi.org/10.1109/tpami.2024.3514654.

17. Qi, Z., et al. “Weakly supervised two-stage training scheme for deep video fight detection model”. *IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI).* Macao, China. 2022. DOI: https://doi.org/10.1109/ictai56018.2022.00105.

18. Peng, S., et al. “Weakly-supervised video anomaly detection via temporal resolution feature learning”. *Applied Intelligence*. 2023. DOI: https://doi.org/10.1007/s10489-023-05072-8.

19. Normatova, T. V. & Mashtalir, S. V. “Road traffic accident classification using a sparse videotransformer and adaptive fragmentation”. *Herald of Advanced Information Technology*. 2025; 8 (4): 464–475, https://www.scopus.com/authid/detail.uri?authorId=36183980100.
DOI: https://doi.org/10.15276/hait.08.2025.29

20. “UBI_Fights dataset”. – Available from: https://www.kaggle.com/datasets/intissarziani/ubi-fightsall. – [Accessed: Jan, 2025].

21. “VFD2000 dataset”. – Available from: https://github.com/Hepta-Col/VideoFightDetection fightsall. – [Accessed: Jan, 2025].

22. “Farneback Algorithm”. – Available from: https://medium.com/pythons-gurus/farneback-algorithm-50682b8aa2eb. – [Accessed: Jan, 2025]

23. Zhuang, Z., Liu, M., Cutkosky, A. & Orabona, F. “Understanding AdamW through Proximal methods and scale-freeness. Transactions on machine learning research”. *arXiv*. 2022. DOI: https://doi.org/10.48550/arXiv.2202.00089.

24. “F1 score in machine learning”. – Available from: https://www.geeksforgeeks.org/machine-learning/f1-score-in-machine-learning. – [Accessed: Jan, 2025].

25. “Micro, Macro & Weighted Averages of F1 Score, Clearly Explained”. – Available from: https://towardsdatascience.com/micro-macro-weighted-averages-of-f1-score-clearly-explained-b603420b292f. – [Accessed: Jan, 2025].

26. “AUC ROC Curve in Machine Learning”. – Available from: https://www.geeksforgeeks.org/machine-learning/auc-roc-curve. – [Accessed: Jan, 2025].

27. “What Is PR AUC?” – Available from: https://arize.com/blog/what-is-pr-auc. – [Accessed: Jan, 2025].

Received 05.01.2026
Received after revision 16.02.2026
Accepted 19.02.2026

DOI: https://doi.org/10.15276/aait.09.2026.03
УДК 004:83

Метод фрагментного просторово-часового аналізу відео для виявлення аномальних подій насильницького характеру

Норматова Тетяна Віталіївна[1)]
ORCID: https://orcid.org/0009-0004-3503-6350; tetiana.normatova@nure.ua
Машталір Сергій Володимирович[1)]
ORCID: https://orcid.org/0000-0002-0917-6622; sergii.mashtalir@nure.ua. Scopus Author ID: 36183980100
[1)] Харківський Національний Університет Радіоелектроніки, пр. Науки 14. Харків, 61166, Україна

АНОТАЦІЯ

У цій роботі ми розглядаємо проблему виявлення та класифікації насильницьких подій у відеопотоках за реалістичних обчислювальних обмежень. Багато критично важливих для безпеки подій, таких як насильницькі взаємодії або аномальна поведінка, характеризуються короткочасними та просторово локалізованими моделями руху, тоді як більшість відеоконтенту залишається статичним або нерелевантним. Традиційні підходи глибокого навчання зазвичай обробляють повні відеокадри або щільні просторово-часові представлення, що призводить до високих обчислювальних витрат та неефективного використання обчислювальних ресурсів. Ми пропонуємо фрагментарний просторово-часовий метод аналізу відео, натхненну принципами відеокодування. Кожен відеокадр розділяється на фрагменти, а активність руху оцінюється за допомогою щільного оптичного потоку. Для подальшої обробки вибираються лише фрагменти, що демонструють значні часові зміни, тоді як статичні області пригнічуються на ранній стадії. Розмір фрагмента адаптивно регулюється відповідно до локальної інтенсивності руху, що дозволяє отримати точнішу просторову роздільну здатність у динамічних областях та грубіше представлення у статичних областях. Вибрані фрагменти утворюють компактне представлення, яке згодом використовується для класифікації подій за допомогою легкої часової агрегації. Зменшуючи просторово-часову надлишковість перед вилученням ознак, запропонований підхід значно знижує обчислювальну складність, зберігаючи при цьому розрізнювальні ознаки руху. Запропонований метод оцінюється на наборі даних UBI-Fights з додатковим доповненням навчальних даних за допомогою набору даних VFD2000. UBI-Fights. Експериментальні результати показують, що метод досягає конкурентоспроможної продуктивності з метрикою, що вимірює площу під кривою помилок до 0,72, та метрикою, що вимірює площу під кривою, яка відображає залежність між точністю (Precision) та повнотою (Recall) при різних порогах до 0,63, та бінарним F1-Score до 0,60, зберігаючи при цьому ефективну швидкість виведення. Ці результати вказують на сприятливий компроміс між точністю та ефективністю порівняно з щільними кадровими базовими лініями, що робить метод придатним для систем аналізу відео в реальному часі та з обмеженими ресурсами.

Ключові слова: відеоаналіз; просторово-часова обробка відео; виявлення подій; виявлення аномальної поведінки; виявлення насильницьких подій; фрагментарне представлення кадру; адаптивна фрагментація; аналіз на основі руху; виявлення відеоаномалій

ABOUT THE AUTHORS

Tetiana Vitaliivna Normatova - PhD student, Informatics Department. Kharkiv National University of Radio Electronics.14, Nauky Ave. Kharkiv, 61166, Ukraine
ORCID: https://orcid.org/0009-0004-3503-6350; tetiana.normatova@nure.ua.
Research field: Image and video processing; data analysis

Норматова Тетяна Віталіївна - аспірантка кафедри Інформатики Харківського національного університету радіоелектроніки, пр. Науки 14.Харків, 61166, Україна

Sergii Volodymyrovych Mashtalir - Doctor of Engineering Science, Professor, Informatics Department. Kharkiv National University of Radio Electronics.14, Nauky Ave. Kharkiv, 61166, Ukraine
ORCID: https://orcid.org/0000-0002-0917-6622; sergii.mashtalir@nure.ua. Scopus Author ID: 36183980100
Research field: Image and video processing; data analysis

Машталір Сергій Володимирович - доктор технічних наук, професор кафедри Інформатики. Харківський національний університет радіоелектроніки, пр. Науки 14.Харків, 61166, Україна