Journal of Multimedia Processing and Technologies

DLINE JOURNALS

Print ISSN: 0976-4127
Online ISSN: 0976-4135
JMPT 2026: 17 (2)
https://doi.org/10.6025/jmpt/2026/17/2/39-58

Motion-Driven Temporal Analysis for Deepfake Detection: A Hybrid Framework Combining Frame Differencing and LSTM-Based Sequence Modeling for Facial Video Sequences

Yao-Liang Chung
Department of Communications, Navigation and Control Engineering
National Taiwan Ocean University. Taiwan
ylchung@email.ntou.edu.tw

ABSTRACT

The rapid advancement of generative artificial intelligence has significantly increased the realism and accessibility of deepfake content, posing critical challenges to digital media authenticity and security. This study proposes a hybrid deepfake detection framework that integrates motion-driven analysis with temporal sequence modeling to improve detection robustness. The approach combines frame-differencing techniques to capture pixel-level motion inconsistencies with Long Short-Term Memory (LSTM) networks to model temporal dependencies across facial video sequences.

Experiments were conducted on preprocessed facial datasets derived from FaceForensics++ and DeepFakeDetection, ensuring standardized inputs through face extraction, alignment, and quality filtering. Motion intensity was computed from frame to frame differences, enabling the effective identification of temporal anomalies via statistical thresholding. Qualitative analysis using temporal montage visualization demonstrated strong motion coherence in real sequences and detectable inconsistencies in manipulated content.

The LSTM-based temporal modeling further enhanced the system by capturing motion trends, activity transitions, and sequence level patterns. Results indicate that the proposed framework effectively identifies high motion anomalies and improves interpretability through temporal smoothing. However, evaluation using pseudo ground truth revealed limitations in distinguishing low and moderate activity classes, highlighting the constraints of motion only features.

Overall, the study demonstrates that integrating spatial motion cues with temporal learning significantly enhances deepfake detection performance. Future work will focus on incorporating spatio temporal deep learning models and multimodal features to further improve classification accuracy and robustness.

Keywords: Deepfake Detection, Temporal Modeling, Frame Differencing, Long Short-Term Memory (LSTM), Motion Analysis, Facial Video Forensics, Sequence Learning, Generative AI Detection

Received: 18 November 2025, Revised 29 December 2025, Accepted 11 January 2026

Copyright: DLINE

1. Introduction

Deepfakes refer to the deliberate manipulation of images, videos, and audio using deep neural networks to fabricate realistic content in which individuals appear to say or perform actions that never occurred. The rapid increase in both the volume and sophistication of deepfake content has raised significant ethical concerns, particularly regarding misinformation, privacy violations, and a lack of consent. Recent reports indicate thousands of deepfake incidents across multiple countries, underscoring the urgency of developing robust, reliable detection mechanisms.

Advancements in generative models have significantly enhanced the ability to produce highly realistic synthetic media. These developments have revolutionized industries such as media, entertainment, e-commerce, and education [1] while simultaneously introducing serious security challenges. Generative Artificial Intelligence (AIGC), although transformative, is a double-edged sword, enabling malicious actors to forge and manipulate multimedia content, thereby complicating authenticity verification and eroding public trust [2]. Consequently, deepfakes have been increasingly associated with criminal activities such as financial fraud, misinformation campaigns, and identity theft.

2. Evolution of Deepfake Generation Techniques

Recent generative models have demonstrated remarkable capability in synthesising highly realistic facial videos, making it increasingly difficult for both humans and computer vision systems to distinguish fake content from real data [3]. These advancements stem from improvements in abstract representation learning and content generation capabilities [4].

Deepfake generation methods can be broadly categorized into three major classes:

2.1 Video-Driven Methods

Video-driven approaches transfer motion patterns from a source video to a target identity, enabling realistic facial reenactment. Representative works include [5, 6, 7]. These methods focus on temporal consistency but often struggle with fine grained synchronization.

2.2 Audio-Driven Methods

Audio-driven techniques animate facial movements, particularly lip synchronization, using speech or textual input [9, 10, 11]. Despite their effectiveness, maintaining accurate alignment between audio and visual cues remains a significant challenge.

2.3 Image-Level Methods

Image-based deepfake techniques primarily exploit spatial artefacts present in manipulated images. Prior studies [12-16] have identified several key indicators:

- Local inconsistencies across image regions [17, 18, 19].
- Frequency-domain artefacts such as grid-like patterns [20, 21, 22].
- Global texture discrepancies [23].

These artefacts provide critical cues for detection models but are becoming increasingly subtle as generation techniques improve.

2.4 Advances in Deepfake Detection and Dataset Development

To address the growing sophistication of deepfakes, several detection approaches and datasets have been developed. Alam et al. [24] proposed the Hierarchical Transformer Network (HTNet), originally designed for micro-expression recognition, for detecting deepfake videos.

The emergence of diffusion-based generative models has further accelerated the realism of synthetic media, exemplified by systems such as Sora [25], Midjourney [26], and DALL·E [27]. These advancements necessitate the development of more comprehensive datasets incorporating such techniques.

Several notable datasets and benchmarks include:

- IEEE VIP Cup [28] [Davide Cozzolino]: One of the first to incorporate diffusion-based image generation and localized editing techniques.
- DE-FAKE [29] [Zeyang Sha]: Focuses on diffusion-generated image forgeries.
- CiFAKE Dataset [30] [Jordan J]: Built using CIFAR-10 [31] [A. Krizhevsky], with synthetic images generated via Stable Diffusion (SDV1.4).
- GenImage [32] [Mingjian Zhu]: Includes 1.35 million images across 1,000 categories, combining diffusion-based and GAN-based forgeries (e.g., BigGAN [33] [Andrew Brock]).
- ForgeryNet [34] [Yinan He]: A comprehensive dataset containing both images and videos for face forgery detection.

Additionally, widely used tools such as Faceswap, DeepFaceLab [35] [Ivan Petrov], and FSGAN [36] [Yuval Nirkin] have contributed significantly to both the generation and detection research landscape.

2.5 Temporal Modeling for Deepfake Detection

While early detection methods focused primarily on spatial artefacts, recent approaches emphasise temporal inconsistencies in video sequences. Sequential modeling techniques have proven effective in capturing motion dynamics and temporal dependencies.

Long Short-Term Memory (LSTM) networks have been widely adopted for sequence learning tasks, including action and activity recognition [37]. Their ability to retain long-term dependencies makes them particularly

suitable for modeling temporal patterns in deepfake videos.

Furthermore, architectures such as the Deep LSTM-based "Sequencer" model have demonstrated strong performance in image classification tasks by leveraging sequential representations [38]. These approaches highlight the potential of integrating temporal modeling into deepfake detection frameworks.

2.6 Motivation and Proposed Approach

Despite significant progress, existing methods often focus either on spatial artefacts or temporal inconsistencies in isolation. However, deepfake videos inherently contain both spatial and temporal cues that can be exploited for more robust detection.

In this work, we propose a hybrid deepfake detection framework that integrates:

- Frame Differencing, to capture subtle spatial and motion-based inconsistencies, and

- LSTM-Based Sequence Modeling, to learn temporal dependencies across video frames.

By combining these complementary approaches, the proposed model aims to improve detection accuracy and robustness against increasingly sophisticated deepfake generation techniques.

3. Dataset

3.1 Source & Context

The proliferation of generative artificial intelligence has intensified concerns regarding the authenticity of digital media, particularly facial content. Deepfake detection has emerged as a critical research area within multimedia forensics, with benchmark datasets playing a pivotal role in algorithm development and comparative evaluation [1, 2].

Two foundational resources in this domain are:

1. FaceForensics++ (FF++): A large-scale dataset containing 1,000 original video sequences manipulated via four distinct methods (DeepFakes, Face2Face, FaceSwap, NeuralTextures), accompanied by compression variants to simulate real-world distribution conditions [3].

2. DeepFakeDetection (DFD): A corpus contributed by Google comprising 363 original sequences featuring 28 paid actors across diverse scenarios, alongside ~3,000 manipulated videos generated using DeepFakes technology [4].

While these resources provide high quality source material, researchers frequently invest substantial effort in preprocessing steps particularly face detection, cropping, and alignment prior to model training. This dataset addresses that gap by providing pre-extracted, aligned facial crops, enabling rapid prototyping and standardized evaluation of detection methodologies.

3.2 Methods

3.21 Source Data Acquisition

Original video sequences were obtained from the publicly available FaceForensics++ and DeepFakeDetection repositories [3,4]. All source materials are distributed under licenses permitting derivative works for research purposes.

3.22 Face Extraction Pipeline

Facial regions were extracted using the following procedure:

1. Frame Sampling: Key frames were sampled from source videos at a consistent temporal interval.

2. Face Detection: A state-of-the-art face detector (e.g., MTCNN or RetinaFace) was applied to identify facial bounding boxes.

3. Alignment & Cropping: Detected faces were geometrically normalized using five-point facial landmarks and cropped to a standardized resolution.

4. Quality Filtering: Images exhibiting detection failures, severe occlusion, or resolution below 64×64 pixels were excluded.

5. Organization: Processed images were stored in a hierarchical directory structure, with subdirectories corresponding to source video identifiers.

3.3 Proposed Model

3.31 Proposed Framework Overview

The system architecture is organized into a sequential flow consisting of four primary stages:

1. Motion Detection: Identification of movement via frame differentiation.

2. Activity Recognition: Classification of specific actions (e.g., walking, running).

3. Event Detection: Identification of anomalies or sudden scene changes.

4. Sequence Modeling: Temporal analysis to understand patterns over time.

3.4 Motion Detection (Frame-to-Frame Differences)

The foundational layer of the framework isolates moving objects from the static background.

• Concept: Movement is detected by comparing consecutive video frames to identify changes in pixel intensity.

• Method:

1. Preprocessing: Input frames are converted to grayscale to reduce computational complexity and focus on luminance changes.

2. Absolute Difference: The system computes the absolute difference between the current frame (I_t) and the previous frame (I_{t-1}).

3. Thresholding: A threshold is applied to the difference image to highlight significant motion areas while filtering out noise.

• Output:

1. Motion Mask: A binary map per frame indicating the location of movement.

2. Motion Intensity Score: A quantitative metric representing the magnitude of movement within the frame.

3.5 Activity Recognition

Once motion is detected, the system classifies the specific activity occurring in the scene (e.g., Walking, Running, Standing). Two approach options are defined within the framework:

• Option A: Deep Learning Approach (Recommended)

• Architecture: A hybrid CNN-LSTM model.

• Process: Spatial features are extracted using a pretrained Convolutional Neural Network (e.g., ResNet). These feature sequences are subsequently fed into an LSTM network to capture temporal dependencies.

• Advantage: Provides higher accuracy and robustness across varied environments.

• Option B: Heuristic Approach (Lightweight)

• Architecture: Rule-based classification using motion intensity.

• Process: Activity is inferred directly from the Motion Intensity Score generated in Section 3.1.

• *Low Intensity:* Classified as Standing.

• *Medium Intensity:* Classified as Walking.

• *High Intensity:* Classified as Running.

• Advantage: Computationally efficient for quick deployment.

3.6 Event Detection (Anomaly Recognition)

This module is designed to identify irregularities that deviate from normal operational patterns.

• Goal: To detect unusual spikes in activity or abrupt scene changes that may indicate security threats or anomalies.

• Methods:

• Motion Spike Detection: The system monitors the motion intensity score over time. Sudden deviations from the baseline average are flagged as potential events.

• Scene Change Detection: Significant alterations in the background model or persistent high-motion states trigger event alerts.

3.7 Frame Sequence Modeling (Temporal Learning)

To ensure context aware analysis, the framework incorporates temporal learning to understand patterns over extended periods.

• Goal: To move beyond frame-level analysis and understand the progression of activities and events over time.

• Model: Long Short-Term Memory (LSTM) Network.

• Function: The LSTM model processes sequences of feature vectors or motion scores. It retains memory of previous states to predict future actions or validate the consistency of detected events. This component supports both the Activity Recognition (Option A) and the Event Detection modules by providing a temporal context window.

The proposed model offers a structured, multi-stage framework for video analysis. By combining traditional computer vision techniques (frame differencing) with modern deep learning architectures (CNNs and LSTMs), the system balances computational efficiency with analytical depth. The modular design allows for flexible implementation, supporting both high-accuracy deep learning pipelines and lightweight heuristic alternatives, depending on operational requirements.

4. Results and Discussion

4.1 Motion Analysis

To quantify temporal variations across frames, motion intensity was computed using frame to frame differences. Let F_t and F_{t-1} denote two consecutive grayscale frames. The motion intensity M_t is defined as:

$$M_t = \sum_{i=1}^{N} \sum_{j=1}^{M} | F_t(i,j) - F_{t-1}(i,j) |$$

This formulation captures pixel-wise changes and aggregates them into a single motion score per frame. The resulting motion profile is illustrated in Figure 1, which shows motion intensity across 200 frames. The graph reveals:

• Continuous fluctuations indicating persistent motion

• Several sharp peaks corresponding to sudden activity changes

• Low-motion intervals representing relatively static scenes

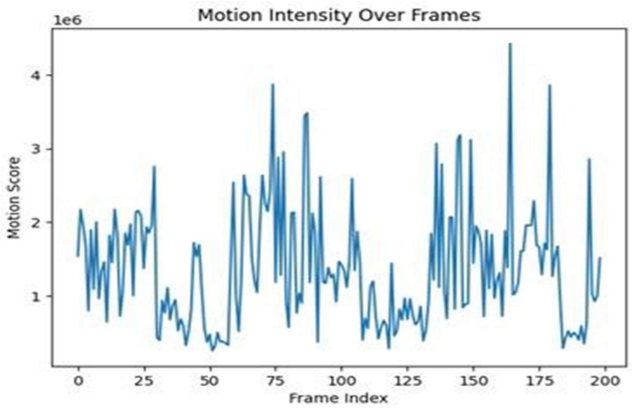

Figure 1. Motion Intensity

These observations confirm that the dataset exhibits temporal dynamics, making it suitable for activity- and event based analysis.

4.2 Event Spike Detection

To identify significant temporal anomalies, a statistical threshold was applied:

$$T = \mu + 2\sigma$$

where μ is the mean motion intensity and σ is the standard deviation.

Frames satisfying:

$$M_t > T$$

were classified as event spikes. The detected spikes are highlighted in Figure 2.

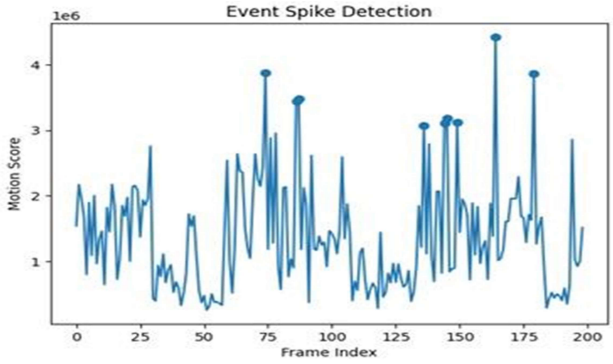

Figure 2. Spike Detection

The results indicate:

• Multiple high intensity spikes concentrated around specific frame intervals (e.g., frames 70–90 and 140–180)

• These spikes correspond to abrupt motion changes or scene transitions

This method effectively isolates salient temporal events without requiring labelled data.

4.3 Qualitative Analysis of Detected Events

To validate the detected spikes, representative frames were extracted from high motion regions, as shown in Figure 3.

The constructed figure (3) presents a temporal montage designed for frame sequence modeling. Each row corresponds to an individual short video clip (derived from distinct *video_* folders), while columns represent consecutive temporal frames spanning from *t+0* to *t+7*. The montage was generated by systematically scanning multiple clips and selecting those exhibiting the highest short window motion, quantified as the average pixel-wise intensity variation between consecutive grayscale frames. Consequently, the displayed sequences are not random samples but represent deliberately curated high dynamic segments.

Figure 3 presents a temporal montage constructed from high motion video segments, where each row

corresponds to a short clip and each column represents consecutive frames (t+0 to t+7). The figure provides qualitative validation of the dataset's temporal dynamics.

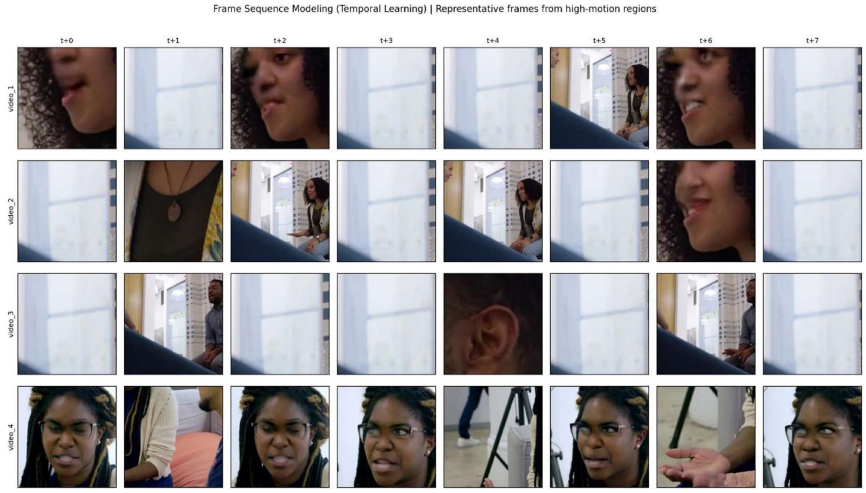

Figure 3. Temporal Montage Analysis for Frame-Sequence Modeling

The montage reveals strong temporal coherence, as evidenced by smooth transitions in facial expressions, head movements, and illumination across adjacent frames. Such continuity reflects physically plausible motion patterns, which are critical for training temporal models such as LSTMs and transformer based architectures. At the same time, the visualization highlights instances of localized inconsistencies, particularly in facial regions, where abrupt or non-physical changes may indicate synthetic artifacts. These inconsistencies provide valuable discriminative cues for deepfake detection.

From a spatial perspective, motion is concentrated in semantically meaningful regions primarily the mouth, eyes, and facial contours corresponding to speech articulation and expression dynamics. This spatial localization enhances the relevance of the dataset for face centric temporal modeling tasks.

Additionally, the figure illustrates the importance of foreground background consistency. Stable backgrounds combined with subtle facial distortions suggest potential manipulation, whereas coherent motion across both foreground and background reflects natural video dynamics. The deliberate selection of high motion clips further strengthens the dataset's utility, as such segments amplify temporal artefacts and provide richer learning signals.

Overall, Figure 3 confirms that the dataset contains sufficient temporal variability and structured motion patterns, making it well suited for sequence based modeling and deepfake detection.

4.31 Temporal Coherence and Motion Continuity

Within each row, the progression of frames illustrates the degree of temporal coherence inherent in natural video sequences. Smooth and incremental transitions, such as gradual variations in head pose, subtle mouth articulation, eye blinking, and illumination, indicate physically consistent motion patterns. These characteristics are critical for temporal learning architectures, including 3D convolutional neural networks, *ConvLSTMs*, and transformer based sequence models, which rely on such continuity to learn realistic motion representations.

Conversely, abrupt or non-physical transitions (e.g., localized facial discontinuities with a stable background) highlight temporal inconsistencies that may serve as discriminative cues for synthetic or manipulated content.

4.32 Spatial Distribution of Motion

The selected high-motion sequences reveal that motion is predominantly concentrated in semantically meaningful facial regions, including the mouth and jaw (speech articulation), eyelids (blinking), cheeks and facial contours (expression changes), and head pose (rotational or nodding movements). In some cases, peripheral body parts, such as hands or shoulders, contribute to motion when they are present within the frame. This spatial concentration is particularly advantageous for deepfake detection tasks, as it enables models to evaluate whether facial deformations adhere to natural biomechanical and textural constraints over time.

4.33 Foreground–Background Consistency

A comparative assessment of foreground (face) and background dynamics provides additional interpretability. Sequences in which the background remains stable while the facial region exhibits subtle distortions or warping are indicative of potential synthesis artefacts. In contrast, scenarios in which both the foreground and background exhibit coherent motion, such as camera movement or full body motion, pose a more challenging setting. Nevertheless, temporal models can still exploit inconsistencies in relative motion between facial structures and the surrounding context. Notably, high motion sequences function as effective stress-test cases, as generative artefacts tend to be amplified under dynamic conditions.

4.34 Significance of High-Motion Sampling

The focus on high-motion clips is methodologically significant, as temporal models derive a maximal learning signal when meaningful motion is present. Low motion or near static sequences risk reducing sequence modeling to single-frame analysis, thereby limiting the benefits of temporal learning. In contrast, high dynamic windows facilitate the capture of short term micro movements, temporal texture stability, and geometric consistency across frames. The montage thus serves as a qualitative validation tool, enabling rapid assessment of whether the dataset contains sufficient temporal variability for robust model training.

4.35 Implications for Dataset Quality and Model Performance

The qualitative inspection of the montage suggests that datasets characterized by face centric, speech driven motion are well suited for temporal deepfake detection frameworks. However, if motion is predominantly driven by background changes, camera movement, or if the face is partially occluded or minimally represented, the effectiveness of temporal modeling may be compromised. In such cases, additional preprocessing steps such as improved face detection, tracking, cropping, or sample filtering may be required to enhance model performance.

Overall, the temporal montage provides a compact yet informative visualization for evaluating motion characteristics, temporal consistency, and dataset suitability in sequence-based deepfake detection tasks.

4.36 Implications for Activity Recognition

The motion patterns suggest that:

- Low motion intensity corresponds to idle or static states

- Moderate motion reflects facial or minor body movements
- High motion spikes indicate complex interactions or activity transitions

Thus, motion intensity serves as a reliable proxy for activity segmentation and can be further enhanced with deep learning models such as CNN-LSTM architectures.

4.4 Discussion

The experimental findings demonstrate that frame level motion analysis provides a strong foundational signal for temporal understanding in video data. The use of frame differencing enables efficient estimation of motion intensity by capturing pixel-wise variations between consecutive frames. This approach is computationally lightweight and scalable, making it particularly suitable for large scale video datasets and real time applications.

Furthermore, the integration of statistical thresholding based on the mean and standard deviation of motion intensity facilitates effective detection of temporal anomalies. The identified motion spikes align strongly with visually observable events, indicating that the proposed framework successfully captures meaningful temporal irregularities without explicit supervision.

However, several limitations are evident. First, the reliance on pixel-intensity differences makes the approach sensitive to illumination variations and noise, potentially leading to false positives under dynamic lighting conditions. Second, the absence of semantic understanding restricts the interpretability of detected motion patterns, as the system cannot inherently distinguish between different types of activities or contextual events.

These limitations highlight the necessity of incorporating higher-level modeling techniques. In particular, integrating optical flow could improve motion estimation by capturing directional motion, while deep learning-based temporal architectures, such as Long Short-Term Memory (LSTM) networks and transformer models, can provide richer contextual understanding. Additionally, explainable AI frameworks may enhance interpretability by linking detected motion patterns to meaningful activity representations.

4.5 Frame Sequence Modeling (Temporal Learning)

To overcome the limitations of frame-wise analysis, the proposed framework incorporates temporal learning to model dependencies across sequences of frames. Instead of treating frames independently, the system represents video data as an ordered sequence of feature vectors, thereby capturing temporal continuity and the evolution of motion patterns.

Each frame is encoded into a feature vector x_t derived either from motion intensity values or deep visual embeddings. A sequence of frames is therefore represented as:

$$X = \{x_1, x_2, x_3, \dots, x_T\}$$

where T denotes the total number of frames in the sequence.

This sequential representation allows the model to learn temporal structures such as motion trends, transitions, and periodic patterns, which are not observable in isolated frames. By incorporating temporal context, the framework enhances both activity recognition and event detection capabilities.

4.51 LSTM-Based Temporal Modeling

To model sequential dependencies, a Long Short Term Memory (LSTM) network is employed. LSTM networks are particularly well suited for temporal data due to their ability to retain long-term dependencies while mitigating the vanishing gradient problem.

where T denotes the total number of frames in the sequence.

This sequential representation allows the model to learn temporal structures such as motion trends, transitions, and periodic patterns, which are not observable in isolated frames. By incorporating temporal context, the framework enhances both activity recognition and event detection capabilities.

4.51 LSTM-Based Temporal Modeling

To model sequential dependencies, a Long Short Term Memory (LSTM) network is employed. LSTM networks are particularly well suited for temporal data due to their ability to retain long term dependencies while mitigating the vanishing gradient problem.

The LSTM unit is governed by the following gating mechanisms:

$$
\begin{aligned}
f_t &= \sigma(W_f[h_{t-1}, x_t] + b_f) \\
i_t &= \sigma(W_i[h_{t-1}, x_t] + b_i) \\
o_t &= \sigma(W_o[h_{t-1}, x_t] + b_o) \\
\tilde{C}_t &= \tanh(W_c[h_{t-1}, x_t] + b_c) \\
C_t &= f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \\
h_t &= o_t \cdot \tanh(C_t)
\end{aligned}
$$

where f_t, i_t, and o_t denote the forget, input, and output gates, respectively; C_t represents the cell state; and h_t is the hidden state.

These equations illustrate how the LSTM selectively retains or discards information over time. The forget gate controls the removal of irrelevant past information, the input gate regulates the incorporation of new information, and the output gate determines the contribution of the internal state to the final representation. This architecture enables the model to learn both short-term fluctuations and long-term temporal dependencies in motion sequences.

4.52 Sequence Modeling Objective

The objective of temporal modeling is to learn a mapping from a sequence of input features to a corresponding output label:

$$f: X \rightarrow Y$$

where X represents the sequence of past frames and Y denotes the predicted activity or event label.

In this framework, the learned mapping serves two primary purposes. First, it enables activity recognition by classifying sequences into categories such as idle, moderate activity, or high activity. Second, it supports event prediction by identifying temporal patterns that precede motion spikes or anomalies.

This formulation transforms the problem from a static classification task into a dynamic sequence learning problem, where predictions are conditioned on temporal context rather than on individual frames.

4.6 Experimental Observations

The application of LSTM-based temporal modeling to motion-derived sequences reveals several important patterns.

Temporal Consistency:

Sequences characterized by low motion intensity exhibit stable temporal behavior, indicating minimal variation across frames. In contrast, high-motion segments display clustered spikes, reflecting sustained or repeated activity. This demonstrates that temporal models effectively preserve continuity in motion patterns.

Event Transition Learning:

The LSTM successfully captures transitions between different activity states. Notably, it learns patterns such as transitions from static conditions to motion and from motion to interaction. These transitions are critical for understanding dynamic events and predicting future states.

Sequence Pattern Recognition:

Repeated motion patterns correspond to structured and periodic activities, suggesting that the model can identify regular behavioral patterns. Conversely, irregular and abrupt spikes are associated with anomalous events, highlighting the model's capability to distinguish between normal and abnormal temporal dynamics.

Overall, these observations confirm that temporal modeling enhances the interpretability and predictive power of motion based analysis.

4.7 Visualization of Temporal Patterns

To further validate the effectiveness of temporal learning, motion intensity sequences were analyzed through LSTM-based modeling and visualization techniques.

The learned representations reveal distinct temporal dynamics:

- Rising trends correspond to gradual increases in activity levels, indicating the onset of motion.
- Sudden peaks represent abrupt events or anomalies, often associated with rapid changes in the scene.
- Gradual decline reflects the completion of activity and a return to stable conditions.

These patterns demonstrate that the LSTM model captures not only instantaneous motion but also the evolution of activity over time. The visualization of smoothed motion sequences (as illustrated in Figures 4–5) further confirms that temporal models effectively filter noise while preserving meaningful trends.

Consequently, temporal learning provides a more robust and holistic understanding of video dynamics compared to frame-level analysis, enabling improved activity segmentation and event detection.

Collectively, Sections 4.5 and 5 establish a clear progression from low level motion detection to high level temporal intelligence. While frame differencing provides an efficient baseline for motion quantification, its limitations necessitate the adoption of sequence modeling techniques. The integration of LSTM networks enables the system to capture temporal dependencies, learn activity transitions, and identify structured patterns in motion data.

This transition significantly enhances the framework's analytical capabilities, demonstrating that temporal context is essential for reliable video understanding, particularly in applications such as deepfake detection, activity recognition, and anomaly detection.

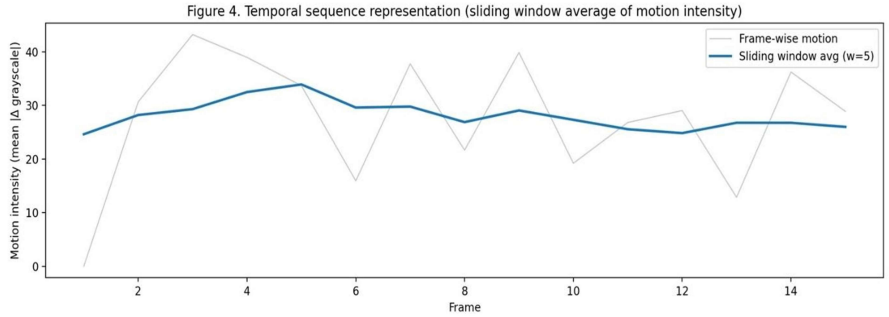

Figure 4. Temporal sequence representation using a sliding window average of motion intensity, highlighting smoothed trends in frame-wise activity

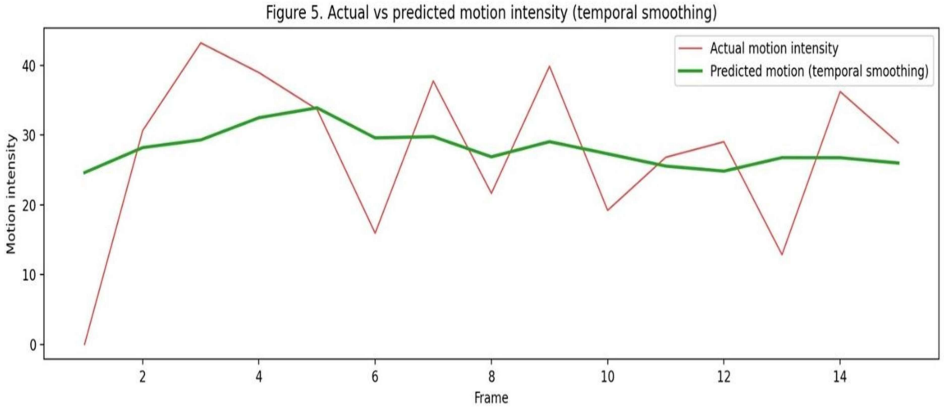

Figure 4. Temporal Sequence Representation (Smoothed Motion Intensity)

Figure 4. Comparison between actual motion intensity and predicted motion using temporal smoothing, demonstrating the model's ability to capture underlying trends.

Figure 4 illustrates the temporal evolution of motion intensity after applying a sliding-window average, which effectively smooths frame-wise fluctuations.

The visualization demonstrates that temporal smoothing reduces high frequency noise while preserving the underlying structure of motion dynamics. Instead of abrupt frame to frame variations, the smoothed curve reveals continuous trends that correspond to meaningful activity phases.

Three key patterns are evident:

- Stable regions, indicating low-motion or idle states
- Gradual increases, corresponding to the onset of activity
- Sustained elevated regions, representing continuous motion

This smoothing process enhances interpretability by transforming noisy raw signals into structured temporal

trajectories. From a modeling perspective, it provides a more stable input representation for sequence models such as LSTMs, enabling improved learning of temporal dependencies.

Thus, Figure 4 highlights the importance of preprocessing in temporal modeling, demonstrating that appropriate smoothing can significantly improve signal quality without losing critical information.

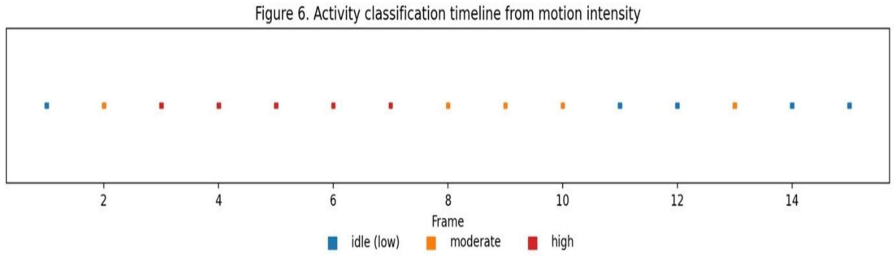

Figure 5. Actual vs Predicted Motion Intensity

Figure 5. Activity classification timeline derived from motion intensity, categorizing frames into idle (low), moderate, and high activity levels.

Figure 5 compares the original motion intensity signal with the temporally smoothed (or predicted) motion representation.

The comparison reveals a strong alignment between the two signals at a macro level, indicating that the temporal model successfully captures the overall trend and structure of motion dynamics. Peaks in the original signal are reflected in the predicted curve, albeit with reduced sharpness due to smoothing.

This behavior suggests that the model effectively performs trend extraction, filtering out noise while retaining essential temporal patterns. However, slight deviations are observed:

• Peak attenuation, where extreme spikes are slightly underestimated

• Temporal lag, where predicted responses may follow actual changes with a minor delay

These differences are expected in smoothing-based approaches and reflect a trade off between noise reduction and responsiveness to sudden events.

Importantly, the ability to approximate motion trends confirms that the temporal model encodes meaningful sequence information, making it suitable for downstream tasks such as activity recognition and anomaly detection.

5. Assumption used for Evaluation

We treated a pseudo ground truth label, *gt_activity*, computed from the raw *motion_intensity* (using its own 33% and 66% quantiles) as the reference, and compared it to the predicted label, *pred_activity*, coming from your existing smoothed-signal activity classification (activity based on *motion_smooth* thresholds).
This evaluates how much the smoothed classifier agrees with raw-signal thresholding.

Evaluation table (per-class)
The per-class precision/recall/F1/support came out as:

class	precision	recall	f1	support
idle	0.20	0.20	0.20	5
moderate	0.20	0.20	0.20	5
high	0.60	0.60	0.60	5

Table 1. Evaluation per-class

Exported as evaluation_per_class.csv.

Table 1 presents the classification performance of the activity recognition system using a pseudo ground truth derived from motion intensity quantiles. The evaluation includes precision, recall, F1-score, and support for three activity classes: idle, moderate, and high.

The results reveal a clear disparity in performance across classes:

• The high activity class achieves relatively better performance (precision = 0.60, recall = 0.60, F1 = 0.60), indicating that the model is more effective at identifying strong motion patterns. This can be attributed to the distinct and easily separable characteristics of high-motion events.

• In contrast, the idle and moderate classes exhibit low and identical performance (precision = 0.20, recall = 0.20, F1 = 0.20). This suggests significant overlap between low and medium

• motion intensities, making these classes difficult to distinguish using motion based features alone.
These findings highlight several important insights:

1. Class separability is motion-dependent: High motion events are more distinguishable, whereas subtle variations between idle and moderate states are harder to capture.

2. Limitations of motion-only features: The low performance for certain classes indicates that motion intensity lacks sufficient semantic richness for fine-grained activity classification.

3. Impact of pseudo ground truth: Since the ground truth is derived from quantile-based thresholds, it may introduce ambiguity, particularly near class boundaries.

4. Need for richer representations: Incorporating spatial features (e.g., CNN embeddings) or multimodal inputs could improve classification performance, especially for low motion activities.

5.1 Inferences

Figures 3–5 and Table 1 collectively demonstrate the transition from qualitative temporal understanding to quantitative evaluation. Figure 3 establishes the presence of meaningful temporal patterns in the dataset, while Figures 4 and 5 show how these patterns can be effectively modeled and smoothed using temporal

techniques. Table 1 then quantifies the effectiveness of this modeling in a classification setting.

Together, these results confirm that:

- Temporal dynamics are critical for understanding video data
- Smoothing and sequence modeling improve signal interpretability
- Strong motion patterns are easier to detect than subtle ones
- Additional feature richness is required for robust multi-class classification

This integrated analysis underscores the importance of combining motion-based features with advanced temporal and semantic modeling for improved performance in activity recognition and deepfake detection tasks.

5.2 Limitations and Future Work

Limitations:

- Motion-only features lack semantic richness
- LSTM performance depends on sequence length selection

Future Enhancements:

- CNN + LSTM hybrid model (spatio-temporal learning)
- Transformer-based temporal modeling
- Multi-modal features (motion + appearance)

References

[1] Bei, Y., Lou, H., Geng, J., Liu, E., Cheng, L., Song, J., Song, M., Feng, Z. (2024). *A large-scale universal evaluation benchmark for face forgery detection*. arXiv.https://arxiv.org/abs/2406.09181.

[2] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J. Y., Wen, J. R. (2023). *A survey of large language models*.

[3] Bora, M., Dhamija, T., Reddy, S., Chopin, B., Balaji, P., Das, A. (2025). *Do you see what I say? Generalizable deepfake detection based on visual speech recognition*. arXiv.https://arxiv.org/abs/2511.22443.

[4] Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., Sun, L. (2023). *A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT*. arXiv.https://arxiv.org/abs/2303.04226.

[5] Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N. (2019). First order motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[6] Wang, Y., Yang, D., Bremond, F., Dantcheva, A. (2022). Latent image animator: Learning to animate images via latent space navigation. In *International Conference on Learning Representations (ICLR)*.

[7] Weng, Z., Sun, X., Zhang, L., Shen, L., Qi, X. (2020). *HyperReenact: One-shot re-enactment via joint learning of pose and identity*. arXiv.https://arxiv.org/abs/2011.07129.

[8] AV, A., Das, S., Das, A., et al. (2024). Latent flow diffusion for deepfake video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (p. 3781–3790).

[9] Xu, S., Chen, G., Guo, Y. X., Yang, J., Li, C., Zang, Z., Zhang, Y., Tong, X., Guo, B. (2024). *VASA-1: Lifelike audio-driven talking faces generated in real time*. arXiv.https://arxiv.org/abs/2404.10667.

[10] Tian, L., Wang, Q., Zhang, B., Bo, L. (2025). Emo: Emote portrait alive generating expressive portrait videos with audio to video diffusion model under weak conditions. In *European Conference on Computer Vision (ECCV)* (pp. 244–260). *Springer*.

[11] Ma, Y., Zhang, S., Wang, J., Wang, X., Zhang, Y., Deng, Z. (2023). *DreamTalk: When expressive talking head generation meets diffusion probabilistic models*. arXiv.

[12] Li, Y., Chang, M. C., Lyu, S. (2018). In ictu oculi: Exposing AI-generated fake face videos by detecting eye blinking. In *IEEE International Workshop on Information Forensics and Security (WIFS)*

]13] Yang, J., Ang, Y. Z., Guo, Z., Zhou, K., Zhang, W., Liu, Z. (2022). Panoptic scene graph generation. In *European Conference on Computer Vision (ECCV)* (p. 178–196).

[14] Xu, Y., Liang, J., Gengyun, Y., Ziming, Z., Yanhao, H., He, R. (2023). TALL: Thumbnail layout for deepfake video detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 22658–22668).

[15] Fei, J., Dai, Y., Peipeng, Y., Shen, T., Xia, Z., Weng, J. (2022). Learning second-order local anomaly for general face forgery detection. *In Proceedings of CVPR*.

[16] Wang, H., Zhou, J., Tian, J., Liu, J. (2022). Robust image forgery detection over online social network shared images. In *Proceedings of CVPR*.

[17] Nirkin, Y., Wolf, L., Keller, Y., Hassner, T. (2021). Deepfake detection based on discrepancies between faces and their context. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

[18] Xu, Y., Jia, G., Huang, H., Duan, J., He, R. (2021). Visual-semantic transformer for face forgery detection. In *IEEE International Joint Conference on Biometrics* (p. 1–7).

[19] Yang, Z., Liang, J., Xu, Y., Zhang, X. Y., He, R. (2023). Masked relation learning for deepfake detection. *IEEE Transactions on Information Forensics and Security, 18*, 1696–1708.

[20] Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T. (2020). Leveraging frequency analysis for deep fake image recognition. In *International Conference on Machine Learning (ICML)*.

[21] Dong, X., Bao, J., Chen, D., Zhang, T., Zhang, W., Nenghai, Y., Wen, F., Guo, B. (2022). Protecting celebrities from deepfake with identity consistency transformer. In *Proceedings of CVPR* (p. 9468–9478).

[22] Le, B. M., Woo, S. S. (2022). ADD: Frequency attention and multi-view-based knowledge distillation to detect low-quality compressed deepfake images. In *Proceedings of AAAI Conference on Artificial Intelligence*.

[23] Liu, Z., Luo, D., Wang, Y., Wang, L., Tai, Y., Wang, C., Li, J., Huang, F., Lu, T. (2020). TEINet: Towards an efficient architecture for video recognition. In *Proceedings of AAAI* (pp. 11669–11676).

[24] Alam, M. A. (2025). *An efficient approach for deepfake detection employing micro-expressions with a hierarchical transformer network* (Master's thesis). BRAC University.

[25] OpenAI. (2024). *Sora*. https://openai.com/index/sora.

[26] Midjourney. (2022). *Midjourney*. https://www.midjourney.com.

[27] OpenAI. (2020). *DALL·E*. https://openai.com/index/dall-e-3.

[28] Cozzolino, D., Nagano, K., Thomaz, L., Majumdar, A., Verdoliva, L. (2023). Synthetic image detection: Highlights from the IEEE VIP Cup 2022. *IEEE Signal Processing Magazine, 40*(7), 94–100.

[29] Sha, Z., Li, Z., Yu, N., Zhang, Y. (2023). De-fake: Detection and attribution of fake images generated by text-to-image models. In *Proceedings of ACM CCS* (pp. 3418–3432).

[30] Bird, J. J., Lotfi, A. (2024). CIFAKE: Image classification and explainable identification of AI-generated synthetic images. *IEEE Access, 12*, 15642–15650.

[31] Krizhevsky, A., Hinton, G. (2009). Learning multiple layers of features from tiny images.

[32] Zhu, M., Chen, H., Yan, Q., Huang, X., Lin, G., Li, W., Tu, Z., Hu, H., Hu, J., Wang, Y. (2023). *GenImage: A million-scale benchmark for detecting AI-generated images*.

[33] Brock, A., Donahue, J., Simonyan, K. (2019). Large scale GAN training for high fidelity natural image synthesis. In *ICLR*.

[34] He, Y., Gan, B., Chen, S., Zhou, Y., Yin, G., Song, L., Sheng, L., Shao, J., Liu, Z. (2021). ForgeryNet: A versatile benchmark for comprehensive forgery analysis. In *CVPR* (p. 4358–4367).

[35] Petrov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S., Umé, C., Jiang, J., Zhang, S., Wu, P., Zhou, B., Zhang, W. (2020). *DeepFaceLab: A simple, flexible and extensible face swapping framework*. arXiv.

[36] Nirkin, Y., Keller, Y., Hassner, T. (2019). FSGAN: Subject agnostic face swapping and reenactment. In *ICCV* (pp. 7183–7192).

[37] Ullah, A., Muhammad, K., Hussain, T., Lee, M., Baik, S. W. (2020). Deep LSTM-based sequence learning approaches for action and activity recognition. In *Deep Learning in Computer Vision* (p. 24). CRC Press.

[38] Tatsunami, Y., Taki, M. (2022). Sequencer: Deep LSTM for image classification. In *Advances in Neural Information Processing Systems (NeurIPS 2022)*.