Received: 12 August 2024 | Revised: 5 November 2024 | Accepted: 13 March 2025 | Published online: 19 March 2025

REVIEW

Journal of Data Science and Intelligent Systems
2025, Vol. 00(00) 1–15
DOI: 10.47852/bonviewJDSIS52024075

Integrative Review of Human Activity Recognition and Violence Detection: Exploring Techniques, Modalities, and Cross-Domain Knowledge Transfer

Paul Turyahabwa[1,*] and Sudi Murindanyi[2]

[1]*Department of Electrical and Computer Engineering, Makerere University, Uganda*

[2]*Department of Computer Science, Makerere University, Uganda*

Abstract: Human Activity Recognition (HAR) is a key topic of research in computer vision having numerous applications in surveillance, healthcare, and robotics. Violence detection (VD) is also an exciting research area under activity recognition, and success in this field would offer the possibilities of reduced crime rate as well as more accurate and scalable error-free monitoring for surveillance applications. Unlike other review papers that generally handle HAR only or VD only as different topics, this review paper carries out a systematic literature review of both HAR and VD with the aim of relating the two and possibly opening up possibilities of knowledge transfer. Our goal is to identify how the best methods and techniques used in HAR can be adapted for VD and vice versa to improve feature extraction, increase accuracies, and create more robust models. In this systematic literature review paper, over 100 papers were studied to obtain information about the most recent methods and techniques being used. Some of the key takeaways this review proposes include the usage of Transformer-based techniques and Graphical Convolutional Network methods that perform well for HAR for VD, more experimentation using Recurrent Neural Network-based methods that perform well for VD for HAR, and also more work needs to be done to improve model performance on datasets in unconstrained environments.

Keywords: human activity recognition, deep learning, violence detection, modalities, datasets, cross-domain

1. Introduction

Human Activity Recognition (HAR) is an exciting field in computer vision since it allows computers to determine what an individual is doing when given a data source like a video or image. This is significant because it provides another degree of cognitive intelligence to computers, allowing them to classify what activity someone is performing rather than simply knowing who is doing it. HAR offers a broad range of practical applications, including robotics, healthcare, and surveillance [1]. Examples of actual use cases in these sectors include enhancing surveillance systems for crime prevention, improving user interactions in smart home settings, and using activity recognition systems for real-time monitoring of the elderly [1].

Although HAR has come a long way, there are still many obstacles to overcome. The field has faced challenges, made significant advancements, and adjusted to new technology during its growth [1]. At first, the biggest obstacle was to attain high accuracy in a variety of intricate real-world situations. Real-time data processing and a thorough comprehension of the nuances of human movement are necessary for this [1]. To tackle this difficulty, scholars have developed inclusive datasets that encompass a broad spectrum of human activities and demographics. Additionally, they concentrated on lowering algorithmic model biases and using multimodal data to improve HAR systems' accuracy and resilience [2, 3].

Violence detection (VD) is one of the main subfields of HAR where computers are used to detect violent activities. In 2019, the United States saw approximately 1.2 million violent incidents, encompassing fights, aggressive acts, and mass shootings [4]. Consequently, the use of surveillance equipment has risen sharply. In recent years, heavily crowded areas like public streets, subways, industrial sites, and banks have been closely monitored by surveillance systems to ensure public safety in smart cities [4]. However, relying on humans to monitor this footage has its drawbacks—it's time-consuming, not scalable, and prone to mistakes. Hence, automated video detection systems are needed [5].

Despite the advancements in HAR and VD, several research gaps persist. For instance, HAR models often struggle with high accuracy in dynamic, unconstrained environments due to complex background variations and occlusions, limiting their robustness in real-world applications [1, 6]. Similarly, VD research faces limitations in dataset availability and domain-specific feature extraction, impacting model generalizability across diverse violent scenarios [4, 5]. Cross-domain knowledge transfer offers a

*Corresponding author: Paul Turyahabwa, Department of Electrical and Computer Engineering, Makerere University, Uganda. Email: paul.turyahabwa@students.mak.ac.ug

promising avenue to address these limitations by adapting techniques from HAR to VD and vice versa, enhancing model resilience and context-aware feature detection.

Our review is unique compared to other review papers in this field because of the following reasons:

1) Provides information on the techniques and methods that have been used in the last 5-6 years since 2018 except for benchmark datasets and benchmark papers. Our aim is to enable the reader of this review paper to get a general sense of the most current techniques being used for HAR and VD.
2) In addition to providing a comprehensive review in line with the principles of writing a systematic literature review (SLR) highlighted in Keele [7], on both HAR and VD, this paper advances the current understanding by highlighting potential pathways for cross-domain knowledge transfer between HAR and VD.
3) While recent works have touched upon aspects of both HAR and VD [8–10], these studies primarily focused on VD implementations using general deep learning (DL) approaches, rather than systematically exploring methodology transfer between the fields. Our work represents the first comprehensive systematic review examining the bidirectional transfer potential between HAR and VD, analyzing over 100 papers to identify specific opportunities where state-of-the-art techniques from each field could benefit the other.
4) For instance, we identify that transformer-based techniques used in HAR could potentially enhance VD by offering robust temporal and spatial feature extraction, which aligns well with VD's demand for high contextual awareness [11]. Additionally, GCN-based methods, which perform well in HAR by capturing relational features, could also improve VD tasks by modeling interactions within violent scenarios [12].
5) Furthermore, this review discusses how RNN-based methods for VD can be adapted to HAR applications, particularly in sequential data settings where temporal dynamics are crucial [13]. This recommendation challenges the current paradigm by suggesting that methodologies traditionally used in one domain can effectively inform the other, supporting a broader and more integrated framework for understanding HAR and VD.
6) Over 100 papers have been reviewed to come up with this review.

Our research contribution shown in Figure 1 can be summarized as follows:

1) Review and categorization of the different features used in the models
2) A review and classification of the latest machine learning (ML) and DL methods used for HAR and VD.
3) Review of the key datasets used for HAR and VD

The remainder of the document is divided into the following sections: Section 2 outlines the procedures, protocols, and criteria utilized to get the research papers, and Section 3 reviews the pertinent literature that was used for VD and HAR. A thorough discussion of the main takeaways from our review is provided in Section 4 and finally a conclusion in Section 5.

2. Materials and Methods

2.1. Review protocol

Before beginning the systematic review, a clear methodology is established, following established guidelines outlined in Keele [7]. Initially, research questions (RQs) are defined and finalized. Subsequently, relevant studies are identified by searching through databases. To ensure thorough coverage and minimize bias, both Artificial Intelligence (AI)-based HAR and VD studies across various domains were considered. These studies are then meticulously filtered and assessed using specific exclusion and quality criteria. Each selected study underwent an additional round of data validation to confirm alignment with review objectives, aiming for replicable and transparent findings. Finally, data pertinent to the RQs are extracted from the selected studies and synthesized to effectively address the research objectives.

2.2. RQs

The following RQs have been defined for this SLR study.

1) RQ1—What are the most effective techniques currently used in HAR and how can they be adapted to improve VD systems?
2) RQ2—What are the most effective techniques currently used in VD and how can they be adapted to improve HAR systems?
3) RQ3—How do the characteristics of training datasets influence the robustness and accuracy of HAR models?

2.3. Search strategy

A comprehensive search strategy was put into place to fully compile papers pertaining to HAR based and VD based on AI technologies. This process entailed methodically searching through a number of electronic databases, including mainly, IEEE Xplore, European Computer Vision Association, Computer Vision Foundation, MDPI, and Google Scholar. Keywords such as "Human Activity recognition", "deep learning", "surveillance", and "Violence detection", were used. To optimize search relevance, we implemented combinations of terms such as ("Human Activity Recognition" AND "Violence Detection"), ("HAR" AND "Deep Learning"), and ("HAR" AND "AI") to refine results.

2.4. Inclusion criteria

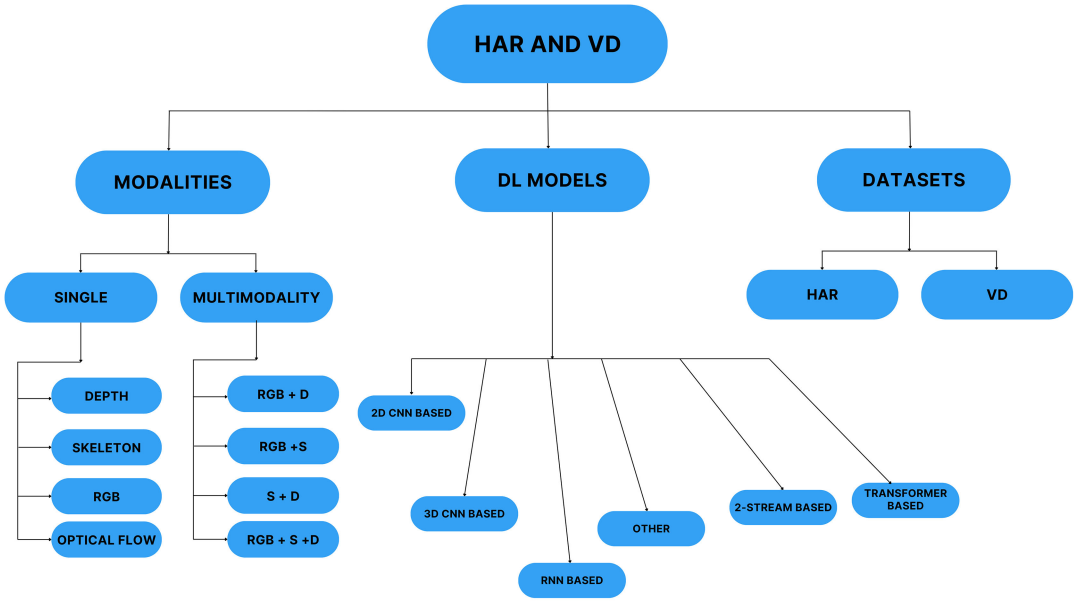
In order to be unique and differentiate ourselves from other review papers as well as to provide quality and relevant research, the inclusion criteria for the papers obtained included the following:

1) English-written Papers published from 2018 or later except for benchmark datasets and benchmark papers.
2) Papers focusing on DL and ML techniques for HAR
3) Studies based on DL and ML techniques specifically for HAR
4) Articles published in peer-reviewed journals or conference proceedings.
5) Studies containing detailed performance metrics, enabling comparative assessment for model efficacy across domains.

2.5. Selection process

A total of 200 papers were obtained from online searches through the electronic databases. From these, 52 duplicates and 45 irrelevant papers were filtered out based on their titles and abstracts to determine their relevance to come up with a total of 103 papers that were used for this review. Following this selection, all articles were screened for completeness of information regarding data preprocessing methods and model

Figure 1
Shows the structure of our review paper



evaluation metrics. Articles lacking transparency in model validation or dataset details were deprioritized to maintain consistency in quality.

3. Literature Review

3.1. Modalities (Data inputs)

Human activities can be captured through multiple data types, including RGB, skeleton, depth, infrared, point cloud, event stream, audio, acceleration, and radar. Each of these modalities offers unique benefits depending on the specific application scenarios [14].

HAR modalities can be divided into visual and non-visual categories [14]. Visual modalities, including RGB, skeleton, depth, infrared, point cloud, and event stream data, offer clear representations of activities and are crucial for applications like surveillance and autonomous navigation. On the other hand, non-visual modalities such as audio, acceleration, radar, and Wi-Fi signals provide privacy-preserving options for detecting activities [14].

3.1.1. Single modality

Single modalities have been intensively investigated. Different modalities have different strengths and weaknesses [15–17]. The following sections explore more into the various modalities and the categories they fall into. However, for the sake of this study, we will only cover the modalities that are often utilized with the most recent deep learning models.

RGB Modality

The term "RGB modality" typically refers to images or videos captured by RGB cameras to mimic human vision [18]. RGB data are usually easy to obtain and offer detailed visual information about the scene context [14]. Applications for RGB-based HAR are numerous and include sports analysis, autonomous navigation [19], and visual surveillance. Due to the differences in backgrounds, perspectives, and lighting, action recognition from RGB data is frequently difficult. Furthermore, simulating the spatiotemporal environment for HAR requires a lot of processing because RGB videos typically have big data volumes [14].

Skeleton Modality

Skeleton data capture joint trajectories and provide insight into human mobility hence making it an appropriate modality for HAR [14]. Pose estimation techniques can extract skeletal data from RGB videos [20, 21] or depth [22], and many recent skeleton-based HAR studies have used data from these sources [23, 24].

Skeleton data are advantageous for HAR as it provides information about body shape and position, offer a straightforward and informative representation, remain invariant to size, and are robust against variations in clothing textures and backgrounds [14]. Skeleton-based HAR is becoming increasingly popular in the research community due to these benefits and the affordability of precise depth sensors.

Depth modality

Depth maps are images with pixel values representing distance between points in a scene from a specific viewpoint [25]. The depth modality is immune to variations in color and texture, providing precise 3D structural and geometric shape information about humans. This makes it highly suitable for HAR [23]. Depth maps are created by converting 3D data into a 2D image [14].

Active sensors work by emitting radiation and detecting the reflected energy to gather depth information, while passive sensors rely on detecting natural energy emitted or reflected by objects in a scene [25]. Compared to active sensors like Kinect and RealSense3D, producing depth maps with passive sensors is more computationally intensive [14]. Furthermore, passive sensors can be less effective in areas that lack texture or in regions with densely repetitive patterns.

Table 1
Showing the advantages and disadvantages of the different single modalities

Modality	Advantages	Disadvantages
RGB	Provides rich appearance information.	Sensitive to viewpoint
	Easy to obtain and operate.	Sensitive to background
	Wide range of applications.	Sensitive to illumination
Skeleton	Simple yet informative	Lack of appearance information
	Insensitive to viewpoint	Lack of detailed shape information
	Insensitive to background	Noisy
Depth	Provide 3D structural information of subject pose	Lack of color and texture information
	Provide Geometric shape information	Limited workable distance

A summary of the advantages and disadvantages of the different single modalities, i.e., RGB, Skeleton, and Depth modalities, is shown in Table 1 [14].

3.1.2. Multimodalities

Humans often see their surroundings through multiple cognitive modes. Multimodal ML is a modeling strategy that processes and correlates sensory information across many modalities [26]. As described in section on single modalities, different modalities may have varying strengths. Combining multiple data modalities can boost HAR effectiveness by leveraging their complementary capabilities [14].

In HAR, two widely used strategies for combining multiple modalities are score fusion and feature fusion. Score fusion [27] involves merging independently made decisions from different modalities, often through methods like weighted averaging [28] or by employing a score fusion model [6], to derive final classification results. On the other hand, feature fusion aggregates features from various modalities to create combined features that are typically highly discriminative and effective for HAR.

Combining RGB with Depth Modalities: It takes advantage of the capabilities of both visual data types to provide a more complete knowledge of human actions [29]. However, this fusion brings with it new obstacles, such as the necessity for exact data stream alignment and synchronization, as well as greater computing complexity [29].

Combining RGB with Skeleton Modalities: The combination of RGB and skeletal modalities in HAR leverages the strengths of both modalities. However, this strategy is not without drawbacks [30] Aligning and integrating two disparate data types can be a complex procedure, and the extra computing burden may have an influence on the system's efficiency [30].

Combining Skeleton with Depth Modalities: The combination of these modalities can result in more accurate and resilient HAR systems by offering a more complete, multidimensional picture of human actions [31]. However, obstacles include the complexity of integrating two different forms of data, as well as the possibility of increased computational needs [31].

Combining RGB, Skeleton. and Depth Modalities: The fusion of these modalities provides increased performance in complex settings where one modality alone may fail, such as when sections of the body are obscured. It also allows you to capture both the appearance and dynamics of the movements [23]. However, there are certain drawbacks, such as increased processing cost and the possibility of overfitting due to the high dimensionality of the combined dataset [23].

3.2. Deep learning models

3.2.1. Two-stream-based methods

Two-stream networks are made up of two concurrent Neural Networks: one processes spatial information from individual frames, while the other gathers motion information using optical flow. This design combines static and dynamic characteristics to provide a more comprehensive knowledge of actions [32, 33].

Because video understanding requires motion information, determining a suitable technique to characterize the temporal relationship between frames is critical to improve the effectiveness of Convolutional Neural Network (CNN)-based video action identification [34].

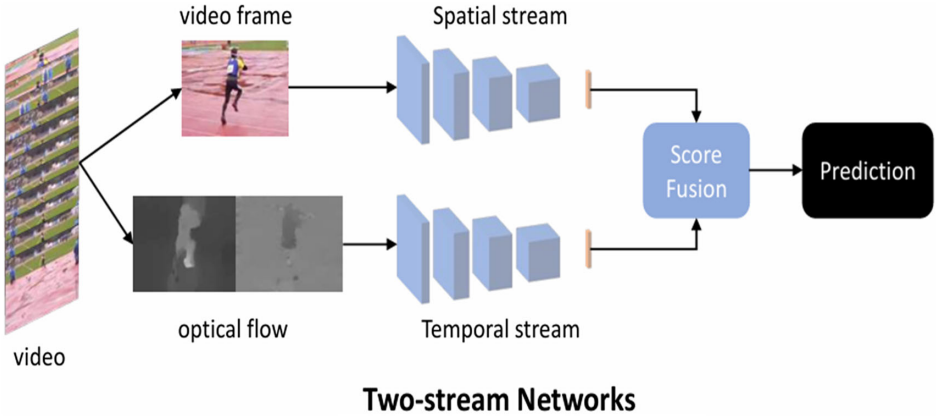
CNNs are a class of deep learning models primarily used for image processing and recognition tasks. They excel in extracting spatial features through convolutional layers, making them suitable for tasks like action recognition in videos [33, 35].

Optical flow is a useful motion model for describing object/scene movement [36]. To be more exact, it is the pattern of apparent motion of objects, surfaces, and edges in a visual scene generated by an observer's relative motion to the scene. The advantage of using optical flow is it delivers orthogonal information compared to the RGB image [36].

As a result, early research [37] introduced two-stream networks that integrate both spatial and temporal streams, as illustrated in Figure 2 [34]. The spatial stream processes raw video frames to capture visual appearance information [37], while the temporal stream uses a stack of optical flow images to capture motion information between frames [37]. The final prediction is derived by averaging the scores from both streams. This approach marked the first time a CNN-based technique achieved performance on par with the best hand-crafted feature, Improved Dense Trajectories on UCF101 (88.0% vs. 87.9%) and HMDB51 datasets [34, 38] (59.4% vs. 61.1%). Two key observations from this study are as follows: firstly, motion information is essential for video action detection, and secondly, CNNs struggle to extract temporal information directly from raw video frames

SlowFast networks for video recognition are introduced by Feichtenhofer et al. [39]. This model consists of two pathways: (i) a Slow pathway that operates at a low frame rate to capture spatial semantics and (ii) a Fast pathway that operates at a high frame

Figure 2
Shows the architecture of two-stream networks



rate to capture detailed motion at a fine temporal resolution. These models demonstrate strong performance in both action classification and detection within videos.

3.2.2. CNN-based methods

Pre-computing optical flow is both computationally intensive and storage-heavy, making it impractical for large-scale training or real-time deployment [34]. Videos can be conceptually viewed as three-dimensional (3D) tensors with two spatial dimensions and one temporal dimension, leading to the use of 3D CNNs for modeling temporal information [34]. However, 3D networks are challenging to optimize, requiring a large and diverse video dataset to train effectively. Fortunately, Sports1M [40], a substantial dataset, is available for training deep 3D networks, but the training process for Convolutional 3D Network (C3D) can take weeks. This challenge contributed to the dominance of two-stream networks based on 2D CNNs in the video action recognition field from 2014 to 2017 [34].

The landscape changed with the introduction of Inflated 3D Convolutional (I3D) Network [41], which processes video clips through stacked 3D convolutional layers. A video clip typically consists of 16 or 32 frames [41]. This innovation addressed the challenge of training 3D CNNs from scratch. With pretraining on the large-scale dataset Kinetics400, I3D achieved scores of 95.6% on UCF101 and 74.8% on HMDB51 [34]. I3D marked a significant shift, ending an era where many algorithms reported results on smaller datasets like UCF101 and HMDB51 [34].

Following the success of I3D, publications were expected to disclose their performance on Kinetics400 or other large-scale benchmark datasets, propelling video action recognition to new heights. Over the next few years, 3D CNNs advanced rapidly, becoming top performers across nearly all benchmark datasets [34].

The superior accuracy of 3D CNNs compared to 2D CNNs in residual learning is highlighted by Tran et al. [42]. Additionally, it shows that breaking down 3D convolutional filters into separate spatial and temporal components significantly boosts accuracy [43]. This research paved the way for the development of a novel spatiotemporal convolutional block, "R(2+1)D", [42]. In this figure, (a) R2D are 2D ResNets; (b) MCx are ResNets with mixed convolutions; (c) rMCx use reversed mixed convolutions; (d) R3D are 3D ResNets; and (e) R(2+1) are (2D spatial + 1D temporal) ResNets. This innovation enables the creation of CNNs that perform at least as well as, and often better than, the state-of-the-art on benchmark datasets such as Sports-1M, Kinetics, UCF101, and HMDB51 [42].

A deep CNN for classification is used in Azmat et al. [44]. For system experimentation and validation, three benchmark datasets were used: UAVGesture, DroneAction, and UAVHuman. The model achieved action recognition accuracies of 0.95, 0.90, and 0.44 on these respective datasets [44].

A novel network called the Four-Stream Adaptive CNN (FSA-CNN) [45]. The FSA-CNN boasts three key features: robustness to spatiotemporal variations, an input-adaptive activation function, and an enhancement of the traditional two-stream approach [45]. Experimental results confirmed the superiority of FSA-CNN on the NTU-RGB+D and ETRI-Activity3D datasets.

For VD:

A fine-tuned MobileNet model applied to video frames from three different violence recognition datasets is presented in Khan et al. [13], optimized with hyperparameters like learning rate, momentum, batch size, and epochs. This model achieved strong results, outperforming state-of-the-art methods on all three VD datasets [13].

In addition, a novel activity recognition method featuring an attention mechanism is introduced in Das et al. [46]. It proposes a pose-driven spatiotemporal attention mechanism using 3D ConvNets. Experimental results demonstrate that this method surpasses state-of-the-art techniques on benchmark datasets, including the Toyota Smarthome dataset [46].

Furthermore, an end-to-end deep learning model based on 3D CNNs is proposed [47]. The network consists of an initial convolutional layer, followed by three dense blocks and a global average pooling layer. The output of the global average pooling layer is fed into a fully connected layer that produces a probability score for violence. Unlike other methods, this model does not rely on hand-crafted features or RNN architectures solely for encoding temporal information. Its performance was validated on three standard datasets, showing superior recognition accuracy compared to other advanced approaches [46].

3.2.3. GCN-based methods

Graph Convolutional Networks (GCNs) apply convolutional operations on graph structures, making them suitable for modeling relationships in non-Euclidean data. In action recognition, GCNs can be employed to analyze human skeletal data or interactions

between objects, providing insights into movement dynamics that traditional CNNs may overlook. This approach is particularly beneficial for recognizing actions that involve multiple interacting entities [32, 33].

Earlier neural network implementations were designed for regular or Euclidean data, but real-world data often have a non-Euclidean graph structure [29]. This has led to advancements in GCNs [29]. While GCNs and CNNs perform similar convolution operations, GCNs are more efficient when applied to graphs [29].

In GCNs, the model learns from adjacent nodes by stacking layers of learned first-order spectral filters, activated by a nonlinear function. Essentially, GCNs take a graph with some labeled nodes as input and generate label predictions for all nodes in the graph [29].

The AutoGCN (Automatic Graph Convolutional Network), a versatile Neural Architecture Search algorithm designed for HAR using GCNs is discussed in Tempel et al. [48]. Traditionally, GCN-based methods are crafted by domain experts for specific datasets, which restricts their broader applicability [48]. AutoGCN overcomes this limitation by using a reinforcement learning controller to explore a wide search space and identify the optimal combination of hyperparameters and architecture [48].

3.2.4. RNN-based methods

Recurrent Neural Networks (RNNs) are neural networks designed to handle sequential data [33]. Given that a video is fundamentally a sequence over time, researchers have explored the use of RNNs for temporal modeling, particularly focusing on Long Short-Term Memory (LSTM) networks [29, 41]. Early efforts utilized LSTM for video action recognition within a two-stream network framework [34].

In this approach, CNN feature maps were fed into a deep LSTM network, which then aggregated frame-level CNN features to make video-level predictions [34]. LSTM was applied to each stream separately, with the final results achieved through late fusion [34]. Despite these efforts, LSTM models did not demonstrate a clear empirical advantage over the two-stream baseline [37].

A framework for activity recognition in surveillance videos is introduced in Ullah et al. [43]. The continuous video stream is initially segmented into key shots using a proposed CNN-based method that focuses on human saliency attributes [43]. Temporal properties of activities within these segments are then extracted using the convolutional layers of a FlowNet2 CNN model. Finally, a multilayer LSTM network is employed to learn long-term sequences from temporal optical flow patterns for effective activity detection [43].

Additionally, a novel approach is presented in Debnath et al. [49], that enhances feature representations from sequences of 3D body joints. This method combines a deep CNN with multi-head attention and a bidirectional LSTM network [49]. Evaluated on three datasets, including the challenging NTU-RGBD dataset, it achieves state-of-the-art results.

The impact of the attention mechanism within ConvLSTM is examined in Zhang et al. [50]. The study keeps the ConvLSTM, Res3D, and MobileNet blocks fixed while modifying the ConvLSTM component to create four variants [50]. Findings indicate that ConvLSTM significantly contributes to temporal fusion through recurrent steps, effectively learning long-term spatiotemporal features when processing spatial or spatiotemporal inputs [50].

Furthermore, a deep learning-based HAR model featuring a 3-dimensional Convolutional Network integrated with multiplicative LSTMs is proposed in Gupta et al. [51]. This model simplifies the understanding of tasks performed by individuals or groups. For real-time object detection, the model incorporates a 3D CNN, an LSTM multiplicative Recurrent Network, and Yolov6 [51]. The proposed model, demonstrates superior performance on the NTU-RGB-D, KITTI, NTU-RGB-D 120, UCF 101, and Fused datasets, achieving accuracy rates of 98.23%, 97.65%, 98.76%, 95.45%, and 97.65%, respectively.

For VD:

A convolutional LSTM (ConvLSTM) is utilized for feature extraction in VD in Ullah et al. [4], leading to the development of VD-Net. This novel approach outperformed existing VD methods by achieving a 3.9% increase in accuracy [4].

Additionally, an innovative end-to-end CNN-LSTM model designed to operate efficiently on low-cost Internet of Things (IoT) devices, such as Raspberry Pi boards, is presented in AlDahoul et al. [52]. The model is trained and evaluated on a comprehensive dataset that included RWF-2000 and RLVS-2000 [52]. It strikes a balance between performance and parameter efficiency, enabling deployment on resource-constrained IoT nodes [52].

Furthermore, a unique architecture depicted in [53] is proposed for VD using video surveillance cameras. The model leverages a U-Net-like network with MobileNet V2 as an encoder for extracting spatial features, coupled with LSTM for temporal feature extraction and classification [53]. Despite its computational efficiency, experiments demonstrated promising results with an average accuracy of 82% ± 2% and precision of 81% ± 3% on a sophisticated real-world security camera dataset derived from RWF-2000 [53].

3.2.5. Transformer-based methods

Transformers have recently emerged as powerful alternatives to RNNs in action recognition tasks [54]. They utilize self-attention mechanisms to weigh the importance of different parts of the input sequence, allowing for parallel processing and capturing long-range dependencies more effectively than RNNs. This shift towards Transformers is driven by their ability to handle complex video data without the limitations of sequential processing [54].

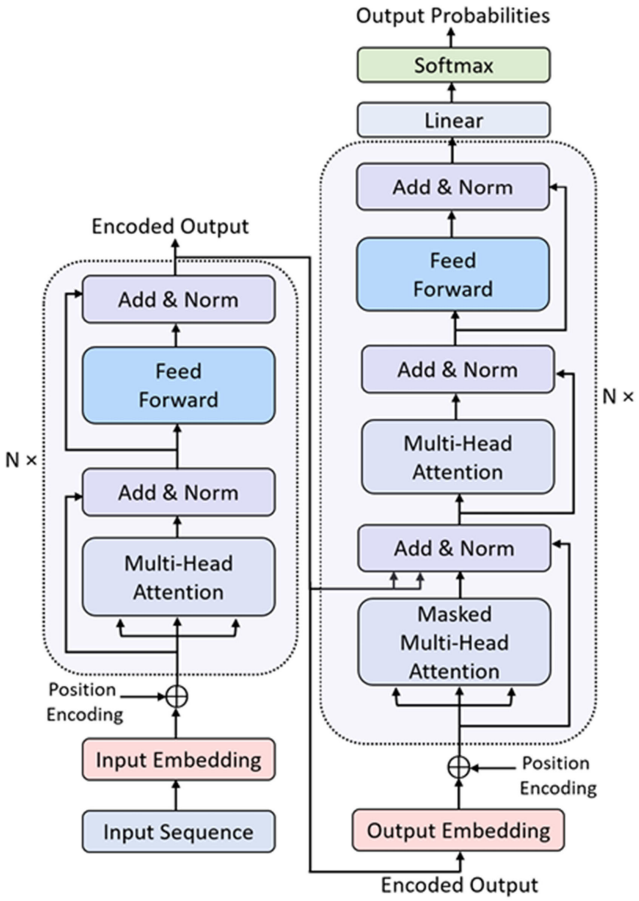
The first step in a Transformer Neural Network (TNN) involves converting input data into a tokenized, learned representation called an embedding [55]. This embedding is a vector where each element signifies a distinct learned concept. For example, in a pixel sequence, a three-element vector could represent the red, green, and blue values [55]. Each input element is transformed into an embedded vector, with each value reflecting the relevance of these concepts to the original element [55].

Transformers are specialized sequence-to-sequence models utilizing a unique form of self-attention known as “scaled dot-product attention”, replacing RNNs in both the encoder and decoder [56]. This design allows for parallel processing of the entire input sequence during encoding, significantly speeding up training and prediction [56].

The attention mechanism in TNNs is “multi-headed”, meaning it employs multiple attention blocks concurrently [55]. Each attention head has its own learned linear projection for queries, keys, and values, enabling it to focus on different parts of the input data [55]. This method enhances the model’s capacity to capture diverse aspects of the input sequence, resulting in more accurate classifications [55]. The transformer architecture is illustrated in Figure 3 [57].

The Recurrent Transformer can replace RNN layers, efficiently training on features extracted by a deep residual CNN, achieving

Figure 3
Shows the transformer architecture



accuracy comparable to traditional LSTM or Gated Recurrent Unit chains [55]. Alternatively, the Vision Transformer can be used to replace the deep residual CNN for feature extraction from frames, subsequently passing the data to the RNN layers [57].

The RGBSformer, depicted in Shi et al. [11], is a novel two-stream framework based purely on Transformers for human action recognition, integrating both RGB and skeletal modalities. This innovative framework surpasses the current state-of-the-art on four benchmark datasets: the widely used Kinetics400, NTU-RGB+D 60, and NTU-RGB+D 120, as well as the detailed FineGym99 dataset [11].

The Action Transformer is introduced in Mazzia et al. [58]. It is a straightforward, fully self-attentional architecture that consistently outperforms more complex networks combining convolutional, recurrent, and attention layers.

Additionally, in Gavrilyuk et al. [59], a transformer is fed with actor-specific static and dynamic features derived from a 2D pose network and a 3D CNN, respectively. The study explores various methods to combine these representations, highlighting their complementary advantages.

Work presented in Saidani et al. [60] leverages a transformer to enhance HAR systems by incorporating data augmentation techniques to generate a robust feature set from activity examples. This feature set is then analyzed by the transformer model, which excels at recognizing activities due to its capability to capture long-range dependencies in the data [60]. The system was tested on the PAMAP2, UCI HAR, and WISDM datasets, achieving 98.2% accuracy on PAMAP2 for 12 activities, 98.6% accuracy on UCI HAR for 6 activities, and 97.3% accuracy on WISDM for 6 activities [60].

Additionally, Saidani et al. [60] introduce UniHCP, a Unified Model for Human-Centric Perceptions, utilizing the plain vision transformer architecture to streamline a variety of human-centric tasks. UniHCP, through large-scale joint training on 33 human-centric datasets, sets new state-of-the-art benchmarks in several tasks: 69.8 mIoU on CIHP for human parsing, 86.18 mA on PA100K for attribute prediction, 90.3 mAP on Market1501 for ReID, and 85.8 JI on CrowdHuman for pedestrian detection, surpassing specialized models designed for each task [60].

3.2.6. Other methods

Works presented in Hachiuma et al. [12] tackle three major issues in traditional skeleton-based action recognition: errors in skeleton detection and tracking, a restricted range of target actions, and recognition tied to specific persons and frames. The paper introduces a point cloud deep learning approach for action recognition and proposes a unified framework featuring a novel deep neural network architecture called Structured Keypoint Pooling.

3.3. Datasets

3.3.1. Datasets used for HAR

The **Kinetics family** is currently the most widely used benchmark dataset for action recognition. It continues to expand, with Kinetics-600 releasing in 2018, containing 480,000 video clips, and Kinetics-700 following in 2019 with 650,000 videos [61, 62]. These datasets offer a diverse, large-scale foundation for training HAR models, addressing RQ3 on how dataset variability affects model generalizability. They test models' capacity to identify a wide range of actions across unconstrained settings [63].

AVA set the precedent as the first large-scale dataset for detecting spatiotemporal actions and has since expanded into AVA-Kinetics. This expanded version includes 352,091 training samples, 89,882 validation samples, and 182,457 testing samples [64].

Introduced in 2018, the **Moments in Time dataset** was designed for large-scale event comprehension and is unique in that it includes not only human actions but also activities involving animals, objects, and natural events [65]. In 2019, it was expanded into Multi-Moments in Time (M-MiT), which features 1.02 million videos, fewer ambiguous classes, and more labels per video [66].

The **HACS dataset,** introduced in 2019, focuses on recognizing and localizing human behaviors in web videos [67]. It includes two types of manual annotations: HACS Clips, with 1.55 million 2-second clip annotations across 504,000 videos, and HACS Segments, with 140,000 complete action segments in 50,000 videos [67]. Both sets are labeled with the same 200 human action classes found in ActivityNet.

Released in 2020, the **HVU dataset** supports multi-label, multi-task video interpretation. It comprises 572,000 videos with 3,142 labels and is split into 481,000 training, 31,000 validation, and 65,000 testing videos [68]. The dataset spans six task categories: scene, object, action, event, attribute, and concept, with each label typically having around 2,112 samples. Video durations vary but do not exceed 10 seconds [68].

Debuting in 2020, the **AViD dataset** was designed for anonymous action recognition. It contains 410,000 training videos and 40,000 testing videos, each lasting between 3 and 15 seconds

and featuring 887 action classes [69]. To minimize data bias, the dataset includes videos from various countries and removes facial identifiers to protect creators' privacy. Consequently, AViD may not be suitable for face-related activity detection [69].

A summary of the state-of-the-art in HAR is shown below in Table 2.

3.3.2. Datasets used for VD

In VD tasks, below are some of the most used public benchmark datasets:

1) **Hockey Fights Dataset** [47, 95]: This dataset comprises 1,000 video clips captured from hockey games, each containing 50

Table 2
Showing the state-of-the-art methods that have been used on the different datasets used for HAR

Dataset (Application)	Type	Model	Acc (%)	Year	Paper
NTU-60 (HAR in Daily Life)	Transformer	Hulk (Finetune, ViT-L)	94.3	2023	[70]
	Two stream, 3D CNNs	PoseC3D [3D Heatmap]	94.1	2021	[71]
	Transformer	Skate Former	93.5	2024	[72]
	GCN	LA-GCN	93.5	2023	[73]
NTU-RGB+D 120 (HAR in Human-Computer Interaction)	Two stream, 3D CNNs	PoseC3D (RGB + Pose)	95.3	2022	[71]
	Transformer	π-ViT (RGB + Pose)	95.1	2023	[74]
	Two stream	MMNet (RGB + Pose)	94.4	2022	[75]
	Transformer	EPP-Net (Parsing + Pose)	92.8	2024	[76]
	Transformer	STAR-Transformer (RGB + Pose)	92.7	2022	[77]
	Transformer	IPP-Net (Parsing + Pose)	91.7	2023	[78]
Kinetics Skeleton (HAR in Surveillance)	2D CNNs, GCN	Structured Keypoint Pooling	52.3	2023	[12]
	3D CNNs	PoseC3D (SlowOnly-346)	49.1	2021	[71]
	GCN	HD-GCN	40.9	2022	[79]
	Two stream	2s-AGCN+TEM	38.6	2020	[80]
Epic Kitchen (HAR in Daily Life and HCI)	3D CNNs	TPN	61.1	2020	[81]
	Transformer	MTV-B (320p)	48.6	2022	[82]
	Transformer	ViViT-L/16x2 FE	44.0	2021	[83]
HMDB51 (HAR in Competitive Sports)	Transformer	ViT-ReT	78.4	2023	[84]
	Transformer	ViT and multilayer LSTM	73.7	2022	[85]
	Transformer	SVT	67.2	2021	[86]
	3D CNNs	MiCT-Net	63.8	2018	[87]
Kinetics 400 (HAR in Competitive Sports)	Transformer	MTV-H (WTS)	89.1	2022	[39]
	Transformer	ViViT-H/16x2 (JFT)	84.9	2021	[63]
	3D CNNs	SlowFast 16×8, R101+NL	79.8	2018	[88]
	3D CNNs	X3D-XL	79.1	2020	[89]
	Transformer	SVT	78.1	2022	[86]
	3D CNNs	IntegralAction ($\lambda = 1.5$)	73.3	2021	[90]
Kinetics 600 (HAR in Competitive Sports)	Transformer	MTV-H (WTS)	89.6	2022	[39]
	Transformer	ViViT-H/16x2 (JFT)	85.8	2021	[63]
	3D CNNs	X3D-XL	81.9	2020	[91]
	2D CNNs	SlowFast 16×8, R101+NL	81.8	2018	[82]
Moments in time (HAR in Daily Life)	Transformer	MTV-H (WTS 280p)	47.2	2022	[82]
	Transformer	ViViT-L/16x2 FE	38.5	2021	[83]
N-UCLA (HAR in Human-Computer Interaction)	RNN	AGC-LSTM	93.3	2019	[79]
	GCN	HD-GCN	89.1	2023	[92]
Something Something (HAR in Daily Life)	Transformer	MTV-B (320p)	68.5	2022	[81]
	3D CNNs	TPN	66.9	2020	[82]
	Transformer	ViViT-L/16x2 FE	65.9	2021	[83]
UCF101 (HAR in Competitive Sports)	Transformer	ViT-ReT	94.7	2023	[93]
	Transformer	SVT	93.7	2021	[91]
	RNN	IP-LSTM	91.4	2019	[84]
	3D CNNs	MiCT-Net	89.1	2018	[85]
	RNN	IP-LSTM	88.5	2019	[86]

(*Continued*)

Table 2
(*Continued*)

Dataset (Application)	Type	Model	Acc (%)	Year	Paper
	3D CNNs	3DCNN	79.9	2022	[94]
	3D CNNs	3D-ShuffleNetV2	77.9	2019	[57]
UCF50 (HAR in	Transformer	ViT-ReT	97.1	2023	[94]
Competitive	Transformer	ViT and multilayer LSTM	96.1	2022	[87]
Sports)	RNN	MobileNet + BILSTM	87.0	2022	[57]

Abbreviations
Acc: Accuracy, ViT: Vision Transformer, AGC-LSTM: Attention-Guided Convolutional Long Short-Term Memory, ViViT: Video Vision Transformer, MTV-H/MTV-B: Multimodal Transformer Variants (H = Higher capacity, B = Baseline), R101+NL: ResNet-101 with Non-Local Block, TPN: Temporal Pyramid Network, SVT: Spatiotemporal Vision Transformer, IP-LSTM: Iterative Pose LSTM, MiCT-Net: Mixed 3D Convolution with Temporal Shift, X3D: Expandable 3D, AGCN: Adaptive Graph Convolutional Network.

frames at a resolution of 720×576 pixels. All videos share a consistent background and depict activities like fights and regular gameplay. Its structured nature provides a controlled environment for VD research, allowing for a comparative baseline that highlights model performance in repetitive, single-action scenes [47]. Such a dataset is valuable for RQ3 as it isolates models' VD accuracy in less complex environments.

2) **Movies Dataset** [47, 53, 95]: Consisting of 200 video clips sourced from action movies, this dataset varies in resolution and content compared to the Hockey Fights dataset.
3) **Violent-Flows Dataset** [47]: This dataset includes 246 movies depicting crowd behavior, scaled to 320 × 240 pixels. It presents greater challenges due to multiple viewpoints, noisy backgrounds, and dense crowds. The dataset encompasses a mix of violent and non-violent scenes.
4) **RW2000 (Real-World Fighting) dataset** [52, 53, 96]: Released in 2019, RW2000 addresses shortcomings in image quality, data quantity, annotation accuracy, and the realism of video sources. It contains 2,000 edited video clips sourced from YouTube and captured by surveillance cameras in real-world environments [96]. Due to its focus on authentic surveillance footage, RWF-2000 allows for the testing of VD models' effectiveness in capturing high-risk actions within complex, real-world scenarios, directly supporting RQ3 on the influence of dataset realism in model.

A summary of the state-of-the-art in VD is shown below in Tables 3 and 4.

Table 3
Showing the state-of-the-art methods that have been used on the different datasets used for VD

Dataset	Type	Model	Accuracy	Year	Paper
RW-2000	2D CNNs, GCN	Structured Keypoint pooling	93.4	2023	[12]
	Reinforcement Learning, 3D CNNs	Semi-Supervised Hard Attention (SSHA)	90.4	2022	[97]
		Human Skeletons + Change Detection	90.2	2022	[98]
	RNN	Separable Convolutional LSTM	89.7	2021	[98]
	3D CNNs	SPIL Convolution	89.3	2020	[99]
	Two stream	Two stream (3D-CNN + 2D-CNN)	88.7	2022	[100]
	Two stream	Flow Gated Network	87.25	2019	[101]
Hockey Fight	RNN	ViolenceNet	99.2	2021	[99]
	Two stream	Two stream (3D-CNN + 2D-CNN)	97.3	2022	[99]
	3D CNNs	SPIL	96.8	2020	[99]
	RNN	VGG+LSTM	95.1	2019	[102]
Movies Fight	RNN	ViolenceNet	100.0	2021	[99]
	3D CNNs	SPIL	98.5	2020	[99]
RLVS	RNN	ViolenceNet	95.6	2021	[102]
	Two stream	Two stream (3D-CNN + 2D-CNN)	92.8	2022	[100]
	RNN	VGG+LSTM Fine-tuned	88.2	2019	[34]
Violent Flow	RNN	ViolenceNet	96.9	2021	[100]
	RNN	VGG+LSTM	90.0	2019	[100]

Abbreviations
SSHA: Semi-Supervised Hard Attention, SPIL: Spatial Pyramid Pooling for Image Localization, VGG: Visual Geometry Group (a type of CNN architecture)

Table 4
Showing the progression of works done in VD overtime

Type	Model	Year	Reference
Two-stream	Two-stream (3D-CNN + 2D-CNN)	2022	[34]
Two-stream	Two-stream (3D-CNN + 2D-CNN)	2022	[34]
Two-stream	Two-stream (3D-CNN + 2D-CNN)	2022	[34]
RNN	ViolenceNet	2021	[100]
RNN	ViolenceNet	2021	[100]
RNN	ViolenceNet	2021	[100]
RNN	ViolenceNet	2021	[100]
3D CNNs	SPIL	2020	[99]
3D CNNs	SPIL	2020	[99]
3D CNNs	SPIL	2020	[102]
3D CNNs	SPIL	2020	[102]
GCN	HL-Net	2020	[102]
RNN	VGG+LSTM	2019	[99]
RNN	VGG+LSTM	2019	[99]
RNN	VGG+LSTM Fine-tuned	2019	[99]
3D CNNs	Flow Gated	2019	[99]

Table 5
Relating the different datasets used in HAR and how they were developed and their top accuracies

Dataset	Environment	Sourcing	Year	Top accuracy
NTU – 60	Constrained	Laboratory	2016	94.3
NTU – 120	Constrained	Laboratory	2019	95.3
Kinetics Skeleton	Unconstrained	Crowdsourcing	2017	52.3
Epic Kitchens	Unconstrained	Crowdsourcing	2018	61.6
HMDB51	Unconstrained	Crowdsourcing	2011	78.4
Kinetics 400	Unconstrained	Crowdsourcing	2017	89.1
Kinetics 600	Unconstrained	Crowdsourcing	2018	89.6
Moments in Time	Unconstrained	Crowdsourcing	2018	47.2
N-UCLA	Constrained	Laboratory	2017	93.3
Something Something	Unconstrained	Crowdsourcing	2017	68.5
UCF101	Unconstrained	Crowdsourcing	2013	94.7

4. Discussion

4.1. Model performance on real-world datasets

For HAR, evidence from Table 5 points towards better performance on datasets in constrained environments and poor performance on datasets in unconstrained environments. Transformer-based HAR methods typically achieved 7–10% higher accuracy on datasets with structured environments. This trend can be attributed to the controlled conditions present in constrained datasets, which allow models to learn more effectively without the noise and variability often found in real-world scenarios. This opens up new research opportunities because, to train robust models that can generalize to real-world scenarios, they need to be trained on datasets that depict real-world environments. Furthermore, exploring data augmentation techniques and incorporating diverse environmental conditions during training could enhance model robustness in unconstrained settings. Another avenue for improvement is on model accuracies for datasets in unconstrained environments.

For VD, evidence from Table 6 shows that most datasets used depict actions occurring in unconstrained environments, and hence, models can easily generalize better. The high accuracy rates observed indicate that VD models benefit from the variability present in these datasets, which helps them learn distinguishing features of violent actions amidst diverse backgrounds and contexts. There is still some room for improvement of model accuracies though smaller than for HAR. Also, less work has been done for VD and this could probably be attributed to the having less datasets saturated with good performing models.

4.2. Common characteristics of good performing methodologies

For HAR, evidence points towards the following as characteristics for the best-performing methods:

1) Transformer-based methods consistently outperform other methods used likely due to their ability to capture long-range dependencies and contextual information effectively. And they even perform better when combined with a recurrent component.
2) 3DCNN-based methods come in second and are only occasionally performed by transformer methods when it comes to datasets in unconstrained environments. This may be due to their spatial-temporal modeling capabilities, which are crucial for action recognition tasks involving movement patterns over time.

For VD, evidence points towards the following as characteristics for the best-performing methods:

Table 6
Relating the different datasets used in HAR and how they were developed and their top accuracies

Dataset	Environment	Sourcing	Year	Top accuracy
RW – 2000	Unconstrained	Crowdsourcing	2019	93.4
Hockey Fights	Unconstrained	Crowdsourcing	2011	99.2
Movies Fight	Unconstrained	Crowdsourcing	2011	100
RLVS	Unconstrained	Crowdsourcing	2019	95.6
Violent Flow	Unconstrained	Crowdsourcing	2012	96.9

1) RNN-based methods outperform other methods and can easily generalize in unconstrained natural environments. This adaptability may be due to their sequential processing capabilities, allowing them to effectively analyze temporal dynamics inherent in violent actions across various contexts.

4.3. Possible avenues for knowledge transfer

HAR to VD

We recommend implementing transformer-based methods for VD, as these have demonstrated superior performance over RNN-based methods by approximately 5–10% on HAR datasets as evidenced in Tables 2 and 5 [70, 72]. To the extent of our knowledge, no transformer-based methods have yet been applied to VD tasks. This approach could introduce advanced feature extraction capabilities into VD systems. This could introduce advanced feature extraction capabilities into VD tasks.

We propose adapting GCN-based methods for VD applications, as these have shown a 5–8% accuracy improvement over 2D CNN-based methods on HAR datasets with relational dynamics according to data in Tables 3 and 5 [12, 73]. To the extent of our knowledge, no GCN-based methods have been applied to VD tasks. Leveraging graph structures could enhance understanding of interactions between entities involved in violent scenarios. Leveraging graph structures could enhance understanding of interactions between entities involved in violent scenarios.

By assessing and recommending cross-domain applications of HAR and VD techniques, this review provides a pathway for addressing real-world challenges in each field, filling critical gaps related to model adaptability, dataset limitations, and feature extraction requirements

One promising proposition is to use pre-trained HAR models in VD applications. For instance, HAR models trained with large-scale datasets, such as Kinetics, can be fine-tuned on VD datasets like RWF-2000 to capture nuanced actions and reduce computational costs. Structured transfer learning experiments can help identify the optimal balance between preserving HAR-learned features and adapting them to recognize violence-specific patterns in new environments (e.g., surveillance settings).

VD to HAR

Experimentation of RNN-based methods which achieve 5–10% higher accuracy than traditional 2D CNN methods on VD datasets, due to their sequential processing capabilities, as shown in Tables 4 and 6 [98, 102]. This could yield valuable insights into temporal dynamics relevant for action recognition tasks; this cross-domain application may help improve accuracy rates by incorporating learned sequential patterns from VD models.

Transformer and RNN-based methods can be improved and fine-tuned to provide better performance on unconstrained datasets.

4.4. Cross-domain knowledge transfer challenges

One of the main challenges in cross-domain knowledge transfer between HAR and VD lies in adapting spatial-temporal feature extraction techniques across different contexts. Specifically, HAR techniques optimized for structured activities in constrained settings may not directly translate to the unstructured, high-variance scenarios typically found in VD [13].

Additionally, the fusion of methodologies, such as using GCNs for relationship modeling in VD from HAR, requires alignment of domain-specific parameters and dataset preprocessing, which can vary significantly [12]. These challenges highlight the need for refined pretraining techniques and flexible model architectures that can accommodate the specificities of each domain without loss of accuracy.

5. Conclusion

This review delivers a comprehensive evaluation of recent advancements in techniques, datasets, and methodologies within HAR and VD. While the field of VD is still in development and is narrower in scope compared to HAR, it benefits from unique application demands and challenges, which are addressed by examining HAR's established techniques.

Our findings suggest that future research in HAR and VD can be enriched through cross-domain adaptation of methods, such as the implementation of transformer and GCN-based models across both domains. Specifically, transformer architectures, which excel in temporal-spatial feature handling in HAR, could significantly benefit VD tasks in surveillance settings where detailed contextual understanding is critical.

Furthermore, our analysis indicates that VD models could leverage RNN-based techniques from HAR to enhance sequential data processing in complex scenarios. These adaptations not only have the potential to improve model performance but also open up possibilities for robust feature extraction and real-time monitoring applications in varied, dynamic environments.

This review underscores the potential of a hybridized framework that combines HAR and VD methodologies, suggesting pathways for creating adaptive models capable of addressing diverse recognition challenges across constrained and unconstrained environments.

Future research can expand on this by focusing on models that handle multimodal data, which may enhance system robustness and accuracy. By doing so, this review serves as a foundation for developing HAR and VD models with the flexibility to operate effectively across domains and application contexts, potentially reshaping the landscape of automated surveillance and safety technology.

Additionally, future research should prioritize evaluating cross-domain models in live, operational settings, where HAR-to-VD adaptation can be practically tested for responsiveness to real-world events. Such implementations are particularly relevant in healthcare monitoring and security systems, where adaptability is crucial. Field tests will help refine cross-domain techniques by revealing context-specific model adjustments, and these insights will pave the way for autonomous, multi-functional systems capable of reliably handling diverse scenarios.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Author Contribution Statement

Paul Turyahabwa: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Writing –

original draft, Writing – review & editing, Visualization, Supervision, Project administration. **Sudi Murindanyi:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

References

[1] Karim, M., Khalid, S., Aleryani, A., Khan, J., Ullah, I., & Ali, Z. (2024). Human action recognition systems: A review of the trends and state-of-the-art. *IEEE Access*, *12*, 36372–36390. https://doi.org/10.1109/ACCESS.2024.3373199

[2] Chen, Y., Wang, J., Hao, S., Peng, X., & Hu, L. (2019). Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*, *119*, 3–11. https://doi.org/10.1016/j.patrec.2018.02.010

[3] Nweke, H. F., Teh, Y. W., Al-Garadi, M. A., & Alo, U. R. (2018). Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: state of the art and research challenges. *Expert Systems with Applications*, *105*, 233–261. https://doi.org/10.1016/j.eswa.2018.03.056

[4] Ullah, F. U. M., Muhammad, K., Haq, I. U., Khan, N., Heidari, A. A., Baik, S. W., & de Albuquerque, V. H. C. (2021). AI-assisted edge vision for violence detection in IoT-based industrial surveillance networks. *IEEE Transactions on Industrial Informatics*, *18*(8), 5359–5370. https://doi.org/10.1109/TII.2021.3116377

[5] Sultani, W., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 6479–6488. https://doi.org/10.48550/arXiv.1801.04264

[6] Wang, L., Ding, Z., Tao, Z., Liu, Y., & Fu, Y. (2019). Generative multi-view human action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6212–6221.

[7] Keele, S. (2007). *Guidelines for performing systematic literature reviews in software engineering*. Technical report, 5. Retrieved from: https://www.researchgate.net/publication/302924724_Guidelines_for_performing_Systematic_Literature_Reviews_in_Software_Engineering

[8] Guchait, P., Nimse, C., Pradhan, C., Fernandes, P., & Kharche, S. (2023). Violence detection using human action recognition. *International Journal of Emerging Technologies and Innovative Research*, *10*(7), a548–a553. http://www.jetir.org/papers/JETIR2307068

[9] Janani, P., Suratgar, A., & Taghvaeipour, A. (2024). Enhancing human action recognition and violence detection through deep learning audiovisual fusion. *arXiv Preprint: 2408.02033.* https://doi.org/10.48550/arXiv.2408.02033

[10] Lopez, D. J. D., & Lien, C. C. (2023). Two-stage complex action recognition framework for real-time surveillance automatic violence detection. *Journal of Ambient Intelligence and Humanized Computing*, *14*(12), 15983–15996. https://doi.org/10.1007/s12652-023-04679-6

[11] Shi, J., Zhang, Y., Wang, W., Xing, B., Hu, D., & Chen, L. (2023). A novel two-stream transformer-based framework for multi-modality human action recognition. *Applied Sciences*, *13*(4), 2058. https://doi.org/10.3390/app13042058

[12] Hachiuma, R., Sato, F., & Sekii, T. (2023). Unified keypoint-based action recognition framework via structured keypoint pooling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22962–22971. https://doi.org/10.1109/CVPR52729.2023.02199

[13] Khan, S. U., Haq, I. U., Rho, S., Baik, S. W., & Lee, M. Y. (2019). Cover the violence: A novel deep-learning-based approach towards violence-detection in movies. *Applied Sciences*, *9*(22), 4963. https://doi.org/10.3390/app9224963

[14] Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., & Liu, J. (2022). Human action recognition from various data modalities: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(3), 3200–3225. https://doi.org/10.1109/TPAMI.2022.3183112

[15] Wang, Y., Xiao, Y., Xiong, F., Jiang, W., Cao, Z., Zhou, J. T., & Yuan, J. (2020). 3DV: 3D dynamic voxel for action recognition in depth video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 511–520. https://doi.org/10.48550/arXiv.2005.05501

[16] Ghosh, R., Gupta, A., Nakagawa, A., Soares, A., & Thakor, N. (2019). Spatiotemporal filtering for event-based action recognition. *arXiv Preprint:1903.07067*. https://doi.org/10.48550/arXiv.1903.07067

[17] Liang, D., & Thomaz, E. (2019). Audio-based activities of daily living (ADL) recognition with large-scale acoustic embeddings from online videos. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 3*(1), 1–18. https://doi.org/10.1145/3314404

[18] Kumar, R., & Kumar, S. (2023). Multi-view multi-modal approach based on 5S-CNN and BiLSTM using skeleton, depth and RGB data for human activity recognition. *Wireless Personal Communications*, *130*(2), 1141–1159. https://doi.org/10.1007/s11277-023-10324-4

[19] Lu, M., Hu, Y., & Lu, X. (2020). Driver action recognition using deformable and dilated faster R-CNN with optimized region proposals. *Applied Intelligence*, *50*(4), 1100–1111. https://doi.org/10.1007/s10489-019-01603-4

[20] Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5693–5703. https://doi.org/10.48550/arXiv.1902.09212

[21] Gong, J., Fan, Z., Ke, Q., Rahmani, H., & Liu, J. (2022). Meta agent teaming active learning for pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11079–11089. https://doi.org/10.1109/CVPR52688.2022.01080

[22] Rahmani, H., Liu, J., Akhtar, N., & Mian, A. (2019). Learning human pose models from synthesized data for robust RGB-D action recognition. *International Journal of Computer Vision*, *127,* 1545–1564. https://doi.org/10.1007/s11263-019-01192-2

[23] Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L. Y., & Kot, A. C. (2019). NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(10), 2684–2701. https://doi.org/10.1109/TPAMI.2019.2916873

[24] Yan, S., Xiong, Y., & Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence, 32*(1). https://doi.org/10.1609/aaai.v32i1.12328

[25] Zollhöfer, M., Stotko, P., Görlitz, A., Theobalt, C., Nießner, M., Klein, R., & Kolb, A. (2018). State of the art on 3D reconstruction with RGB-D cameras. In *Computer*

Graphics Forum, 37(2), 625–652. https://doi.org/10.1111/cgf.13386

[26] Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(2), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607

[27] Dhiman, C., & Vishwakarma, D. K. (2020). View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics. *IEEE Transactions on Image Processing*, *29*, 3835–3844. https://doi.org/10.1109/TIP.2020.2965299

[28] Rani, S. S., Naidu, G. A., & Shree, V. U. (2021). Kinematic joint descriptor and depth motion descriptor with convolutional neural networks for human action recognition. *Materials Today: Proceedings*, *37*, 3164–3173. https://doi.org/10.1016/j.matpr.2020.09.052

[29] Shaikh, M. B., & Chai, D. (2021). RGB-D data-based action recognition: A review. *Sensors*, *21*(12), 4246. https://doi.org/10.3390/s21124246

[30] Liu, J., Shahroudy, A., Xu, D., Kot, A. C., & Wang, G. (2017). Skeleton-based action recognition using spatio-temporal LSTM network with trust gates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(12), 3007–3021. https://doi.org/10.1109/TPAMI.2017.2771306

[31] Tang, Y., Tian, Y., Lu, J., Li, P., & Zhou, J. (2018). Deep progressive reinforcement learning for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 5323–5332. https://doi.org/10.1109/CVPR.2018.00558

[32] Silva, D., Manzo-Martínez, A., Gaxiola, F., Gonzalez-Gurrola, L., & Ramírez-Alonso, G. (2022). Analysis of CNN architectures for human action recognition in video. *Computacióny Sistemas*, *26*(2), 623–641. https://doi.org/10.13053/cys-26-2-4245

[33] Alomar, K., Aysel, H. I., & Cai, X. (2024). RNNs, CNNs and transformers in human action recognition: A survey and a hybrid model. *arXiv Preprint: 2407.06162*. https://arxiv.org/html/2407.06162v2

[34] Zhu, Y., Li, X., Liu, C., Zolfaghari, M., Xiong, Y., Wu, C., . . . , & Li, M. (2020). A comprehensive study of deep video action recognition. *arXiv Preprint: 2012.06567*. https://doi.org/10.48550/arXiv.2012.06567

[35] Jingfei Wang, L. Z. (2024). Basketball technique action recognition using 3D convolutional neural networks. *Scientific Reports*, *14*, 13156. https://doi.org/10.1038/s41598-024-63621-8

[36] Horn, B. K. P., & Schunck, B. G. (1981). Determining optical flow. *Artificial Intelligence*, *17*(1–3), 185–203. https://doi.org/10.1016/0004-3702(81)90024-2

[37] Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. *Advances in Neural Information Processing Systems*, *27*. https://doi.org/10.48550/arXiv.1406.2199

[38] Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011). HMDB: A large video database for human motion recognition. In *2011 International Conference on Computer Vision,* 2556–2563. https://doi.org/10.1109/ICCV.2011.6126543

[39] Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6202–6211. https://doi.org/10.48550/arXiv.1812.03982

[40] Varadarajan, B., Toderici, G., Vijayanarasimhan, S., & Natsev, A. (2015). Efficient large scale video classification. *arXiv Preprint: 1505.06250*. https://doi.org/10.48550/arXiv.1505.06250

[41] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. http://dx.doi.org/10.1162/neco.1997.9.8.1735

[42] Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., & Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6450–6459. https://doi.org/10.48550/arXiv.1711.11248

[43] Ullah, A., Muhammad, K., Del Ser, J., Baik, S. W., & de Albuquerque, V. H. C. (2018). Activity recognition using temporal optical flow convolutional features and multilayer LSTM. *IEEE Transactions on Industrial Electronics*, *66*(12), 9692–9702. https://doi.org/10.1109/TIE.2018.2881943

[44] Azmat, U., Alotaibi, S. S., Abdelhaq, M., Alsufyani, N., Shorfuzzaman, M., Jalal, A., & Park, J. (2023). Aerial insights: Deep learning-based human action recognition in drone imagery. *IEEE Access*, *11*, 83946–83961. https://doi.org/10.1109/ACCESS.2023.3302353

[45] Jang, J., Kim, D., Park, C., Jang, M., Lee, J., & Kim, J. (2020). ETRI-activity3D: A large-scale RGB-D dataset for robots to recognize daily activities of the elderly. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems,* 10990–10997. https://doi.org/10.1109/IROS45743.2020.9341160

[46] Das, S., Dai, R., Koperski, M., Minciullo, L., Garattoni, L., Bremond, F., & Francesca, G. (2019). Toyota smarthome: Real-world activities of daily living. In *Proceedings of the IEEE/CVF International Conference on Computer Vision,* 833–842. https://doi.org/10.1109/ICCV.2019.00092

[47] Li, J., Jiang, X., Sun, T., & Xu, K. (2019). Efficient violence detection using 3D convolutional neural networks. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance*, 1–8. https://doi.org/10.1109/AVSS.2019.8909883

[48] Tempel, F., Ihlen, E. A. F., & Strümke, I. (2024). AutoGCN-towards generic human activity recognition with neural architecture search. *IEEE Access*, *12*, 39505–39516. https://doi.org/10.1109/ACCESS.2024.3377103

[49] Debnath, B., O'Brient, M., Kumar, S., & Behera, A. (2021). Attention-driven body pose encoding for human activity recognition. In *2020 25th International Conference on Pattern Recognition,* 5897–5904. https://doi.org/10.1109/ICPR48806.2021.9412487

[50] Zhang, L., Zhu, G., Mei, L., Shen, P., Shah, S. A. A., & Bennamoun, M. (2018). Attention in convolutional LSTM for gesture recognition. *Advances in Neural Information Processing Systems, 31*.

[51] Gupta, C., Gill, N. S., Gulia, P., Yadav, S., Pau, G., Alibakhshikenari, M., & Kong, X. (2023). A real-time 3-dimensional object detection based human action recognition model. *IEEE Open Journal of the Computer Society*, *5*, 14–26. https://doi.org/10.1109/OJCS.2023.3334528

[52] AlDahoul, N., Karim, H. A., Datta, R., Gupta, S., Agrawal, K., & Albunni, A. (2021). Convolutional neural network-long

short term memory based IOT node for violence detection. In *2021 IEEE International Conference on Artificial Intelligence in Engineering and Technology*, 1–6. https://doi.org/10.1109/IICAIET51634.2021.9573691

[53] Vijeikis, R., Raudonis, V., & Dervinis, G. (2022). Efficient violence detection in surveillance. *Sensors*, *22*(6), 2216. https://doi.org/10.3390/s22062216

[54] Shaikh, M. B., Chai, D., Islam, S. M. S., & Akhtar, N. (2024). From CNNs to transformers in multimodal human action recognition: A survey. *ACM Transactions on Multimedia Computing, Communications and Applications*, *20*, 1–24. https://doi.org/10.48550/arXiv.2405.15813

[55] Wensel, J. (2022). *Transformer neural networks for human activity recognition.* Doctoral Dissertation, Kansas State University.

[56] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . , & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*(2017).

[57] Wensel, J., Ullah, H., & Munir, A. (2023). ViT-ReT: Vision and recurrent transformer neural networks for human activity recognition in videos. *IEEE Access*, *11*, 72227–72249. https://doi.org/10.1109/ACCESS.2023.3293813

[58] Mazzia, V., Angarano, S., Salvetti, F., Angelini, F., & Chiaberge, M. (2022). Action transformer: A self-attention model for short-time pose-based human action recognition. *Pattern Recognition*, *124*, 108487. https://doi.org/10.48550/arXiv.2107.00606

[59] Gavrilyuk, K., Sanford, R., Javan, M., & Snoek, C. G. (2020). Actor-transformers for group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 839–848. https://doi.org/10.48550/arXiv.2003.12737

[60] Saidani, O., Alsafyani, M., Alroobaea, R., Alturki, N., Jahangir, R., & Jamel, L. (2023). An efficient human activity recognition using hybrid features and transformer model. *IEEE Access*, *11*, 101373–101386. https://doi.org/10.1109/ACCESS.2023.3314492

[61] Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C., & Zisserman, A. (2018). A short note about kinetics-600. *arXiv Preprint: 1808.01340*. https://doi.org/10.48550/arXiv.1808.01340

[62] Carreira, J., Noland, E., Hillier, C., & Zisserman, A. (2019). A short note on the kinetics-700 human action dataset. *arXiv Preprint: 1907.06987*. .https://doi.org/10.48550/arXiv.1907.06987

[63] Feichtenhofer, C. (2020). X3D: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 203–213. https://doi.org/10.48550/arXiv.2004.04730

[64] Li, S., Neupane, A., Paul, S., Song, C., Krishnamurthy, S. V., Chowdhury, A. K. R., & Swami, A. (2018). Adversarial perturbations against real-time video classification systems. *arXiv Preprint:1807.00458*. https://doi.org/10.14722/ndss.2019.23202

[65] Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S. A., Yan, T., . . . , & Oliva, A. (2019). Moments in time dataset: One million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(2), 502–508. https://doi.org/10.1109/TPAMI.2019.2901464

[66] Monfort, M., Pan, B., Ramakrishnan, K., Andonian, A., McNamara, B. A., Lascelles, A., . . . , & Oliva, A. (2021). Multi-moments in time: Learning and interpreting models for multi-action video understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(12), 9434–9445. https://doi.org/10.1109/TPAMI.2021.3126682

[67] Zhao, H., Torralba, A., Torresani, L., & Yan, Z. (2019). HACS: Human action clips and segments dataset for recognition and temporal localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8668–8678. https://doi.org/10.1109/ICCV.2019.00876

[68] Diba, A., Fayyaz, M., Sharma, V., Paluri, M., Gall, J., Stiefelhagen, R., & Van Gool, L. (2020). Large scale holistic video understanding. In *Computer Vision–ECCV 2020: 16th European Conference*, 593–610. https://doi.org/10.1007/978-3-030-58558-7_35

[69] Piergiovanni, A. J., & Ryoo, M. (2020). Avid dataset: Anonymized videos from diverse countries. *Advances in Neural Information Processing Systems*, *33*, 16711–16721. https://doi.org/10.48550/arXiv.2007.05515

[70] Wang, Y., Wu, Y., Tang, S., He, W., Guo, X., Zhu, F., . . . , & Ouyang, W. (2023). Hulk: A universal knowledge translator for human-centric tasks. *arXiv Preprint: 2312.01697*. https://doi.org/10.48550/arXiv.2312.01697

[71] Duan, H., Zhao, Y., Chen, K., Lin, D., & Dai, B. (2022). Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2969–2978. https://doi.org/10.1109/CVPR52688.2022.00298

[72] Do, J., & Kim, M. (2025). SkateFormer: Skeletal-temporal transformer for human action recognition. In *European Conference on Computer Vision*, 401–420. https://doi.org/10.48550/arXiv.2403.09508

[73] Xu, H., Gao, Y., Hui, Z., Li, J., & Gao, X. (2023). Language knowledge-assisted representation learning for skeleton-based action recognition. *arXiv Preprint: 2305.12398*. https://doi.org/10.48550/arXiv.2305.12398

[74] Reilly, D., & Das, S. (2024). Just add?! Pose induced video transformers for understanding activities of daily living. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 18340–18350. https://doi.org/10.48550/arXiv.2311.18840

[75] Bruce, X. B., Liu, Y., Zhang, X., Zhong, S. H., & Chan, K. C. (2022). Mmnet: A model-based multimodal network for human action recognition in RGB-D videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(3), 3522–3538. https://doi.org/10.1109/TPAMI.2022.3177813

[76] Liu, J., Ding, R., Wen, Y., Dai, N., Meng, F., Zhang, F. L., . . . , & Liu, M. (2024). Explore human parsing modality for action recognition. *CAAI Transactions on Intelligence Technology*, *29*, 1623–1633. https://doi.org/10.48550/arXiv.2401.02138

[77] Ahn, D., Kim, S., Hong, H., & Ko, B. C. (2023). Star-transformer: A spatio-temporal cross attention transformer for human action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3330–3339. https://doi.org/10.1109/WACV56688.2023.00333

[78] Ding, R., Wen, Y., Liu, J., Dai, N., Meng, F., & Liu, M. (2023). Integrating human parsing and pose network for

human action recognition. In *CAAI International Conference on Artificial Intelligence*, 182–194. https://doi.org/10.1007/978-981-99-8850-1_15

[79] Lee, J., Lee, M., Lee, D., & Lee, S. (2023). Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision,* 10444–10453. https://doi.org/10.1109/ICCV51070.2023.00958

[80] Obinata, Y., & Yamamoto, T. (2021). Temporal extension module for skeleton-based action recognition. In *2020 25th International Conference on Pattern Recognition*, 534–540. https://doi.org/10.1109/ICPR48806.2021.9412113

[81] Yang, C., Xu, Y., Shi, J., Dai, B., & Zhou, B. (2020). Temporal pyramid network for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 591–600. https://doi.org/10.1109/ICPR48806.2021.9412113

[82] Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). ViViT: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision,* 6836–6846. https://doi.org/10.1109/ICCV48922.2021.00676

[83] Yan, S., Xiong, X., Arnab, A., Lu, Z., Zhang, M., Sun, C., & Schmid, C. (2022). Multiview transformers for video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 3333–3343. https://doi.org/10.1109/CVPR52688.2022.00333

[84] Yu, S., Xie, L., Liu, L., & Xia, D. (2019). Learning long-term temporal features with deep neural networks for human action recognition. *IEEE Access*, *8*, 1840–1850. https://doi.org/10.1109/ACCESS.2019.2962284

[85] Zhou, Y., Sun, X., Zha, Z. J., & Zeng, W. (2018). MiCT: Mixed 3D/2D convolutional tube for human action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 449–458. https://doi.org/10.1109/CVPR.2018.00054

[86] Ranasinghe, K., Naseer, M., Khan, S., Khan, F. S., & Ryoo, M. S. (2022). Self-supervised video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2874–2884. https://doi.org/10.48550/arXiv.2112.01514

[87] Hussain, A., Hussain, T., Ullah, W., & Baik, S. W. (2022). Vision transformer and deep sequence learning for human activity recognition in surveillance videos. *Computational Intelligence and Neuroscience*, *2022*(1), 3454167. https://doi.org/10.1155/2022/3454167

[88] Moon, G., Kwon, H., Lee, K. M., & Cho, M. (2021). IntegralAction: Pose-driven feature integration for robust human action recognition in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 3339–3348. https://doi.org/10.1109/CVPRW53098.2021.00372

[89] Li, T., Zhang, R., & Li, Q. (2020). Multi scale temporal graph networks for skeleton-based action recognition. *arXiv Preprint: 2012.02970*. https://doi.org/10.48550/arXiv.2012.02970

[90] Plizzari, C., Cannici, M., & Matteucci, M. (2021). Skeleton-based action recognition via spatial and temporal transformer networks. *Computer Vision and Image Understanding*, *208–209*, 103219. https://doi.org/10.48550/arXiv.2008.07404

[91] Kopuklu, O., Kose, N., Gunduz, A., & Rigoll, G. (2019). Resource efficient 3D convolutional neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. https://doi.org/10.48550/arXiv.1904.02422

[92] Si, C., Chen, W., Wang, W., Wang, L., & Tan, T. (2019). An attention enhanced graph convolutional LSTM network for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 1227–1236. https://doi.org/10.48550/arXiv.1902.09130

[93] Vrskova, R. H. (2022). Human activity classification using the 3DCNN architecture. *Applied Sciences*, *12*(2), 931. https://doi.org/10.3390/app12020931

[94] Xing, Z., Dai, Q., Hu, H., Chen, J., Wu, Z., & Jiang, Y. G. (2023). SVFormer: Semi-supervised video transformer for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18816–18826. https://doi.org/10.48550/arXiv.2211.13222

[95] Huszar, V. D., Adhikarla, V. K., Négyesi, I., & Krasznay, C. (2023). Toward fast and accurate violence detection for automated video surveillance applications. *IEEE Access*, *11*, 18772–18793. https://doi.org/10.1109/ACCESS.2023.3245521

[96] Ming Cheng, K. C. (2019). *Activity recognition on RWF-2000*. Retrieved from: paperswithcode.com.; https://paperswithcode.com/sota/activity-recognition-on-rwf-2000

[97] Mohammadi, H., & Nazerfard, E. (2023). Video violence recognition and localization using a semi-supervised hard attention model. *Expert Systems with Applications*, *212*, 118791. https://doi.org/10.1016/j.eswa.2022.118791

[98] Islam, Z., Rukonuzzaman, M., Ahmed, R., Kabir, M. H., & Farazi, M. (2021). Efficient two-stream network for violence detection using separable convolutional LSTM. In *2021 International Joint Conference on Neural Networks,* 1–8. https://doi.org/10.1109/IJCNN52387.2021.9534280

[100] Su, Y., Lin, G., Zhu, J., & Wu, Q. (2020). Human interaction learning on 3D skeleton point clouds for video violence recognition. In *Computer Vision–ECCV 2020: 16th European Conference*, 74–90. https://doi.org/10.1007/978-3-030-58548-8_5

[100] Rendón-Segador, F. J., Álvarez-García, J. A., Enríquez, F., & Deniz, O. (2021). ViolenceNet: Dense multi-head self-attention with bidirectional convolutional LSTM for detecting violence. *Electronics*, *10*(13), 1601. https://doi.org/10.3390/electronics10131601

[101] Cheng, M., Cai, K., & Li, M. (2021). RWF-2000: An open large scale video database for violence detection. In *2020 25th International Conference on Pattern Recognition*, 4183–4190. https://doi.org/10.1109/ICPR48806.2021.9412502

[102] Soliman, M. M., Kamal, M. H., Nashed, M. A. E. M., Mostafa, Y. M., Chawky, B. S., & Khattab, D. (2019). Violence recognition from videos using deep learning techniques. In *2019 Ninth International Conference on Intelligent Computing and Information Systems*, 80–85. https://doi.org/10.1109/ICICIS46948.2019.9014714

How to Cite: Turyahabwa, P., & Murindanyi, S. (2025). Integrative Review of Human Activity Recognition and Violence Detection: Exploring Techniques, Modalities, and Cross-Domain Knowledge Transfer. *Journal of Data Science and Intelligent Systems*. https://doi.org/10.47852/bonviewJDSIS52024075