REVOLUTIONIZING VIDEO ANALYTICS: A REVIEW OF ACTION RECOGNITION USING 3D

Yunusa Mohammed Jeddah[1*], Aisha Hassan A.H.[1], Othman O. Khalifa[1], Adamu A. Ibrahim[2]

[1]Department of Electrical and Computer Engineering, International Islamic University, Malaysia
[2]Department of Computer Science, International Islamic University Malaysia

yunusmj2@hotmail.com

Article history

Received: 27 Dec 2024

Received in revised form: 31 Dec 2024

Accepted: 31 Dec 2024

Published online: 31 Dec 2024

*Corresponding author: yunusmj2@hotmail.com

ABSTRACT

As an interdisciplinary field, the 3D video action recognition model interprets human actions in three-dimensional video data using approaches such as pose-based, volumetric, motion-based, and hybrid methods. 3D video action recognition finds application in surveillance, sports analysis, human-computer interaction, healthcare, robotics, and augmented/virtual reality. This paper provides an overview of recent research in 3D video action recognition, concentrating on different deep learning architectures, self-supervised learning, graph-based methods, few-shot and zero-shot learning, cross-modal action understanding, and model interpretability. It also addresses the practicalities of implementing action recognition algorithms in real-world situations, which include tools like deep learning frameworks, pre-trained models, open-source libraries, cloud services, GPU acceleration, and evaluation metrics. Several case studies demonstrate the transformative influence of action recognition in surveillance, human-computer interaction, sports analysis, industrial automation, healthcare, and retail, with applications in autonomous vehicles, healthcare monitoring, retail analytics, crowd management, video content filtering, and manufacturing quality control. Findings show that integrating 3D video with action recognition algorithms augments accuracy and detail while challenges in video action recognition such as long video features capturing, temporal context maintenance, computational costs management, variability handling, inadequate training data, domain adaptation, and benchmarking methods are addressed. Researchers have contributed novel techniques, architecture, and datasets in an attempt to advance the field and improve the performance and application of video action recognition through research.

Keywords: *Video Analytics, Action Recognition, 3D Video, Deep Learning, Computer Vision*

1.0 INTRODUCTION

Action recognition is crucial in computer vision, focusing on algorithms that interpret human actions in images or videos. It involves preprocessing data, extracting features, and classifying actions using models like support vector machines or convolutional neural networks, with

performance evaluated using metrics like accuracy and precision. However, challenges like lighting variations, occlusions, complex backgrounds, and changes in viewpoint persist. Figure 1 illustrates the general flow of an action recognition system.

Figure 1: Action Recognition Block Diagram

Recognizing actions in two-dimensional (2D) format, such as images or videos, particularly challenging due to the loss of depth information, which can distort skeleton proportions as subjects vary in distance and orientation relative to the camera[1]. Data normalization, scaling, and quantization also substantially impact on the accuracy of 2D recognition [1]. Other challenges include handling occlusions, irregular movements, changes in the background, and changes in viewing angles [2]. These limitations are more pronounced in images than in videos due to the lack of temporal continuity [3].

To overcome the limitations, three-dimensional (3D) methods have emerged. These methods utilize 3D Convolutional Neural Networks (3D-CNNs) to extract spatiotemporal features, offering better performance in recognizing actions under challenging conditions [4] [5]. For example, 3D methods are robust against pose estimation noise and can handle multiple-person scenarios with no additional computational costs [6]. Attention mechanisms [7] integrated into some 3D methods improve focus on key features, enhancing recognition outcomes.

The advent of 3D video technology [8] [9] has further revolutionized action recognition by providing depth information, enabling a more comprehensive understanding of human actions [10]. This advancement has led to substantial progress in research disciplines, like deep learning, machine learning, computer vision, and pattern recognition, addressing key challenges and unlocking innovative potential. 3D video action recognition focuses on designing algorithms capable of accurately identifying and interpreting human actions in three-dimensional video sequences [11]. Approaches include pose-based [12], motion-based [13], and hybrid methods [14], each utilizing depth information and machine/deep learning techniques for performance enhancement.

3D video action recognition application spans a broad range of domains, including human-computer interaction, surveillance, sports analysis, healthcare, robotics, and augmented/virtual reality. By understanding and interpreting human actions in 3D video, applications benefit from advanced video analysis, automation, and improved decision-making capabilities. Notably, 3D methods have demonstrated better performance on standard action

recognition benchmarks, making them a preferred choice for addressing the complexities of action recognition [6] [5].

This paper provides a review of 3D video action recognition, stressing recent advances in areas like deep learning, self-supervised learning, cross-modal action understanding, few-shot and zero-shot learning, graph-based methods, and model interpretability. Practical challenges in real-world implementation, as well as technologies and tools available to researchers are also discussed. We also present case studies highlighting the transformative impact of 3D video action recognition across various fields.

Section two explores 3D and the way it works with action recognition. An overview of 3D video action recognition is elaborated in subsection three. In section four, we look at case studies of 3D video action recognition. Subsections five, six, and seven focus on implementation, challenges, and conclusion respectively.

2.0 3D VIDEO AND HOW IT WORKS WITH ACTION RECOGNITION

Unlike traditional 2D videos, which offer information about the colour and motion in a scene only, 3D videos integrate an additional dimension, depth, that improves the accuracy and level of detail in action recognition tasks. 3D video can simply be regarded as the representation of video data that captures the spatial information of a scene, allowing for the perception of depth and three-dimensional structure. It is imperative to consider the key components involved in understanding how 3D videos work with action recognition

2.1 Depth Information

Depth information is one of the important components involved in understanding how 3D video works in action recognition [15]. Its main focus is to capture the relative distance between objects and the camera to provide an understanding of the scene's three-dimensional structure. Depth can be obtained using several sensing technologies, such as time-of-flight, structured light, or stereovision [16]. These techniques measure the time it takes for light to travel to and from the objects in the scene or triangulate depth from multiple camera viewpoints. The ensuing depth maps or point clouds offer valuable cues for action recognition algorithms.

2.2 Spatiotemporal Representation

3D video action recognition requires the extraction of meaningful spatiotemporal features that capture both appearance and motion information over time [17]. Features of appearance describe the visual appearance of objects or body parts, whereas motion features capture the dynamic changes in scenes. Spatiotemporal representation combines these appearance and motion cues, enabling action recognition algorithms to analyze and interpret human actions accurately. Methods such as 3D CNNs are commonly used to learn spatiotemporal features from 3D video data[18].

2.3 Action Recognition Algorithms:

Action recognition algorithms in 3D video leverage the spatiotemporal representation to recognize and classify human actions. These algorithms focus on learning discriminative patterns and temporal dependencies from input data. They employ machine learning techniques like deep learning, to automatically extract features and make predictions. Popular architectures for 3D action recognition comprise Inflated 3D (I3D) networks, Two-Stream Convolutional Networks, SlowFast networks, and X3D architectures [19] [20]. These models are trained on large-scale datasets and fine-tuned for specific action recognition tasks.

2.4 Multimodal Fusion

Multimodal fusion enhances action recognition robustness and accuracy by combining data from corresponding modalities, each contributing unique information to increase the overall understanding of human actions. RGB video provides rich visual details like colours and textures but lacks depth or detailed motion cues. Optical flow captures temporal motion patterns, essential for identifying dynamic activities, which RGB alone may not discern. Skeletal data offers precise spatial information about body joints and poses, improving the granularity of action understanding.

Exploiting these diverse inputs, fusion addresses challenges like occlusions, variations in lighting, or missing data from any single modality. For example:

- Early fusion integrates multiple modalities at the input stage, allowing the model to learn shared feature representations.
- Late fusion combines predictions from individual modalities, ensuring robustness by compensating for weaknesses in any one modality.

Researches have always established that multimodal fusion improves model performance, leading to more accurate and resilient action recognition outcomes, particularly in complex situations [21].

2.5 Dataset Availability

A vital role in training, evaluating, and advancing action recognition algorithms in 3D video is played by Benchmark datasets. These Datasets include but are not limited to NTU RGB+D, Kinetics-400, and UCF101-3D [22] [23] [24] which provide labelled video sequences of various human actions, allowing researchers to develop and benchmark their algorithms. The datasets comprise a wide range of actions made by different persons in numerous circumstances, ensuring the robustness and generalization of action recognition models.

2.6 Applications

Action recognition through 3D video has extensive applications as mentioned above. Like in surveillance, 3D video analysis can improve security systems and personnel through

automating the detection and classification of suspicious activities. 3D action recognition aids in detailed performance analysis, player tracking, and event detection in sports analysis. 3D video understanding can assist in gesture recognition and immersive virtual experiences in human-computer interaction. In the area of healthcare, it can support patient movements and rehabilitation exercise monitoring. In the robotics field, 3D action recognition enables robots themselves.

3.0 OVERVIEW OF 3D VIDEO ACTION RECOGNITION

Here, we overview and highlight core components and techniques in 3D video action recognition, emphasizing deep learning architectures, graph-based methods, self-supervised learning, cross-modal action understanding, model interpretability, and few-shot and zero-shot learning.

3.1 Deep Learning Architectures

Deep learning architecture plays an essential role in evolving 3D video action recognition. Researchers have utilised CNNs to handle 3D video data spatiotemporal information. A popular approach for CNN-based action recognition is a two-stream architecture (see figure 2), which uses stacked optical frames and RGB as motion information and appearance, respectively; it proves that the combination of two streams improves action recognition accuracy[25]. These architectures control techniques such as temporal pooling, 3D convolutions, and feature fusion to extract features from 3D video sequences [26], [27].

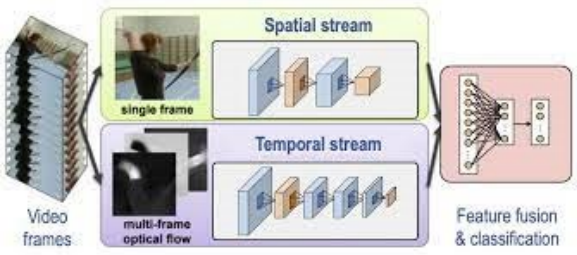

Figure 2: Two-stream Architecture [28]

3.2 Self-supervised Learning

This learning technique has developed to be a valuable approach in the field of 3D video action recognition. Two competing standards exist for self-supervised learning in action recognition from 3D [29], [30]. [31]. This learning method leverages the inherent structure or content of the data itself, instead of relying on labeled data to learn meaningful representations. Models can learn powerful representations that transfer well to downstream action recognition tasks by formulating pretext tasks like predicting temporal order, temporal alignment, or motion segmentation. The technique, as can be seen in figure 3 enables leveraging large volumes of unlabeled data, reducing the reliance on annotated expensive datasets.

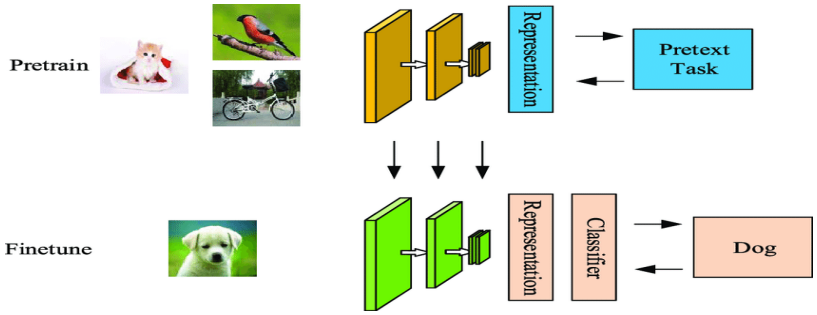

Figure 3: Structure of Self-Supervised Learning [32]

3.3 Graph-based Methods

These methods gained attention in 3D video action recognition for capturing temporal and spatial relationships between actions and body parts. [33] and [34] proposed an innovative end-to-end model based on a Graph Convolutional Network (GCN) for 3D skeleton-based video action recognition. GCNs model the human skeleton as a graph and exploit the networks between human joints to learn novel representations. The methods exploit graph-based (see figure 4) convolutions to gather information from adjoining joints, collect contextual dependencies, and improve the performance of action recognition.

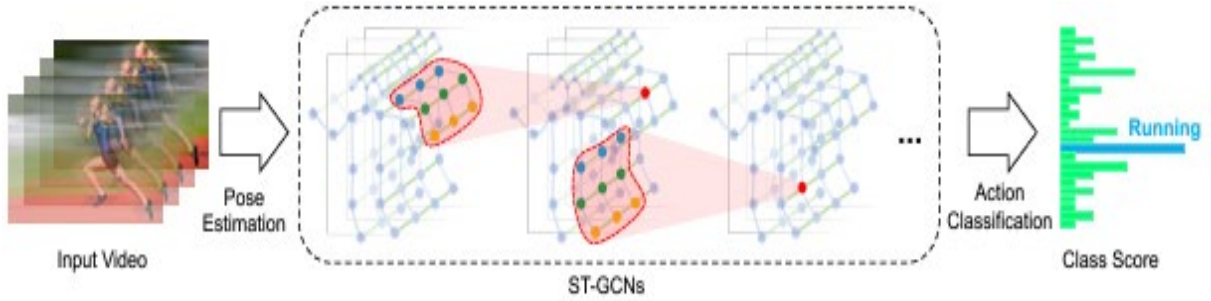

Figure 4: Spatial-Temporal Graph on Skeleton Sequences [35]

3.4 Few-shot, One-Shot, and Zero-shot Learning

These learning methods focus on addressing the challenge of limitations on labelled data in 3D video action recognition. While zero-shot learning extends the ability to identify invisible actions that were absent during training, few-shot learning identifies actions using a few labelled examples. The one-shot learning approach is where a model is trained to identify objects using only a few examples, allowing it to generalize from minimal instances. These methods leverage transfer learning and domain adaptation techniques to take a broad view of knowledge from seen actions to unseen actions. Meta-learning, attribute-based representations, and generative model approaches have been capitalized on to allow effective few-shot and zero-shot action recognition. [36-37].

3.5 Cross-modal Action Understanding

In cross-modal action understanding the main focus is to integrate multiple modalities (such as RGB, audio, and depth) to improve action recognition [38] [39]. Through the combination of complementary information from diverse modalities, models can capture the spatiotemporal dynamics of actions better. Fusion methods like early fusion, late fusion, and multi-stream architecture allow cross-modal representation learning, enabling enhanced action recognition robustness and accuracy.[40], [41].

3.6 Model Interpretability

This is a crucial method for understanding and trusting 3D video action recognition systems [42]. Interpretable models offer insights into the process of decision-making and contribute to the results' explainability of action recognition. Methods like attention mechanisms, saliency maps, and visualisation methods help identify the important spatiotemporal signs and body sections that contribute to action recognition [43]. Model interpretability augments accountability, transparency, and user acceptance of 3D video action recognition systems.

4.0 APPLICATIONS OF 3D VIDEO ACTION RECOGNITION (CASE STUDIES)

In this section, we highlight real-world cases and applications that demonstrate the impact of action recognition on video analytics. By comparing 3D video action recognition techniques with existing or previous methods, we illustrate how this technology has transformed various fields, enabling advanced video analysis, automation, and decision-making capabilities.

4.1 Surveillance and Security

Traditional video surveillance systems rely heavily on manual monitoring, often leading to delayed or missed detections due to human error. In contrast, action recognition enhances surveillance systems with automated real-time detection and classification of actions. Unlike conventional motion-detection algorithms, which are limited to identifying simple movements, modern action recognition systems can discern complex human behaviors, such as aggressive actions, crowding, or unattended objects, with higher accuracy and robustness [44], [45] [46]. For instance, systems integrating 3D action recognition are deployed in airports and critical infrastructure to provide early warnings and facilitate informed decision-making by security personnel, a significant improvement over prior methods that required retrospective video analysis.

4.2 Human-Computer Interaction (HCI)

Compared to earlier interaction methods like keyboard and mouse inputs, action recognition has significantly advanced human-computer interaction by enabling natural, intuitive interfaces. Gesture recognition and action-based commands allow users to control systems effortlessly, improving accessibility and user experience. For example, conventional gesture-

based systems relied on 2D video data, which often struggled with depth perception and pose ambiguity. 3D action recognition resolves these challenges by incorporating depth information, allowing for more accurate and robust recognition of gestures in applications such as sign language interpretation, gaming, and virtual reality environments [26] [45].

4.3 Sports Analysis

Traditional sports analysis methods involved manual tagging of player movements, which was time-consuming and prone to inaccuracies. In contrast, action recognition automates the analysis of athletes' movements, providing precise metrics such as player positioning and motion dynamics [47]. Unlike earlier tracking systems, which relied on single-camera views and lacked depth information, 3D action recognition offers detailed analysis from multi-camera setups or depth sensors, enabling trainers to optimize performance, prevent injuries, and make tactical decisions with unprecedented precision.

4.4 Industrial Automation

Conventional industrial automation systems often relied on fixed programming for robotic operations, limiting their adaptability to dynamic environments. Action recognition transforms industrial settings by enabling robots to understand and respond to human actions in real-time, allowing safe human-robot collaboration [48] [49]. For example, compared to earlier vision-based systems that required predefined gestures, 3D action recognition systems can interpret nuanced human actions, enhancing productivity and safety in manufacturing processes.

4.5 Healthcare and Assistive Technologies

Traditional rehabilitation and eldercare monitoring systems often required direct human supervision or used basic motion sensors with limited capabilities. Action recognition, by contrast, offers non-invasive, detailed monitoring of patient activities such as walking, sitting, or falling, enhancing the precision of postoperative care and eldercare [50], [51]. For assistive technologies, traditional systems using basic gesture controls have been replaced by 3D action recognition-based interfaces, enabling more accurate and seamless interaction for individuals with disabilities. In surgical training, the analysis of surgeon movements using 3D action recognition offers a more objective and comprehensive evaluation compared to earlier observational methods.

4.6 Retail and Customer Analytics

Previous retail analytics relied heavily on transaction data or rudimentary video analytics for customer behavior insights. In contrast, 3D action recognition enables detailed analysis of customer movements, such as browsing behaviors and product interactions[52]. Unlike older systems that lacked contextual understanding, these advanced techniques provide deeper insights into customer preferences, optimizing store layouts, personalizing marketing strategies, and enhancing overall customer experiences.

4.7 Others

Traditional video analytics methods in areas such as content moderation or autonomous vehicles relied on simple pattern recognition techniques, which often failed to handle complex scenarios. By integrating 3D action recognition, these systems are now able to achieve greater accuracy and reliability, enabling applications in autonomous navigation, virtual/augmented reality, and robotics.

5.0 IMPLEMENTING ACTION RECOGNITION ALGORITHMS IN REAL-WORLD SITUATIONS

In real-world situations, the implementation of action recognition algorithms requires thoughtful consideration of several factors and the application of existing tools and technologies. Here, we attempt to give an overview, emphasizing the importance of deep learning frameworks, open-source libraries, pre-trained models, GPU acceleration, cloud services, and evaluation metrics in the implementation process.

5.1 Deep Learning Frameworks

Deep learning frameworks act as a foundation for developing and deploying action recognition algorithms. They offer a means to build and train models that can identify actions in videos. TensorFlow, PyTorch, and Keras are some popular deep learning frameworks for action recognition [53]. These frameworks offer comprehensive tools and APIs for building deep learning models. They provide a variety of functionalities, which include support for 3D convolutions, visualization tools, and model optimization techniques. Deep learning frameworks empower developers and researchers to implement and experiment with different architectures and algorithms for action recognition efficiently.

5.2 Pre-trained Models

Pre-trained models are valuable resources for the implementation of action recognition. They offer the means to build and train models that can identify actions in videos without the need to start from scratch. I3D, TSN, and SlowFast are some of the popular action recognition pre-trained models. [54] [55]. The pre-trained models are trained on large-scale datasets and capture wide-ranging knowledge about various actions and visual features [56] [54] Pre-trained models let practitioners and researchers gain from transfer learning, wherein the pre-trained learned representations are fine-tuned on specific datasets or tasks. Models, like those in the Kinetics dataset [23] or the ImageNet dataset, offer an easy start and can speed up the development and deployment of action recognition systems significantly.

5.3 Open-Source Libraries

Open-source libraries offer exciting functionalities and resources for action recognition algorithm implementation. Libraries such as OpenCV, sci-kit-learn, and NumPy offer tools for

feature extraction, image and video processing, and data manipulation. Popular action recognition open-source libraries include TensorFlow, PyTorch, and Keras [57] [58] [59]. They offer a means for building and training models that can identify actions in videos without the need to start from the beginning. In addition, domain-specific libraries such as OpenPose, which focuses on human pose estimation, can be employed to extract essential features for action recognition. Open-source libraries save effort and time by offering ready-to-use implementations and functions which facilitate the integration and development of action recognition algorithms.

5.4 Cloud Services

Cloud services offer a flexible and scalable infrastructure for the deployment of action recognition models in real-world scenarios. The services provide a means for storing and processing large data, which is significant for training and testing models. Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure are some of the popular cloud services for machine learning [57] [55] [60]. Cloud service platforms such as Amazon Web Services (AWS), Microsoft Azure, and Google Cloud offer computing resources, storage, and machine learning services. These cloud platforms provide for efficient model training, hosting, and inference at scale. The platforms are predominantly suitable for dealing with huge datasets, resource-intensive computations, and real-time applications wherein scalability and availability are key.

5.5 GPU Acceleration

Graphics Processing Unit (GPU) acceleration is significant in the implementation of action recognition algorithms in real-world circumstances for it offers faster processing of large quantities of data. GPUs are specifically designed for parallel handling of large amounts of data, which makes them ideal for machine learning tasks such as action recognition [57] [26]. The computational need for action recognition algorithms can be drastically reduced by employing GPUs for parallel computing. They outshine fast-tracking deep learning, comprising convolutions and matrix operations, ensuring faster training and inference times. TensorFlow and PyTorch frameworks provide GPU support, enabling experts to leverage GPU acceleration and accomplish momentous improvements in performance in action recognition tasks.

5.6 Evaluation Metrics

Evaluation metrics are indispensable in implementing action recognition algorithms in the real world for the reason that they allow for the measurement of the performance of the models. Accuracy, precision, recall, and F1-score are some common evaluation metrics for action recognition [61][62][55]. They are crucial in computing the performance of action recognition algorithms. The effectiveness of models in classifying actions correctly is done using metrics like accuracy, precision, recall, F1-score, and mean Average Precision (mAP) measure. Evaluation metrics, like Top-1 accuracy in the Kinetics dataset, provide a

standardized way to compare and benchmark different algorithms are dataset-specific evaluation metrics. Employing suitable evaluation metrics guarantees reliable performance assessment and gives room for meaningful comparisons amongst diverse action recognition models and techniques.

Researchers can effectually implement action recognition algorithms in real-world situations by capitalizing on deep learning frameworks, pre-trained models, cloud services, open-source libraries, GPU acceleration, and evaluation metrics. These tools and technological advancement offer the necessary resources, efficiency, scalability, and performance evaluation abilities to create and deploy accurate and robust action recognition systems in several domains and applications.

6.0 CHALLENGES IN VIDEO ACTION RECOGNITION

Some of the Challenges in video action recognition include the following.

6.1 Capturing Long Video Features

It's obvious that videos often contain long action sequences and catching informative features throughout the entire video is a big challenge [63][24]. Traditional methods for capturing informative features that process fixed-length video clips are characterised by the tendency to miss important temporal context or suffer information loss at the margins of the clips. Developing approaches that can effectively capture long-range dependencies and temporal dynamics is vital for accurate action recognition.

6.2 High Computational Costs

Action recognition in videos entails processing large quantities of data, which can be computationally expensive. Deep learning models, such as 3D CNNs, commonly used for video action recognition, need substantial computational resources for training and inference [24]. Real-time or near real-time applications, mostly on resource-constrained devices can be hampered due to the high computational costs.

6.3 Variability in Action Execution

Actions can be executed with variations in speed, viewpoint, scale, lighting conditions, and other factors, leading to intra-class variations [64]. The challenging task here is in handling the variabilities and also developing models that are robust to the variations. Good models should be able to generalize well across diverse instances of the same action and have the ability to distinguish between faint differences in action execution.

6.4 Lack of Adequate and Diverse Training Data

Deep learning model training for video action recognition entails large-scale labelled datasets. Nevertheless, collecting and annotating the datasets can be expensive, time-consuming, and challenging [65]. Likewise, making sure diversity in the datasets, including an extensive range of action categories, action execution variations, and different environmental conditions, is vital to designing robust models. The limitation in the availability of diverse and correctly annotated datasets remains a challenge in the field.

6.5 Benchmarking and Methods Comparing

The comparison and evaluation of different action recognition methods can be regarded as challenging due to the lack of homogenous evaluation protocols and benchmarks[66]. Dissimilar datasets, experimental setups, and evaluation metrics make it tough to directly compare the performance of distinct algorithms. Creating common benchmarks and evaluation protocols can go a long way in facilitating fair and meaningful evaluations between different methods.

6.6 Domain Adaptation and Generalization

Deployment of action recognition models in real-world situations often requires acclimating the models to new domains or concealed environments [67]. Acclimatizing models trained on a dataset to execute well on another dataset with unlike characteristics is challenging. Models should be capable of generalizing across different domains, handling domain shifts, and effectually transferring knowledge across different datasets.

7.0 CONCLUSION

Undoubtedly, the integration of 3D video technology with action recognition has revolutionized the video analytics field. This integration exploits the potential of depth information, advanced deep learning architectures, spatiotemporal representation, graph-based methods, cross-modal understanding, self-supervised learning, and model interpretability. These components provide accurate and detailed analysis of human actions in the context of three-dimensional video data.

The transformative impact of 3D video action recognition transcends theoretical boundaries and provides practical applications in numerous domains. Some notable examples as indicated in this paper include improving security through comprehensive surveillance systems, allowing thorough performance analysis in sports, improving user experiences utilizing immersive human-computer interactions, supporting patient care in healthcare situations, and driving automation, in retail analytics and industrial automation.

Realizing the potential of this technology in real-world situations demands a well-extensive approach. Deep learning frameworks, such as Keras, PyTorch, and TensorFlow, act

as the foundation for developing and experimenting with action recognition models. The presence of pre-trained models has significantly improved the model development process, allowing for the leveraging of transfer learning, and benefiting from the knowledge derived from large-scale datasets.

Open-source libraries, such as OpenCV and Scikit-learn, provide researchers and other experts with feature extraction tools, data manipulation, and model development, reducing the required time and effort for integration. Cloud services, provided by platforms such as GCP, AWS, and Azure, provide the scalability and computational resources needed for managing large datasets and resource-intensive computations, particularly essential in real-time applications.

The GPU acceleration role, which is intended for parallel processing, cannot be underestimated. It significantly enhances the processing speed, making real-time analysis possible even in the face of large amounts of data. In addition, standardized evaluation metrics, such as accuracy, precision, recall, and dataset-specific measures, ensure reliable performance assessment, and meaningful comparisons amid numerous models and techniques.

Acknowledgement
The authors would like to express their appreciation to all those who contributed in this paper.

REFERENCES

[1] Elias, P., Sedmidubsky, J., & Zezula, P. (2021). Understanding the limits of 2D skeletons for action recognition. *Multimedia Systems, 27*(3), 547–561.

[2] Belluzzo, B., & Marana, A. N. (2022). Human action recognition based on 2D poses and skeleton joints. In J. C. Xavier-Junior & R. A. Rios (Eds.), *Springer International Publishing* (pp. 71–83).

[3] Singh, P. K., Kundu, S., Adhikary, T., Sarkar, R., & Bhattacharjee, D. (2022). Progress of human action recognition research in the last ten years: A comprehensive survey. *Archives of Computational Methods in Engineering, 29*(4), 2309–2349.

[4] Stroud, J. C., Ross, D. A., Sun, C., Deng, J., & Sukthankar, R. (2020). D3D: Distilled 3D networks for video action recognition. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 614–623).

[5] Carreira, J., & Zisserman, A. (2017). Quo Vadis, action recognition? A new model and the kinetics dataset. *Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, 4724–4733.

[6] Duan, H., Zhao, Y., Chen, K., Lin, D., & Dai, B. (2022). Revisiting skeleton-based action recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2959–2968).

[7] Shrestha, L., Dubey, S., Olimov, F., Rafique, M. A., & Jeon, M. (2022). 3D convolutional with attention for action recognition. *arXiv Preprint arXiv:2206.02203*.

[8] Fernando, A., Worrall, S. T., & Ekmekcioğlu, E. (2013). *Processing and transmission of 3D video signals*.

[9] Zhao, R., Ali, H., & Van Der Smagt, P. (2017). Two-stream RNN/CNN for action recognition in 3D videos. *Proceedings*, 4260–4267.

[10] Li, X., Huang, Q., Wang, Z., & Yang, T. (2023). Real-time 3D human action recognition based on hyperpoint sequence. *IEEE Transactions on Industrial Informatics, 19*(8), 8933–8942. https://doi.org/10.1109/TII.2022.3223225

[11] Gu, F., Chung, M. H., Chignell, M., Valaee, S., Zhou, B., & Liu, X. (2022). A survey on deep learning for human activity recognition. *ACM Computing Surveys, 54*(8). https://doi.org/10.1145/3472290

[12] Boualia, S. N., & Ben Amara, N. E. (2019). Pose-based human activity recognition: A review. In *2019 15th International Wireless Communications and Mobile Computing Conference (IWCMC)* (pp. 1468–1475). https://doi.org/10.1109/IWCMC.2019.8766694

[13] Liu, Q., Xing, D., Tang, H., Ma, D., & Pan, G. (2021). Event-based action recognition using motion information and spiking neural networks. In *IJCAI International Joint Conference on Artificial Intelligence* (pp. 1743–1749). https://doi.org/10.24963/ijcai.2021/240

[14] Shuvo, M. M. H., Ahmed, N., Nouduri, K., & Palaniappan, K. (2020). A hybrid approach for human activity recognition with support vector machine and 1D convolutional neural network. *Proceedings - Applied Imagery Pattern Recognition Workshop, 2020-October*. https://doi.org/10.1109/AIPR50011.2020.9425332.

[15] Lee, M. J., Lee, J. W., & Lee, H. K. (2011). Perceptual watermarking for 3D stereoscopic video using depth information. *Proceedings of the 7th International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP 2011)*, 2, 81–84. https://doi.org/10.1109/IIHMSP.2011.83

[16] Horaud, R., Hansard, M., Evangelidis, G., & Ménier, C. (2016). An overview of depth cameras and range scanners based on time-of-flight technologies. *Machine*

Vision and Applications, 27(7), 1005–1020. https://doi.org/10.1007/s00138-016-0784-4

[17] Xu, H., & Saenko, K. (2017). R-C3D: Region Convolutional 3D Network for Temporal Activity Detection. *Proceedings of the IEEE International Conference on Computer Vision*, 5783–5792.

[18] Agyeman, R., Rafiq, M., Shin, H. K., Rinner, B., & Choi, G. S. (2021). Optimizing spatiotemporal feature learning in 3D convolutional neural networks with pooling blocks. *IEEE Access, 9,* 70797–70805. https://doi.org/10.1109/ACCESS.2021.3078295

[19] Kumawat, S., Verma, M., Nakashima, Y., & Raman, S. (2022). Depthwise spatio-temporal STFT convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(9), 4839–4851. https://doi.org/10.1109/TPAMI.2021.3076522

[20] Feichtenhofer, C., & Ai, F. (2020). X3D: Expanding architectures for efficient video recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 203–213.

[21] Duhme, M., Memmesheimer, R., & Paulus, D. (2021). Fusion-GCN: Multimodal action recognition using graph convolutional networks. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13024, 265–281. https://doi.org/10.1007/978-3-030-92659-5_17

[22] Shahroudy, A., Liu, J., Ng, T. T., & Wang, G. (2016). NTU RGB+D: A large scale dataset for 3D human activity analysis. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-December, 1010–1019. https://doi.org/10.1109/CVPR.2016.115.

[23] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P. and Suleyman, M., (2017). The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

[24] Jiang, G., Jiang, X., Fang, Z., & Chen, S. (2021). An efficient attention module for 3D convolutional neural networks in action recognition. *Applied Intelligence, 51*(10), 7043–7057. https://doi.org/10.1007/s10489-021-02195-8

[25] Paper, I. (2021). Recent advances in video action recognition with 3D convolutions. *Applied Intelligence, 51*(6), 846–856.

[26] Zhu, Y., Li, X., Liu, C., Zolfaghari, M., Xiong, Y., Wu, C., Zhang, Z., Tighe, J., Manmatha, R. and Li, M., (2020). A comprehensive study of deep video action recognition. *arXiv preprint arXiv:2012.06567*.

[27] Karácsony, T., Loesch-Biffar, A. M., Vollmar, C., Rémi, J., Noachtar, S., & Cunha, J. P. S. (2022). Novel 3D video action recognition deep learning approach for near real-time epileptic seizure classification. *Scientific Reports, 12*(1), 1–13. https://doi.org/10.1038/s41598-022-23133-9

[28] Xiong, Q., Zhang, J., Wang, P., Liu, D., & Gao, R. X. (2020). Transferable two-stream convolutional neural network for human action recognition. *Journal of Manufacturing Systems, 56*, 605–614. https://doi.org/10.1016/j.jmsy.2020.04.007

[29] Ben Tanfous, A., Zerroug, A., Linsley, D., & Serre, T. (2022). How and what to learn: Taxonomizing self-supervised learning for 3D action recognition. In *Proceedings - 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 2888–2897). https://doi.org/10.1109/WACV51458.2022.00294

[30] Yang, S., Liu, J., Lu, S., Hwa, E. M., Hu, Y., & Kot, A. C. (2023). Self-supervised 3D action representation learning with skeleton cloud colorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[31] Schiappa, M. C. (2022). Self-supervised learning for videos: A survey. *ACM Computing Surveys*, 1–33.

[32] V7 Labs. (2024). Self-supervised learning: Definition, tutorial & examples. Retrieved December 16, 2024, from https://www.v7labs.com/blog/self-supervised-learning-guide

[33] Gao, X., et al. (2019). 3D skeleton-based video action recognition by graph convolution network. In *Proceedings of the 2019 IEEE International Conference on Smart Internet of Things (SmartIoT)* (pp. 500–501). https://doi.org/10.1109/SmartIoT.2019.00093

[34] Li, M., & Leung, H. (2017). Graph-based approach for 3D human skeletal action recognition. *Pattern Recognition Letters, 87*, 195–202. https://doi.org/10.1016/j.patrec.2016.07.021

[35] 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*(7). https://doi.org/10.1609/aaai.v32i1.12328

[36] Kawamoto, K. (2021). Zero-shot action recognition with three-stream graph convolutional networks.

[37] Guo, M., Chou, E., Huang, D. A., Song, S., Yeung, S., & Fei-Fei, L. (2018). Neural graph matching networks for fewshot 3d action recognition. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 653-669).

[38] Kim, S., Ahn, D., & Ko, B. C. (2022). Cross-modal learning with 3D deformable attention for action recognition.

[39] Yang, L., Huang, Y., Sugano, Y., & Sato, Y. (2022). Interact before align: Leveraging cross-modal knowledge for domain adaptive action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 14702–14712). https://doi.org/10.1109/CVPR52688.2022.01431

[40] Ghosh, S., Aggarwal, T., Hoai, M., & Balasubramanian, N. (2023). Text-derived knowledge helps vision: A simple cross-modal distillation for video-based action anticipation. In *Findings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023)* (pp. 1837–1852). https://doi.org/10.18653/v1/2023.findings-eacl.141

[41] Wu, W., et al. (2023). A large cross-modal video retrieval dataset with reading comprehension.

[42] Meng, L., et al. (2019). Interpretable spatio-temporal attention for video action recognition. In *Proceedings of the 2019 International Conference on Computer Vision Workshop (ICCVW 2019)* (pp. 1513–1522). https://doi.org/10.1109/ICCVW.2019.00189

[43] Huang, X., & Cai, Z. (2023). A review of video action recognition based on 3D convolution. *Computers and Electrical Engineering, 108*, 108713. https://doi.org/10.1016/J.COMPELECENG.2023.108713

[44] Sanchez, J., Neff, C., & Tabkhi, H. (2022). Real-world graph convolution networks (RW-GCNs) for action recognition in smart video surveillance.

[45] Pham, H. H., Khoudour, L., Crouzil, A., Zegers, P., & Velastin, S. A. (2022). Video-based human action recognition using deep learning: A review (pp. 1–25).

[46] Admin. (2023). The transformative impact of AI on video surveillance and security systems. Retrieved October 2, 2023, from https://www.securens.in/blog/the-transformative-impact-of-ai-on-video-surveillance-and-security-systems/

[47] Wu, F., et al. (2022). A survey on video action recognition in sports: Datasets, methods, and applications. *IEEE Transactions on Multimedia, PP*, 1–25. https://doi.org/10.1109/TMM.2022.3232034

[48] Edison, A., & Jiji, C. V. (2019). Automated video analysis for action recognition using descriptors derived from optical acceleration. *Signal, Image and Video Processing, 13*(5), 915–922. https://doi.org/10.1007/s11760-019-01428-1

[49] Kulbacki, M., et al. (2023). Intelligent video analytics for human action recognition: The state of knowledge. *Sensors, 23*(9), 1–31. https://doi.org/10.3390/s23094258

[50] Pareek, P., & Thakkar, A. (2021). A survey on video-based human action recognition: Recent updates, datasets, challenges, and applications. *Artificial Intelligence Review, 54*(3). Springer Netherlands. https://doi.org/10.1007/s10462-020-09904-8

[51] Jalloul, N., Poree, F., Viardot, G., L'Hostis, P., & Carrault, G. (2018). Activity recognition using complex network analysis. *IEEE Journal of Biomedical and Health Informatics, 22*(4), 989–1000. https://doi.org/10.1109/JBHI.2017.2762404

[52] Qualtrics. (n.d.). The customer behavior analysis guide. Retrieved from https://www.qualtrics.com/au/experience-management/customer/customer-behaviour-analysis/

[53] Parvat, A., & Dev, S. (2017). A survey of deep-learning frameworks (pp. 1–7).

[54] Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, A., Zhang, L. and Han, W., (2021). Pre-trained models: Past, present and future. *AI Open*, *2*, 225-250.

[55] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications, and research directions. *SN Computer Science, 2*(3), 1–21. https://doi.org/10.1007/s42979-021-00592-x

[56] Zhang, L., & Corporation, L. (2011). Deep learning for sentiment analysis - A survey. *A+U-Architecture and Urbanism, 487*, 121.

[57] Nagaraj, A., Sood, M., Sureka, C., & Srinivasa, G. (2022). Real-time action recognition for fine-grained actions and the hand wash dataset. *arXiv preprint arXiv:2210.07400*.

[58] Suárez-Hernández, A., Segovia-Aguas, J., Torras, C., & Alenyà, G. (2021). Online action recognition. *35th AAAI Conference on Artificial Intelligence, AAAI 2021, 13B*(Snyder), 11981–11989. https://doi.org/10.1609/aaai.v35i13.17423

[59] Su, Y., Li, Y., & Liu, A. (2019). Open-view human action recognition based on linear discriminant analysis. *Multimedia Tools and Applications, 78*(1), 767–782. https://doi.org/10.1007/s11042-018-5657-6

[60] Kong, Y., & Fu, Y. (2022). Human action recognition and prediction: A survey. *International Journal of Computer Vision, 130*(5), 1366–1401. https://doi.org/10.1007/s11263-022-01594-9

[61] Gupta, N., Gupta, S. K., Pathak, R. K., Jain, V., Rashidi, P., & Suri, J. S. (2022). Human activity recognition in artificial intelligence framework: A narrative review. *Artificial Intelligence Review, 55*(6). Springer Netherlands. https://doi.org/10.1007/s10462-021-10116-x

[62] Wang, L. (2021). Analysis and evaluation of Kinect-based action recognition algorithms.

[63] Chang, Y. L., Chan, C. S., & Remagnino, P. (2021). Action recognition on continuous video. *Neural Computing and Applications, 33*(4), 1233–1243. https://doi.org/10.1007/s00521-020-04982-9

[64] Shao, L., Liu, L., & Yu, M. (2016). Kernelized multiview projection for robust action recognition. *International Journal of Computer Vision, 118*(2), 115–129. https://doi.org/10.1007/s11263-015-0861-6

[65] Verburg, M., & Menkovski, V. (2019). Micro-expression detection in long videos using optical flow and recurrent neural networks. *Proceedings of the 14th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2019)* (pp. 1–6). https://doi.org/10.1109/FG.2019.8756588

[66] Liu, C., Hu, Y., Li, Y., Song, S., & Liu, J. (2017). PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding.

[67] Bao, W., Yu, Q., & Kong, Y. (2021). Evidential deep learning for open set action recognition. *Proceedings of the IEEE International Conference on Computer Vision* (pp. 13329–13338). https://doi.org/10.1109/ICCV48922.2021.01310