Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 20 No. 5 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijim.v20i05.58381

PAPER

Multimodal Human Action Recognition for Ubiquitous Systems: Cross-Attention of Skeleton and Audio

Mounir Boudmagh[1] (✉), Adlen Kerboua[2], Mohamed Redjimi[2]

[1]Badji Mokhtar Annaba University, Annaba, Algeria

[2]University of 20 August 1955, Skikda, Algeria

mounir.boudmagh@univ-annaba.dz

ABSTRACT

Human action recognition (HAR) systems are foundational for mobile educational technologies, such as gesture-based learning analytics and remote skill acquisition. However, current systems often fail in real-world settings due to visual occlusion and the neglect of the rich contextual information provided by the acoustic modality, particularly in visual-centric datasets such as NTU RGB+D 60 and MSR Daily Activity 3D. By manually producing action-relevant audio streams for these datasets, we propose a multimodal approach that fuses skeleton and audio modalities through a cross-attention mechanism. Our framework processes skeleton data by integrating joints and limbs into an $H \times W \times 31$ spatial feature map, which is then fed into a ResNet50 backbone. Log-Mel spectrograms are encoded using a ConvNeXt-T architecture. A cross-attention mechanism is employed to fuse these features, effectively learning inter-modal dependencies. Evaluations demonstrate significant gains: 94.7% on NTU RGB+D X-SUB (up from 90.5% using only skeleton data) and 97.9% on MSR Daily Activity 3D (compared to 89.8%). These results quantitatively establish the critical role of audio in enabling robust, real-time feedback loops that are essential for smart learning environments and interactive mobile coaching, where visual data alone is unreliable.

KEYWORDS

human action recognition (HAR), artificial intelligence, computer vision, skeleton, audio, cross-attention

1 INTRODUCTION

The human perception is formed by integrating diverse pieces of information in the brain, whether watching a video, listening to a conversation, or engaging in sports. At various stages, the brain senses and combines inputs from multiple sensory organs, including vision, hearing, touch, and smell, to construct distinct perceptions of the world. In the modern era of ubiquitous computing, recreating a human-like perceptual ability in technology is essential for developing next-generation

Boudmagh, M., Kerboua, A., Redjimi, M. (2026). Multimodal Human Action Recognition for Ubiquitous Systems: Cross-Attention of Skeleton and Audio. *International Journal of Interactive Mobile Technologies (iJIM)*, 20(5), pp. 70–86. https://doi.org/10.3991/ijim.v20i05.58381

Article submitted 2025-08-22. Revision uploaded 2025-12-19. Final acceptance 2025-12-21.

interactive and mobile applications, particularly in fields such as mobile health, remote rehabilitation, and interactive training systems.

While early HAR studies relied heavily on surveillance and broad-scene understanding [1, 2], recent research focuses on ubiquitous computing spaces, including real-time mobile health monitoring [3, 4], personalized remote rehabilitation [5, 6], and smart environments utilizing data from smart devices [7, 8]. The proliferation of smartphones and portable devices equipped with advanced sensing capabilities allows for sophisticated analysis in real-world, uncontrolled environments. Early work focused on unimodal representations (vision, inertial sensing) [9–11]. However, advances in deep learning now enable the extraction of detailed skeleton data directly from commodity mobile cameras or augmented reality (AR) applications. This capability is particularly vital for embodied learning scenarios, such as remote physical therapy or technical skill training (machinery repair via AR). In these contexts, a system must verify both the user's posture (skeleton) and the acoustic outcome of their action (audio), such as the click of a tool or the rhythm of a movement, to provide valid educational feedback.

Unimodal data are inherently limited. As defined by Lahat et al. [12], data fusion across different sensors mitigates uncertainty and improves representation quality beyond what single modalities can achieve. Integrating audio and vision is advantageous; auditory information compensates for low visual quality (e.g., occlusion and low light), while visual data is robust to acoustic interference [13–20]. However, challenges remain, including modality of misalignment and non-discriminative audio signals. For ubiquitous mobile deployment, relying solely on heavy visual processing introduces latency and battery drain. Therefore, integrating lightweight audio cues is essential to maintain high recognition confidence when visual processing is throttled or occluded.

This paper addresses these issues by proposing a novel system designed as a robust foundation for interactive, portable HAR applications. Unlike previous works focused solely on benchmarks, we position our cross-attention framework as an enabler for ubiquitous learning, capable of resolving semantic ambiguities that hinder automated mobile tutoring. The remainder of this paper is organized as follows: Section 2 reviews vision-based, inertial, and audio-visual fusion techniques. Section 3 describes the proposed methodology. Section 4 presents the experiments and results, and Section 5 concludes with implications and future research directions.

2 RELATED WORK

2.1 Vision-based recognition approach

Vision-based activity recognition algorithms are typically used to analyze third-person videos and can be categorized into two main approaches: object-based and motion-based. Motion-based techniques leverage temporal patterns in bodily movements to recognize activities. Recent advancements emphasize robust feature extraction and spatiotemporal modeling.

For RGB video, Xing et al. [21] proposed MS-HARA, a multiscale RGB model designed to jointly identify and predict human activities using temporal modeling and attention mechanisms with strong performance on UCF101 and HMDB51. In-depth-based approaches, Rao et al. [22] introduced the enhanced depth motion map (EDMM), which improves upon conventional depth motion maps (DMMs)

by utilizing per-pixel motion disambiguation and a 9-layer CNN. They tested the EDMM on two well-known datasets, MSR Action3D and UTD-MHAD. For skeleton-based action recognition, Ibh et al. [23] proposed TemPose. This transformer model enhances fine-grained motion recognition in badminton by eliminating padded tokens in attention maps to reduce noise and jointly encoding player and ball positions for action classification. Plizzari et al. [24] introduced a spatial-temporal transformer network (ST-TR) utilizing spatial self-attention to correlate joints within a frame and temporal self-attention to track dynamics across frames.

2.2 Inertial sensing (IMU) recognition approach

To address vision-only limitations such as occlusion and privacy, researchers often utilize an inertial measurement unit (IMU), primarily consisting of accelerometers and gyroscopes. It is highly valued in ubiquitous computing and HCI for its low power consumption and privacy preservation. Alam et al. [25] demonstrated that fusing wrist and smartphone sensor data improves accuracy using a bidirectional LSTM. Cao et al. [26] introduced a graph-LSTM (G-LSTM) to model the spatial topology among multiple IMU sensors explicitly. By incorporating a multi-task classification model for joint action and identity prediction. However, IMU systems are constrained by their inability to capture external environmental cues or human-object interactions, leading to semantic ambiguity.

2.3 Multimodal fusion approach

Multimodal approaches combine various data streams, such as auditory and sensor data, to classify human actions and analyze behavior. In the context of interactive mobile technologies, research has explored multimodal fusion at both the sensor level and the model level. Fang et al. [27] employed model-level fusion in resource-constrained VR using a dual-path decoder with attention-based rescoring to fuse streaming (CTC) and non-streaming speech recognition outputs, balancing low latency and high accuracy for continuous speech and interaction behavior analysis. Zellers et al. [28] proposed MERLOT-Reserve, a multimodal transformer that utilizes cross-modal attention to jointly learn representations of video, audio, and text, thereby gaining temporal and semantic context to understand audio-visual events. Bock et al. [29] introduced the WEAR dataset and demonstrated temporal action localization models that fuse egocentric video and inertial data via simple concatenation for robust outdoor sports human activity recognition. While Chadha et al. [30] used a CNN-LSTM hybrid for model-level fusion of mobile sensor data, achieving 98.41% transitional activity accuracy by combining feature extraction with temporal sequence modeling.

3 THE PROPOSED APPROACH

Our activity recognition pipeline involves two stages: extracting audio-visual feature representations and fusing these features to capture cross-modal dependencies. The framework is shown in Figure 1.

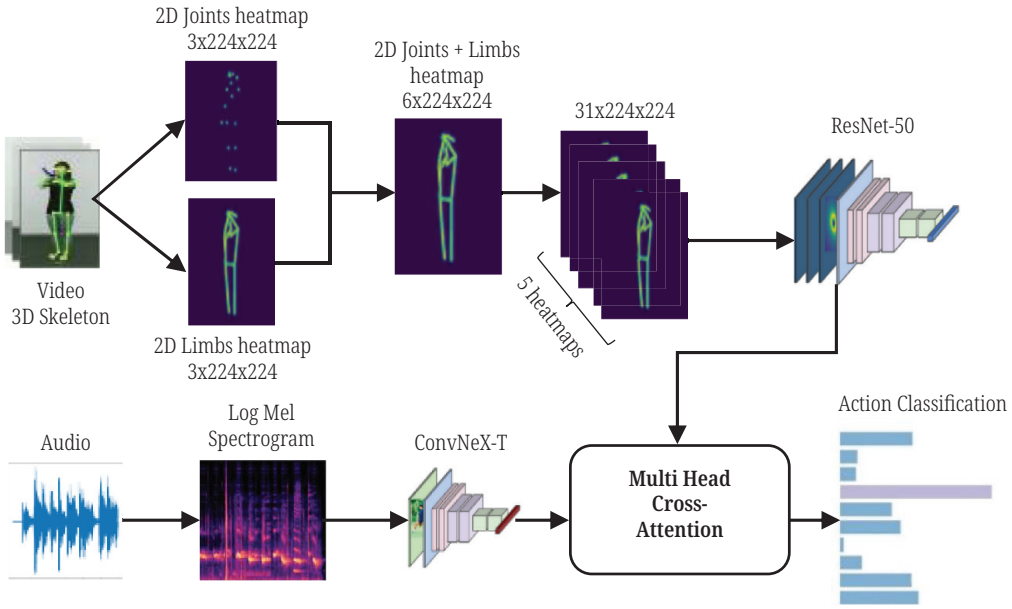

Fig. 1. Our audiovisual activity recognition diagram

3.1 Visual stream architecture (skeleton processing)

Inspired by PoseConv3D [31], we reformulate 3D joint sequences into spatiotemporally enriched heatmap representations. The process begins with a cleaning phase where pseudo-skeletons with low confidence scores (lower than 0.2) are eliminated, followed by normalization relative to the torso and centering on the pelvis joint to ensure spatial invariance. These cleansed 3D joint coordinates are then projected to the XY (frontal), XZ (top-down), and YZ (side) planes to generate associated heatmaps. For each view, we generate two complementary maps: Joint Heatmaps, which model joint presence using a Gaussian kernel, and Limb Heatmaps, which encode the connectivity between joints using a Gaussian distribution along the limb segment (see Appendix A, Eq. 1, 2).

These maps are concatenated channel-wise, resulting in a fused heatmap tensor of size $H \times W \times 6$, which we stacked across five frames to $H \times W \times 30$ to capture motion dynamics, then augmented with a motion-specific channel to produce the final $H \times W \times 31$ representation. Uniform Temporal Sampling is applied to systematically capture motion evolution across the video duration before the data is fed into a modified ResNet-50 backbone, where the architecture is designed to match the exact standard input sizes used by ResNet models [32]. The model is shown in Figure 2.

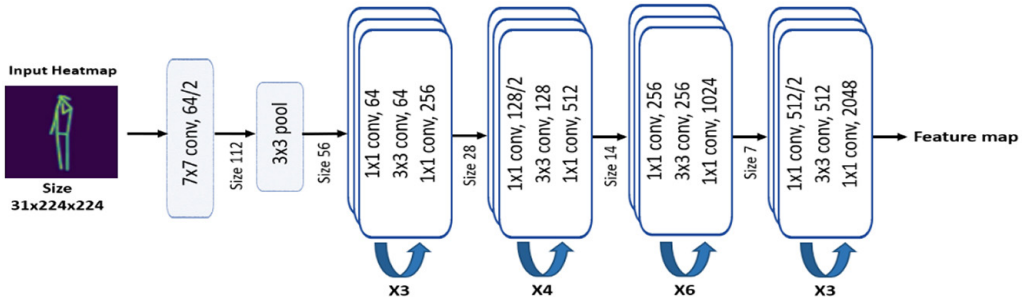

Fig. 2. ResNet-50 model architecture

3.2 Audio stream architecture

Each raw audio sample is normalized to a consistent length (by padding or trimming) and enhanced by utilizing a random time shift to improve the model's robustness to misalignment in time. To improve generalization against real-world temporal misalignments, we apply random time shifting during training. We transform the signal into the frequency domain using the short-time Fourier transform (STFT) to derive log-Mel spectrograms, which effectively capture energy distribution across perceptually relevant frequency bands. Crucially, to encode the temporal dynamics of acoustic events, we compute the first-order (Δ) and second-order (Δ^2) derivatives of the spectrogram. These are concatenated with the base log-Mel coefficients to form a rich 3-channel pseudo-image tensor. This input is processed by a ConvNeXt-Tiny architecture [33], which utilizes large kernel depth-wise convolutions and inverted bottleneck blocks to efficiently extract hierarchical acoustic features (see Appendix A Eq.: 3–5). Figure 3 shows the architecture of the ConvNeXt-Tiny model.

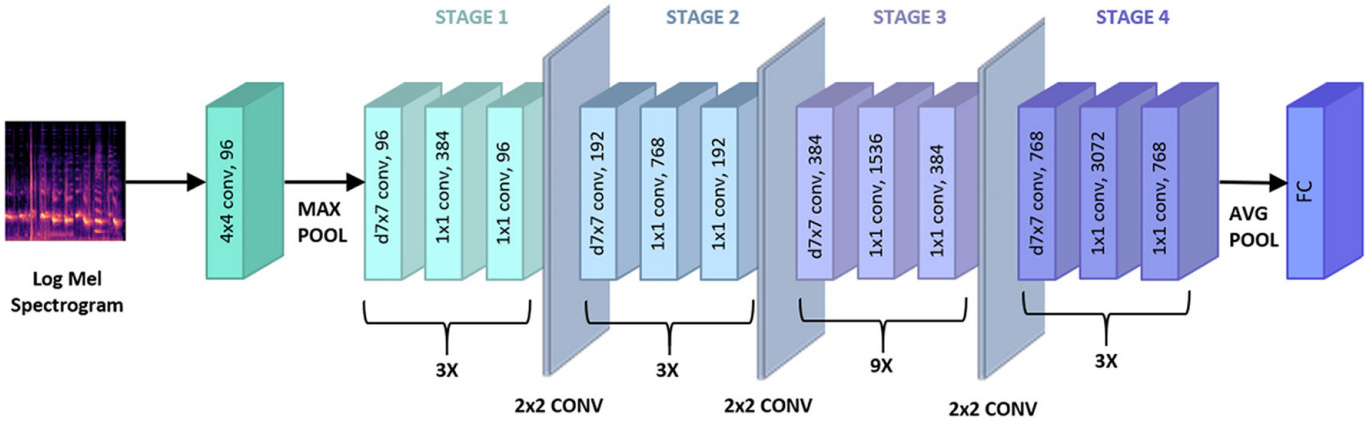

Fig. 3. ConvNeXt-T model architecture

3.3 Cross-attention fusion mechanism

To fuse these heterogeneous modalities, we first project the high-dimensional features from the visual (ResNet-50) and audio (ConvNeXt-T) streams into a common 256-dimensional latent space using learned linear layers. This dimensionality reduction is strategic, balancing representational capacity with computational tractability. The core fusion mechanism is a 4-head cross-attention module [34]. In this configuration, the audio features serve as queries (Q_A), effectively "scanning" the visual features, which act as Keys (K_V) and Values (V_V). This allows the model to learn inter-modal dependencies, dynamically weighing the importance of specific visual cues based on the acoustic context (and vice versa). The attention output is refined through residual connections, layer normalization, and dropout regularization to ensure training stability (detailed derivation in Appendix A, Eq. 6–8). Figure 4 demonstrates the multi-head cross-attention mechanism:

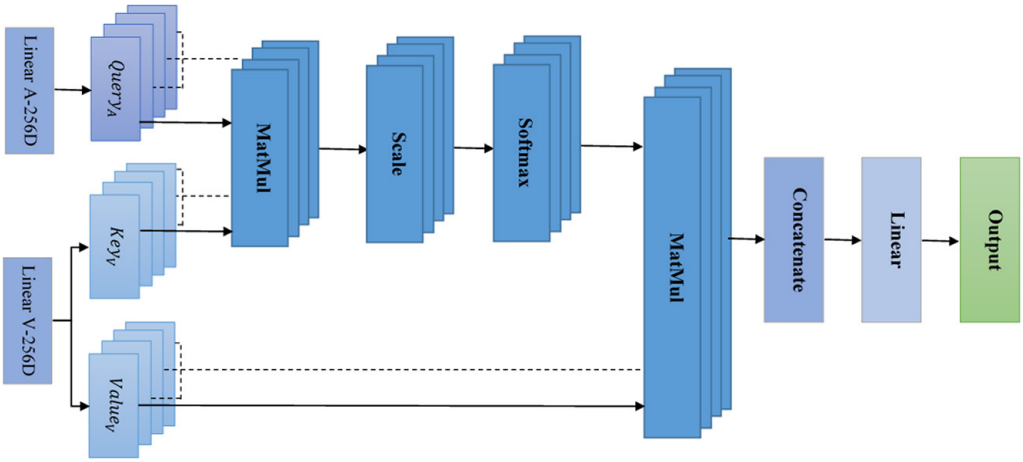

Fig. 4. Multi-head cross-attention mechanism

4 EXPERIMENTAL EVALUATION

4.1 The datasets

We utilize two datasets, NTU RGB+D and MSR Daily Activity 3D, for our experiments. We create an auditory stream for each of the videos in these two datasets.

NTU RGB+D.

Visual branch: This benchmark dataset [35] comprises 56,880 RGB+D video samples encompassing 60 actions, captured from three camera angles using Microsoft Kinect v.2. It contains RGB videos, depth map sequences, 3D skeletal data, and infrared (IR) videos for each sample. The resolutions of RGB videos are 1920x1080; depth maps and IR videos are all in 512x424; and 3D skeletal data contains the 3D coordinates of 25 body joints at each frame.

Aural branch (NTU RGB+D Audio dataset): The NTU RGB+D audio dataset was synthetically generated using AudioGen (Meta AudioCraft) [36], a large language model designed for text-to-audio generation. We created 56,880 audio samples (WAV, 32 kHz, mono) corresponding to the visual events.

MSR Daily Activity 3D.

Visual branch: This dataset [37] contains 320 Kinect-recorded daily activity examples across 16 activities performed in two positions (sitting on a sofa and standing), presenting significant challenges: high noise, occlusion, complex human-object interactions, and substantial proportions of complicated human-object interactions that add further complexity for analysis and validation.

Aural Branch: We employed a human-in-the-loop annotation process. Two designated actors performed the acoustic components of the actions specifically to match the semantic context of the visual clips. While this dubbing approach ensures semantic alignment between modalities, we acknowledge that it may not capture the micro-temporal synchronization of the original visual performance. However, this suffices for the proposed semantic disambiguation task.

4.2 Implementation details

The model was implemented using the TensorFlow framework. We utilized standard cross-subject (X-SUB) and cross-view (X-VIEW) protocols for NTU RGB+D and

a 60/20/20 split for the smaller MSR Daily Activity 3D dataset to ensure robust evaluation. The ResNet-50 was initialized with ImageNet weights, while the fusion module was trained with the backbones frozen to preserve the integrity of the pre-trained feature representations. We employed the Adam optimizer with decoupled weight decay across all streams. To prevent overfitting, we applied extensive data augmentation (including spatial rotation, frame skipping, and SpecAugment) and a regularization strategy involving label smoothing and dropout. Detailed hyperparameters, including learning rates, batch sizes, and augmentation values, are provided in Appendix B.

4.3 Model complexity and deployment feasibility

To comprehensively assess suitability for ubiquitous systems, we analyzed the computational complexity, storage footprint, and runtime latency for mobile deployment, the results are summarized in Table 1.

Table 1. Model complexity and suitability for real-time mobile and Edge applications

Model	Parameters (M)	FLOPs (G)	Model Size (Storage)	Peak Memory (RAM)	Latency/FPS
Visual Stream	25.6 M	4.1	~102.4 MB	~180 MB	~31 ms (32 FPS)
Audio Stream	28.0 M	4.5	~112 MB	~200 MB	~34 ms (29 FPS)
Total Fused Model	53.6 M	8.6	~214.4 MB	~450 MB	~55 ms (18 FPS)

As shown, the fused model requires approximately 450 MB of peak memory, fitting comfortably within the hardware limits of modern smartphones. However, the uncompressed storage footprint (~214.4 MB) and estimated throughput (~18 FPS) indicate that while the system is viable for interactive feedback on high-end mobile NPUs (e.g., Apple Neural Engine, Snapdragon Hexagon), widespread deployment on low-power edge devices requires immediate quantization (INT8) and pruning (as detailed in Section 5.2). These optimizations are crucial for reducing application size, preventing thermal throttling, and ensuring sustained fluid performance.

4.4 Experimental results

The performance results of the video-only model trained with the ResNet-18, ResNet-34, and ResNet-50 architectures; the audio-only model trained with ConvNeX-T; and the cross-attention fusion approach are shown in Table 2.

Table 2. Accuracy for NTU RGB+D (X-SUB, X-VIEW) and MSR Daily Activity 3D

	Model	Params	MSR Daily Activity	NTU RGB+D (X-Sub)	NTU RGB+D (X-View)
Video Only	**ResNet-18**	11.7 M	81.4%	79.4%	84.3%
	ResNet-34	21.8 M	88.7%	85.4%	88.5%
	ResNet-50	25.6 M	91.8%	90.5%	94.6%
Audio Only	**ConvNeXt-T**	28 M	87.8%	81.7%	81.7%
Video + Audio	**Cross-Attention**	53.5 M	**97.9%**	**94.7%**	**98.6%**

As shown, video-only models outperform audio-only models across all datasets, confirming the stronger discriminative power of visual cues for HAR. However, the cross-attention fusion model consistently surpasses both unimodal baselines, improving accuracy by +6.1% on MSR Daily Activity and up to +4.1% (X-Sub)/+4.0% (X-View) on NTU RGB+D over the strongest video-only architecture (ResNet-50). These results demonstrate that integrating skeleton-based visual features with audio information enables the model to resolve ambiguities present in unimodal streams and significantly boosts overall recognition performance.

Figure 5 shows the confusion matrix of the 20 least accurate actions in the NTU RGB+D dataset, and Figure 6 shows the confusion matrix for the MSR Daily Activity 3D dataset.

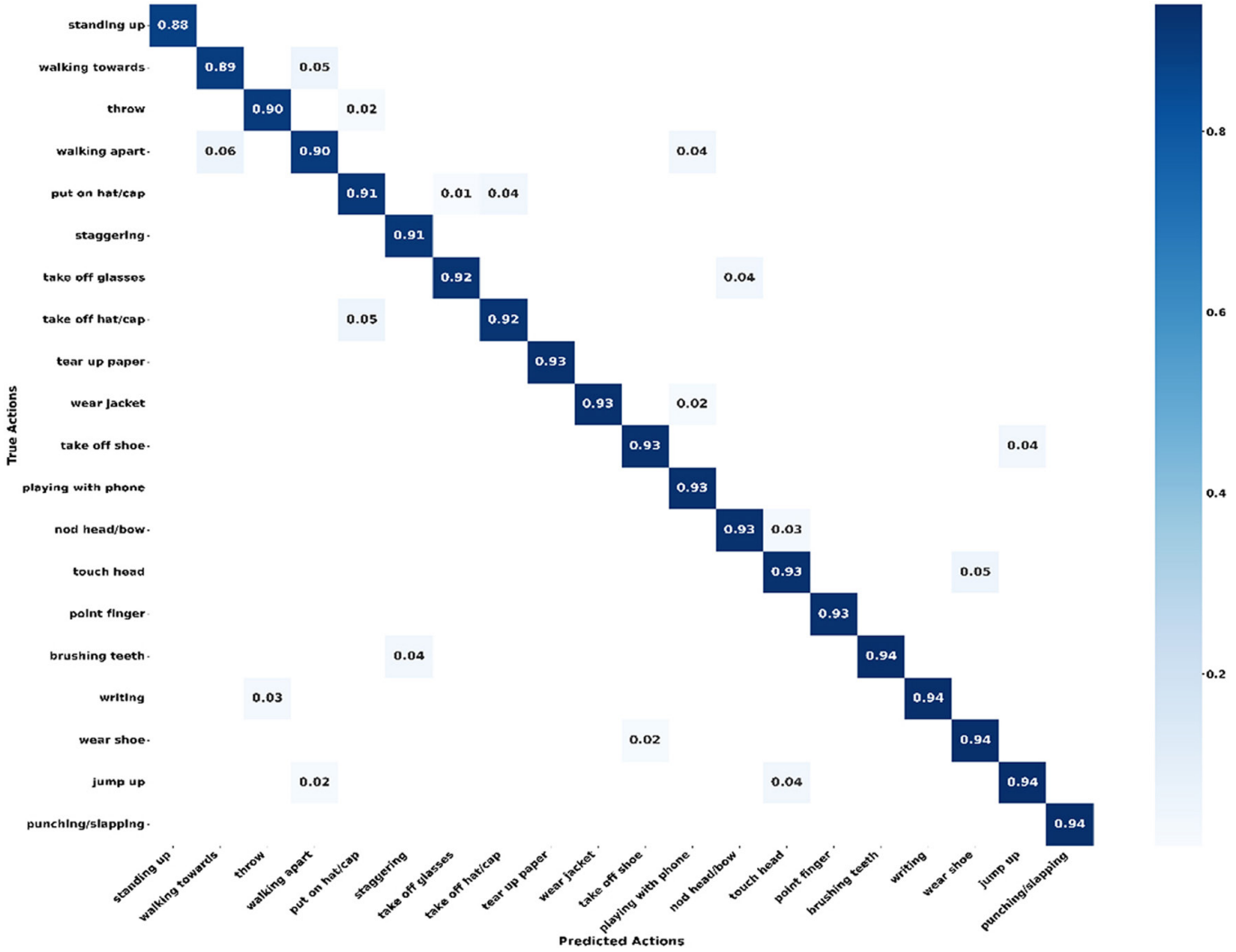

Fig. 5. The NTU RGB-D 60 confusion matrix for the lowest 20 actions

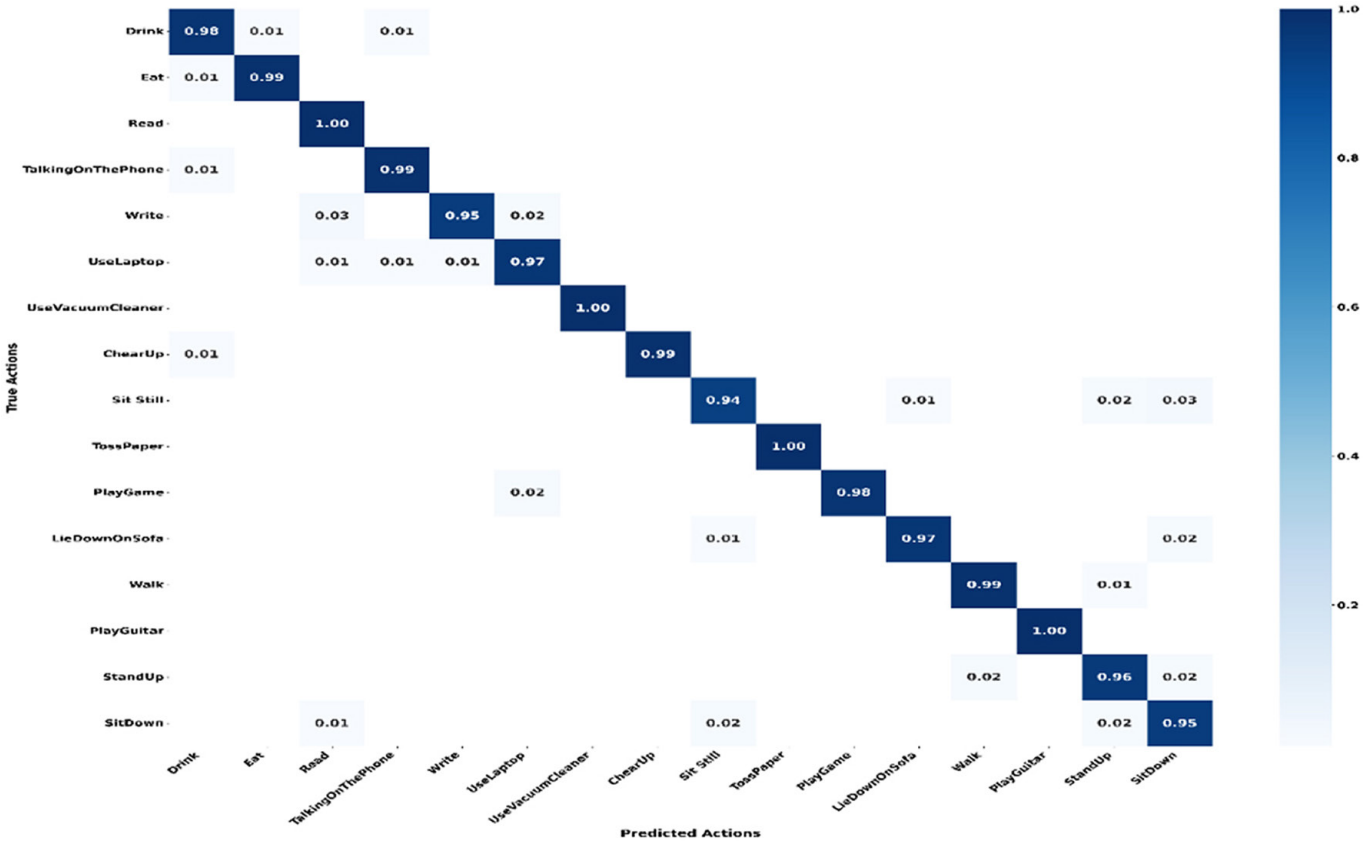

Fig. 6. The MSR Daily Activity 3D confusion matrix

We compare our models with state-of-the-art methods on the NTU RGB+D 60 and MSR Daily Activity 3D datasets, as shown in Tables 3 and 4, respectively.

Table 3. Comparison of the accuracy (%) with state-of-the-art methods on NTU RGB-D 60

Approach	Modality	Accuracy		
		X-Sub	X-View	Average
DSTA-Net [38]	Skeleton	91.5	96.4	93.9
MS-G3D ++ [39]	Skeleton	92.2	96.2	94.2
POSECONV3D [31]	Skeleton	94.1	97.1	95.6
VPN (I3D) [40]	Skeleton + RGB	93.5	96.2	94.6
TCEM-MMNet [41]	Skeleton + RGB	94.3	98.8	96.6
Cross-Attention	**Skeleton + Audio**	**94.7**	**98.6**	**96.7**

Table 4. Comparison of the accuracy (%) with state-of-the-art methods on MSR Daily Activity

Approach	Modality	Accuracy
Jiagang et al. [42]	RGB + Skeleton	93.0
Yang et al. [43]	Skeleton	94.7
Kerboua et al. [44]	Skeleton	95.3
DSSCA-SSLM [45]	Depth	96.3
Cross-Attention	**Skeleton + Audio**	**97.9**

Both tables consistently demonstrate that our framework achieves state-of-the-art performance on both the NTU RGB+D 60 and MSR Daily Activity datasets. This strong performance reflects the strength of combining acoustic cues with skeletal data to supplement the model's ability to disambiguate actions in the video that are visually similar, which the skeleton-only model might miss. Figure 7 compares skeleton-only vs. multimodal accuracy for complex actions on NTU RGB+D 60 (X-Sub) and MSR Daily Activity 3D.

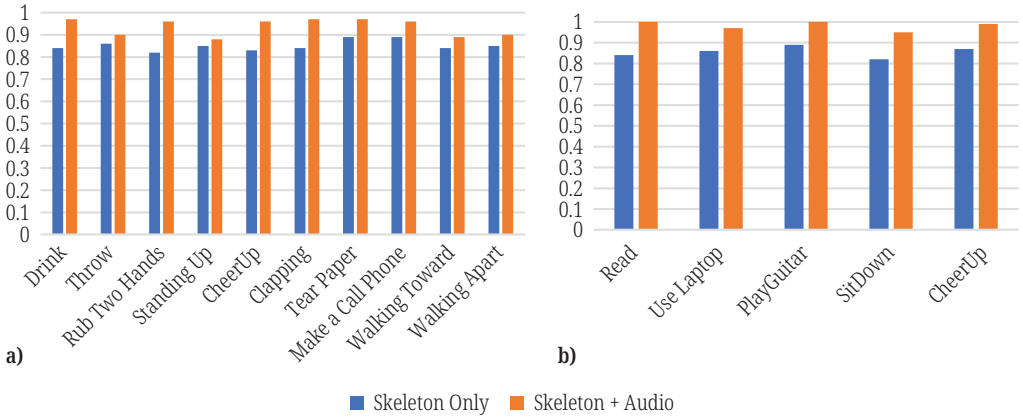

Fig. 7. Classification accuracy for top-10 challenging actions on across-subject protocol of NTU RGB+D 60 (a) and top-5 of MSR Daily Activity 3D (b) datasets

Figure 7 demonstrates that audio consistently enhances action recognition, improving both dynamic and distinct actions, such as "Drink" and "Throw," where audio associates sounds, and static or subtle actions such as "Read" and "Use Laptop," where skeletal motion provides limited discriminative information; audio provides critical complementary acoustic details (e.g., page turning and person voice reading and typing sounds), thereby resolving ambiguity, as qualitatively illustrated by the corrected predictions in Figure 8.

SKELETON PREDICTION	AUDIO-VISUAL PREDICTION	DATASET
WALKING APART	WALKING TOWARDS	NTU RGB+D
HAND WAVING	CHEER UP	NTU RGB+D
STAND UP	SIT DOWN	MSR DAILY ACTIVITY 3D
TALKING ON THE PHONE	USE LAPTOP	MSR DAILY ACTIVITY 3D

Fig. 8. Qualitative evaluation of skeleton only and skeleton + audio models on NTU RGB+D 60 and MSR Daily Activity 3D

The analysis of Figure 8 reveals that the skeleton-only approach occasionally misclassifies actions, resulting in limited accuracy compared to the audio-visual approach. The model's performance is greatly enhanced by adding audio-visual cues, which accurately recognize actions that were previously misclassified or had lower accuracy.

5 DISCUSSION AND CONCLUSION

This study proposes a new and reliable multimodal human action recognition framework that creatively combines skeleton and audio features through a cross-attention mechanism. Our main contribution lies in enriching datasets through the manual creation of semantically relevant audio streams for the NTU RGB+D and MSR Daily Activity 3D datasets, as well as the design of a robust fusion architecture. Our unique modality fusion approach demonstrated meaningful performance benefits of up to +6.1% on MSR and +4.2% on NTU compared to unimodal approaches and even showed the discriminative power of the final audio (87.8%, with 81.7% audio-only accuracy on MSR and NTU RGB+D, respectively). These results illustrate the role of audio in distinguishing nuances related to actions and providing context for dynamic movements and further developing towards more robust and complete action recognition systems.

5.1 Implications for mobile learning and interactive systems

The proposed framework supports ubiquitous and mobile applications through real-time multimodal feedback, offering critical advantages for educational technology and interactive systems:

1. Gesture-Based Learning Analytics: In mobile tutoring, simple correctness scores are often insufficient, our model adds granularity by cross-referencing skeletal pose with audio intensity. The cross-attention mechanism resolves visual ambiguity by prioritizing audio cues, enabling the system to accurately validate the semantic category of the action.
2. Embodied Learning in AR/VR: For remote skill acquisition, the fusion model verifies that physical actions produce the correct environmental sounds, confirming successful object interaction rather than just movement. This closes the feedback loop during visual occlusion, enabling the system to detect execution errors (e.g., a failed vacuum seal) via acoustic signatures even when the user's hands block the camera view.
3. Smart Learning Environments and Ubiquitous Monitoring: In crowded classrooms, the system maintains analytics by detecting acoustic signatures (e.g., typing) despite frequent visual occlusion. This reliability extends to mHealth, where fusing skeletal fall detection with audio impact sounds significantly reduces false alarms in daily activity tracking.

5.2 Challenges and future work

Deploying this architecture on mobile or edge devices poses key challenges:

1. **Model Efficiency:** The current computational cost (8.6 GFLOPs) and uncompressed storage footprint (~214 MB) risk high latency and excessive app size on mid-range mobile devices. With an estimated throughput of ~18 FPS, the system

approaches the lower bound of interactivity but lacks the fluidity required for seamless feedback. Future work must prioritize model quantization (INT8) and channel pruning to drastically reduce the 214 MB footprint and energy consumption, ensuring sustained real-time performance without sacrificing the semantic precision of the audio stream.

2. **Data Validity:** While the high performance on MSR Daily Activity 3D validates the cross-attention mechanism against ecologically valid, live-recorded audio, the reliance on synthetic audio for NTU RGB+D introduces a domain gap. The next critical step is validation against unconstrained, noisy mobile microphone data to ensure real-world robustness.
3. **Extension, Integration, and Generalizability:** Future work involves expanding the framework with additional modalities (objects, RGB, and inertial) and developing temporal compression methods for real-time acceleration. Crucially, achieving global scalability requires explicitly addressing cultural diversity in action execution and validating robustness under diverse environmental acoustic conditions (reverberation).

In summary, the proposed system advances multimodal human action recognition toward more accurate, mobile-ready, and ecologically valid real-world applications.

6 DATA AND CODE AVAILABILITY

To support replicability and open science, the source code for the Cross-Attention architecture, along with the scripts used for synthetic audio generation, will be made publicly available via a GitHub repository upon publication.

7 ETHICAL AND GENERATIVE AI STATEMENT

All datasets used in this study comply with ethical research standards. The human-recorded audio data were collected from consenting adult participants performing predefined actions without speech or identifiable information. The synthetic audio was generated using AudioGen under the Apache License 2.0 and used solely for academic research. No personal, biometric, or sensitive data were processed in this study. No generative AI tools were used for the writing or editing of this manuscript.

8 ACKNOWLEDGEMENTS

The authors acknowledge the use of AudioGen (Meta AI, 2023) for generating the synthetic audio data used in this study, which is provided under the Apache License 2.0 (royalty-free with attribution). License details are available at https://www.apache.org/licenses/LICENSE-2.0.

9 REFERENCES

[1] M. Liu, F. Meng, C. Chen, and S. Wu, "Novel motion patterns matter for practical skeleton-based action recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, pp. 1701–1709, 2023. https://doi.org/10.1609/aaai.v37i2.25258

[2] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, Kuala Lumpur, Malaysia, 2015, pp. 579–583. https://doi.org/10.1109/ACPR.2015.7486569

[3] A. R. Rasa, "Artificial intelligence and its revolutionary role in physical and mental rehabilitation: A review of recent advancements," *BioMed Research International*, vol. 2024, p. 9554590, 2024. https://doi.org/10.1155/bmri/9554590

[4] S. Klakegg *et al.*, "CARE: Context-awareness for elderly care," *Health Technol.*, vol. 11, pp. 211–226, 2021. https://doi.org/10.1007/s12553-020-00512-8

[5] A. Hisam, S. Zia-ul-Haq, S. Aziz, P. Doherty, and J. Pell, "Effectiveness of mobile health augmented cardiac rehabilitation (MCard) on health-related quality of life among post-acute coronary syndrome patients: A randomized controlled trial," *Pak. J. Med. Sci.*, vol. 38, no. 3, pp. 716–723, 2022. https://doi.org/10.12669/pjms.38.3.4724

[6] A. Farsi, G. Cerone, D. Falla, and M. Gazzoni. "Emerging applications of augmented and mixed reality technologies in motor rehabilitation: A scoping review," *Sensors*, vol. 25, no. 7, p. 2042, 2025. https://doi.org/10.3390/s25072042

[7] T. S. Qureshi *et al.*, "A systematic literature review on human activity recognition using smart devices: Advances, challenges, and future directions," *Artif. Intell. Rev.*, vol. 58, p. 276, 2025. https://doi.org/10.1007/s10462-025-11275-x

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

[9] Q. Zhao, C. Zheng, M. Liu, and C. Chen, "A single 2D pose with context is worth hundreds for 3D human pose estimation," in *Proc. 37th Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2023. https://doi.org/10.5555/3666122.3667315

[10] S. Yao, S. Hu, Y. Zhao, A. Zhang, and T. Abdelzaher, "DeepSense: A unified deep learning framework for time-series mobile sensing data processing," in *Proc. 26th Int. Conf. World Wide Web (WWW '17)*, Geneva, Switzerland, 2017, pp. 351–360. https://doi.org/10.1145/3038912.3052577

[11] C. Feichtenhofer, "X3D: Expanding architectures for efficient video recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 200–210. https://doi.org/10.1109/CVPR42600.2020.00028

[12] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015. https://doi.org/10.1109/JPROC.2015.2460697

[13] S. Papadakis, S. H. Lytvynova, S. M. Ivanova, and I. A. Selyshcheva, "Advancing lifelong learning with AI-enhanced ICT: A review of 3L-Person 2024," in *CEUR Workshop Proceedings*, Lviv, Ukraine, 2023.

[14] C. Xu, R. Panda, A. Nagrani, J. Lin, and R. Feris. "Audio-visual slowfast networks for video recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10457–10467. https://arxiv.org/abs/2001.08740

[15] M. Boudmagh, M. Redjimi, and A. Kerboua, "Video activity recognition based on objects detection using recurrent neural networks," in *Innovations in Smart Cities Applications*, vol. 4, 2021, pp. 555–565. https://doi.org/10.1007/978-3-030-66840-2_65

[16] P. Wang, F. Guo, F. Gu, M. Li, and X. Long, "MobileHAR: A lightweight and efficient human activity recognition model based on inverted residual inception block," in *2024 20th International Conference on Mobility, Sensing and Networking (MSN)*, Harbin, China, 2024, pp. 834–841. https://doi.org/10.1109/MSN63567.2024.00116

[17] G. Sawadwuthikul *et al.*, "Visual goal human-robot communication framework with few-shot learning: A case study in Robot Waiter system," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 3, pp. 1883–1891, 2022, https://doi.org/10.1109/TII.2021.3049831

[18] W. Yang, Q. Xiao, and Y. Zhang, "HAR2bot: A human-centered augmented reality robot programming method with the awareness of cognitive load," *J. Intell. Manuf.*, vol. 35, no. 5, pp. 1985–2003, 2024. https://doi.org/10.1007/s10845-023-02096-2

[19] E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen, "EPIC-fusion: Audio-visual temporal binding for egocentric action recognition," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), 2019, pp. 5491–5500. https://doi.org/10.1109/ICCV.2019.00559

[20] S. Papadakis, A. M. Striuk, H. M. Kravtsov, M. P. Shyshkina, M. V. Marienko, and H. B. Danylchuk, "Embracing digital innovation and cloud technologies for transformative learning experiences," in *Proc. 11th Workshop Cloud Technol. Educ. (CTE 2023)*, Kryvyi Rih, Ukraine, 2024.

[21] Y. Xing, S. Golodetz, A. Everitt, A. Markham, and N. Trigoni, "Multiscale human activity recognition and anticipation network (MS-HARA)," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 7679–7692, 2022. https://doi.org/10.1109/TNNLS.2022.3167824

[22] D. S. Rao, L. K. Rao, V. Bhagyaraju, and G. K. Meng, "Enhanced depth motion maps for improved human action recognition from depth action sequences," *Traitement du Signal*, vol. 41, no. 3, pp. 1461–1472, 2024. https://doi.org/10.18280/ts.410334

[23] M. Ibh, S. Grasshof, D. Witzner, and P. Madeleine, "TemPose: A new skeleton-based transformer model designed for fine-grained motion recognition in badminton," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 5198–5207. https://doi.org/10.1109/CVPRW59228.2023.00548

[24] C. Plizzari, M. Cannici, and M. Matteucci, "Spatial-temporal transformer network for skeleton-based action recognition," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR) Workshops*, 2021, pp. 694–701. https://arxiv.org/abs/2008.07404

[25] A. Alam, A. Das, M. S. Tasjid, and A. A. Marouf, "Leveraging sensor fusion and sensor-body position for activity recognition for wearable mobile technologies," *Int. J. Interact. Mob. Technol.*, vol. 15, no. 17, pp. 141–155, 2021. https://doi.org/10.3991/ijim.v15i17.25197

[26] J. Cao, Y. Wang, H. Tao, and X. Guo, "Sensor-based human activity recognition using graph LSTM and multi-task classification model," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 18, no. 3s, pp. 1–19, 2022. https://doi.org/10.1145/3561387

[27] L. Fang, X. Wang, and L. Zhang, "Design of a virtual reality-supported immersive English learning environment and interaction behavior analysis," *Int. J. Interact. Mob. Technol.*, vol. 19, no. 21, pp. 184–198, 2025. https://doi.org/10.3991/ijim.v19i21.58853

[28] R. Zellers *et al.*, "MERLOT RESERVE: Neural script knowledge through vision and language and sound," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 2022, pp. 16354–16366. https://doi.org/10.1109/CVPR52688.2022.01589

[29] M. Bock, H. Kuehne, K. Van Laerhoven, and M. Moeller, "WEAR: An outdoor sports dataset for wearable and egocentric activity recognition," in *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 8, no. 4, 2024. https://doi.org/10.1145/3699776

[30] J. Chadha, A. Jain, Y. Kumar, and N. Modi, "Hybrid deep learning approaches for human activity recognition and postural transitions using mobile device sensors," *SN Computer Science*, vol. 5, 2024. https://doi.org/10.1007/s42979-024-03300-7

[31] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, USA, 2022, pp. 2959–2968. https://doi.org/10.1109/CVPR52688.2022.00298

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90

[33] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 2022, pp. 11966–11976. https://doi.org/10.1109/CVPR52688.2022.01167

[34] A. Vaswani *et al.*, "Attention is all you need," in *Proc. 31st Conf. Neural Information Processing Systems (NeurIPS 2017)*, Long Beach, USA, 2017. https://arxiv.org/abs/1706.03762

[35] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3d human activity analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. https://doi.org/10.1109/CVPR.2016.115

[36] J. Copet *et al.*, "Simple and controllable music generation," in *Proc. 37th Conf. Neural Information Processing Systems*, New Orleans, LA, USA, 2023. https://arxiv.org/abs/2306.05284

[37] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012, pp. 1290–1297. https://doi.org/10.1109/CVPR.2012.6247813

[38] D. S. Park *et al.*, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617. https://arxiv.org/abs/1904.08779

[39] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 140–149, https://doi.org/10.1109/CVPR42600.2020.00022

[40] S. Das, S. Sharma, R. Dai, F. Bremond, and M. Thonnat, "VPN: Learning video-pose embedding for activities of daily living," in *Proceedings Computer Vision– ECCV, 16th European Conference*, Glasgow, UK, Part IX 16, Springer, 2020, pp. 72–90. https://doi.org/10.1007/978-3-030-58545-7_5

[41] D. Liu *et al.*, "Temporal cues enhanced multimodal learning for action recognition in rgb-d videos," *Neurocomputing*, vol. 594, 2024. https://doi.org/10.1016/j.neucom.2024.127882

[42] J. Zhu *et al.*, "Action machine: Rethinking action recognition in trimmed videos," *arXiv preprint arXiv:1812.05770,* 2018. https://arxiv.org/abs/1812.05770

[43] H. Yang, D. Sun, Y.-J. Cai, J. Yang, X.-Y. Si, and S.-M. Zhou, "Learning topological representation of 3D skeleton dynamics with persistent homology for human activity recognition," *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Istanbul, Turkiye, 2023, pp. 2709–2716. https://doi.org/10.1109/BIBM58861.2023.10385684

[44] A. Kerboua and M. Batouche, "3D skeleton action recognition for security improvement," *International Journal of Intelligent Systems and Applications*, vol. 11, no. 3, pp. 42–52, 2019. https://doi.org/10.5815/ijisa.2019.03.05

[45] A. Shahroudy, T.-T. Ng, Y. Gong, and G. Wang, "Deep multimodal feature analysis for action recognition in RGB+D videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 5, pp. 1045–1058, 2018. https://doi.org/10.1109/TPAMI.2017.2691321

10 APPENDIX A: MATHEMATICAL DEFINITIONS

Visual Heatmap Generation: The joint heatmaps (J_k) and limb heatmaps (L_l) are defined as:

$$J_{kij} = \exp\left(-\frac{(i-u_k)^2 + (j-v_k)^2}{2\sigma^2}\right) \quad (1)$$

Where J_{kij} represents the weight at spatial position (i, j) for the k-th key point, i and j are the coordinates of a location in the spatial grid, u_k and v_k are the horizontal and vertical coordinates of the k-th key point's center, and σ is the standard deviation controlling how quickly the influence decreases as the distance from the key point center increases.

$$L_{lij} = \exp\left(-\frac{D((i,j),seg[a,b])^2}{2\sigma^2}\right) \tag{2}$$

Where $D((i, j), seg[a, b])$ is the perpendicular distance from pixel (i, j) to the segment connecting the two joints, similar to joints, the limb maps are summed across all limbs in each projection.

ConvNeXt Architecture Details: The depthwise convolution and normalization operations are computed as:

$$Y_{c,i,j} = \sum_{p=0}^{k-1}\sum_{q=0}^{k-1} X_{c,i+p,j+q} \cdot K_{c,p,q} \tag{3}$$

Where $Y_{c,i,j}$ is the value of the output feature map at channel c and spatial coordinates (i, j), $X_{c,i+p,j+q}$ is the value of the input feature map at channel c and spatial coordinates $(i+p, j+q)$, $K_{c,p,q}$ is the value of the convolutional kernel (filter) at channel c and coordinates (p, q), k is the size of the square kernel, p and q are iterators for the kernel's height and width. They also employ inverted bottleneck blocks for computational efficiency:

$$F(X) = Conv_{1\times1}(DW\,Conv_{1\times1}(X)) \tag{4}$$

Furthermore, the use of GELU activation and Layer Normalization (which normalizes features across channels ensures stable training and robust feature learning:

$$y = \frac{x - E[x]}{\sqrt{Var[x]+\varepsilon}} \cdot \gamma \cdot \beta \tag{5}$$

Where x is the input feature for a given layer, $E[x]$ is the mean of the input features over all channels for a single sample, $Var[x]$ is the variance of the input features over all channels, ε is a small constant added for numerical stability to avoid division by zero, and γ and β are learnable scaling and bias parameters.

Cross-Attention Fusion: The fusion mechanism utilizes Multi-Head Attention (MHA), defined as:

$$head_i = softmax\left(\frac{Q_A K_V^T}{\sqrt{d_k}}\right)V_V \tag{6}$$

$$\text{MHA}\,(Q_A, K_V, V_V) = \text{Concat}\,(head_1 \ldots, head_H)W_O \tag{7}$$

Where Q_A = audio queries, K_V, V_V = video keys/values, and $W_O \in R^{256\times256}$ projects concatenated heads. The final fused representation (F) is computed after the MHA stage by incorporating residual connections, Layer Normalization, and Dropout regularization to prevent overfitting:

$$\text{F} = \text{LayerNorm}\,(\text{A} + \text{Dropout}\,(\text{MHA}\,(Q_A, K_V, V_V))) \tag{8}$$

11 APPENDIX B: IMPLEMENTATION AND TRAINING PARAMETERS

Visual Network Settings:

- **Augmentation:** Random rotations (±15°), horizontal flipping (p = 0.5), frame skipping (stride = 2), joint dropout (10%), Gaussian noise (N (0, 0.01)).
- **Optimizer:** Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\lambda = 1 \times 10^{-4}$).
- **Learning Rate:** Initial 3×10^{-4}, plateau-based reduction (factor = 0.1, patience = 5).
- **Training:** Batch size 32, 100 epochs, early stopping (patience = 10).

Aural Network Settings:

- **Preprocessing:** 25-ms frames with 10-ms hop size.
- **Augmentation:** SpecAugment (time warping, freq/time masking), Pitch shift (±2 semitones), Time stretch (±10%).
- **Training:** Batch size 5, Initial LR 3×10^{-4}, (Cosine Annealing).

Fusion Network Settings:

- **Regularization:** Dropout (p = 0.5), Label Smoothing ($\alpha = 0.1$), Gradient Clipping (max norm = 1.0).
- **Training:** Batch size 16, Initial LR $\times 10^{-4}$, (Cosine Decay).

12 AUTHORS

Mounir Boudmagh is a PhD Student in the Embedded Systems Laboratory (LASE), Department of Computer Science at Badji Mokhtar-Annaba University, Algeria. His research focuses on applied Artificial Intelligence, particularly computer vision, object recognition, and human action prediction (E-mail: mounir.boudmagh@univ-annaba.dz).

Adlen Kerboua is a researcher in computer science and intelligent systems, with expertise in numerical optimization, robotics, and data-driven modeling. His research focuses on optimization algorithms, sensitivity analysis, and computational methods applied to robotic mechanisms and engineering systems. He has contributed to peer-reviewed international journals in the fields of robotics and intelligent applications.

Mohamed Redjimi is full professor in computer science department at 'université 20 Août 1955', Skikda, Algeria. He obtained the PhD from 'Université des sciences et des techniques', Lille 1, France 1984, then the post doctorate degree ('Habilitation universitaire') from Badji Mokhtar University, Annaba, Algeria, 2007. He has held several high-ranking positions in university administration and research: Dean of Faculty, Head of Department, Director of Research Laboratory, etc. His current research domain focuses on applied artificial intelligence and emerging systems, particularly routing strategies and protocols in wireless sensor networks.