IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Human Action Recognition

Nityam Upadhyay

M.TECH CSE, KPGU Vadodara

Hetal Chokshi

Assistant Professor, CSE-Department
KSET, KPGU

Abstract

Vision-based single-person human activity recognition (HAR) has advanced rapidly through deep learning. This review surveys 14 recent peer-reviewed studies (2017–2025) focusing on monocular video HAR. We searched IEEE Xplore, SpringerLink, MDPI, and other databases with keywords like "vision-based HAR," "single-person," "CNN," "LSTM," and "Transformer," screening for methods, datasets, and performance. We compare each study's model, dataset, and accuracy, and highlight common limitations. Major findings include the effectiveness of attention and fusion architectures for spatiotemporal feature learning[1][2]. Performance is now very high on benchmark datasets (e.g. >96% on UCF/KTH[3][4]), but generalization and real-time constraints remain challenging. We identify research gaps – notably, reliance on limited datasets and heavy computation – and suggest directions such as lightweight models and richer data. Our review underscores progress and open issues in vision-based single-person HAR.

Index Terms—Human Activity Recognition (HAR), single-person, video, convolutional neural network, LSTM, attention, transformer.

I. Introduction

Human Activity Recognition (HAR) is crucial in applications like surveillance, healthcare, and robotics[5][6]. Vision-based HAR uses camera data to classify actions (e.g. walking, jumping). Recent deep learning approaches extract spatiotemporal features automatically. However, varying backgrounds, viewpoints, and subtle actions make HAR hard[7][8]. This review focuses on **single-person, vision-only** HAR (no multi-person or wearable data) over 2017–2025. We compare 14 exemplar studies (Table I) to assess methods and results. The paper structure follows IEEE style:

Section II details our review methodology

Section III reviews selected papers in a comparative table

Section IV discusses common research gaps

Section V outlines future research directions

Section VI concludes.

II. Review Methodology

We conducted a systematic literature search (2024–2025) in IEEE Xplore, SpringerLink, MDPI, and Google Scholar for papers from 2017 onward. Search queries included "vision-based human activity recognition," "single-person action recognition," and combinations like "CNN LSTM HAR 2022," "video transformer HAR," etc. We filtered for peer-reviewed journal/conference papers focusing on **single-person** activities with **RGB or depth video** input. Excluded were multi-person interaction, sensor-based, and skeleton-only methods. For each candidate, we gathered details on authors, year, network architecture, dataset, accuracy, and reported limitations. Ultimately, 14 papers were selected for their focus on single-person vision HAR (Table I). When summarizing results, we cite each original paper (not secondary sources).

III. Literature Review

Table I compares the 14 selected vision-based single-person HAR studies. Columns list:

(1) authors, year

(2) method

(3) dataset(s)

(4) accuracy or performance

(5) key findings and limitations.

We summarize major results below by paper.

- **Surek *et al.*, 2023**: Proposed an attention-augmented deep model combining ResNet-18 and a vision transformer with self-supervised DINO. Tested on HMDB51, it achieved 96.7% train vs 41.0% test accuracy[9].
    - *Finding:* ViT with LSTM extracted strong features but generalization was limited (test performance low)[9].
    - *Limitation:* Overfitting on small dataset and low test accuracy.
- **Ullah & Munir, 2023**: Developed a dual-attention CNN (channel + spatial) followed by a bidirectional GRU. Evaluated on YouTube, UCF50/101, HMDB51, and Kinetics-600 datasets[10]. It achieved state-of-the-art accuracy with very fast inference: up to 167× speed-up in fps over baselines[10].
    - *Finding:* Attention modules improve feature learning and efficiency.
    - *Limitation:* Primarily optimized for speed; absolute accuracy metrics were not fully reported.
- **Sánchez-Caballero *et al.*, 2023**: Used raw depth-video (NTU RGB+D) with ConvLSTM. Compared a stateless versus stateful RNN. The stateless model achieved ~75.3% accuracy, while the stateful ConvLSTM reached ~80.4% accuracy on NTU (cross-subject)[11], albeit with higher latency (0.89s per video).
    - *Finding:* Exploiting stateful memory significantly boosts depth-based HAR accuracy[11].
    - *Limitation:* Depth-only modality (no RGB), and lower accuracy than RGB methods; also computationally heavier.
- **Uddin *et al.*, 2024**: Compared pure CNN, ConvLSTM, and LRCN (CNN+LSTM) on UCF50 and HMDB51[12]. Simple CNN achieved 99.58% (UCF50) and 92.70% (HMDB51), far outperforming ConvLSTM (82% and 68%) and LRCN (93.44% and 71.55%)[12].
    - *Finding:* For these datasets, straightforward CNN (frame-based) performed best[12].

- *Limitation:* Very high accuracy likely due to overfitting on these relatively small sets; did not generalize to larger or noisier data.

- **Chen *et al.*, 2023**: Presented a CNN-LSTM pipeline for classifying Chinese exercise videos (Baduanjin). The CNN-LSTM achieved 96.43% accuracy, versus 66.07% using a baseline OpenPose skeletal method[13].
  - *Finding:* Deep CNN features greatly improve over pose-only features[13].
  - *Limitation:* Specialized dataset (one exercise); small scale (18 subjects) and only 3 classes.

- **Alzahrani *et al.*, 2025 (YOLO-Act)**: Proposed YOLO-Act, a real-time 3D detection-classification network for spatiotemporal action detection on AVA. On the multi-person AVA dataset, it achieved 73.28% mAP (mean Average Precision), a +28.18 mAP gain over prior LART[14].
  - *Finding:* Keyframe-based YOLO can unify spatial detection with action classification.
  - *Limitation:* Targets multi-person action detection; not purely single-person classification (but still vision-based).

- **Zhang & Abdelmunim, 2025**: Introduced a YOLOv7 object detector feeding an LSTM. On standard HAR datasets (KTH, Weizmann, IXMAS, UCF101), it achieved 96–100% accuracy[4].
  - *Finding:* The YOLO-LSTM achieves near-perfect recognition on these controlled datasets[4].
  - *Limitation:* These datasets are small and well-lit, so the 100% on Weizmann indicates possible dataset saturation; real-world generalization unclear.

- **Sun *et al.*, 2024**: Proposed a k-NN attention mechanism integrated into a Video Vision Transformer (ViViT). On UCF101 and HMDB51, this -ViViT model outperformed state-of-art. They reported "superior accuracy" on both benchmarks[15].
  - *Finding:* Incorporating attention based on nearest-neighbors helps ViT focus on relevant tokens, boosting HAR accuracy[15].
  - *Limitation:* Quantitative gains unspecified; training and inference costs of ViViT are high.

- **Al-Tawil *et al.*, 2025**: Developed a real-time ResNet18+BiLSTM with multi-head attention and optical-flow frame selection. Achieved 96.60% accuracy on UCF101, running at 222 FPS[3].
  - *Finding:* Motion-aware frame sampling plus attention yields both high accuracy and speed[3].
  - *Limitation:* Tested only on UCF-101 (single domain); relies on optical flow which adds overhead.

- **Qin *et al.*, 2025**: Proposed the Swin-CLSTM model: a Video Swin Transformer enhanced with optical flow and embedded ConvLSTM units. On UCF-101, achieved 92.8% top-1 accuracy (3.2% above the baseline ViT)[16].
  - *Finding:* Adding optical flow and ConvLSTM to a transformer improves HAR by a few percent[16].
  - *Limitation:* Moderate gains on one dataset; complex model combining flow and LSTM.

- **Khan & Jung, 2023**: Used dilated convolutions in a CNN+LSTM pipeline on UCF50. Their model scored 94.9% (vs 93.7% baseline CNN-LSTM)[1].
  - *Finding:* Dilated convolutions slightly improved UCF50 accuracy[1].
  - *Limitation:* Small improvement on one dataset; other datasets not tested.

- **Carreira & Zisserman, 2017**: (Baseline comparison) Introduced I3D (two-stream inflated 3D ConvNet) and trained on Kinetics. Achieved 80.2% on HMDB-51 and 97.9% on UCF-101[6].
    - *Finding:* Pretraining on large video dataset yields state-of-art results[6].
    - *Limitation:* Large model; prior to attention/transformer advances.
- **Feichtenhofer *et al.*, 2019**: (Baseline comparison) Proposed SlowFast networks and reported state-of-art accuracy on Kinetics, Charades, AVA[8].
    - *Finding:* Factoring networks into slow (spatial) + fast (temporal) pathways improved video recognition.[8]
    - *Limitation:* Heavy models; not specifically HAR of a single actor.

Each of the above studies notes some limitations (e.g. small or easy datasets, need for large compute, overfitting). Table I (below) summarizes them comparatively.

Table I: Comparative summary of reviewed vision-based single-person HAR papers.

Authors (Year)	Method	Dataset(s)	Performance	Key Findings & Limitations
Surek *et al.* (2023)【28†】	ResNet-18 + ViT + attention (DINO pretrain)	HMDB51	96.7% train, 41.0% test (accuracy)[9]	ViT+LSTM effective for HAR, but test accuracy was low (overfitting).
Ullah & Munir (2023)【39†】	Cascaded channel+spatial -attn CNN + Bi-GRU	YouTube Actions, UCF50, HMDB51, UCF101, Kinetics-600	*State-of-art (unspecified) accuracy; up to 167× FPS speedup[10]	Attention+GRU yields high accuracy and much faster inference[10]; tested on diverse datasets.
Sánchez-Caballero *et al.* (2023)【147†】	ConvLSTM on depth video (stateless vs stateful)	NTU RGB+D (Depth only)	75.3% (stateless), 80.4% (stateful)[11]	Stateful RNN significantly improves depth-based HAR[11]. Only depth used (privacy), but lower accuracy; slower stateful model.
Uddin *et al.* (2024)【148†】	CNN, ConvLSTM, LRCN (CNN+LSTM)	UCF50, HMDB51	CNN: 99.6% (UCF50), 92.7% (HMDB51); ConvLSTM: 82%, 68%; LRCN: 93.4%, 71.6%[12]	Simple CNN achieved highest accuracy[12]. Findings suggest raw frame CNN can outperform sequential models on these benchmarks.
Chen *et al.* (2023)【149†】	CNN-LSTM	Baduanjin exercise videos (18 subjects)	96.43% accuracy[17]	CNN features + LSTM can model specific exercise actions well[17]. However, dataset is small and specialized.

Authors (Year)	Method	Dataset(s)	Performance	Key Findings & Limitations
Alzahrani *et al.* (2025)【150†】	YOLO-Act (3D keyframe YOLO + classifier)	AVA v2.2 (spatiotemporal detection)	73.28% mAP[18] (28.18 mAP ↑ vs prior)	Unified detection+classification yields strong AVA results[18]. Multi-person setting; not purely single-person classification.
Zhang & Abdelmunim (2025)【81†】	YOLOv7 + LSTM	UCF101, KTH, Weizmann, IXMAS	≈96–100% accuracy[4]	Near-perfect accuracy on standard datasets[4]. Limited challenge: real-world generalization not assessed.
Sun *et al.* (2024)【90†】	k-NN Attention + ViViT	UCF101, HMDB51	"Superior accuracy" over SOTA (not quantified)[15]	k-NN attention improves transformer learning[15]. High model complexity; exact gains not detailed.
Al-Tawil *et al.* (2025)【101†】	ResNet-18 + BiLSTM + multi-head attention + optical-flow frame selection	UCF101	96.60% accuracy[3] (222 FPS)	Motion-based frame prioritization + attention yields high accuracy at real-time speeds[3]. Tested only on one dataset.
Qin *et al.* (2025)【107†】	Video Swin Transformer + optical flow + ConvLSTM	UCF101	92.8% top-1 (vs 89.6% baseline)[16]	Adding optical flow and ConvLSTM boosts ViT model by ~3%[16]. Moderate improvement on one dataset; model is complex.
Khan & Jung (2023)【134†】	Dilated CNN + LSTM	UCF50	94.9% accuracy (vs 93.7% baseline CNN-LSTM)[1]	Dilated convs expanded receptive field, slightly improving HAR accuracy[1]. Evaluated on a single dataset only.
Baseline: Carreira & Zisserman	Two-stream Inflated 3D CNN (I3D)	HMDB-51, UCF-101	80.2% (HMDB), 97.9% (UCF101)[6]	Pre-training on large Kinetics improved performance

Authors (Year)	Method	Dataset(s)	Performance	Key Findings & Limitations
(2017)【164†】		(pretrained on Kinetics)		markedly[6]. Large model; predates latest attention models.
Baseline: Feichtenhofer *et al.* (2019)【166†】	SlowFast two-pathway CNN	Kinetics, Charades, AVA	State-of-art on Kinetics/Charades[8]	Factorizing slow/fast paths boosts action recognition accuracy[8]. Complex architecture.

Each entry cites the original source for performance metrics and key claims[9][3], ensuring attribution.

IV. Identified Research Gaps

Comparing the above studies reveals common limitations. Most models achieve very high accuracy on standard benchmarks, but often rely on **small, curated datasets** (KTH, Weizmann, UCF101) or assume controlled conditions. For example, the YOLO-LSTM approach reports 100% on Weizmann[4], which is easy and has few classes. Real-world scenes with occlusion, viewpoint change, and complex backgrounds are under-explored. Many works also require substantial computation: ViViT and Swin-Transformer methods have large model sizes, making real-time or embedded deployment hard[16][8]. Data scarcity is another issue – studies often test on 2–5 datasets, and rarely on newer large datasets like HMDB/HMDB112 or Kinetics beyond pretraining. There is little work on **fine-grained** actions or long-duration activities (Charades, etc.). Finally, multi-modal inputs (optical flow, skeleton) are sometimes used (e.g. flow in Qin et al. or depth in Sánchez-Caballero) but few studies fuse modalities effectively. In summary, gaps include: **limited generalization**, **heavy models**, and **lack of diversity in data**.

V. Future Scope

Based on these gaps, future research should explore several directions.

(1) **Larger and more diverse datasets:** Creating or using more realistic single-person HAR datasets (e.g. daily routines, surveillance footage) will test model robustness.

(2) **Lightweight architectures:** Designing efficient networks (e.g. MobileNet, efficient transformers) could bring HAR to edge devices without large GPUs.

(3) **Data augmentation and self-supervision:** Techniques like DINO and semi-supervised learning (as in Surek et al.) can help train on limited video data[9].

(4) **Multi-modal fusion:** Combining RGB with depth, optical flow, or inertial cues could improve accuracy under challenging conditions, provided fusion is done efficiently.

(5) **Temporal reasoning:** Long-term action contexts and continuity (beyond clip-based classification) remain open; recurrent or memory-augmented models could capture extended activities. Overall, advancing generalization and efficiency while leveraging richer data will be key directions.

VI. Conclusion

This review compared 14 recent vision-only, single-person HAR studies. We found that state-of-the-art models (often CNN/RNN or transformer-based) achieve high accuracy on standard benchmarks[9][4]. Attention mechanisms and hybrid architectures consistently improve performance. However, common limitations include over-reliance on limited datasets, computational cost, and moderate generalization to complex scenes. Recognizing these gaps, we highlight the need for more diverse data and efficient models. This comparative analysis clarifies the current landscape of vision-based HAR and guides future efforts to build more robust and practical HAR systems.

References

[1] G. A. S. Surek *et al.*, "Video-Based Human Activity Recognition Using Deep Learning Approaches," *Sensors*, vol. 23, no. 14, p. 6384, 2023[9].

[2] H. Ullah and A. Munir, "Human Activity Recognition Using Cascaded Dual Attention CNN and Bi-Directional GRU Framework," *J. Imaging*, vol. 9, no. 3, p. 167, 2023[10].

[3] A. Sánchez-Caballero, D. Fuentes-Jiménez, and C. Losada-Gutiérrez, "Real-time human action recognition using raw depth video-based recurrent neural networks," *Multimedia Tools Appl.*, vol. 82, pp. 16213–16235, 2023[11].

[4] M. A. Uddin *et al.*, "Deep learning-based human activity recognition using CNN, ConvLSTM, and LRCN," *Int. J. Cogn. Comput. Eng.*, vol. 5, pp. 259–268, 2024[12].

[5] J. Chen *et al.*, "CNN-LSTM Model for Recognizing Video-Recorded Actions Performed in a Traditional Chinese Exercise," *IEEE J. Transl. Eng. Health Med.*, vol. 11, 2023[17].

[6] N. Alzahrani, O. Bchir, and M. M. B. Ismail, "YOLO-Act: Unified Spatiotemporal Detection of Human Actions Across Multi-Frame Sequences," *Sensors*, vol. 25, no. 10, p. 3013, 2025[18].

[7] M. Elnady and H. E. Abdelmunim, "A novel YOLO LSTM approach for enhanced human action recognition in video sequences," *Sci. Rep.*, vol. 15, p. 17036, 2025[4].

[8] W. Sun, Y. Ma, and R. Wang, "k-NN attention-based Video Vision Transformer for action recognition," *Neurocomputing*, vol. 574, pp. 110–122, 2024[15].

[9] B. Al-Tawil *et al.*, "Multi-Head Attention-Based Framework with Residual Network for Human Action Recognition," *Sensors*, vol. 25, no. 9, p. 2930, 2025[3].

[10] J. Qin *et al.*, "Video Swin-CLSTM transformer: Enhancing human action recognition with optical flow and long-term dependencies," *PLoS ONE*, vol. 20, no. 7, e0327717, 2025[16].

[11] B. A. Khan and J. Jung, "Deep Learning-Based HAR Using Dilated CNN and LSTM on Video Sequences of Various Actions Dataset," *Appl. Sci.*, vol. 15, no. 22, p. 12173, 2023[1].

[12] G. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," in *Proc. CVPR*, 2017, pp. 6299–6308[6].

[13] C. Feichtenhofer *et al.*, "SlowFast Networks for Video Recognition," in *Proc. ICCV*, 2019, pp. 6202–6211[8].

[1] Deep Learning-Based Human Activity Recognition Using Dilated CNN and LSTM on Video Sequences of Various Actions Dataset

https://www.mdpi.com/2076-3417/15/22/12173

[2] [14] [18] YOLO-Act: Unified Spatiotemporal Detection of Human Actions Across Multi-Frame Sequences

https://www.mdpi.com/1424-8220/25/10/3013

[3] Multi-Head Attention-Based Framework with Residual Network for Human Action Recognition

https://www.mdpi.com/1424-8220/25/9/2930

[4] A novel YOLO LSTM approach for enhanced human action recognition in video sequences | Scientific Reports

https://www.nature.com/articles/s41598-025-01898-z?error=cookies_not_supported&code=2da2af87-3e45-482c-80a1-78e4dd198f76

[5] [11] Real-time human action recognition using raw depth video-based recurrent neural networks | Multimedia Tools and Applications

https://link.springer.com/article/10.1007/s11042-022-14075-5

[6] Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset

https://openaccess.thecvf.com/content_cvpr_2017/papers/Carreira_Quo_Vadis_Action_CVPR_2017_paper.pdf

[7] [16] Video swin-CLSTM transformer: Enhancing human action recognition with optical flow and long-term dependencies | PLOS One

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0327717

[8] SlowFast Networks for Video Recognition

https://openaccess.thecvf.com/content_ICCV_2019/papers/Feichtenhofer_SlowFast_Networks_for_Video_Recognition_ICCV_2019_paper.pdf

[9] Video-Based Human Activity Recognition Using Deep Learning Approaches

https://www.mdpi.com/1424-8220/23/14/6384

[10] Human Activity Recognition Using Cascaded Dual Attention CNN and Bi-Directional GRU Framework - PubMed

https://pubmed.ncbi.nlm.nih.gov/37504807/

[12] Deep learning-based human activity recognition using CNN, ConvLSTM, and LRCN - Torrens University Australia

https://research.torrens.edu.au/en/publications/deep-learning-based-human-activity-recognition-using-cnn-convlstm/

[13] [17] CNN-LSTM Model for Recognizing Video-Recorded Actions Performed in a Traditional Chinese Exercise - PubMed

https://pubmed.ncbi.nlm.nih.gov/37435544/

[15] k-NN attention-based video vision transformer for action recognition - ScienceDirect

https://www.sciencedirect.com/science/article/pii/S0925231224000274