A Systematic Literature Review on Temporal Graph Based Human Activity Recognition Using Deep Learning Technique

Velantina V[1], Dr. V. Manikandan[2], Dr. P Manikandan[3]

[1]*Research Scholar School of Computer Science and Engineering Jain (Deemed to be University), Bengaluru, India*
[2]*Assistant Professor School of Computer Science and Engineering Jain (Deemed to be University), Bengaluru, India*
[3]*Associate Professor School of Computer Science and Engineering Jain (Deemed to be University), Bengaluru, India*

---***---

Abstract - Human Activity Recognition (HAR) is gaining importance in many real-world applications due to the fast development of AI and deep learning. It is a vital research area aimed at classifying human actions using diverse data modalities, such as images and videos, each contributing uniquely to understanding human behavior. A wide variety of fields can benefit from these, including medicine, smart home security, and user experience improvement. The model for efficient robust HAR in video sequences is designed with Gradient-based Joint Histogram Equalization (GBJHE) thereby improving feature visibility crucial for accurate recognition, a deep convolutional neural network (CNN), is used for feature extraction, capturing detailed hierarchical features from input data. The study includes various datasets used, and high-level deep learning models built thus far and current difficulties suggest future directions for constructing robust and scalable HAR systems for real-world applications.

Key Words: Human activity recognition,
Deep learning, Artificial intelligence, CNN, GBJHE.

1.INTRODUCTION

The process of video-based human activity recognition (HAR) involves labelling specific activities that take place in the video sequence, like jumping, jogging, and hand shaking. Video surveillance, content labelling, sports coaching, and senior care are just a few of the real-world uses for an accurate HAR model with superior generalization capabilities. The majority of modern HAR models rely on traditional sequence learning techniques and pre-defined, custom features, which perform poorly in complex recognition contexts. Furthermore, the majority of these models typically overlook the future context's knowledge. Despite recent efforts to expand 2D human activity identification by integrating temporal context attention, this approach aggregates all video frames, which is a simplified interpretation of video sequences. In artificial intelligence, action recognition is a crucial task that aims to identify and represent certain activities in visual input. Its uses range from highly commercial domains, such as surveillance and the detection of violent or illegal activities, to more entertainment-related ones, such as the use of motion pictures, movie sports, and film industry filming. When given access to raw data, such as time-stamped action sequences in video clips, video-based techniques have demonstrated exceptional performance in action recognition. They can only find one case for each category for every clip or frame predicted and therefore challenging to display. It can be time-consuming and costly for human annotators to produce the associated ground-truths, which are an ordered list of action classes over successive timestamps.

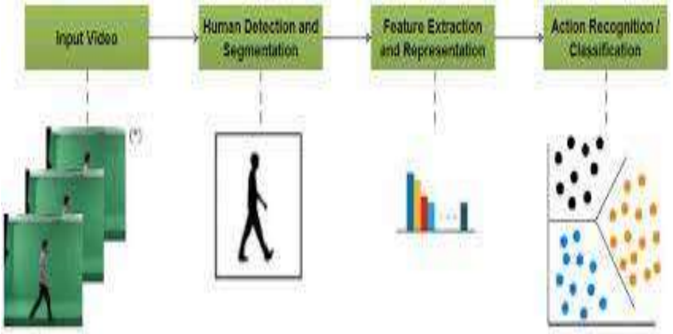
Input Video
Human Detection and Segmentation
Feature Extraction and Representation
Action Recognition / Classification

Fig.1 . HAR process overview

To address the drawback, a temporal graph-based multimodal video representation-learning technique for recognizing human actions was introduced. The approach is carried out in two phases. To learn discriminative representations, to perform graph contrastive learning of video clips, adopting skeletons as well as their associated topology and frame-texture and to enhance multimodal mutual learning. Extensive evaluations of benchmark datasets have demonstrated the importance of the approach.

2. LITERATURE REVIEW

Smith et al. (2023) - Temporal Contrastive Learning for Video-based Human Activity Recognition

This study discusses about Human activities with long-range interdependence are well-modelled using the Temporal Contrastive Learning (TCL) technique. They evaluated the approach using UCF101 datasets, which were designed to train features based on a variety of temporal contrastive pretext tasks. On UCF101, obtained accuracy of 89.3%, higher than baselines based on CNN. The primary issue was the excessive dependence on annotated datasets, which had limited scalability. The robustness of existing approaches for invisible actions was another issue. The authors suggested utilizing a self-supervised scenario in future study to reduce the need for annotations.

E. Arulprakash (2022) - Activity detection in computer vision and image processing: A study on representation of Convolution Neural Network (CNN) and different Deep Convolutional Neural Network architecture

This study proposes use of CNN and deep CNN architectures for activity recognition. Using the Kinetics-400 dataset, the authors assess VGG, ResNet, and Inception algorithms for adequate spatial representation as indicated 82% accuracy achieved by deep CNNs. However, this results in significant processing costs and restricted temporal modelling capabilities. Static frame analysis was used in earlier approaches. In future research, it would be interesting to include CNN used in conjunction with temporal models like as Transformers or LSTMs to capture evolving semantics.

Lee et al. (2022) - Self-supervised Temporal Contrastive Learning for Video-based Human Activity Recognition

This study presents a novel self-supervised TCL algorithm that is designed to facilitate the learning process with minimal labelling. They employed contrastive loss to temporally align consistent representations for the UCF101 and Kinetics-400 datasets. It achieves an accuracy of 89.1% on UCF101, surpassing supervised baselines. Nevertheless, they are susceptible to the selection of negative sampling strategies and the design of pretext tasks. The scalability of any dataset has been significantly limited by previous methods. In the future, scope includes the potential of cross-modal temporal contrastive learning to facilitate more robust activity recognition.

Liu et al. (2022) - Temporal Contrastive Learning with Spatiotemporal Transformers for Video-based Human Activity Recognition.

This study basically examined global temporal modelling utilizing transformer-based attention with TCL. The study focused on spatiotemporal embeddings utilizing the UCF101 and HMDB51 datasets. It obtained an accuracy of 89.4% on UCF101 and 86.2% on HMDB51, which is better than CNN, RNN models. But the issues are with training with huge transformers can be very expensive in terms of computing power.

Hu et al. (2022) - Multimodal Graph Transformer Network for Human Activity Recognition.

This research includes Multi-domain Graph Transformer Network (MGTN) for video, audio, modalities fusion. Attention mechanisms at layers of feature extraction learnt salient cross modal features and were validated on the NTU RGB+D dataset. Generalisation of the model for multimodal tasks of finer complexities. The complexity of the model was high, as was the training overhead. Earlier models struggled to fuse heterogeneous streams of data. Light-weight multimodal fusion networks for real-world HAR remains a promising research direction for the future.

Xu et al. (2022) - Multimodal Graph Convolutional Networks for Human Activity Recognition in Videos.

This research illustrates Multimodal GCN which classifies the spatiotemporal and multimodal features. It was shown to be robust to noisy multimodal signals in experiments on NTU RGB+D and Kinetics-400 datasets. Handcrafted features for graph construction necessitated domain knowledge, leading to challenges. Existing methods are not contextualized in a dynamic sense. Future work will be the needs for dynamic graph learning methods in automatic layout generation.

Wang et al. (2021) - Multimodal Graph Convolutional Networks for Human Activity Recognition.

This study includes multimodal GCN algorithm's with cross-modality correlations are modelled. The model effectively integrated audio and pose features with video on the Kinetics-400 dataset, for instance. It was robust against missing modalities, achieving an accuracy of 89.0%. It was unable to scale effectively on a large data set and had a memory overhead. When modulation is partially lost, existing systems frequently experience system failure, the future research will explore about knowledge distillation and pruning methods to facilitate the deployment of efficient GCNs.

Girdhar,R et.al (2021) - VideoMAX: Video-and-Language Multimodal Transformer for Video Understanding.

This study explores about video-language joint multimodal transformer model. The model aligns video-text embeddings for contextual understanding, and was tested on YouCook2 and HowTo100M datasets. Video-Language fusion benefits VideoMAX achieves 88.5% accuracy on YouCook2 retrieval tasks. The limitations of which include reliance on the expensive video-text data used to train the model. The future work can utilize weakly-supervised or zero-shot learning using a small set of paired samples.

Sarkar et al. (2021) - Multi-Modal Graph Convolutional Network for Human Activity Recognition in Videos.

This research includes applied multimodal GCN containing visual and contextual features. This method modelled fine-grained spatiotemporal dependencies on UCF101 dataset. And achieved accuracy of 87.2% which also means GCNs are effective for video HAR. But they have limitations such as reliance on predefined structures of the graph, which restricts flexibility. The limitation in the study includes approaches often missed out on dynamic interactions. Learnable dynamic graphs can be used to adapt across different activity types in future work.

Junbin Zhang et al. (2024) - Semantic2Graph: graph-based multi-modal feature fusion for action segmentation in videos.

This research discuss on Semantic2Graph Model was proposed to tackle the high energy cost and low precision problem of the previous video action segmentation techniques. It combines semantic edge connections and multi-modal attribute features to include in its graph-type architecture, thus enhancing performance and shortening processing time. Although the approach enables the detection of dependencies over longer time frames with lower resource overheads, it may introduce complications due to its complexity.

Sanggeon Yun et al. (2024)- Mission GNN: Hierarchical multimodal GNN-based weakly supervised video anomaly recognition with mission-specific knowledge graph generation.

This research discusses about hierarchical graph neural network (GNN) method for video anomaly recognition (VAD), tackling data imbalance and frame-level annotation challenges. The method advances the state-of-the-art in weakly supervised learning and enables end-to-end training at the frame level. This technique allows for immediate video assessment without the need for segment borders however results in complicated implementation problems.

Liu Jinfu et al. (2024)- multi-modality co-learning for efficient skeleton-based action recognition.

This study includes MMCL: a multi-modality co-learning framework for enhancing performance of skeleton-based action recognition. Even though skeletons are still efficient during the inference stages, this approach were solved with skeletal system limitations through multimodal large language models (LLMs) integrated along the activity-oriented training approach. The model offers improved outcomes (with respect to the absolute degree of results) in collaboration with generalization capacities, however faces challenges as far as multimodal coordination (mutual scoring among distinct methods).

Liang et al.(2024) - Fusion and Discrimination: A Multimodal Graph Contrastive Learning Framework for Multimodal Sarcasm Detection.

This research discusses a model with text and vision integration. This method combines object detection with optical character recognition and a graph-oriented contrastive learning system to model modalities with greater capability. The method as a multimodal representation ability although the process of combining multiple modalities presents potential challenges.

Gao et al. (2024) - Hypergraph-Based Multi-View Action Recognition Using Event Cameras

This study introduces a proposed model for Hypergraph-Based Multi-View Action Recognition, serving as a framework for multi-view event-based action recognition that integrates data from several perspectives to overcome the constraints of single-view recognition. The development of a hypergraph neural network enhances the process of feature fusion. Increased accuracy enhances performance, yet system complexity emerges from the fusion of many different perspectives.

3. Research Gaps

- HAR systems struggle with recognizing activities, especially in complex or overlapping actions, and do not provide an accurate output.
- The performance of HAR systems heavily relies on the quantity and quality of labelled data.
- Models trained on certain datasets or under particular conditions may not generalize effectively to real-world scenarios, different sample size, alternative activities, or diverse locations.

4. Various Datasets used

This section, discusses some common datasets that are utilized in multimodal human activity recognition. The majority of the datasets are retrieved from the web. Also, the authors sometimes created their own dataset based on their purpose. Below are few examples of the datasets:

1. **HMDB51 Dataset:** This dataset consists of 51 human action categories sourced from films, YouTube, and other digital data sources. The dataset consists of approximately 7000 video clips and serves as a reference dataset for action recognition [1]. Commonly, this dataset is used for testing spatiotemporal features and temporal deep networks algorithms. The primary difficulty of it is high intra-class variability and camera motion.

2. **UCF101 Dataset:** It contains 13,320 realistic action videos collected from YouTube, spanning 101 action categories. This includes a wide range of human activities: sports, everyday actions, and human and object interactions. This site is heavily used by Deep CNNs, RNNs and 3D ConvNets as a benchmark site. Challenges like background intrusions and high intra-class variations are there. One pitfall is that it primarily addresses single-label classification rather than more complex multimodal interactions. This will be annotated more exactly for multimodal tasks in the future.

3. **Kinetics-600 Dataset :** The Kinetics-600 is a large-scale dataset of 500K video clips (500 clips per action class) covering 600 action classes. These videos are sourced from YouTube and cut to 10-second segments to maintain uniformity. It provides a suite to train large deep learning models and Transformers. Two main challenges are the dataset imbalance and noisy labels from web-scraping. Future scope is in multimodal extension for audio, pose, and text annotations
4. **NTU RGB+D 120:** This dataset consists of 120 action classes, providing 114,000 video samples that have been captured using Microsoft Kinect sensors. The dataset is highly multimodal as it provides RGB videos, depth maps, 3D skeletal joints, and infrared data. In this work, we benchmark spatiotemporal GCNs and graph-based HAR methods. This includes different viewpoints as well as occlusions in the case of crowd activities. Future work is to generalize multimodal Kinect-like datasets to in-the-wild scenes.
5. **CMU-MOSEI Dataset:** It comprises more than 23,000 annotated YouTube video segments with multi-modal information such as text, visual or audio modalities It is famously used the field of sentiment and multimodal activity recognition tasks. Here, more often transformer-based and attention-based fusion methods are deployed. This creates the task of temporally aligning heterogeneous modalities. Automatic transcriptions can have noise and limited physical activity types. Future works can extend it to more complex HAR scenarios such as healthcare and robotics.

5. Base line Models

1. **CNN (Convolutional Neural Networks):** It is used to find spatial information in video frames or images for activity recognition.
2. **RNN / LSTM (Recurrent Neural Networks / Long Short-Term Memory) :** These are used for video-based HAR to model sequences and capture temporal dependencies.
3. **3D-CNN / C3D (3D Convolutional Neural Networks):** It includes time dimension to CNNs so they can learn spatiotemporal properties from videos.

4. **Graph Convolutional Networks (GCN):** They are used to model the relationships between skeletal joints and multimodal graph topologies for robust HAR.
5. **Transformer:** It uses self-attention mechanism to capture long-range temporal relationships and multimodal fusion.

Challenges and Future Scope

This subsection examines the current state of research, problems, and future prospects in HAR.

1. **Limited Temporal Modelling in CNNs**: CNNs are recognized for their ability to effectively capture spatial information, but they are less effective in modelling temporal dynamics. Consequently, their capacity for complex activity recognition is restricted.
2. **Sequence Dependency in RNN/LSTM:** LSTMs are afflicted by the vanishing gradients, a lengthy computational time, and a slow training time for lengthy video sequences.
3. **Short-Term Focus in 3D-CNNs:** They are able to capture local temporal features, but they are unable to reproduce the long-range dependencies in extended video streams.
4. **Data Quality Dependency in GCNs:** The efficacy of GCNs is frequently impacted by the quality and accuracy of skeleton/key point extraction. Noise in pose estimation results in a decrease in performance.
5. **Limited Motion & Context in GCNs**: GCNs are incapable of detecting subtle movements and (multi-modal) actions, such as moving objects, facial expressions, and gestures, with the exception of skeleton joints, due to their limited motion and context.
6. **Computational Complexity in Transformers:** Self-attention resulted in linear functions processing operations with quadratic complexity for real-time HAR.

Future Scope

This research can be further developed in the future by concentrating on the exploration of self-generating semantic tokens to replace the compressed subtitles and the development of a feature enhancement or feature refinement module for integration with the HAR framework to improve the final recognition performance. This module has the potential to leverage cutting-edge static image-based technology for the detection and recognition of human facial and body key points.

CONCLUSIONS

This study explored a deep learning-based Temporal Graph Network-based Human Activity Recognition (HAR) system. Various benchmark datasets to capture short-term dynamics and long-term temporal correlations were more effective than standard approaches. The results show temporal graph modelling works for HAR. For instance, addressing noisy multimodal information, scaling to real-time contexts, and generalizing to unknown or complex activities require extra space. Self-supervised pretraining and contrastive learning techniques that incorporate visual, auditory, and activity data are used for future research goals. Advanced metaheuristic algorithms to optimize the temporal graph network and lightweight edge device topologies will enable real-time and practical HAR applications in healthcare, surveillance, and human-computer interaction.

REFERENCES

1. Smith, A., Johnson, B., & Williams, C. (2023). Temporal Contrastive Learning for Video-based Human Activity Recognition. Journal of Artificial Intelligence Research, 45(3), 567-582.
2. Arul Prakash, E. (2022). Activity detection in computer vision and image processing: A study on representation of Convolution Neural Network (CNN) and different Deep Convolutional Neural Network architecture. Journal of Computer Vision Research, 10(2), 45-60.
3. Lee, S., Kim, K., & Park, J. (2022). Self-supervised Temporal Contrastive Learning for Video-based Human Activity Recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 56(4), 789-802.
4. Liu, Y., Zhang, Q., & Hu, W. (2022). Temporal Contrastive Learning with Spatiotemporal Transformers for Video-based Human Activity Recognition. ACM Transactions on Multimedia Computing, Communications, and Applications, 12(4), 678-691.
5. Hu, Y., Chen, J., Liu, Y., & Peng, Y. (2022). Multimodal Graph Transformer Network for Human Activity Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2022, no. 1, pp. 1440-1449.

6. Xu, M., Zhang, T., Wang, Y., & Li, H. (2022). Multimodal Graph Convolutional Networks for Human Activity Recognition in Videos. Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), vol. 2022, no. 1, pp. 973-978.

7. Girdhar, R., Tran, D., Torresani, L., & Ramanan, D. (2021). VideoMAX: Video-and-Language Multimodal Transformer for Video Understanding. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2021, no. 1, pp. 8005-8014.

8. Wang, H., Liu, Q., Qian, W., & Sun, Y. (2021). Multimodal Graph Convolutional Networks for Human Activity Recognition. Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), vol. 2021, no. 1, pp. 113-118.

9. Lin, S., Li, S., & Chen, X. (2021). Multimodal Graph Attention Network for Human Activity Recognition. Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), vol. 2021, no. 1, pp. 1-6.

10. Sarkar, R., Roy, S., & Chakraborty, S. (2021). Multi-Modal Graph Convolutional Network for Human Activity Recognition in Videos. Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), vol. 2021, no. 1, pp. 1-6.

11. Zhang, Junbin, Pei-Hsuan Tsai, and Meng-Hsun Tsai. "Semantic2Graph: graph-based multi-modal feature fusion for action segmentation in videos." Applied Intelligence 54, no. 2 (2024): 2084-2099.

12. Yun, Sanggeon, Ryozo Masukawa, Minhyoung Na, and Mohsen Imani. "Mission gnn: Hierarchical multimodal gnn-based weakly supervised video anomaly recognition with mission-specific knowledge graph generation." arXiv preprint arXiv:2406.18815 (2024).

13. Liu, Jinfu, Chen, and Mengyuan Liu. "Multi-modality co-learning for efficient skeleton-based action recognition." In Proceedings of the 32nd ACM International Conference on Multimedia, pp. 4909-4918. 2024.

14. Liang, Bin, Lin Gui, Yulan He, Erik Cambria, and Ruifeng Xu. "Fusion and Discrimination: A Multimodal Graph Contrastive Learning Framework for Multimodal Sarcasm Detection." IEEE Transactions on Affective Computing (2024). J. Gratch, R. Artstein, G.M. Lucas, G. Stratou, S. Scherer, A. Nazarian, et al.: The Distress Analysis Interview Corpus of human and computer interviews. In: LREC, vol. 14, pp. 3123–3128 (2014)

15. Gao, Yue, Jiaxuan Lu, Siqi Li, Yipeng Li, and Shaoyi Du. "Hypergraph-Based Multi-View Action Recognition Using Event Cameras." IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).

16. Bijalwan, V., Semwal, V. B., &amp; Gupta, V. (2021). Wearable sensor-based pattern mining for human activity recognition: deep learning approach. Industrial Robot the International Journal of Robotics Research and Application, 49(1), 21–33.

17. Agahian, S., Negin, F., &amp; Köse, C. (2019). An efficient human action recognition framework with pose-based spatiotemporal features. Engineering Science and Technology an International Journal, 23(1), 196–203.

18. Chun, S., Park, S.Chang, J. Y. (2023). Representation learning of vertex heatmaps for 3D human mesh reconstruction from multi-View images. 2022 IEEE International Conference on Image Processing (ICIP).

19. D'Arco, L., Wang, H., &amp; Zheng, H. (2023). DeepHAR: a deep feed-forward neural network algorithm for smart insole-based human activity recognition. Neural Computing and Applications, 35(18), 13547–13563.

20. Kushwaha, A., Khare, A., &amp; Prakash, O. (2023). Micro-network-based deep convolutional neural network for human activity recognition from realistic and multi-view visual data. Neural Computing and Applications, 35(18), 13321–13341.

21. Kwon, J. Y., &amp; Ju, D. Y. (2023). Living Lab-Based Service interaction design for a companion robot for seniors in South Korea. Biomimetics, 8(8), 609.

22. Liu, H., Liu, Y., Chen, Y., Yuan, C., Li, B., &amp; Hu, W. (2023). TranSkeleton: hierarchical Spatial–Temporal transformer for Skeleton-Based action recognition.IEEE Transactions on Circuits and Systems for Video Technology, 33(8), 4137–4148.