Research Proposal: Integrated Motion Recognition and Generation Using Skeleton Data

Runze Zhao
The Chinese University of Hong Kong, Shenzhen
runzezhao@link.cuhk.edu.cn

1 INTRODUCTION

In the realm of digital technology, motion recognition and prediction are paramount, with wide-ranging applications in video surveillance [1], human-computer interaction [2], sports analysis [3], and pedestrian tracking [4]. The employment of 3D skeleton data sequences presents a sophisticated alternative to traditional video processing, effectively mitigating issues arising from background interference, lighting conditions, and camera angle variations.

Central to this research is the dual objective of enhancing action recognition and pioneering long-term motion prediction. Though existing deep neural network (DNN) structures like 3D Convolutional Neural Networks (3D-CNN), Graph Convolutional Networks (GCN), and transformer-based architectures have significantly advanced motion recognition, they encounter limitations in terms of computational efficiency, accuracy, and generalizability, particularly in predicting future actions from current motion data. This challenge is particularly pertinent in applications demanding rapid and accurate predictions, such as in remote-controlled robotics or dynamic interactive systems. Additionally, while many existing works utilize traditional statistical methods or DNN structures for predicting human motion [5], the focus has primarily been on short-term predictions. There is a significant gap in long-term forecasting capabilities, which is essential for applications extending beyond immediate future predictions.

This research proposes to bridge this gap by developing an integrated framework that enhances action recognition through advanced graph-based models and neural networks, and pioneers long-term motion prediction. Leveraging the latest AI advancements, this framework aims to intuitively understand and predict movements. The goal is to create a system that excels in both accurate recognition and future action forecasting, paving the way for breakthroughs in remote robotic surgery, autonomous vehicles, and interactive AI systems.

2 PROBLEM STATEMENT

Motivation:

1. Despite significant advancements in action recognition, the integration of this technology with motion generation in a unified system remains underexplored. Such an integrated approach could revolutionize medical and surgical robotics, leading to the development of sophisticated medical surgery assistants and versatile everyday robotic aides. This research aims to explore and capitalize on this potential, addressing a critical gap in the field.
2. Current research in action recognition predominantly focuses on single-modal inputs. However, the success of large language models like ChatGPT and Google Gemini in handling multimodal inputs suggests a promising direction for robotic systems. This project proposes the development of a system that leverages multimodal inputs (voice, ambient sounds, video, etc.) to enhance the effectiveness of robotic assistants in diverse environments.

Research Questions:

1. How can a multitasking framework be designed to effectively utilize multimodal inputs for simultaneous motion prediction and action recognition, particularly in robotic applications?
2. What are the key factors in building a robust and computationally efficient framework for action recognition and prediction, and how can these be implemented in practice?

3 LITERATURE REVIEW

3.1 Motion Recognition

CNN and RNN based action recognition models. Traditional algorithms in skeleton-based action recognition have primarily utilized hand-crafted feature-based methods. Notable examples include [6], which employed covariance matrices of joint trajectories, and [7], focusing on the dynamics of joints' relative positions. Recent advances mostly utilize deep learning frameworks,particularly Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), due to their effectiveness in handling complex data structures. For instance, [8, 9] demonstrated the conversion of skeleton sequence coordinates into pseudo-images for CNN processing. Other noteworthy developments include the use of 3D-CNNs, as evidenced by the stacking of pseudo-images of distance matrices [10, 11] and the integration of 3D skeletons into cuboids [12]. Additionally, [13] proposed an innovative two-stream 3D CNN framework using 3D heatmap volumes to represent skeleton data. RNN-based models have also been prominent, with [14, 15] effectively capturing temporal dependencies between consecutive frames.

Graph-based models. The optimization of graph construction strategies for extracting spatial and temporal information has been a significant focus in recent research, exemplified by [16]'s use of Graph Convolutional Networks (GCNs). Subsequent studies have enhanced GCNs through methods like adjacency powering for multi-scale modeling [17, 18] and integrating self-attention mechanisms [19, 20]. A novel approach was introduced by [21], representing the human body as a hypergraph in a semi-dynamic hypergraph neural network, offering richer information capture compared to traditional GCNs. This concept was further expanded by [22], which captured both spatio-temporal information and higher-order dependencies, thus advancing skeleton-based action recognition.

Transformer-based models. The introduction of transformer-based models in action recognition has marked a significant advancement in the field. These models, such as the self-supervised video transformer by [23], utilize self-attention mechanisms to

align features from different views. [24] also employed transformer self-attention mechanisms, processing both spatial and temporal dimensions. Predominantly, these transformer-based methods have been applied to video frames as input tokens [25, 26], demonstrating their adaptability in processing complex data structures.

3.2 Motion Prediction

The domain of motion prediction has seen a predominance of Deep Neural Network (DNN) based frameworks. Pioneering work by [27] introduced two frameworks: the Encoder-Recurrent-Decoder (ERD) and a LSTM-based recurrent neural network (LSTM-3LR). [28] developed a scalable, jointly trainable stacked RNN based on LSTMs (SRNN), which marked a significant advancement in the field. Further, [29] adopted a convolutional sequence-to-sequence network for long-term motion prediction. [30]'s Skeleton Temporal Network (Skel-TNet) focused on learning spatial and temporal dependencies for human motion prediction. Notably, [31] explored the use of Generative Adversarial Networks (GANs) to learn the joint distribution of body poses and global motion, demonstrating the capability to hypothesize large sections of the input 3D tensor with missing data.

4 METHODOLOGY

4.1 Dataset

The experiments will be conducted using established datasets, outlined as follows:

1. **NTU-RGB+D** [32]: This comprehensive dataset includes 56,880 video sequences, featuring up to two subjects per sequence and capturing 25 joint skeletons. Evaluation protocols are (i) Cross-Subject (X-Sub) and (ii) Cross-View (X-View), providing diverse and challenging testing conditions.
2. **NTU RGB+D 120 (NTU-120)** [33]: An extension of NTU-RGB+D, NTU-120 encompasses 120 action classes across 114,480 RGB+D video samples. This dataset is characterized by its large-scale subject variety and multiple camera viewpoints, following two evaluation protocols: (i) Cross-Subject (X-Sub) and (ii) Cross-Setup (X-Set).
3. **Human 3.6M (H3.6M)** [34]: A significant motion capture dataset, H3.6M features seven subjects performing 15 action classes, represented through 32 joints. For our analysis, we downsample sequences by half and focus on training with six subjects, using specific clips of the 5th subject for testing. Joint locations are converted from angle space to exponential maps, emphasizing the 21 joints with substantial data.
4. **Northwestern-UCLA** [35]: This dataset, comprising 1,494 video clips from three Kinect cameras, offers multiple viewpoints for each of the 10 actions performed by 10 subjects. We adhere to the original evaluation protocol as mentioned in [35], training on footage from the first two cameras and testing on clips from the third.

4.2 Research Route

1. **Development of a Basic Motion Recognition-Prediction Framework:** This framework consists of two primary modules:
   - *Recognition Module:* Designed for motion class prediction, utilizing advanced algorithms to accurately classify movements.
   - *Prediction Module:* Focused on forecasting the movement trajectory of a virtual robotic arm. This module incorporates environmental constraints like table boundaries and surface avoidance to ensure operational feasibility and safety.

   I have initially trained a toy model using self-recorded videos through an LSTM network. The preliminary outcomes of this training can be viewed at [here].
2. **Comprehensive Literature Review:** Our review will cover:
   - The application of 3D-CNN, GCN, and transformers in action recognition using skeleton data.
   - A detailed analysis of both short-term and long-term motion prediction algorithms for skeleton sequences, such as RNN, diffusion models, and transformers.
   - Examination of transformer models proficient in processing multimodal inputs.
3. **Designing an Advanced Engineering Framework for Multimodal input:** This system will enable a robotic arm to plan trajectories based on multimodal input, such as environmental cues and verbal commands.

This approach aims to integrate and advance the fields of robotics and artificial intelligence, creating a robust framework for innovative human-machine interactions.

5 SIGNIFICANCE

Advancing Human-Machine Interaction and Robotic Assistance: This research marks a pivotal advancement in human-machine interaction, particularly in medical and surgical robotics. By integrating motion recognition and prediction with multimodal input, it sets a new paradigm in robotic assistance, offering significant benefits in healthcare and daily living.

Setting New Benchmarks in AI and Predictive Modeling: The project aims to overcome current computational efficiency and accuracy limitations in AI and predictive modeling. By applying neural network architectures inspired by large language models, it seeks to enhance the anticipation and understanding of human behavior. The implications extend beyond medical robotics to areas such as autonomous vehicles, sports analytics, and interactive entertainment, heralding a new era of technologically advanced, safe, and efficient environments.

6 FUTURE WORK

In the next phase of this research, I aim to integrate and leverage the capabilities of fine-tuned Large Language Model (LLM) platforms to enhance our motion recognition and prediction framework.This future work will build upon our existing research in motion recognition, utilizing the strengths of LLMs in understanding and generating human-like responses to enhance the interaction between robotic systems and their human counterparts. The expected outcome is a more intuitive and responsive system, capable of understanding complex commands and executing tasks with greater autonomy.

7 REFERENCES

[1] Utkarsh Gaur, Yingying Zhu, Bi Song, and A Roy-Chowdhury. A "string of feature graphs" model for recognition of complex activities in natural videos. In *2011 International conference on computer vision*, pages 2595–2602. IEEE, 2011.

[2] Liang-Yan Gui, Kevin Zhang, Yu-Xiong Wang, Xiaodan Liang, José MF Moura, and Manuela Veloso. Teaching robots to predict human motion. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 562–567. IEEE, 2018.

[3] Magnus Ibh, Stella Grasshof, Dan Witzner, and Pascal Madeleine. Tempose: A new skeleton-based transformer model designed for fine-grained motion recognition in badminton. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5198–5207, 2023.

[4] De-An Huang and Kris M Kitani. Action-reaction: Forecasting the dynamics of human interaction. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*, pages 489–504. Springer, 2014.

[5] Rui Zhao and Qiang Ji. An adversarial hierarchical hidden markov model for human pose modeling and generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[6] Mohamed E Hussein, Marwan Torki, Mohammad A Gowayyed, and Motaz El-Saban. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In *Twenty-third international joint conference on artificial intelligence*, 2013.

[7] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1290–1297. IEEE, 2012.

[8] Carlos Caetano, Jessica Sena, François Brémond, Jefersson A Dos Santos, and William Robson Schwartz. Skelemotion: A new representation of skeleton joint sequences based on motion information for 3d action recognition. In *2019 16th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–8. IEEE, 2019.

[9] Diogo C Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5137–5146, 2018.

[10] Zeyi Lin, Wei Zhang, Xiaoming Deng, Cuixia Ma, and Hongan Wang. Image-based pose representation for action recognition and hand gesture recognition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 532–539. IEEE, 2020.

[11] Alejandro Hernandez Ruiz, Lorenzo Porzi, Samuel Rota Bulò, and Francesc Moreno-Noguer. 3d cnns on distance matrices for human action recognition. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1087–1095, 2017.

[12] Hong Liu, Juanhui Tu, and Mengyuan Liu. Two-stream 3d convolutional neural network for skeleton-based action recognition. *arXiv preprint arXiv:1705.08106*, 2017.

[13] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2969–2978, 2022.

[14] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1110–1118, 2015.

[15] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 816–833. Springer, 2016.

[16] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[17] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3595–3603, 2019.

[18] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 143–152, 2020.

[19] Bin Li, Xi Li, Zhongfei Zhang, and Fei Wu. Spatio-temporal graph routing for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8561–8568, 2019.

[20] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12026–12035, 2019.

[21] Shengyuan Liu, Pei Lv, Yuzhen Zhang, Jie Fu, Junjin Cheng, Wanqing Li, Bing Zhou, and Mingliang Xu. Semi-dynamic hypergraph neural network for 3d pose estimation. In *IJCAI*, pages 782–788, 2020.

[22] Xiaoke Hao, Jie Li, Yingchun Guo, Tao Jiang, and Ming Yu. Hypergraph neural network for skeleton-based action recognition. *IEEE Transactions on Image Processing*, 30:2263–2275, 2021.

[23] Kanchana Ranasinghe, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan, and Michael S Ryoo. Self-supervised video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2874–2884, 2022.

[24] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Skeleton-based action recognition via spatial and temporal transformer networks. *Computer Vision and Image Understanding*, 208:103219, 2021.

[25] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Yuanjie Shao, Zhengrong Zuo, Changxin Gao, and Nong Sang. Oadtr: Online action detection with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7565–7575, 2021.

[26] Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Xia, Zhuowen Tu, and Stefano Soatto. Long short-term transformer for online action detection. *Advances in Neural Information Processing Systems*, 34:1086–1099, 2021.

[27] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE international conference on computer vision*, pages 4346–4354, 2015.

[28] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 5308–5317, 2016.

[29] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5226–5234, 2018.

[30] Xiao Guo and Jongmoo Choi. Human motion prediction via learning local structure representations and temporal dependencies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2580–2587, 2019.

[31] Alejandro Hernandez, Jurgen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7134–7143, 2019.

[32] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

[33] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2684–2701, 2019.

[34] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013.

[35] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning and recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2649–2656, 2014.