HOIMotion: Forecasting Human Motion During Human-Object Interactions Using Egocentric 3D Object Bounding Boxes

Zhiming Hu[1], Zheming Yin[1], Daniel Häufle[3,4,2], Syn Schmitt[1,2], Andreas Bulling[1,2]

[1]University of Stuttgart

[2]Bionic Intelligence Tuebingen Stuttgart

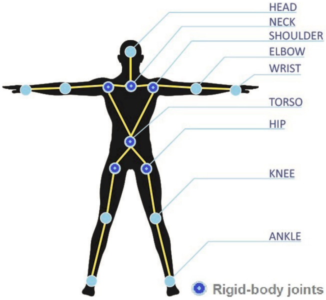
[3]Heidelberg University

[4]University of Tuebingen

Table of Contents

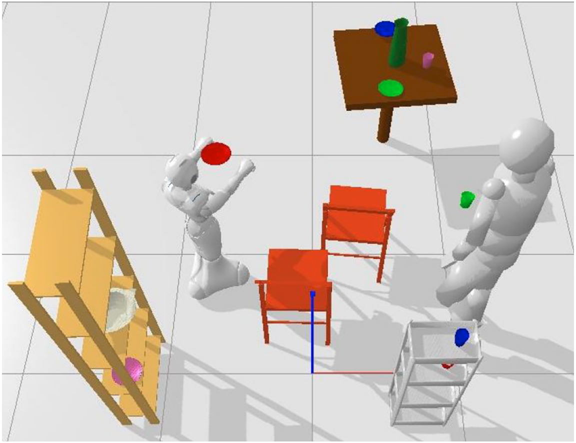
Research Background

- **Human pose:** 3D positions of human joints (e.g. wrist, elbow, shoulder, knee, ankle)
- **Motion forecasting:** predict future human poses from historical poses



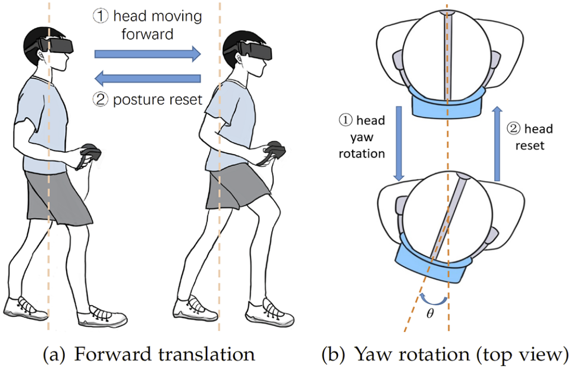
Human pose
[Alexiadis TCSVT'16]

Applications of human motion forecasting



Human-agent collaboration
[Le RHIC'21]

Applications of human motion forecasting

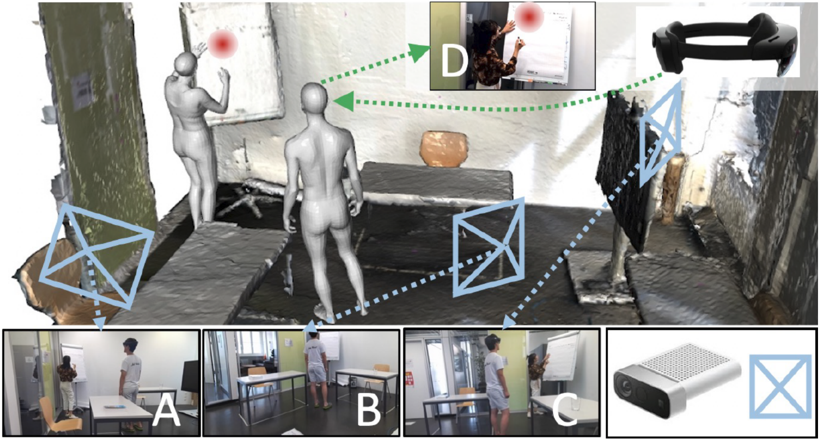


Redirected walking in XR environments
[Lin TVCG'22]

Applications of human motion forecasting

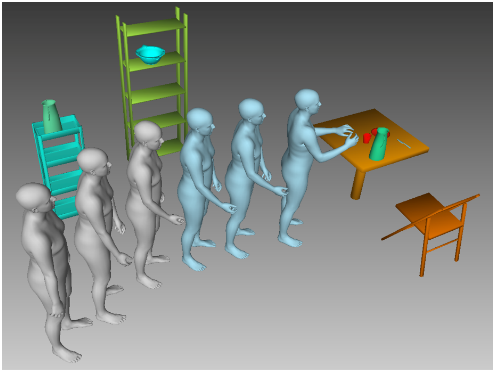
Low-latency and precise interaction in XR
[Belardinelli IROS'22]

Applications of human motion forecasting



Safe and comfortable interaction in XR
[Zhang ECCV'22]

Coordination of human body motion and scene environment



Human body movements in daily pick and place activities

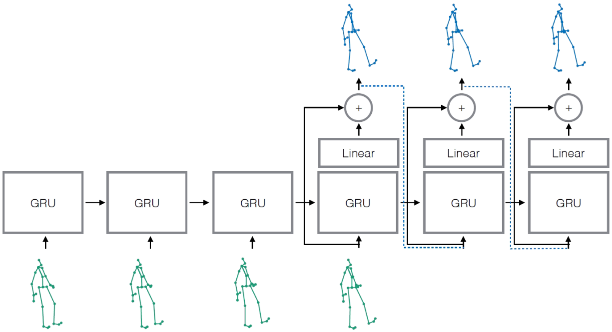
Use scene object information to guide human motion forecasting

Contributions

- Demonstrate the effectiveness of **egocentric 3D object bounding boxes** for human motion forecasting
- Propose a novel **GCN-based** method to **forecast human motions** from **body pose and egocentric object** features
- Conduct extensive experiments on **two public datasets** and report a **user study** to show the **superiority** of our method

Res-RNN: residual recurrent neural network

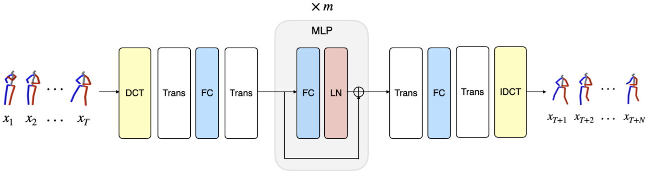
- Sequence-to-sequence architecture
- Residual architecture



[Martinez CVPR'17]

siMLPe: simple multi-layer perceptrons

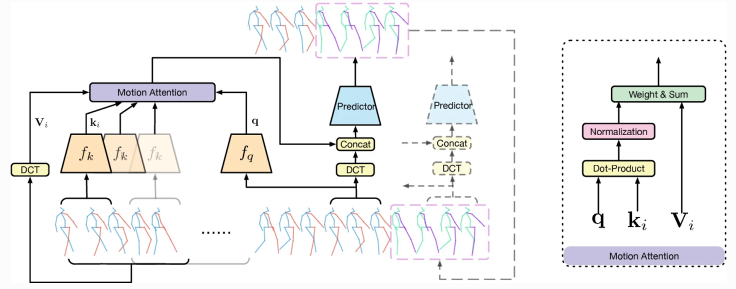
- Fully connected layers, layer normalisation, and transpose operations
- Residual architecture



[Guo WACV'23]

HisRep: human motion prediction via motion attention

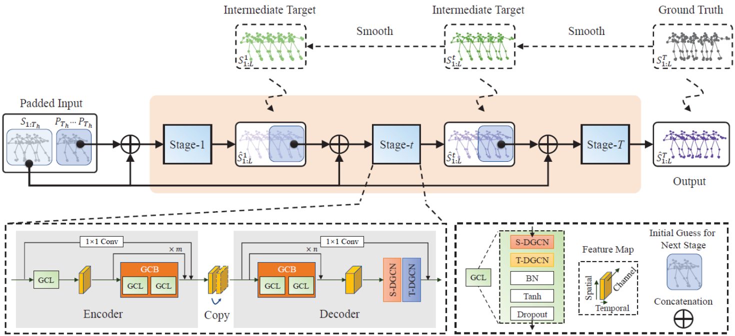
- Sequence-to-sequence architecture
- Attention-based architecture



[Mao ECCV'20]

PGBIG: progressively generating better initial guesses

- Multi-stage human motion prediction framework
- Spatial and temporal dense graph convolutional networks



[Ma CVPR'22]

Related Work

Traditional methods

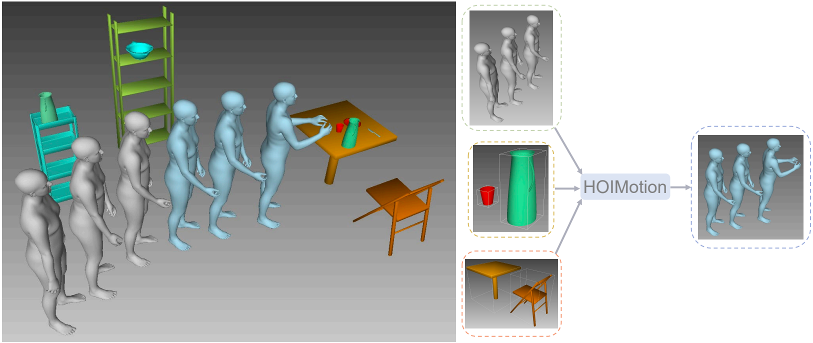
- Predict future poses from historical poses

Our method

- Extract features from **scene objects**
- Predict future poses from **past pose and scene object** features

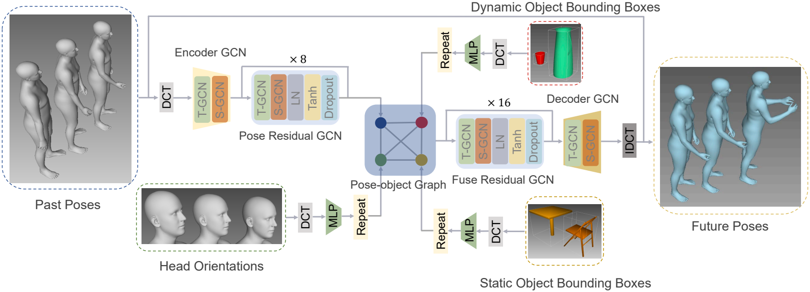
Problem formulation

- Daily **human-object interaction** activities
- Use **egocentric 3D object bounding boxes** to forecast human motion



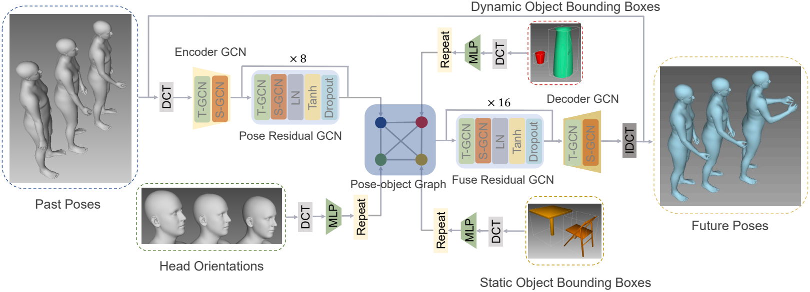
HOIMotion method

- Pose-object feature extraction
- Pose-object fusion
- Motion forecasting



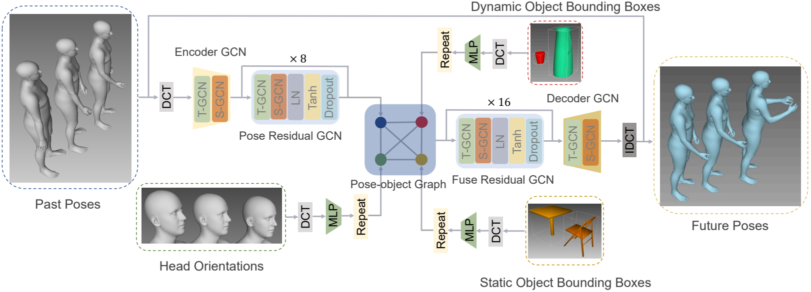
HOIMotion method: Pose-object feature extraction

- Past poses, head orientations, static and dynamic objects
- DCT, spatio-temporal GCN, and MLP



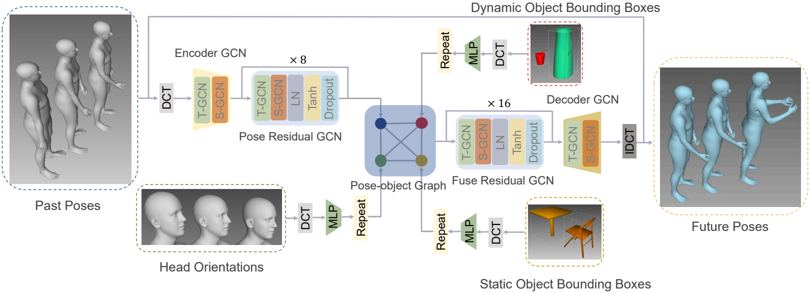
HOIMotion method: Pose-object fusion

- Treat scene objects and body joints as **nodes** in a graph
- Fully-connected spatio-temporal graph



HOIMotion method: Motion forecasting

- Spatio-temporal GCN
- Fuse residual GCN and decoder GCN



Evaluation settings

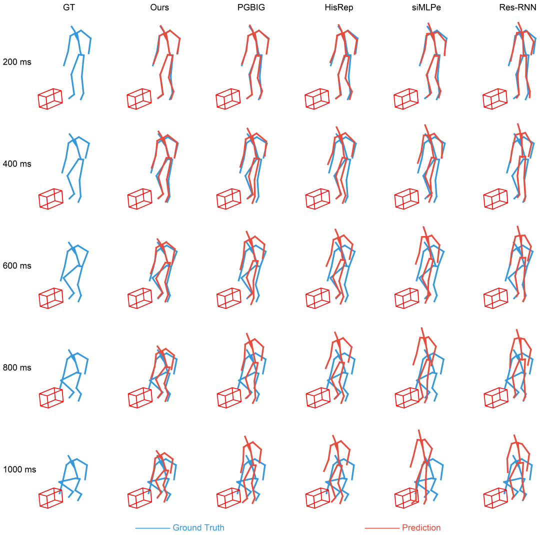
- Datasets: **ADT** [Pan ICCV'23] and **MoGaze** [Kratzer RAL'20]
- Metric: mean per joint position error (MPJPE)
- Input: 10 frames in the past
- Output: 30 frames in the future

Motion forecasting performance

Dataset	Method	100 ms	200 ms	300 ms	400 ms	500 ms	600 ms	700 ms	800 ms	900 ms	1000 ms	Average
ADT	*Res-RNN* [Martinez CVPR'17]	23.7	33.9	44.8	56.8	68.6	80.8	93.1	105.7	118.3	131.1	72.3
	siMLPe [Guo WACV'23]	26.6	30.4	37.8	46.8	57.5	68.2	79.7	92.5	105.3	119.5	63.2
	HisRep [Mao ECCV'20]	8.3	15.4	22.6	30.2	38.4	47.2	56.6	66.6	76.8	87.8	42.0
	PGBIG [Ma CVPR'22]	8.9	15.5	22.4	29.6	37.4	46.0	55.0	64.7	75.0	86.2	41.3
	Ours *pose only*	<u>5.8</u>	<u>11.9</u>	<u>18.8</u>	<u>26.4</u>	<u>34.8</u>	<u>43.9</u>	<u>53.6</u>	<u>63.9</u>	<u>74.7</u>	<u>85.8</u>	<u>39.1</u>
	Ours	**5.5**	**11.4**	**18.1**	**25.6**	**33.7**	**42.5**	**52.0**	**61.8**	**72.0**	**82.5**	**37.7**
MoGaze	*Res-RNN* [Martinez CVPR'17]	38.5	53.1	71.1	91.3	113.2	136.8	161.7	187.5	214.0	240.8	124.3
	siMLPe [Guo WACV'23]	28.8	40.6	55.5	72.0	89.4	108.8	130.2	152.6	176.3	201.0	99.5
	HisRep [Mao ECCV'20]	17.1	31.4	45.4	60.5	77.1	95.4	115.0	135.3	156.4	177.9	85.3
	PGBIG [Ma CVPR'22]	16.0	29.4	43.0	57.7	74.1	92.0	110.8	130.7	151.1	171.5	82.0
	Ours *pose only*	<u>14.3</u>	<u>26.9</u>	<u>40.4</u>	<u>55.0</u>	<u>71.2</u>	<u>88.8</u>	<u>107.5</u>	<u>126.9</u>	<u>147.0</u>	<u>167.3</u>	<u>79.0</u>
	Ours	**13.2**	**25.6**	**38.6**	**52.9**	**68.7**	**85.7**	**103.9**	**122.7**	**142.0**	**161.3**	**76.1**

Our method (Ours and Ours *pose only*) **consistently outperforms** prior methods at different time intervals

Motion forecasting performance

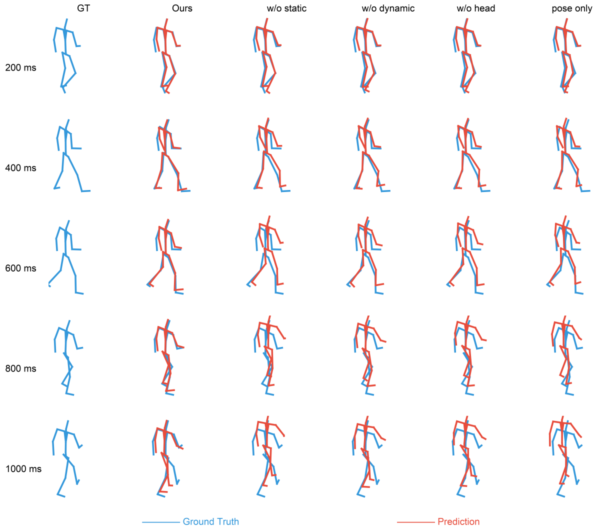


Ablation study

Method	100 ms	200 ms	300 ms	400 ms	500 ms	600 ms	700 ms	800 ms	900 ms	1000 ms	Average
w/o *static*	13.8	26.3	39.7	54.3	70.2	87.2	105.3	124.1	143.4	162.6	77.3
w/o *dynamic*	13.8	26.2	39.6	54.1	69.9	86.9	105.0	123.9	143.2	162.4	77.1
w/o *static+dynamic*	13.9	26.6	40.0	54.5	70.5	87.8	106.0	124.9	144.3	163.9	77.8
w/o *head*	13.7	26.2	39.5	54.2	70.1	87.2	105.2	124.1	143.6	163.0	77.2
w/o *static+dynamic+head*	14.3	26.9	40.4	55.0	71.2	88.8	107.5	126.9	147.0	167.3	79.0
Ours	**13.2**	**25.6**	**38.6**	**52.9**	**68.7**	**85.7**	**103.9**	**122.7**	**142.0**	**161.3**	**76.1**

Our method significantly outperforms the ablated versions

Ablation study



User study

- Stimuli: 20 randomly selected motion forecasting samples
- Participants: 20 users (10 males and 10 females)
- Procedure: rank different methods according to ***precision*** (*align with the ground truth*) and ***realism*** (*physically plausible*)

User study

		Ours	*PGBIG* [Ma CVPR'22]	*HisRep* [Mao ECCV'20]
Precision	Median	**1.0**	2.0	3.0
	Mean	**1.2**	2.3	2.5
	SD	0.5	0.6	0.6
Realism	Median	**1.0**	2.0	2.0
	Mean	**1.3**	2.2	2.3
	SD	0.6	0.7	0.7

Our method outperforms prior methods in terms of both *precision* and *realism*

Table of Contents

Limitations

- **Long-term** motion forecasting performances are not as good as **short-term** performances
- Designed for **human-object interactions** and may not work well for **human-human interactions**

Future work

- Explore other **scene object-related** information such as **object shape** for human motion forecasting
- Add some **physical constraints** for the predicted human poses to make them more **physically plausible**
- Integrate our method into motion-related applications such as **redirected walking** and **human-agent collaboration**

Main contributions

- Validate the effectiveness of **egocentric 3D object bounding boxes** for human motion forecasting
- Propose a novel method consisting of three components: **pose-object feature extraction**, **pose-object fusion**, and **motion forecasting**
- Demonstrate the **superiority** of our method through experiments on **two public datasets** and a **user study**

Code available at **zhiminghu.net/hu24_hoimotion**

Acknowledgement

Thank you!

Alexiadis TCSVT'16. An integrated platform for live 3d human reconstruction and motion capturing. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(4):798–813, 2016.

Belardinelli IROS'22. Intention estimation from gaze and motion features for human-robot shared-control object manipulation. In *Proceedings of the 2022 IEEE International Conference on Intelligent Robots and Systems*, pages 9806–9813. IEEE, 2022.

Guo WACV'23. Back to mlp: A simple baseline for human motion prediction. In *Proceedings of the 2023 IEEE Winter Conference on Applications of Computer Vision*, pages 4809–4819, 2023.

Kratzer RAL'20. Mogaze: A dataset of full-body motions that includes workspace geometry and eye-gaze. *IEEE Robotics and Automation Letters*, 6(2):367–373, 2020.

Le RHIC'21. Hierarchical human-motion prediction and logic-geometric programming for minimal interference human-robot tasks. In *Proceedings of the 2021 IEEE International Conference on Robot and Human Interactive Communication*, pages 7–14. IEEE, 2021.

Lin TVCG'22. Intentional head-motion assisted locomotion for reducing cybersickness. *IEEE Transactions on Visualization and Computer Graphics*, 2022.

Ma CVPR'22. Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In *Proceedings of the 2022 IEEE Conference on Computer Vision and Pattern Recognition*, pages 6437–6446, 2022.

Mao ECCV'20. History repeats itself: Human motion prediction via motion attention. In *Proceedings of the 2020 European Conference on Computer Vision*, pages 474–489. Springer, 2020.

Martinez CVPR'17. On human motion prediction using recurrent neural networks. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2891–2900, 2017.

References ii

Pan ICCV'23. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20133–20143, 2023.

Zhang ECCV'22. Egobody: Human body shape, motion and social interactions from head-mounted devices. In *Proceedings of the 2022 European Conference on Computer Vision*, 2022.