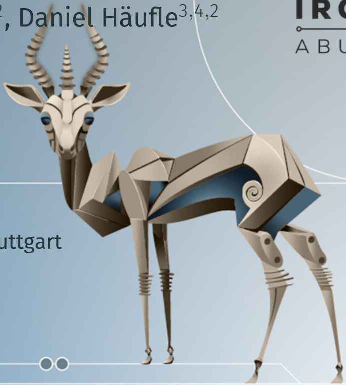
GazeMotion: Gaze-guided Human Motion Forecasting

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

IROS '24
ABU DHABI

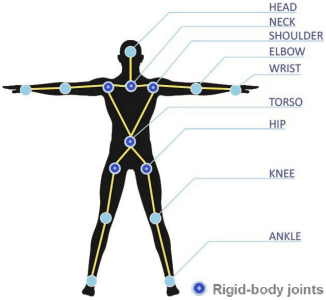


[1]University of Stuttgart
[2]Bionic Intelligence Tuebingen Stuttgart
[3]Heidelberg University
[4]University of Tuebingen

Table of Contents

Research Background

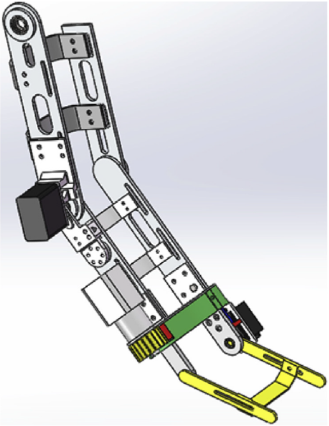
- **Human pose:** 3D positions of human joints (e.g. wrist, elbow, shoulder, knee, ankle)
- **Motion forecasting:** predict future human poses from historical poses



Human pose
[Alexiadis TCSVT'16]

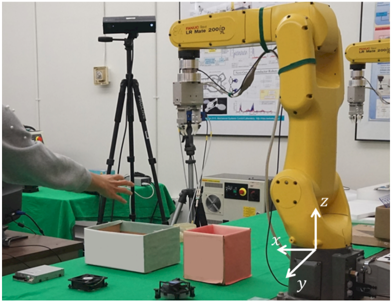
Applications of human motion forecasting

Wearable arm exosuit
[Lotti RAM'20]

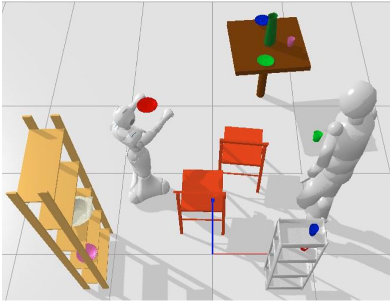


Upper limb exoskeleton
[Zhang BSPC'19]

Applications of human motion forecasting



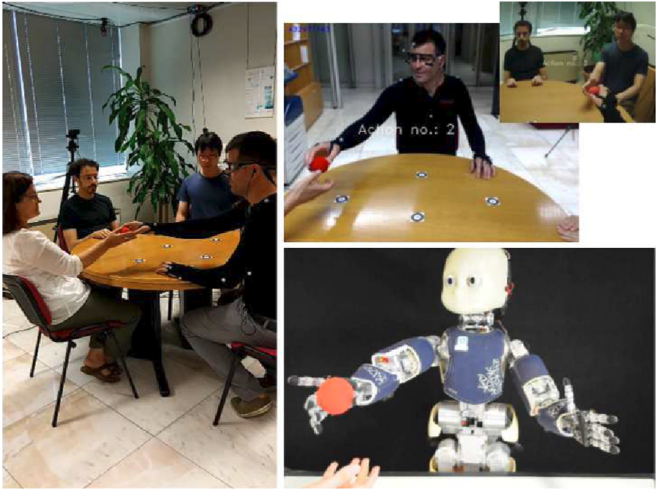
Human-robot collaboration
[Landi IRS'19]



Human-robot collaboration
[Le RHIC'21]

Research Background

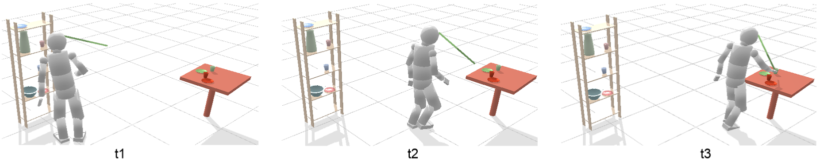
Applications of human motion forecasting



Human-human and human-robot interaction
[Duarte RAL'18]

Eye-body coordination

- Eye-head coordination [Hu TVCG'19; Hu TVCG'20; Hu TVCG'21]
- Eye-hand-head coordination [Emery ETRA'21]
- Eye-head-torso coordination [Sidenmark ToCHI'19]



Eye and body movements in daily pick and place activities

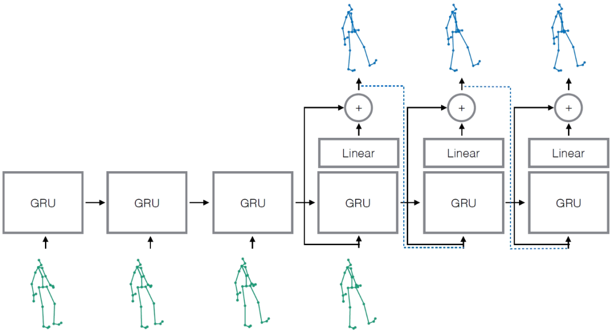
Use eye gaze information to guide human motion forecasting

Contributions

- A novel method that first **predicts future eye gaze from past gaze** and then **forecasts future poses** using the predicted gaze and past poses through a **spatio-temporal GCN**
- Experiments on **three public datasets** that demonstrate **significant performance improvements** over prior methods
- A **user study** that validates our method **outperforms** prior methods in both **precision** and **realism**

Res-RNN: residual recurrent neural network

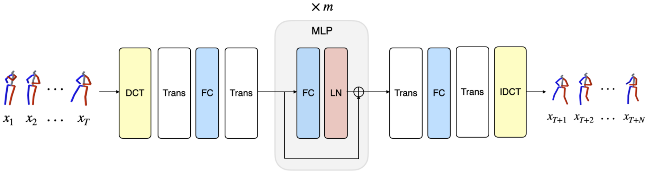
- Sequence-to-sequence architecture
- Residual architecture



[Martinez CVPR'17]

siMLPe: simple multi-layer perceptrons

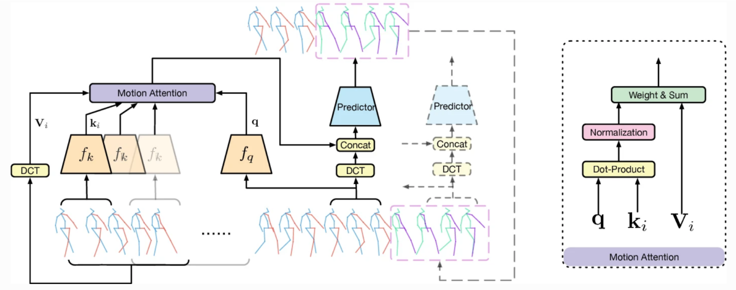
- Fully connected layers, layer normalisation, and transpose operations
- Residual architecture



[Guo WACV'23]

HisRep: human motion forecasting via motion attention

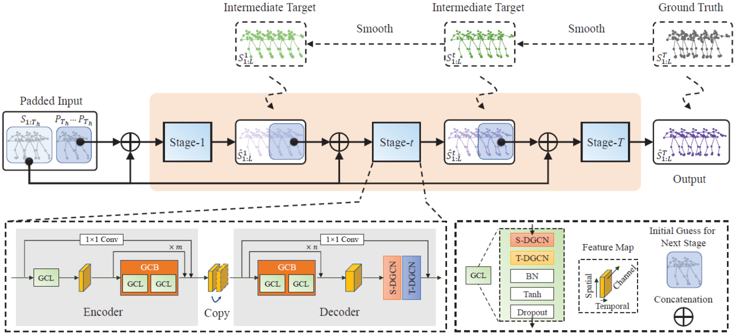
- Sequence-to-sequence architecture
- Attention-based architecture



[Mao ECCV'20]

PGBIG: progressively generating better initial guesses

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



[Ma CVPR'22]

Traditional methods

- Predict future poses from historical poses

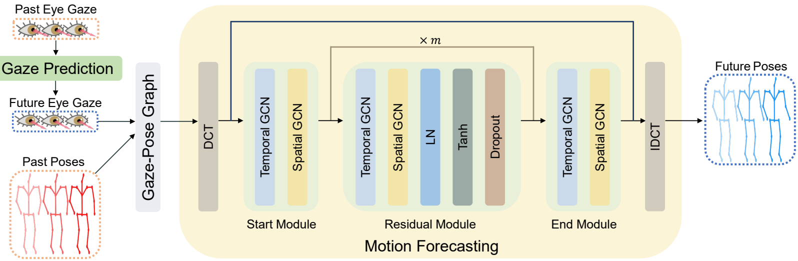
Our method

- **Predict future eye gaze** from historical gaze
- Predict future poses from **past poses and the predicted gaze**

Table of Contents

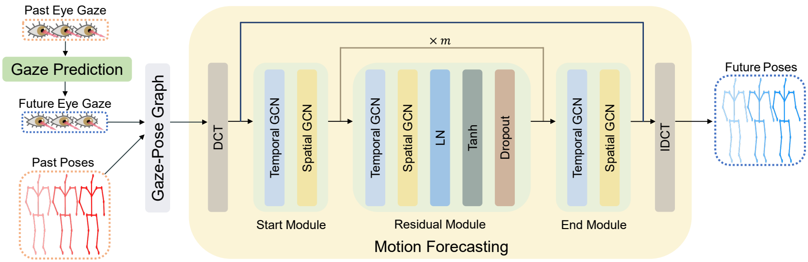
GazeMotion method

- Eye gaze prediction
- Gaze-pose fusion
- Motion forecasting



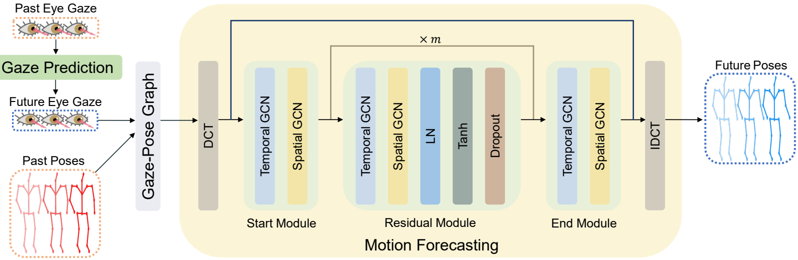
GazeMotion method: Eye gaze prediction

- 1D convolutional neural network



GazeMotion method: Gaze-pose fusion

- Treat eye gaze and body joints as **nodes** in a graph
- Fully-connected spatio-temporal graph



GazeMotion method: Motion forecasting

- Spatio-temporal graph convolutional network
- Start module, residual module, end module

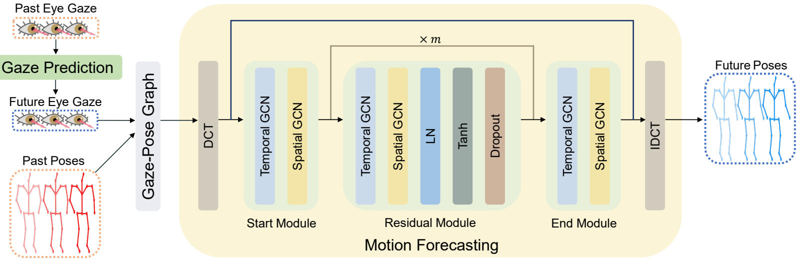


Table of Contents

Evaluation settings

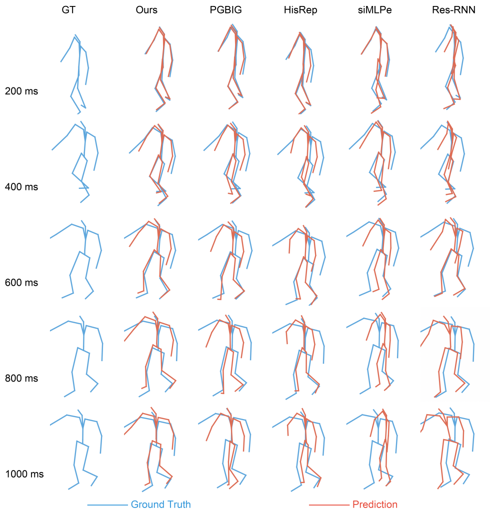
- Datasets: **MoGaze** [Kratzer RAL'20], **ADT** [Pan ICCV'23], **GIMO** [Zheng ECCV'22]
- Metric: mean per joint position error (MPJPE)
- Input: 10 frames in the past
- Output: 30 frames in the future

Motion forecasting performance

Dataset	Method	200 ms	400 ms	600 ms	800 ms	1000 ms	Average
MoGaze	*Res-RNN* [Martinez CVPR'17]	53.1	91.3	136.8	187.5	240.8	124.3
	siMLPe [Guo WACV'23]	40.6	72.0	108.8	152.6	201.0	99.5
	HisRep [Mao ECCV'20]	31.4	60.5	95.4	135.3	177.9	85.3
	PGBIG [Ma CVPR'22]	29.4	57.7	92.0	130.7	171.5	82.0
	Ours w/o *gaze*	27.2	55.3	88.9	126.9	167.1	79.0
	Ours	**25.8**	**53.3**	**85.8**	**122.0**	**160.0**	**75.9**
ADT	*Res-RNN* [Martinez CVPR'17]	35.6	55.7	77.8	100.0	122.5	70.1
	siMLPe [Guo WACV'23]	29.9	48.3	69.1	93.8	120.7	63.8
	HisRep [Mao ECCV'20]	15.5	30.5	47.6	66.8	88.2	42.3
	PGBIG [Ma CVPR'22]	14.5	28.7	45.4	64.4	85.8	40.6
	Ours w/o *gaze*	12.0	26.6	44.0	63.8	85.3	39.1
	Ours	**11.7**	**25.8**	**42.8**	**62.1**	**82.8**	**38.0**
GIMO	*Res-RNN* [Martinez CVPR'17]	82.6	126.4	170.2	212.9	255.4	152.8
	siMLPe [Guo WACV'23]	42.8	78.3	114.6	150.7	188.5	100.3
	HisRep [Mao ECCV'20]	41.8	78.1	115.0	152.7	192.4	100.2
	PGBIG [Ma CVPR'22]	38.0	68.6	101.9	136.1	172.2	89.2
	Ours w/o *gaze*	33.7	66.1	99.7	134.4	170.4	86.8
	Ours	**32.6**	**64.1**	**97.0**	**130.0**	**162.4**	**83.8**

Our method (Ours and Ours w/o *gaze*) **consistently outperforms** prior methods at different time intervals

Motion forecasting performance



Ablation study

Method	200 ms	400 ms	600 ms	800 ms	1000 ms	Average
w/o *spatial GCN*	30.9	62.1	96.3	133.8	173.1	84.7
w/o *temporal GCN*	46.6	74.0	107.9	147.0	188.0	99.3
w/o gaze	27.2	55.3	88.9	126.9	167.1	79.0
past gaze	26.3	54.3	87.2	123.8	162.0	77.1
Ours	**25.8**	**53.3**	**85.8**	**122.0**	**160.0**	**75.9**

Our method **consistently outperforms** the ablated versions

User study

- Stimuli: 24 randomly selected motion forecasting samples
- Participants: 20 users (12 males and 8 females)
- Procedure: rank different methods according to ***precision*** (*align with the ground truth*) and ***realism*** (*physically plausible*)

User study

		Ours	*PGBIG*	*HisRep*	*siMLPe*	*Res-RNN*
Precision	Mean	**1.6**	3.2	3.2	3.3	3.7
	SD	0.9	1.2	1.2	1.3	1.3
Realism	Mean	**1.9**	3.3	3.1	3.3	3.5
	SD	1.3	1.2	1.3	1.3	1.4

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
- Ignore the **stochastic nature** of human motions

Future work

- Integrate more **context** information such as user's **goal** or **task** into human motion forecasting
- Explore other important body signals such as **hand gestures** for motion forecasting
- Integrate our method into motion-related applications such as **assistive devices**

Table of Contents

Main contributions

- A novel method consisting of three components: **eye gaze prediction**, **gaze-pose fusion**, and **motion forecasting**
- Experiments on **three public datasets** that demonstrate the **superiority** of our method over prior methods
- A **user study** that validates the **precision** and **realism** of our predictions

Code available at **zhiminghu.net/hu24_gazemotion**

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.

Duarte RAL'18. Action anticipation: Reading the intentions of humans and robots. *IEEE Robotics and Automation Letters*, 3(4):4132–4139, 2018.

Emery ETRA'21. Openneeds: A dataset of gaze, head, hand, and scene signals during exploration in open-ended vr environments. In *Proceedings of the 2021 ACM Symposium on Eye Tracking Research and Applications*, pages 1–7, 2021.

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.

Hu TVCG'19. Sgaze: a data-driven eye-head coordination model for realtime gaze prediction. *IEEE Transactions on Visualization and Computer Graphics*, 25(5):2002–2010, 2019.

Hu TVCG'20. Dgaze: Cnn-based gaze prediction in dynamic scenes. *IEEE Transactions on Visualization and Computer Graphics*, 26(5):1902–1911, 2020.

Hu TVCG'21. Fixationnet: forecasting eye fixations in task-oriented virtual environments. *IEEE Transactions on Visualization and Computer Graphics*, 27(5):2681–2690, 2021.

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.

Landi IRS'19. Prediction of human arm target for robot reaching movements. In *Proceedings of the 2019 IEEE International Conference on Intelligent Robots and Systems*, pages 5950–5957. IEEE, 2019.

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.

Lotti RAM'20. Adaptive model-based myoelectric control for a soft wearable arm exosuit: A new generation of wearable robot control. *IEEE Robotics and Automation Magazine*, 27(1):43–53, 2020.

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.

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.

Sidenmark ToCHI'19. Eye, head and torso coordination during gaze shifts in virtual reality. *ACM Transactions on Computer-Human Interaction*, 27(1):1–40, 2019.

Zhang BSPC'19. An upper limb movement estimation from electromyography by using bp neural network. *Biomedical Signal Processing and Control*, 49:434–439, 2019.

Zheng ECCV'22. Gimo: Gaze-informed human motion prediction in context. In *Proceedings of the 2022 European Conference on Computer Vision*, 2022.