



University of Stuttgart
Germany



香港科技大学 (广州)
THE HONG KONG
UNIVERSITY OF SCIENCE AND
TECHNOLOGY (GUANGZHOU)

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



bits Bionic Intelligence
Tübingen Stuttgart

SIGGRAPH 2025

The Premier Conference & Exhibition on
Computer Graphics & Interactive Techniques

HOIGaze: Gaze Estimation During Hand-Object Interactions in Extended Reality Exploiting Eye-Hand-Head Coordination

Zhiming Hu, Daniel Häufle, Syn Schmitt, Andreas Bulling

zhimingu.net/hu25_hoigaze 



Research Background

Related Work

Method

Results

Discussion

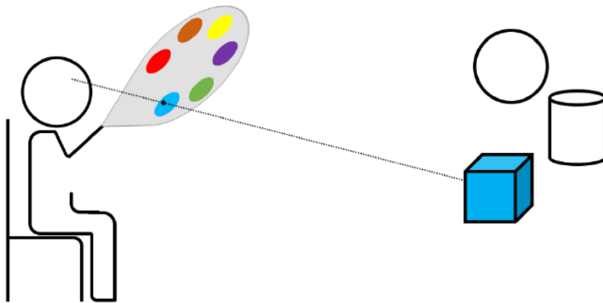
Conclusion

Applications of human eye gaze in XR



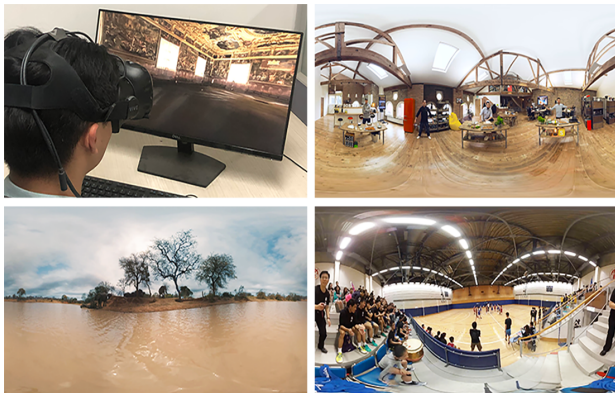
Gaze-contingent rendering [Hu TVCG'20]

Applications of human eye gaze in XR



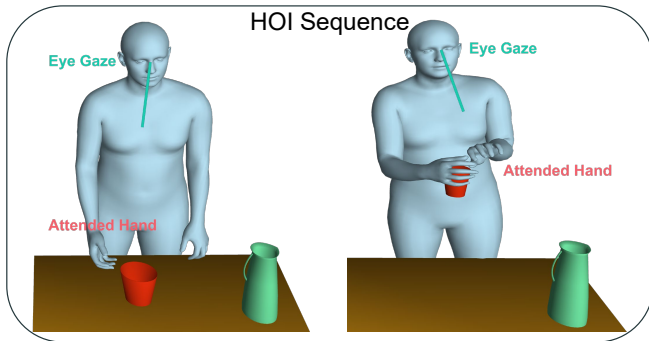
Gaze-based interaction [Mardanbegi IEEE VR'19]

Applications of human eye gaze in XR



Gaze-based activity recognition [Hu TVCG'22]

Eye-hand-head coordination during hand-object interactions



Estimate eye gaze during HOIs using eye-hand-head coordination

Research Background

Related Work

Method

Results

Discussion

Conclusion

Eye-head coordination during gaze shift in XR



[Sidenmark ToCHI'19]

Eye-hand coordination during object manipulation in XR



[Belardinelli IROS'22]

Previous works

- Only explore correlation between eye gaze and **hand trajectories**

Our work

- Focus on coordination of eye gaze and **hand gestures**
- Estimate eye gaze using **eye-hand-head coordination**

Research Background

Related Work

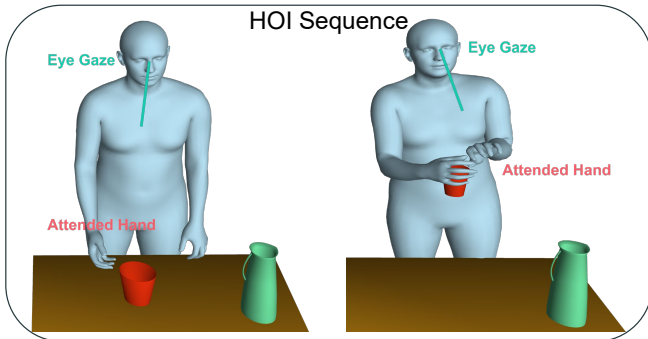
Method

Results

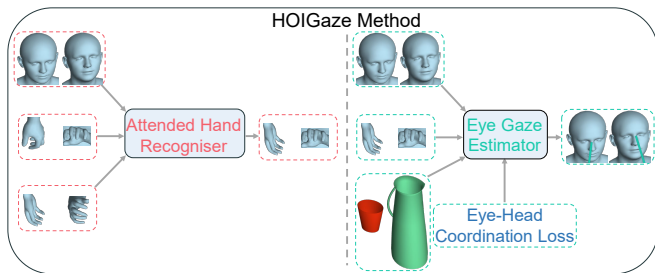
Discussion

Conclusion

Attended hand: the hand that is closer to eye gaze in terms of angular distance

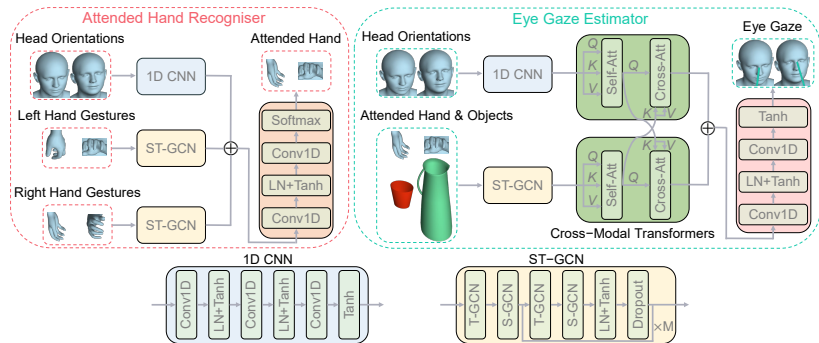


A novel **hierarchical framework** that first **recognises attended hand** and then **estimates eye gaze** based on the attended hand



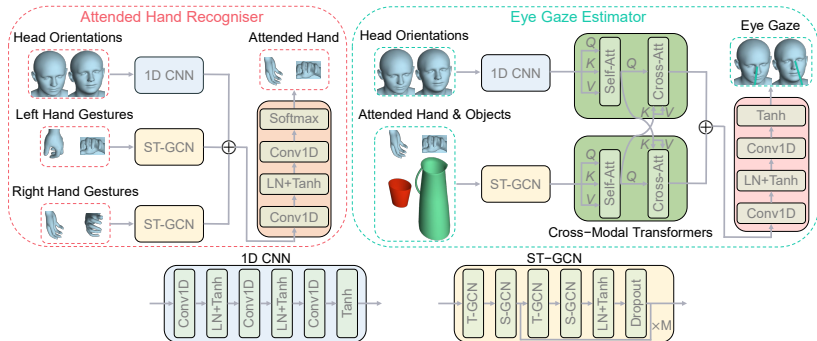
Attended hand recogniser

- 1D CNN for head orientations
- Spatio-temporal GCN for hand gestures



Eye gaze estimator

- 1D CNN for head orientations
- Spatio-temporal GCN for hand gestures and scene objects
- Cross-modal transformers for fusing features



Eye-head coordination loss that increases the weights of the loss assigned to eye-head-coordinated samples:

$$\ell_i = \begin{cases} f_{eh} * (g_i - \hat{g}_i)^2, & \text{if } g_i \cdot h_i > \text{Cos}_{eh} \\ (g_i - \hat{g}_i)^2, & \text{otherwise} \end{cases}$$

Insight: eye-head coordinated samples are **more important** than the samples with little eye-head correlation

Research Background

Related Work

Method

Results

Discussion

Conclusion

Gaze estimation performance

	HOT3D (Cross-User)				HOT3D (Cross-Scene)				ADT			
	$\{P1, P2, P3\}$	$\{P9, P10, P11\}$	$\{P12, P14, P15\}$	Average	Room	Kitchen	Office	Average	Work	Decoration	Meal	Average
Head Direction	23.24°	28.00°	17.85°	23.20°	23.69°	22.83°	23.16°	23.20°	22.88°	18.44°	25.23°	22.25°
DGaze [Hu TVCG'20]	12.17°	15.08°	14.87°	14.29°	13.37°	12.98°	11.39°	12.81°	8.84°	10.53°	10.77°	9.92°
FixationNet [Hu TVCG'21]	11.90°	14.60°	14.78°	14.00°	12.78°	12.84°	11.34°	12.53°	8.82°	10.50°	10.83°	9.92°
Pose2Gaze [Hu TVCG'24]	10.69°	10.73°	11.80°	11.10°	9.79°	9.73°	9.96°	9.80°	8.25°	9.71°	10.43°	9.34°
Ours	9.23°	9.16°	9.69°	9.37°	8.55°	8.69°	8.69°	8.64°	7.81°	9.46°	9.41°	8.78°

Our method **significantly outperforms** prior methods for both **cross-user** and **cross-scene** evaluations

Gaze estimation performance



Our method achieves **good estimation accuracy** during HOIs

Ablation study

	HOT3D (Cross-User)				HOT3D (Cross-Scene)				ADT			
	<i>{P1, P2, P3}</i>	<i>{P9, P10, P11}</i>	<i>{P12, P14, P15}</i>	Average	Room	Kitchen	Office	Average	Work	Decoration	Meal	Average
Ours w/o attended hand	9.89°	11.24°	10.57°	10.67°	9.71°	9.32°	9.16°	9.43°	8.26°	9.97°	9.87°	9.25°
Ours w/o Transformers	9.60°	10.07°	10.24°	10.02°	8.87°	8.85°	9.17°	8.92°	8.03°	9.74°	9.96°	9.12°
Ours w/o eye-head coord. loss	9.83°	9.48°	9.70°	9.64°	8.79°	8.71°	8.84°	8.76°	7.87°	9.49°	9.71°	8.90°
Ours	9.23°	9.16°	9.69°	9.37°	8.55°	8.69°	8.69°	8.64°	7.81°	9.46°	9.41°	8.78°

Each component contributes to our method's performance

Downstream task of gaze-based activity recognition

GT	Ours	Pose2Gaze	FixationNet	DGaze	Head Direction	Chance
72.9%	<u>71.8%</u>	68.7%	66.6%	66.0%	47.1%	33.3%

Our method achieves **higher recognition accuracies** than other methods

Research Background

Related Work

Method

Results

Discussion

Conclusion

Limitations

- Evaluation datasets only contain interactions with **real physical objects**
- Ignore the influence of **image/texture features** on eye gaze

Future work

- Evaluate on interactions with **both real and virtual objects**
- Integrate **image/texture features** into our method to further boost its performance

Research Background

Related Work

Method

Results

Discussion

Conclusion

Main contributions

- A novel **hierarchical** framework, a new gaze estimator that combines **GCN** and **cross-modal Transformers**, and a novel **eye-head coordination loss**
- Extensive experiments on **two public datasets** that demonstrate the effectiveness of our method
- Experiments on the application of **gaze-based activity recognition** that validate the usefulness of our method



- 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.
- 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.
- Hu TVCG'22. Ehtask: recognizing user tasks from eye and head movements in immersive virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 2022.
- Hu TVCG'24. Pose2gaze: Eye-body coordination during daily activities for gaze prediction from full-body poses. *IEEE Transactions on Visualization and Computer Graphics*, 2024.
- Mardanbegi IEEE VR'19. Eyeseethrough: unifying tool selection and application in virtual environments. In *Proceedings of the 2019 IEEE Conference on Virtual Reality and 3D User Interfaces*, pages 474–483, 2019.
- 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.