Human trajectory forecasting in 3D contexts with simulated visual impairments

Franz Franco, Hui-Yin Wu, Lucile Sassatelli

Human trajectory forecasting in 3D contexts with simulated visual impairments

Franz Franco [1] Hui-Yin Wu [1] Lucile Sassatelli[2]

[1]Centre Inria d'Université Côte d'Azur [2]Université Côte d'Azur, CNRS, I3S and Institut Universitaire de France

ANR CREATTIVE3D project

The ANR CREATTIVE3D project aims to understand human behavior and the impact of low-vision conditions using Virtual Reality (VR) technologies, and using this understanding for training and rehabilitation protocols.

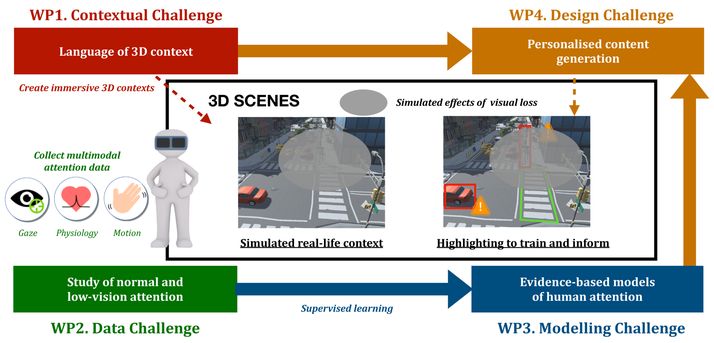

Figure 1. ANR CREATTIVE3D contains 4-core work packages: (1) a domain specific language (DSL) of 3D contexts, (2) the study of multi-modal attention under normal and simulated central vision loss conditions, (3) establishing models of user attention and behavior with the goal of (4) creating adapted 3D content for training and rehabilitation.

Data Collection and Multimodal features

Rich human-centered data is essential for effective trajectory forecasting. Virtual environments, such as Virtual or Augmented Reality, offer controlled scenarios. In our approach, we utilize the GusT-3D framework[1] to collect diverse data, including motion, gaze, physiological response, and scene context logs from 40 users navigating a virtual environment with diverse scenarios, including road-crossing tasks with varying intents, levels of interaction, and crucially, different levels of visual impairments.

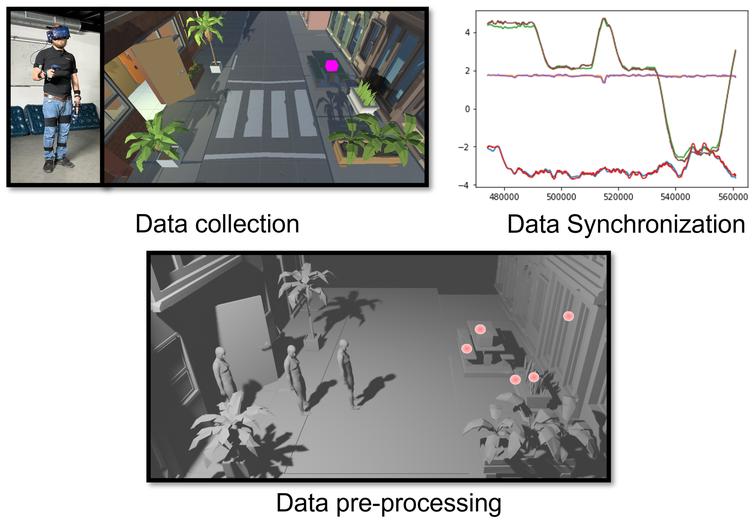

Figure 2. The data workflow consists of (1) data collection using a VR headset, motion capture suit, and physiology sensors, (2) synchronization of the multivariate data, and (3) pre-processing for visualization and input for ML models.

- **Data Collection:** The recorded data includes motion capture data from the XSens MVN, head and gaze tracking from the HTC Vive Pro Eye headset, and skin conductance and heart rate from Shimmer3 GRS+ sensors.
- **Data Synchronization:** This task encompasses the precise temporal and spatial synchronization of gathered data within the 3D scene reference framework.
- **Data pre-processing:** In this stage, we calculate the Point of Regard (POR), determine SMPL-X poses, and generate scene point clouds.
- **Public dataset in zenodo:** https://zenodo.org/records/8269109

Motivation

Virtual reality (VR) provides a unique solution to improve current trajectory forecasting by capturing data in controlled, interactive scenarios, ensuring precision and inclusivity. Our work collects rich multimodal human data under VR scenarios considering the impact of low vision. We propose a novel model capable of handling multimodal data to enhance trajectory forecasting accuracy.

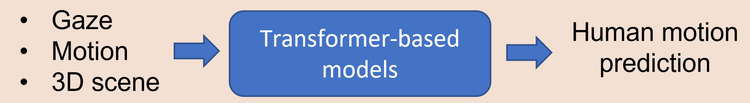

Figure 3. Proposed model: integrating motion, gaze, and 3D scene inputs into a Transformer-based model for improved prediction of future movement.

Highlights

- The ANR CREATTIVE3D project aims to develop ways to understand human behavior and the effects of **low-vision conditions**.
- Using **Virtual environments** facilitates the collection of behavioral data in controlled, life-like, and varied interactive scenarios.
- Collecting human-centered multivariate data utilizing the **GusT-3D framework** featuring 40 users navigating virtual environments and varying levels of **vision impairments**.
- Synchronizing data is essential when dealing with multi-modal data collected in virtual environments because various devices are used to gather the data.
- Investigation of learning tasks such as trajectory prediction taking into account **human intent, visual attention, and context**.
- Simulated low-vision scenarios to diversify user profiles in trajectory prediction tasks for trajectory forecasting.
- Benchmarking existing models, and proposing a novel model investigating the **impact of low vision** and capable of handling multivariate data

Motion Forecasting using Multi-modal data

Addressing the challenge of navigating the uncertainty in human behavior, particularly within dynamic environments, we put forth a model that takes as input multivariate data including gaze, motion, and 3D scene from the CREATTIVE3D dataset.

We benchmark existing transformer architectures, and establish baseline models that can be subsequently adapted to new learning tasks and richer contextual data.

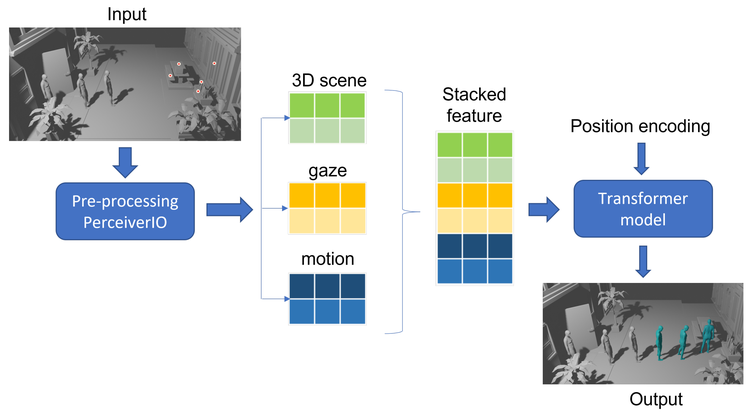

Figure 4. Baseline Model Overview: Using pre-processed multivariate GUsT-3D data, we extract distinctive features from each modality. These features are then collectively stacked to robustly predict future motion patterns.

- Recent advancements in human motion prediction have integrated crucial elements, including gaze, motion, and scene context, gathered from real indoor environments. [2]
- Baseline model is based on the **PerceiverIO architecture**. [3]
- To extract both local and global scene features PointNet++ and PointNet are used respectively. [4]

Perspective

- This study goes beyond data collection to explore the influence of low vision on the generalization error of models trained on normal-vision data for predicting the motion of individuals with low vision. It contributes to both advancing the understanding of human motion and meeting the crucial requirement for inclusive trajectory prediction models. This ensures that advancements in autonomous systems account for the diversity of individuals, including those with visual impairments.
- Leverage virtual scene annotations to comprehensively grasp and analyze the spatial layout, objects, and contextual elements within the environment. Subsequently, integrate this enhanced understanding into the prediction model to refine and optimize its performance.

References

[1] Florent Robert, Hui-Yin Wu, Lucile Sassatelli, Stephen Ramanoël, Auriane Gros, and Marco Winckler. An integrated framework for understanding multimodal embodied experiences in interactive virtual reality. IMX '23, page 14–26, New York, NY, USA, 2023. Association for Computing Machinery.

[2] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, C. Karen Liu, and Leonidas J. Guibas. Gimo: Gaze-informed human motion prediction in context. *ECCV*, 2022.

[3] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver io: A general architecture for structured inputs outputs. *ICLR*, 2022.

[4] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5099–5108, 2017.