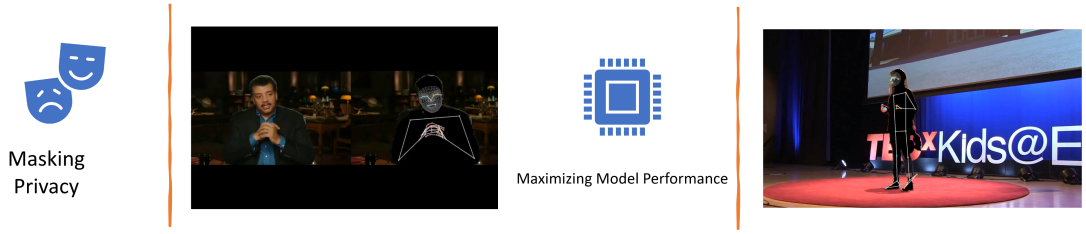
Integrating Privacy and Multimodal Behaviour Analytics for Audio Visual Recordings

Master Project Summer 2023 – AI and Intelligent Systems Group

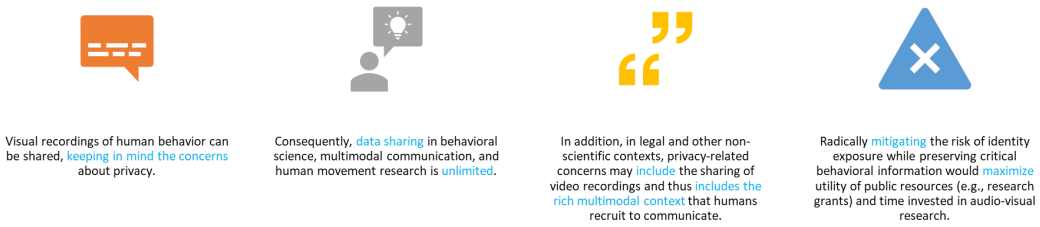
Abstract

Multimodal data analysis has become increasingly important in various fields such as human–computer interaction, healthcare, and psychology. However, such data can contain sensitive personal identity information, which may need to be protected. Masking is a technique that can be used to hide or replace sensitive information, while preserving the performance of the task or application. This master project aims to investigate the use of multimodal masking and analytics techniques to preserve privacy in videos while enabling us to gain important insights about human behavior. The project is a collaboration between the HPI AI and Intelligent Systems group and two partners in Nijmegen, Netherlands, the Donders Institute and the Max Planck Institute for Psycholinguistics (Multimodal Language Department). We plan to have at least one exchange visit, enabling the students to benefit from feedback on how to leverage multimodal data, including videos, audio, images, gestures, and affect data. The students will develop a set of masking and multimodal behavioral feature extraction techniques that can be applied to each modality of data and evaluate the performance in a task or application, such as speech recognition, video understanding, or affect recognition, on the original dataset and the masked dataset.

The students will also analyze **the trade-off between privacy and performance and study the effect of different masking strategies on human performance and engagement**.



The broader goal of this project is to develop new multimodal masking and behavioral analysis techniques that can preserve and summarize human behavior while increasing privacy, which will enable data sharing in science and promoting common behavior analysis standards (e.g., summarizing eye gaze direction, hand gesture use, and semantic content of speech).



Project Approach

This project involves the following:

- Exploring different approaches to visual and audio/voice de-identification while preserving behavioral information, such as investigating face detection tools and approaches for tracking facial gestures, and investigating replacement techniques that generate new appearances.
- Further developing and optimizing the existing proof-of-concept implementation of a de-identification tool, including improving the usability, model accuracy, robustness, and speed of the de-identification process for three modes (face, audio, and gestures)
- Exploring multimodal behavioral analytic techniques on a variety of audiovisual recordings, i.e., automatic analyses and categorization of behaviors (e.g., facial and hand gestures)
- Collaborating in an interdisciplinary workshop in Nijmegen (sponsored visit) with cognitive science researchers to understand their needs and incorporate their feedback into the development process.
- Writing a technical report outlining the approach taken, results, and future directions for the project, with the goal of publishing it as a scientific research paper.
- Improving the user-friendliness of the tool, including options for adjusting the level of de-identification and previewing the results, and contributing to the open source community by releasing the tool as open source software.

Project Outline

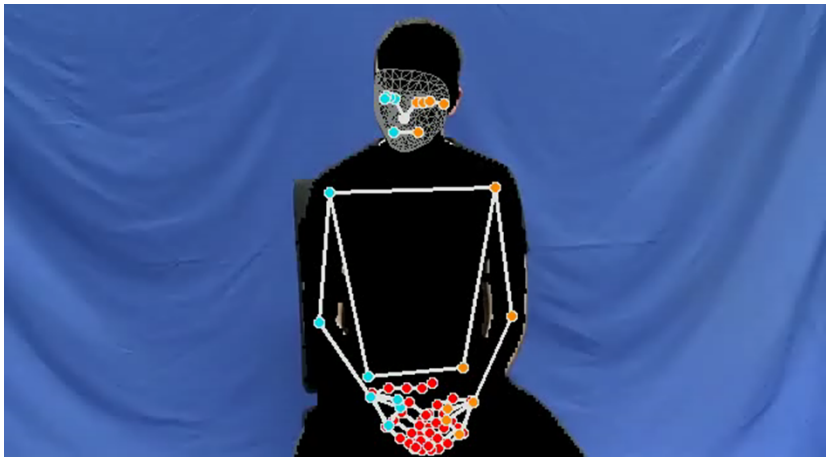
The project will follow a multi-step approach to address the research questions and achieve the objectives of the project. The project approach is somewhat divided into the following stages with iterations expected:

- Introductory Meeting/Workshop: The project will begin with an initial exchange workshop with colleagues at MPI, students will be supported to travel and co-organize a

three day workshop in Nijmegen to gain multimodal research insights from behavioral scientists.

- Data collection: The project will then begin by accessing and assessing a collection of datasets (multimodal data), including images, videos, audio, gestures, and affect data. The data is (will be) collected from various sources such as publicly available datasets and datasets present at the partner institutes Donders and MPI.
- Preprocessing and annotation: The collected data will be preprocessed to ensure that it is suitable for the analysis.
- Masking techniques: The project will develop a set of masking techniques that can be applied to each modality of data. These techniques will be designed to hide or replace sensitive information while preserving the performance of the task or application.
- Analysis and results evaluation: The performance of a specific masking or behavioral analysis task or application will be studied, which may involve speech recognition tasks, emotion/gesture recognition, or deidentification, and evaluated on the original dataset and other test datasets.
- Discussion and conclusion: The project will conclude with a discussion of the key findings and results, and a summary of the contributions and achievements of the project and how they can benefit the stakeholders in the collaboration (computer science and cognitive science). (This will result in a joint publication either for a top behavioral research journal or a computer science outlet)

This is a general outline, and the actual project approach may be adjusted depending on the specific research question and methods addressed by the students.



Additionally, the students would consult with the HPI and MPI supervisors to ensure that the project approach is feasible and aligns with the project's objectives.

To summarize, sample research questions for the project are as follows:

- What are the most effective masking techniques for each modality of data (images, videos, audio, gestures, affect) in terms of preserving privacy and performance for a specific task or application?

- How does the combination of different modalities of human behavior be made insightful by automatic analyses and categorization (e.g., what is said, and the 'non-verbal' behavior) ?
- Can we find a balance between preserving privacy and maximizing performance in multimodal behavioral data analysis?
- How do we evaluate the privacy preserving properties of different masking techniques on multimodal data?

Related Work

Owoyele, B., Trujillo, J., de Melo, G., & Pouw, W. (2022). Masked-piper: Masking personal identities in visual recordings while preserving multimodal information. https://doi.org/10.31234/osf.io/bpt26
https://github.com/WimPouw/TowardsMultimodalOpenScience

Pouw, W., Dingemanse, M., Motamedi, Y., & Ozyurek, A. (2021). A systematic investigation of gesture kinematics in evolving manual languages in the lab. Cognitive Science, 45(7): e13014. doi:10.1111/cogs.13014

Pouw, W., Tobin, S. J., & Trujillo, J. P. (2022-01-03). Gesture Networks Module 1: Computing kinematic distances using dynamic time warping. [the day you viewed the site]. Retrieved from: https://wimpouw.github.io/EnvisionBootcamp2021/gesturenetworks_module1.html

Pouw, W., Trujillo, J. P., & Dixon, J. A. (2020). The quantification of gesture–speech synchrony: A tutorial and validation of multimodal data acquisition using device-based and video-based motion tracking. Behavior Research Methods, 52, 723-740. doi:10.3758/s13428-019-01271-9

Rasenberg, M., Özyürek, A. and Dingemanse, M. (2020), Alignment in Multimodal Interaction: An Integrative Framework. Cogn Sci, 44: e12911. https://doi.org/10.1111/cogs.12911

Lin, Z., Geng, S., Zhang, R., de Melo, G., Wang, X., Dai, J, Qiao, Y, Gao, P., Li, H. (2022), Frozen CLIP Models are Efficient Video Learners. Proceedings of ECCV 2022. https://arxiv.org/abs/2208.03550

Zeng, H., Wang, X., Wang, Y., Wu, A., Pong, T. C., & Qu, H. (2022). GestureLens: Visual Analysis of Gestures in Presentation Videos. *IEEE Transactions on Visualization and Computer Graphics*. Paper https://doi.org/10.48550/arXiv.2204.08894

Contact

babajide.owoyele@hpi.de, gerard.demelo@hpi.de

This project will be jointly supervised by the HPI Chair for Artificial Intelligence and Intelligent Systems (Babajide Owoyele, Prof. Gerard de Melo) and its partners in the Netherlands: the Donders Institute and the Max Planck Institute for Psycholinguistics Multimodal Language Department (Dr. Wim Pouw, Dr. James Trujillo, Prof. Asli Ozyurek).