SPRING

Deliverable D4.4: Multi-modal behaviour recognition in relevant environments

Due Date: 31/05/2023

Main Author: UNITN

Contributors: INRIA, CVUT

Dissemination: Public Deliverable

This project has received funding from the European Union’s Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

This project has received funding from the European Union's Horizon 2020
Research and Innovation Programme under Grant Agreement No. 871245.

DOCUMENT FACTSHEET

Deliverable	D4.4: Multi-modal behaviour recognition in relevant environments
Responsible Partner	UNITN
Work Package	WP4: Multi-Modal Human Behaviour Understanding
Task	Result of T4.2 on Broca Data.
Version & Date	31/05/2023
Dissemination	Public Deliverable

CONTRIBUTORS AND HISTORY

Version	**Editor**	**Date**	**Change Log**
1	UNITN	26/05/2023	First Draft
Final	UNITN	02/06/2023	Filal draft including reviewer comments

APPROVALS

Authors/editors	UNITN
Task Leader	UNITN
WP Leader	UNITN

This project has received funding from the European Union’s Horizon 2020
Research and Innovation Programme under Grant Agreement No. 871245.

Contents

Executive Summary

This document, D4.4, is part of WP4 of the H2020 SPRING project. It aims to present the quantitative and qualitative outcomes of a fully functional framework for analyzing the human face and body from visual data related to T4.2 "*Individual and Group Behaviour Recognition*". The framework includes tools for multi-target body pose estimation and face analysis, which have been tested using the corresponding data. All evaluation data presented in this deliverable was collected at Broca Hospital in Paris by the SPRING project.

This document presents evaluations of various methods, including: **a) face mask detection b) biometric recognition**, **c) gaze target detection**, **d) monocular depth estimation**, and **e) multi-target body pose estimation**. It is important to note that none of the developed methods (a)-(e) were fine-tuned on the target dataset, demonstrating the effectiveness of the proposed methods in unseen scenarios and domains.

The source code of the modules presented in this deliverable is available in the SPRING repository[1]. As per European Commission requirements, the repository will be available to the public for a duration of at least four years after the end of the SPRING project.

[1] https://gitlab.inria.fr/spring

1 Introduction

This document, **D4.4**, is a part of **WP4** of the H2020 SPRING project. It presents the outcomes of T4.2, which involves a fully functional framework for analyzing the human face and body from visual data, including tools for multi-target body pose estimation and face analysis that have been tested on the dataset collected at the Broca Hospital in Paris by the SPRING project

In this context, we mainly present a qualitative evaluation of:

- the models related to human face analysis: face mask detection and biometric recognition;
- the model related to human actions: gaze target detection;
- monocular depth estimation method;
- multi-party body pose estimation method.

Furthermore, we conducted a manual annotation of 9952 frames, allowing us to quantitatively evaluate our mask detection and gender estimation models.

The dataset used in this deliverable was collected by the SPRING partners at the Broca Hospital in Paris. The rest of this document is structured as follows: first, we present qualitative results on face modules, including face mask detection, biometric recognition, and gaze target detection. Meanwhile, we also provide confusion matrices for face mask detection and gender estimation. Then, we demonstrate the qualitative performance of monocular depth estimation and multi-target body pose estimation tested on rectified images collected by the front fish-eye camera. Finally, we conclude this document with a summary of the results of the proposed methods on the Broca dataset.

2 Individual & Group Behaviour Recognition

2.1 Face Mask Detection

The description of the architecture of the face mask detection is available in D4.1. It is important to note that we did not perform fine-tuning on the Broca dataset, demonstrating the effectiveness of our model in unseen scenarios. The qualitative results are shown in Figure 2.1.

We provide the quantitative results in terms of confusion matrix computed considering a subset of 9952 frames that have been manually annotated. The corresponding results are given in Table 2.1. Qualitative results of the face mask detection module are presented in Figure 2.1 It can be observed that our approach in the vast majority of cases is able to detect if a person wears a mask. Few failure cases can occur and mostly corresponds to cases where the person wears a mask in an incorrect way (e.g. showing the nose).

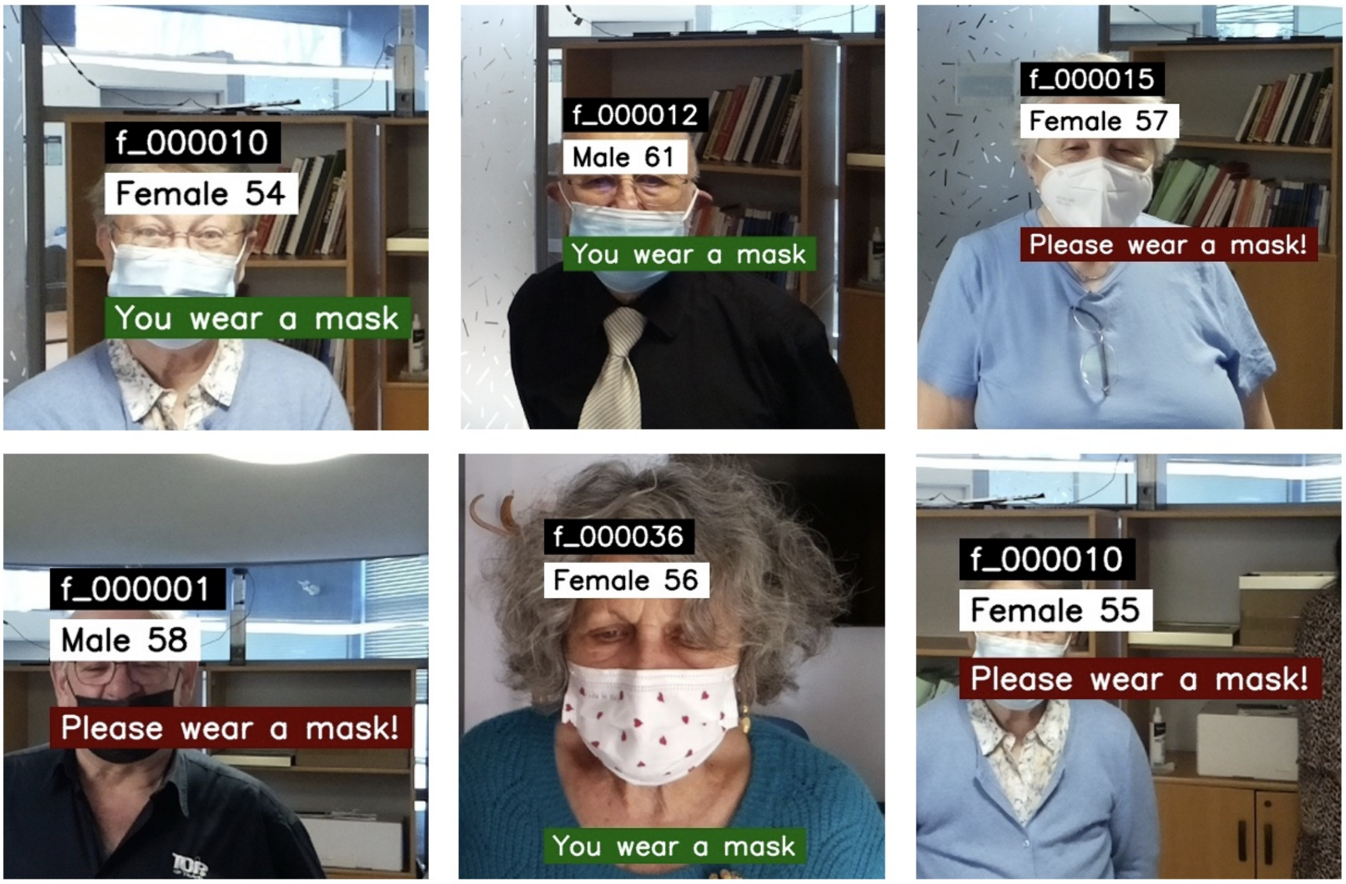

Figure 2.1: A sample of correct and wrong predictions of the face mask detector.

2.2 Biometric Recognition

The architecture of the biometric recognition module is described in D4.1. In this document, we mainly provide a qualitative evaluation of the biometric recognition due to a lack of annotations on the latest Broca dataset. Some

	Mask	No mask
Mask	**9822**	0
No mask	130	**0**

Table 2.1: Confusion matrix of our face mask detection model. Rows are predictions, columns are ground-truth.

qualitative results are presented in Figure 2.2. The results demonstrate that the proposed biometric recognition module is usually able to estimate the gender of the volunteers correctly. However, it sometimes fails to estimate their age within the acceptable range. This may be due to an unbalanced age distribution in the original training data. In addition, the quantitative results, confusion matrices, showing the performance of the gender estimation model are given in Table 2.2.

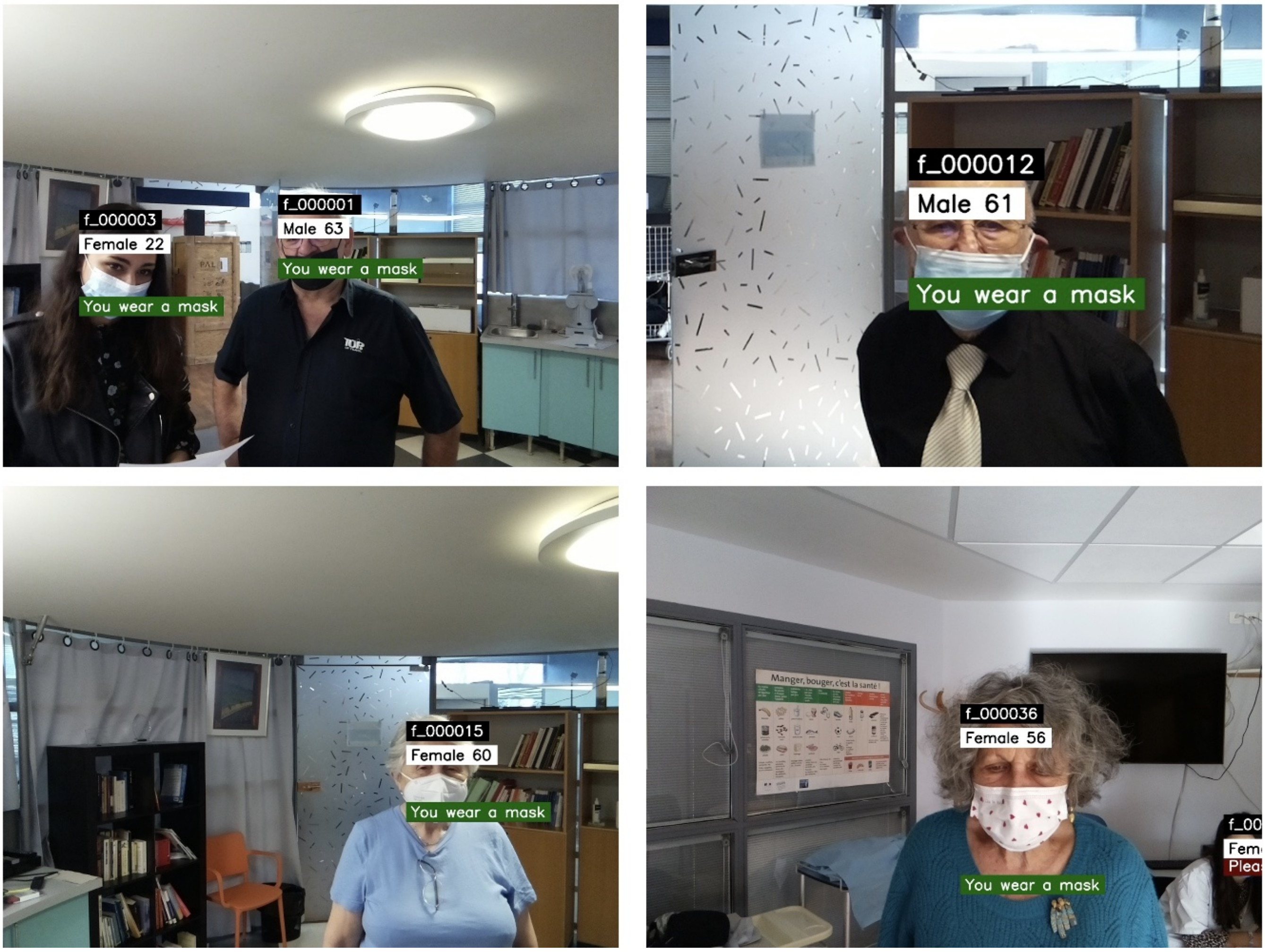

Figure 2.2: Evaluation of the biometric recognition model.

	Male	Female
Male	**4963**	43
Female	138	**2370**

Table 2.2: Confusion matrix of our gender estimation model. Rows are predictions, columns are ground-truth.

2.3 Gaze Target Detection

The gaze target detection module, described in detail in D4.3 has been trained on diverse, real-world datasets for gaze following [3][1]. To evaluate its performance on the Broca dataset, we conducted a qualitative assessment, with the results shown in Figure 2.3.

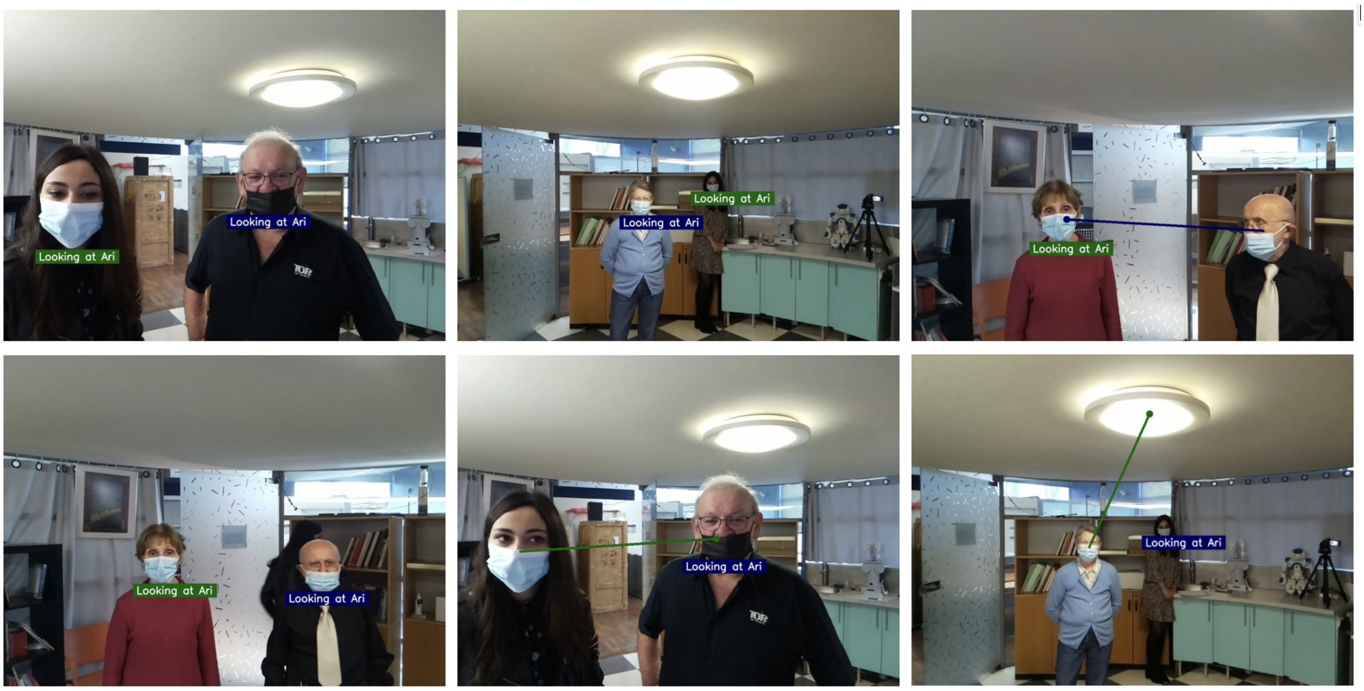

Figure 2.3: Evaluation of the gaze target detection module on multiple scenarios.

During the evaluation process, we identified areas for improvement that are currently being addressed. Notably, we observed sub-optimal predictive capabilities of the module when determining whether the observed person is looking at Ari or outside the frame. This aspect requires further refinement to enhance accuracy. Additionally, the presence of masks in the Broca dataset presented a challenge as it adversely affected the accuracy of the predicted gaze. This phenomenon can be attributed to a notable data shift between the datasets used for training the module, which does not include individuals wearing masks, and the target domain of the Broca dataset, where masks are prevalent.

To address these issues, we plan to incorporate egocentric videos from [2] into our training data. By doing so, we aim to improve the module's accuracy specifically in scenarios where the person is looking at Ari. This additional data source is expected to enrich the module's understanding of the "looking at Ari" context, enhancing its predictive capabilities in such scenarios. Furthermore, we intend to augment the training data by including images that feature people wearing masks. This augmentation strategy aims to redirect the module's focus toward the upper part of the face, even in the presence of masks. By explicitly training the module on masked individuals, we anticipate an improvement in its ability to accurately detect gaze targets, irrespective of whether masks are present.

In addition to these technical enhancements, we are committed to enhancing the evaluation process. To achieve this, we plan to provide comprehensive annotations for the Broca dataset. These annotations will enable a more nuanced evaluation, facilitating a thorough assessment of the module's performance and enabling valuable insights into its strengths and limitations.

2.4 Monocular Depth Estimation

In our evaluation, we examined the performance of the monocular depth estimation model by utilizing a subset of images extracted from the SPRING project's dataset in Broca. One distinguishing feature of this model is its capability to generate a depth map using either the head camera or the fisheye camera. To evaluate the fisheye camera, we first rectified the captured images by applying a calibration matrix obtained in our laboratory, ensuring accurate depth estimation results.

Figure 2.4 serves as a visual representation of the monocular depth estimation model's output, showcasing a sample result obtained from rectified frames captured by the fisheye camera.

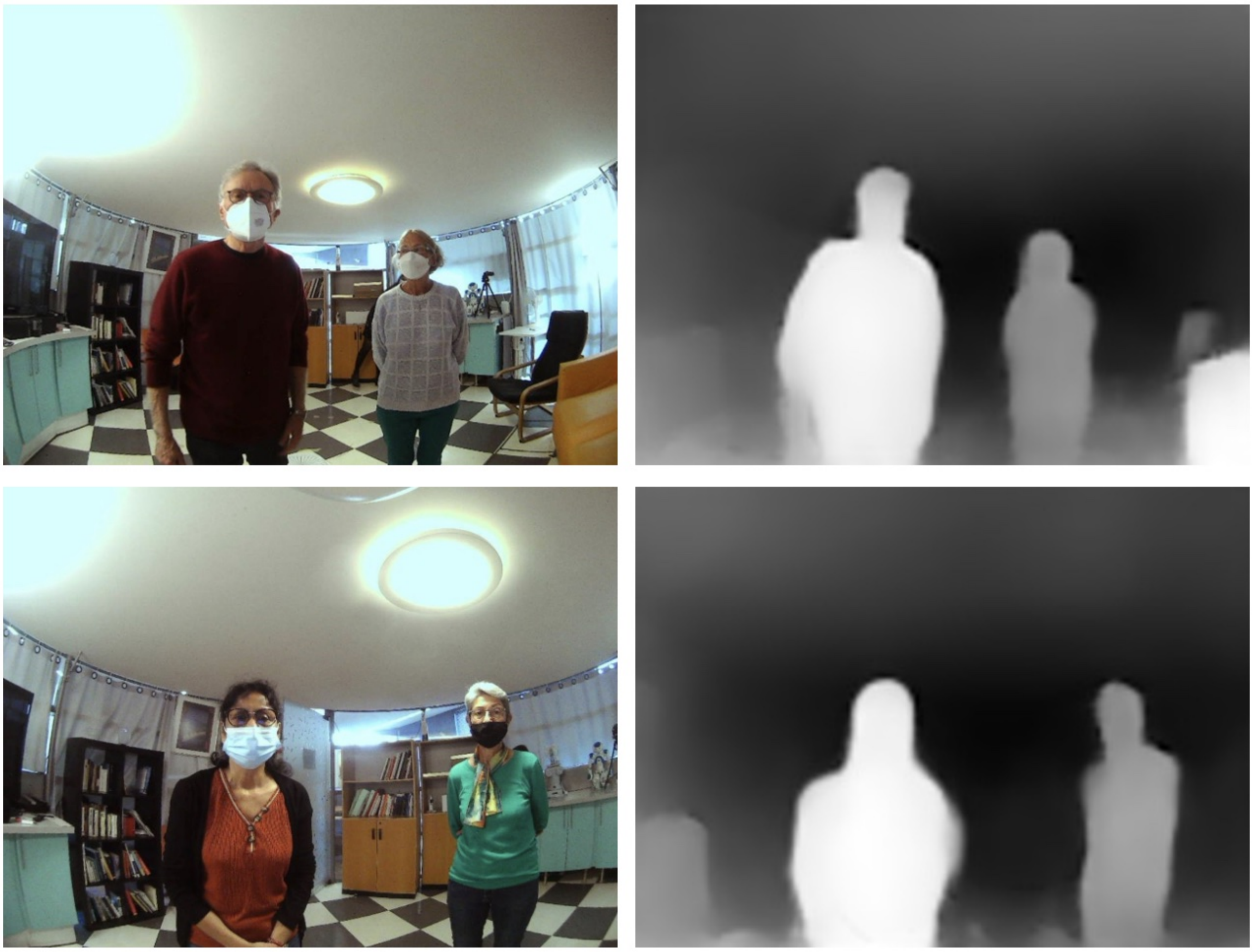

Figure 2.4: The monocular depth estimation model running on the rectified images of the front fisheye camera of ARI.

2.5 Human Pose Estimation

The details regarding the architecture of the multi-target body pose estimator can be found in D4.1.[1] The qualitative evaluation of the multi-target body pose estimation was carried out on the rectified fisheye images taken from the SPRING project dataset in Broca. Specifically, we corrected the fisheye distortion using a calibration matrix obtained in our laboratory. An example output of the multi-target body pose estimation on the Broca dataset is illustrated in Figure 2.5.

[1] https://spring-h2020.eu/wp-content/uploads/2022/02/SPRING_D4.1_Human-description-in-realistic-environments_VFinal_30.06.2021.pdf

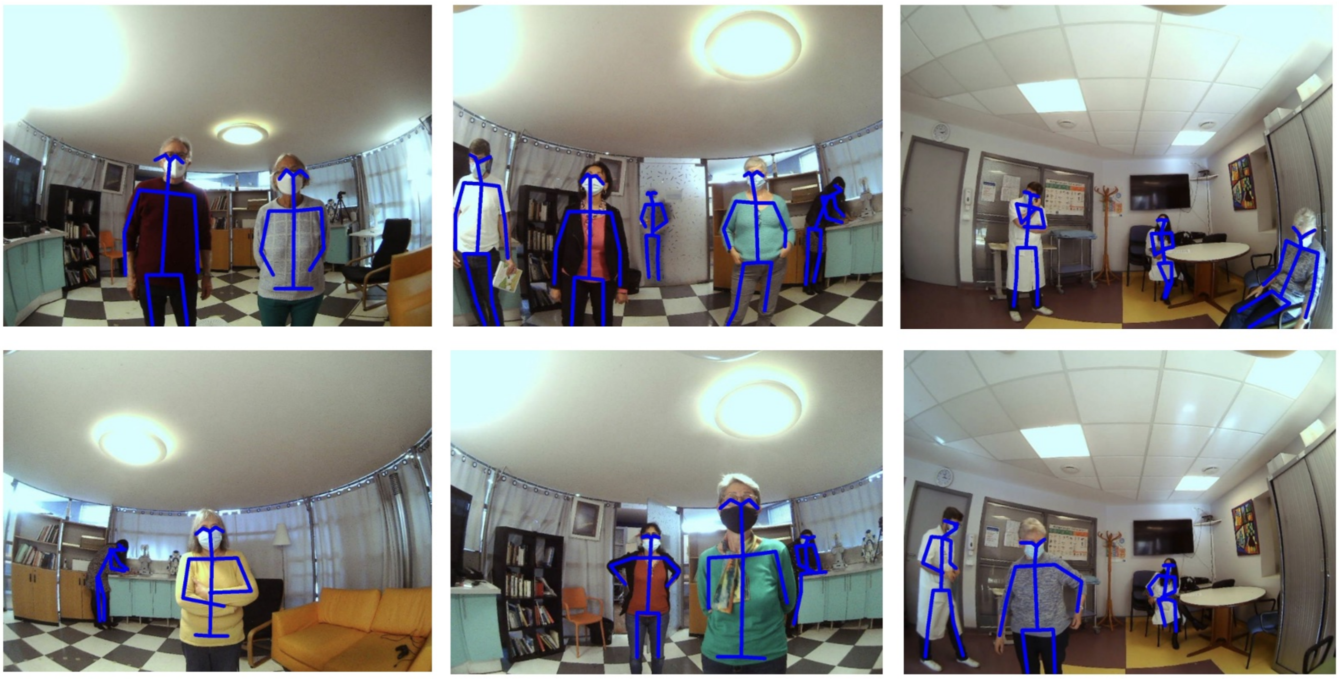

Figure 2.5: Evaluation of the multi-target body pose estimation model on Broca scenarios.

3 Conclusions

This deliverable presents the outcomes of T4.2, which is a fully functional framework for the human face and body analysis from visual data. We demonstrated the qualitative evaluation of several models related to the human face, such as **a) face mask detection**, **b) biometric recognition**, **c) gaze target detection** as well as **d) monocular depth estimation**, and **e) multi-target body pose estimation**. It is worth noting that none of the models were fine-tuned using the Broca dataset, which indicates the robustness of our solutions in unfamiliar scenarios. Furthermore, in this deliverable, we mainly provide the qualitative results of each model due to the lack of annotations in the Broca dataset. Besides, for models that are easy for humans to direct know the ground truth (*i.e.*, mask detection, gender estimation), we also provide the corresponding quantitative confusion matrices.

Bibliography

[1] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5396–5406, 2020.

[2] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[3] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.