Deliverable D4.2: Human description in relevant environments

Due Date: 30/06/2022

Main Author: UNITN

Contributors: UNITN

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	D4.2: Human description in relevant environments
Responsible Partner	UNITN
Work Package	WP4: Multi-Modal Human Behaviour Understanding
Task	Result of T4.1.
Version & Date	30/06/2022
Dissemination	Public Deliverable

CONTRIBUTORS AND HISTORY

Version	Editor	Date	Change Log
1	UNITN	23/06/2022	First Draft
2	BIU	30/06/2022	BIU added their contribution
3	UNITN	30/06/2022	UNITN added the link for Youtube
3	UNITN	04/07/2022	UNITN sent the deliverable to reviewers
4	Matthieu Py (INRIA)	07/07/2022	First cycle review was supplied
5	UNITN	13/07/2022	UNITN addressed the corrections / suggestions of Matthieu Py
6	BIU	25/07/2022	BIU addressed the corrections / suggestions of Matthieu Py

APPROVALS

Authors/editors	UNITN, BIU
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

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

Executive Summary

This deliverable, namely, D4.2 is a part of WP4 of the H2020 SPRING project. The aim of this document is to present the results of the fully functional framework for human face and body analysis from visual data, including the tools for multi-target body pose estimation and face analysis tested in relevant environments. Audio-based speaker recognition is also incorporated into this deliverable. Specifically, we present quantitative and qualitative results of the developed approaches related to T4.1 "*Describing Humans*". More importantly, the approaches were tested on environments and scenarios relevant to the SPRING project. The quantitative and qualitative analysis were made using the data collected by the SPRING project in Broca hospital as well as the dataset collected in the Robotics Laboratory of The University of Trento.

These approaches tested using the aforementioned datasets regard: **a) face mask detection**, **b) biometric recognition**, **c) gaze target recognition**, **d) monocular depth estimation**, **e) multi-target body pose estimation**, and **f) audio-based speaker recognition**. It is important to mention that the developed models: (a)-(e) were not fine-tuned on the evaluation datasets, allowing us to show the effectiveness of the proposed methodologies in unseen scenarios.

Another important aspect in "*Describing Humans*" is the recognition of a person based on his voice signature. This will allow ARI to communicate with people in a natural way. We have adopted NVIDIA's speaker identification software and wrapped it in a ROS package together with the voice activity detection. This package is incorporated into the audio pipleline. At this stage only a preliminary evaluation was carried out, demonstrating promising results.

The source code of the modules presented in this deliverable are available in the SPRING repositories[1]. As per European Commission requirements, the repositories 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`

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

1 Introduction

This deliverable **D4.2** is part of **WP4** of the H2020 SPRING project, presenting the results of T4.1: Fully functional framework for human face and body analysis from visual data. It includes tools for multi-target body pose estimation and face analysis tested in relevant environments. The framework also incorporates information from audio-based speaker recognition.

In this context we present the qualitative and/or quantitative evaluation of:

- the models related to human face analysis: face mask detection, biometric recognition, and gaze target recognition;
- monocular depth estimation method;
- multi-party body pose estimation method;
- audio-based speaker recognition implementation

on the datasets captured in relevant environments. Two datasets were used for evaluation purposes in this deliverable. These are collected by (a) the SPRING project in Broca hospital in Paris and (b) The University of Trento.

The rest of this deliverable is structured as follows: first, we describe the collection and annotation process of the dataset collected by the University of Trento as well as the rectification and post-processing applied on the dataset collected in Broca. Second, we present quantitative and qualitative results on face modules like face mask detection, biometric recognition, and gaze target recognition. Third, we qualitatively demonstrate the performance of the monocular depth estimation and the multi-target body pose estimation tested on rectified images collected by the front fish-eye camera. Following that, a description of the audio-based speaker recognition is presented. We conclude this deliverable with a summary of the results of the proposed methods on the relevant data.

2 Human Face Analysis

The models regarding human face analysis were quantitatively and qualitatively evaluated on the video clips recorded in the Robotics Laboratory of the University of Trento. During this data collection, we used a total of 8 participants (3 female, 5 male) and gave them some scenarios, which resulted in unconstrained multi-party interactions between each other as well as the ARI robot. The scenarios include introducing themselves to each other, introducing ARI to the other human agent, and gazing at objects inside and outside the field-of-view of ARI. After collecting the aforementioned dataset, we proceeded with the annotation process. We relied on external observers' annotations, i.e., used such annotations as the ground-truth data. An alternative could be using the self-assessments of the participants, however, we followed the more frequent application aka using external observers, allowing us to obtain more reliable annotations rather than relying on a single person's judgements.

Two annotators annotated the recorded videos, i.e. each frame has two independent annotations. They did the annotations regarding age, gender, and whether or not a person wears a mask. The two annotations per frame, were further compared and the frames not having consensus were discarded, i.e., were not used for the evaluation of the models.

2.1 Face Mask Detection

The description of the architecture of the face mask detection is available in D4.1[1]. In total, we collected and annotated 3.5K frames. Upon post-processing, i.e. keeping frames with matching annotations on the "wear a mask" attribute only, we evaluated the model on 2.7K frames. When evaluated on benchmark datasets, our face mask detector obtained an accuracy of 91%. When evaluated on the dataset collected by the University of Trento, our detector obtained an accuracy of 80%. Notice that we did not perform fine-tuning on the recorded data, thus showing the effectiveness of our model to unseen scenarios. The confusion matrix of the model is given in Table 2.1, and some qualitative results are presented in Figure 2.1.

	Mask	No mask
Mask	**967**	354
No mask	194	**1114**

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

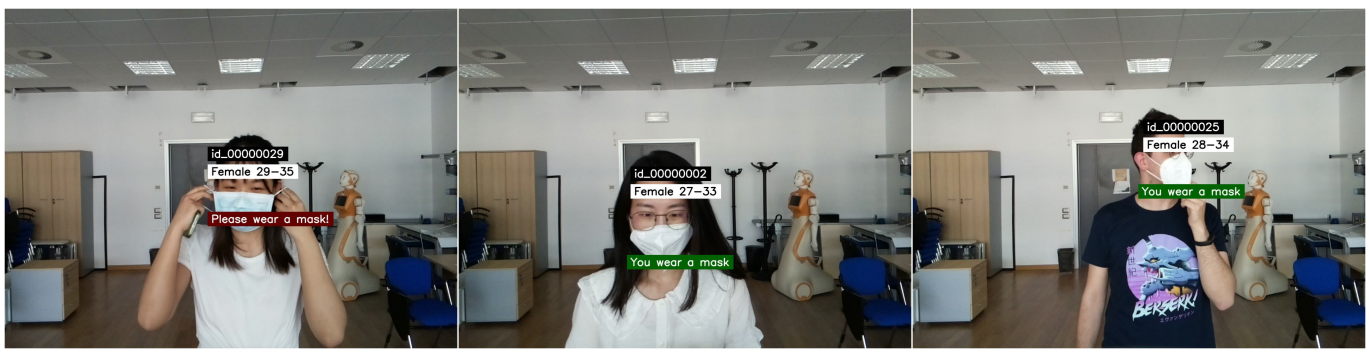

Figure 2.1: A sample of correct and wrong predictions of the face mask detector.

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

2.2 Biometric Recognition

The description of the architecture of the biometric recognition module is available in D4.1[2]. To quantitatively evaluate the biometric recognition module (i.e., gender and age estimation), likewise applied for evaluating the face mask detection model, we discarded the frames having conflicting annotations on age and gender, which resulted in 28K frames out of 35K frames. When asking for annotations regarding the age, we propose an age interval to the annotators.

The confusion matrices showing the performances of age and gender estimations models are given in Table 2.2 while Figure 2.2 presents some qualitative results. The biometric recognition module obtained an accuracy of 93% for gender estimation, and an accuracy of 70% for age estimation.

Gender	Male	Female
Male	**17343**	1135
Female	895	**7862**

Age	25 - 29	30 - 34	35 - 39	40 - 44
25 - 29	**18571**	0	7568	343
30 - 34	0	**0**	0	0
35 - 39	331	0	**422**	0
40 - 44	0	0	0	**0**

Table 2.2: Confusion matrices of gender and age estimation models of the biometric recognition module.

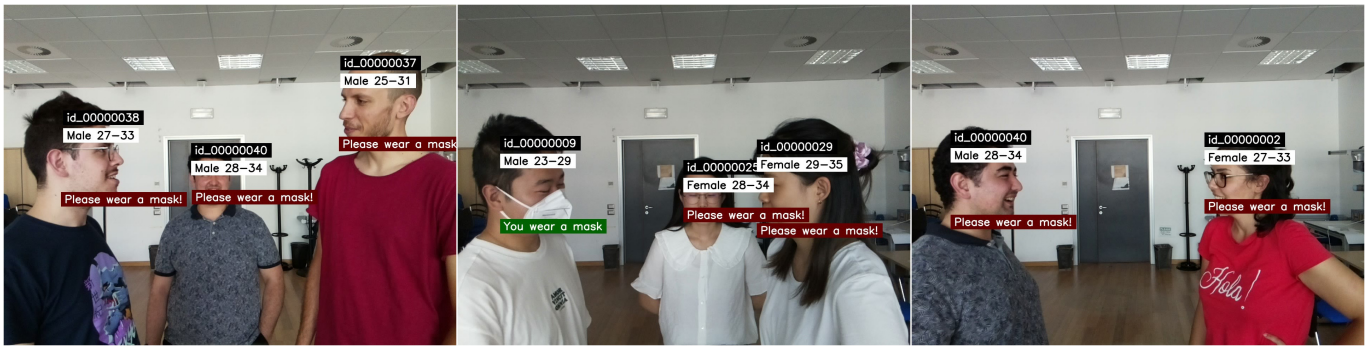

Figure 2.2: Evaluation of the biometric recognition model. The age intervals have been increased for visualisation purposes only.

2.3 Gaze Target Recognition

Figure 2.3 shows the results of our gaze target recognition module trained on an in-the-wild dataset for gaze following: GazeFollow [1]. To qualitatively evaluate this model, we used a sample of the dataset collected in the Robotics Laboratory at The University of Trento. At this stage, no fine-tuning on relevant data has been performed.

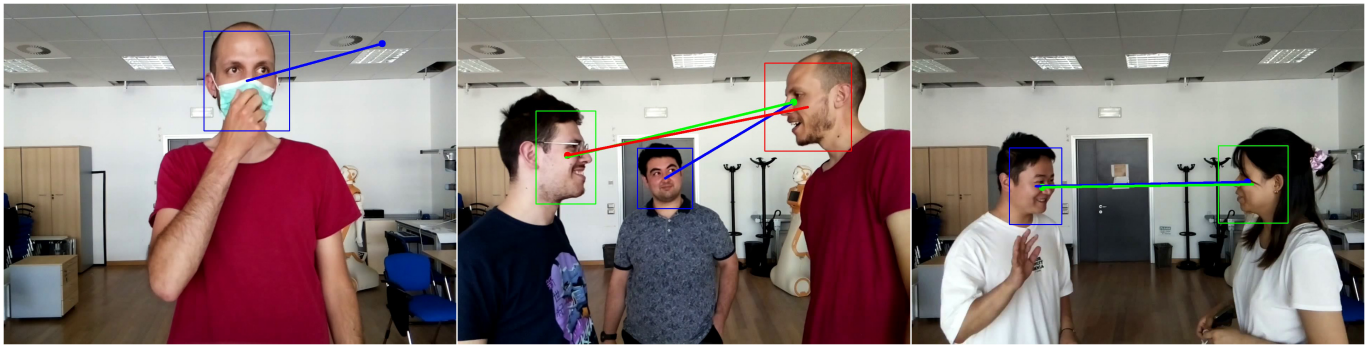

Figure 2.3: Evaluation of the gaze target recognition module on multiple scenarios.

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

3 Monocular Depth Estimation

We evaluated the monocular depth estimation model on a sample of images extracted from the dataset collected by the SPRING project in Broca. To improve the results of the model, we first rectified the images of the front fisheye camera using a calibration matrix obtained in our lab. A sample output of our monocular depth estimation model is shown in Figure 3.1. Furthermore, a video containing the output of the model on the Broca dataset is available on the YouTube channel of the project[1].

Figure 3.1: The monocular depth estimation model running on the rectified images of the front fisheye camera of ARI.

[1] https://youtu.be/8Z1-ZGnfDDU

4 Multi-target Body Pose Estimation

Figure 4.1: Evaluation of the multi-target body pose estimation model on multiple scenarios.

A description of the architecture of the multi-target body pose estimator is available in D4.1[1]. We performed a qualitative evaluation of the multi-target body pose estimation on the rectified fisheye images collected by the SPRING project in Broca. More specifically, we corrected the distortion of the fisheye using a calibration matrix obtained in our lab. A sample output of the multi-target body pose estimation on the Broca dataset is shown in Figure 4.1, and a video of the module in action is available on the YouTube channel of the project[2].

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

[2] https://youtu.be/yyDeair-qfE

5 Audio-based Speaker Recognition

Identifying the speaker is important for proper handling of the interaction between humans and ARI. Speakers can be identified based on the visual appearance but also by their voices. For example, if a person wears a mask only during a part of the interaction and removes it in other parts, the visual appearance may be confusing. Moreover, in multi-party interactions, when the speakers in the visual scene take roles in talking, it is important to keep track of the identity of the active speakers based on their voice signature. Finally, if the speaker id will be applied after the speaker separation module, the order of the speakers in each output channel may be arbitrary. Voice-based speaker id will sort out this permutation ambiguity.

The Speaker Identification ROS package (`speaker_id`) consists of two main parts:

1. Speaker embedding by NVIDIA-nemo:[1] A model trained to create a 192-dimensional voice embedding vector, representing a single-channel audio utterance at least 3s long. We used a pre-trained model.
2. Voice activity detection (VAD) algorithm by WebRTC:[2] using Gaussian Mixture Model (GMM), The package returns a Boolean value corresponding to the activity of an audio segment 30ms long (roughly refers to 1 ROS audio message with 512 samples at sampling rate of 16KHz).

The input to the `speaker_id` ROS package is the enhanced audio after MoDE ROS package processing. The output is a unique label for each active speaker. Currently, each speaker is attributed with a speaker identifier corresponding to the time they first spoke to the robot. When a new speaker utters speech, his/her voice embedding is saved in a simple database. No audio samples are saved. The algorithm runs locally, without using any cloud services.

The matching between a speaker and his voice embedding is carried out by measuring the similarity between the current speaker's embedding and any other voice embedding in the database, followed by a threshold. Currently we use a threshold of 0.75 for the similarity, and a threshold of 0.5 for the mean VAD score (averaged over 3 seconds).

The outcome of the package is published to the ROS topic `/human/speaker_id`. Code is available in SPRING repository.[3] The block diagram of the package is depicted in Fig.5.1. Preliminary results with real people speaking to

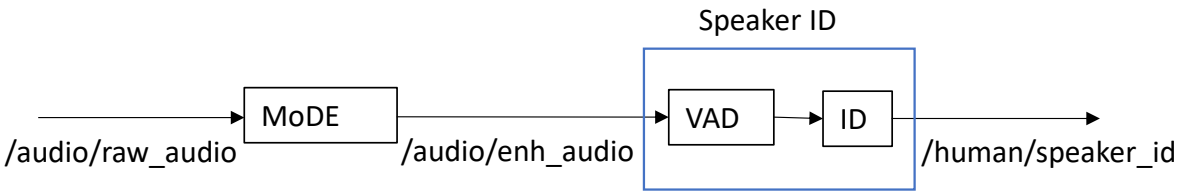

Figure 5.1: Block diagram of the ROS package `speaker_id`.

ARI in reverberating environment are promising. A more elaborated performance analysis will be carried out shortly.

[1] `https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/titanet_large`
[2] `https://webrtc.org/`
[3] `https://gitlab.inria.fr/spring/wp4_behavior/non-integrated-contributions/speaker_identification`

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

6 Conclusions

This deliverable presented the results of T4.1: Fully functional framework for human face and body analysis from visual data. We showed quantitative and qualitative evaluation of models related to human face such as **a) face mask detection**, **b) biometric recognition**, **c) gaze target recognition**, as well as **d) monocular depth estimation**, **e) multi-target body pose estimation**, and **f) audio-based speaker recognition**. The evaluations were made using the data collected by the SPRING project in Broca hospital and in the Robotics Laboratory of The University of Trento. In particular, models (a)-(e) were not fine-tuned on the mentioned datasets, thus showing the robustness of our solutions in unknown scenarios. We quantitatively evaluated the face mask detection and biometric recognition models with promising results: the face mask detection, gender, and age estimation models achieved accuracies of 80%, 93%, and 70%, respectively. Furthermore, we evaluated the performance of the gaze target recognition, monocular depth estimator, and multi-target body pose estimator and provided videos of the models running on Broca data on the YouTube channel of the project[1].

For the speaker identification task we use commercial software by NVIDIA together with voice activity detection based on WebRTC by Google. These modules were wrapped in a ROS package and incorporated into the audio pipleline. The preliminary results, tested with English utterances, are promising. Next, we will carry out a comprehensive evaluation with several speakers and languages in diverse acoustic conditions, including higher reverberation and low SNR. For that, we will record relevant data at BIU acoustic lab. After finalizing the speaker separation module we will further test the audio pipeline in complex environments including overlapping speakers. Data recorded at BROCA will be used as well.

[1] `https://www.youtube.com/channel/UC2n8oh--6_hylEhy1EviARg`

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

Bibliography

[1] 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.