The MuMMER Data Set for Robot Perception in Multi-party HRI Scenarios

Olivier Canévet[†], Weipeng He[†‡], Petr Motlicek[†] and Jean-Marc Odobez[†‡]

***Abstract*— This paper presents the MuMMER data set, a data set for human-robot interaction scenarios that is available for research purposes[1]. It comprises 1 h 29 min of multimodal recordings of people interacting with the social robot Pepper in entertainment scenarios, such as quiz, chat, and route guidance. In the 33 clips (of 1 to 4 min long) recorded from the robot point of view, the participants are interacting with the robot in an unconstrained manner.**

The data set exhibits interesting features and difficulties, such as people leaving the field of view, robot moving (head rotation with embedded camera in the head), different illumination conditions. The data set contains color and depth videos from a Kinect v2, an Intel D435, and the video from Pepper.

All the visual faces and the identities in the data set were manually annotated, making the identities consistent across time and clips. The goal of the data set is to evaluate perception algorithms in multi-party human/robot interaction, in particular the re-identification part when a track is lost, as this ability is crucial for keeping the dialog history. The data set can easily be extended with other types of annotations.

We also present a benchmark on this data set that should serve as a baseline for future comparison. The baseline system, IHPER[2] (Idiap Human Perception system) is available for research and is evaluated on the MuMMER data set. We show that an identity precision and recall of ~80% and a MOTA score above 80% are obtained.

I. Introduction

Human-Robot Interaction (HRI) requires robots to have an accurate perception of their environment to enable a continuous and natural interactions with people. Three main components are involved in the perception: visual tracking, to detect where the humans are around the robot, as well as re-identification, when a human re-enters the field of view or when a track is lost because of motion; speech localization, to first discriminate between speech and noise, and then eventually detect who is speaking; and non-verbal cues detection like nodding [1], gaze and attention [2], emotions, addressees [3] of a speech utteranace, or engagement [4]

Our work takes place in the framework of a humanoid robot (based on the Pepper platform) that interacts with the general public in a shopping mall. The robot should be able to naturally engage and entertain customers, by chatting with them, telling jokes, asking quizzes, or giving information about the shops around.

† Idiap Research Institute, Martigny, Switzerland

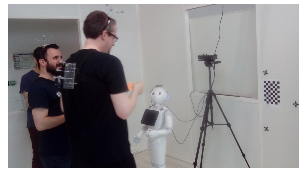
‡ École Polytechnique Fédérale de Lausanne, Switzerland.

Corresponding author: odobez@idiap.ch

This research was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement no. 688147 (MultiModal Mall Entertainment Robot, MuMMER, www.mummer-project.eu).

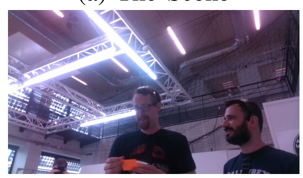
[1]https://www.idiap.ch/dataset/mummer

[2]https://www.idiap.ch/software/ihper/

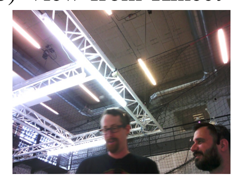


(a) The Scene

(b) View from Kinect v2



(c) View from Intel D435



(d) View from the robot

Fig. 1: An example scene from the data set and the views from the three cameras.

The robot perception in such environments faces many challenges. The illumination conditions are bad and change frequently depending on the time of the day and on how the robot is oriented; for audio perception, there are various background noise and strong reverberation due to the large space; natural and unconstrained human behaviors are difficult to interpret, and some people may act abnormally in order to fool the system; many people can be in the field of view, even if the robot is only interacting with one or two of them. Moreover, as the sensors are embedded on the humanoid robot head, the ego-motion causes images to be blurry, and people to frequently "move" in and out of the field of view, while they are static in front of the robot, which may lead to identity switches, or track loss.

In such multi-party HRI situations, the person re-identification is crucial. Identity switches are problematic because they invalidate the whole dialogue history and the interaction no longer makes sense. This issue is usually circumvented by heuristics, such as assuming the person of interest is the closest one, or assuming a constrained location of the humans (always left and right when only two persons), which no longer holds when the robot is in the field, in a shop or a shopping mall.

To facilitate the research in robot perception, in particular for the evaluation of perception approaches, we have created the MuMMER data set, which consists of audio-visual recordings of people interacting with the robot in an open environment using a Wizard-of-Oz (WoZ) approach. In each scenario, two to three participants interact with the robot (chat, quiz, guidance task) while additional persons are simply spectating. All face bounding boxes and identities were manually annotated. Extra labels, such as begin/end of interaction or user-engagement, can be easily added based on the annotated faces.

We also present a real-time audio-visual tracking system which has been developed to address the multi-party HRI task. The system efficiently tracks the faces, re-identifies them when re-entering the field of view, detects if a person is speaking, and recognizes non-verbal cues. We evaluate this system on the MuMMER data set and show that it performs well in such non-constrained scenarios.

The main contributions of this paper are (i) a new HRI data set consisting of videos and audio of a humanoid robot (Pepper) interacting with humans in various entertainment scenarios, including the manual annotation of all face locations and identities, and (ii) a tracking system along with re-identification and its benchmarking on the data set. The data set and the system are both publicly released.

II. Related data sets

In this section, we briefly present other HRI related data sets and explain the need for a new data set that covers different aspects of HRI research, especially those under more challenging conditions.

The Vernissage Corpus [5] is a data set in which two persons interact with the NAO robot in the context of an art exhibition, during which NAO presents paintings and asks the humans questions about them. Several cues are annotated such as the 3D location of the persons, the visual focus of attention, and the nodding events. The conditions in this data set are challenging because the camera is moving as the robot speaks and exhibits several patterns.

The AVDIAR data set [6] is a data set of unstructured informal meetings (27 minutes in total) where people stand and move in front of the camera. The data set contains annotations of the faces, identities, upper-bodies, and speaking activities but does not exhibit challenging situations of people leaving the field of view or occlusion.

The MHRRI data set [7] was collected to study attention and engagement in human-human and human-robot (Nao) dyadic interactions. It contains multi-modal data of participants, such as the video placed on the forehead and biosensor data. It includes people speaking about themselves and asking pre-defined questions. However, it does not contain images shot from the robot.

The UE-HRI data set [8] was collected to study the engagement of users in spontaneous HRI scenarios. It was recorded with a Pepper robot which was located in a public place and the users were free to start the interaction and to end it when they wanted. Interaction comprised different phases like consent form agreement, open questions, explanations about the robot's human detection capacity, interaction survey. The data set was manually annotated to characterize different engagement cues: sign of engagement decrease, early sign of future engagement breakdown, engagement breakdown, and temporary disengagement.

The first-person video data set [9] was collected to study interactive activity recognition, such as "shaking hands", "hugging the observer", "waving a hand". A camera was mounted on the forehead of a humanoid model (a teddy bear) placed on a rolling chair that could be moved by a human, thus simulating a moving robot.

TABLE I: Main figures of the data set

Number of participants	28
Number of protagonists	22
Number of clips	33
Shortest clip	1 min 6 s
Longest clip	5 min 6 s
Total duration	1 h 29 min
Maximum number of persons in one frame	9
Kinect color frames	80,488
Kinect depth frames	80,865
Intel color frames	80,346
Intel depth frames	80,310
Robot color frames	47,023
Robot depth frames	23,450
Number of annotated faces	506,713

As we have seen, all the data sets are limited to interactions with one or two participants, with a controlled background in which there is no people or very little (people passing by far away). In our context of a robot in a crowded public place such as a shopping mall, there is a need for a more challenging data set because we are interested in evaluating perception algorithms from a robot camera, in real situations of multiparty interaction, where the robot head is moving, when the human re-enters the field of view, and also when there are non-interacting humans in the field of view.

With this in mind, we present in the next section the MuMMER data set.

III. The MuMMER data set

The context of the data collection is an entertainment humanoid robot to be deployed in a shopping mall for several hours to interact with the customers. The envisioned use cases are among others, chatting with the customers, telling jokes, playing quiz, telling the news, and giving directions. These scenarios imply multi-party dialogue between the robot and several persons, passerby in the background, troublemakers trying to grab the attention of the robot, people leaving the field of view when the robot indicates a direction, and potentially, people coming back after a while to tell the robot about the early direction or recommendation it made earlier. These are all features can be found in the data set we present.

A. Setup and sensors

The data was gathered over two days in an open lab in which several signs of shops were displayed to be more realistic. 28 people participated in the collection, 22 of them acting as protagonists (people interacting with the robot). The recordings were performed in sequences of 1 to 5 minutes long, with either two or three protagonists and several passerby (people farther away, in the background, not speaking with the robot) in the background. In total, there are 33 short clips. Table I summarizes the main figures of the data set.

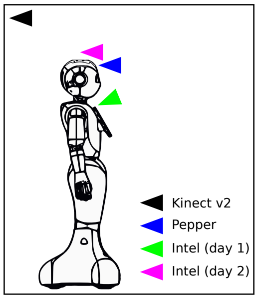


Fig. 2: Location of the cameras

We used the Pepper[3] robot. It is equipped with one frontal color camera and one Asus Xtion depth camera. These asynchronous cameras can work at 8 (resp. 5) frames per second (fps) at a resolution of 640×480. In addition, we used an Intel D435 camera placed on the top of the tablet during the first day of recording, and on the top of the head during the second day[4] which is more realistic. Finally, we put a Kinect v2 camera behind the robot to shoot the entire scene (see Fig. 2). This camera was static on a tripod. The D435 and Kinect were set to 15 fps, at a resolution 1280×720 and 960×540 respectively. Fig. 1 shows synchronized views of the 3 sensors along with a photo taken at the same time, which gives an idea of the setup.

On the technical side, all the data was recorded on one single laptop connected to the robot through an Ethernet cable and the two other cameras through USB cables. The robot operating system (ROS) framework was used, which provided inherently synchronized sensor streams with the timestamp of the machine. The data from Pepper were accessed via the NAOqi driver, kinect via IAI Kinect2[5], and Intel via realsense[6]. We recorded the video streams (color and depth), the audio, the joint states of the robot, and the 3D locations of the protagonists via a motion capture system. These streams were stored as ROS bag files (thus inherently synchronizing all data ith the timestamp of the machine). The color video streams were compressed.

This data set has the advantage of both static and moving cameras: both the robot sensors and the Intel ones are moving as the robot is moving, while the Kinect is static. The audio streams of the microphone arrays with four channels were recorded with both the Kinect and the robot at frame rate of 48 kHz.

B. Scenarios

Our use cases are interactions with customers in a shopping mall. To this end, we have designed several scenarios of interest that were played in a Wizard-of-Oz setting in an unconstrained manner. A human not visible by the participants was controlling the robot through a graphical user interface with pre-defined buttons that triggered one action of the robot (nodding, pointing, looking at a particular person, etc).

[3]https://www.softbankrobotics.com/emea/en/pepper
[4]The manufacturer of the robot forbade us to put the camera on the head at first.
[5]https://github.com/code-iai/iai_kinect2
[6]https://github.com/intel-ros/realsense

The robot interacts with the participants by first inquiring them (e.g. How are you?, Where are you from?), and then acting different scenarios in sequences like asking if it can do anything (participant replies a pre-determined answer like go for burgers, go for a coffee) or other questions that the participants were not told about, to make the conversation more natural. In all scenarios, passerby were asked to behave like curious people, who want to see what is going on, to talk with each other, to make signs to the robot, to take pictures of the scene, and to simply walk behind the protagonists. Also, in addition to actions triggered by the WoZ to conduct the interactions, including nodding, speech, or pointing, the robot was constantly moving its head in a social manner (as it is implemented as the AnimatedSpeech from NAOqi), which render the data set more challenging because of camera motion (Robot and Intel cameras). The set of scenarios that were used is provided below.

Interaction. The robot inquires about the participants (their name, where they are from, what they bought, if they are having a nice time);

Satisfaction study. Get the feedback of the customer in the shopping mall. The robot displays the usual three buttons (red, yellow, and green) on its tablet that the participant is invited to select to rate its experience in the mall. This causes the participant to come closer to the robot, touch it, and get back to his/her original location, which creates interesting for robot's perception;

Directions. The human wants to know the location of a particular place where to go: to have a coffee, a burger, to buy shoes. The robot indicates where the corresponding shop is. Sometimes, the robot asks the participant to move a little bit, so that the robot can correctly point at the place, or so that the participant can better see the direction or the target. This renders the scenario very challenging because the illumination changes, the person may go outside the field of view. Inherently, the robot moves its head a lot (i.e. the camera is moving) in this scenario;

Questions. The robot asks general and funny questions about artificial intelligence and robots, which often causes the participants to laugh and triggers gestures;

Treasure hunt. This is a small game. The robot asks the participants to get a piece of paper stuck on a wall nearby, to take it back, and read its content. This scenario causes the participants to leave the field of view, and the illumination changes as well when the human re-enters it;

Quiz. The robot asks questions like in "Who wants to be a millionaire" and the participants are asked to give to correct answer. The questions were made very hard on purpose to trigger surprise and contempt, and the participants always discussed between them about which answer to select.

C. Annotations

To properly benchmark tracking and perception algorithms, we have annotated all the faces and identities that appear in the three color video streams. This annotation process was done in two main steps: a pre-automatic one, and a manual one.

In the pre-automatic phase, we used the single-stage headless face detector [10] to detect the faces in the images. The parameters of the detector were loosened to reduce the miss detection rate (thus increasing the number of false positive detections). Then, a basic tracker was applied to group the face detections of adjacent frames in tracklets. The tracker first estimated the camera movement with the CMotion2D software[7] [11] to cancel the visual motion due to the robot's head rotations, then performed association only based on a tight intersection-over-union threshold to avoid any wrong identity merge. This process led to pure trackets.

In the manual phase, a human was asked to merge the tracklets together. To this purpose, the human was presented five representative images of two tracklets at a time, and was asked to select whether the two trackets belonged to the same person, to different persons, or if the tracklet was made of false positive detections (tracklet to remove from the data set). The tracklets were presented in a decreasing order of "probability matching": Using OpenFace [12] as a feature extractor, we presented the next couple of tracklets that contained the minimum pair-wise distance between the OpenFace features, which corresponds to two faces that were close in terms of features. This strategy enabled the annotator to click on "merge" most of the time, thus facilitating fast annotation process. Finally, when all the tracklets were merged, all the annotations were checked with a modified version of the LabelMe[8] software, to add or remove additional bounding boxes. The identities are consistent across all the recordings: a participant has the same identity in all the frames (all sensors, all sessions).

Annotation statistics. In this workflow, the 518,294 predetected faces were grouped in 30,872 pure tracklets, and were merged and validated in roughly 60 hours, yielding 506,713 faces at the end (from all 3 sensors). Figure 3 shows a histogram of the number of faces per frame for each sensor. Most frames have 2 persons while 25% of them have at least 1 spectator not interacting with the robot. The annotations are in the MOT challenge format, making it straightforward to use and evaluate with their evaluation code.

This data set can easily be further annotated to study other HRI elements such as user engagement, turn taking, begin and end of interaction [8], engagement willingness [7]. These new annotations could be events (beginning, end) and can be easily obtained as the most difficult part of the annotation process (face detection and identity naming) is already done.

IV. Audio-Visual perception for HRI

This section presents the modules of our perception system[9] that is used as a baseline. The system consists of six main components:

- Body joint detection,
- Head pose estimation,
- Head pose tracking,
- Face re-identification,
- Sound source localization,
- Fusion of all the previous modules.

[7]http://www.irisa.fr/vista/Motion2D/index.html
[8]https://github.com/wkentaro/labelme
[9]https://www.youtube.com/watch?v=Cfsc0zXAMVU

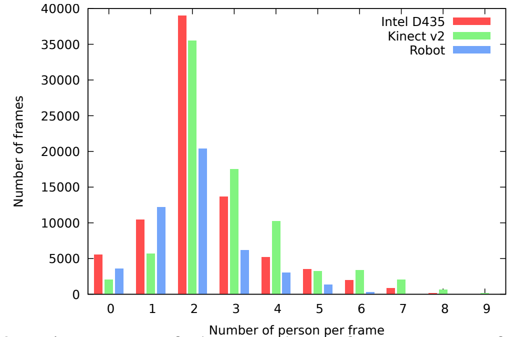


Fig. 3: Histogram of the number of person per frame for each sensor

The code[2] is available for research purpose and is platform independent. It can run on a simple RGB-Depth (RGBD) camera accessed via ROS. The system is analog to other systems like [13], [14], and [15].

A. Visual face tracking

The first part of the system is the detection of people using the convolutional pose machines (CPM) [16] which outputs the body joints. This detector is very robust at the distance concerned here: less than 2 meters for the interaction, and up to 5 meters before the interaction starts, when the robot tries to grab the attention of the detected persons.

When a person is detected, the head pose is estimated by leveraging the output activation maps of the CPM as described in [17]. Given the location of the nose and eyes, the activation maps are cropped and pass to an estimator which provides the roll, pitch, and yaw angles of the head, with an error of less than 7 degrees.

The face locations and the head-poses are used as input to the multi-person head pose tracker described in [18] which tracks the faces by combining a priori texture and color models, and manages creation and deletion of tracklets based on a sound probabilistic framework [19]. The tracker provides consistent identities of faces across consecutive frames, and as long as a person remains in the field of view of the robot. When a person re-enters the field of view, or after a tracklet was lost (due to fast movements of the head for instance), (s)he is assigned a new identity which has to be managed by the face re-identifier described below.

So when a new frame arrives, faces are detected and the visual tracker either extends current tracklets (and their current identity) to the new face, or create a new tracklet (with a new identity).

B. Face re-identification

When a track is lost (for instance when the robot moves its head too fast when speaking, or when it turns its head for pointing), the person gets a new track identity when the tracking is resumed. There is then a need to associate the new identity with the previous one, so that the correct history can be maintained.

The re-identification is done the following way. At time t, when a face f_j is tracked according to the visual tracker (and is associated with the identity label y_j), we compute the OpenFace [12] features x_j of this face and compare it (Euclidean distance) to the features of all presenteds faces encountered so far. When the distance between face f_j (therefore represented by (x_j, y_j)) and a face (x_i, y_i) of the gallery is lower than a re-identification threshold, we consider it as a match and increment the match counter C_{y_i,y_j} between identities y_j and y_i. When this counter reaches a certain amount of matches, the face and its associated history (current tracklet and overal track) with id y_j are re-indentified with the identity y_i, and a bookkeeping step of the gallery and counter is performed. Otherwise, if none of the existing id y_i are such that C_{y_i,y_j} is above a threshold, the face remains with its current id y_j.

Algorithm 1 summarizes more formally this re-identification procedure. Let $x_j \in \mathbb{R}^{128}$ be the OpenFace features computed on face f_j, and y_j the identity of this face according to the visual tracker. At time t, a tracklet $T_j = \{x_1, \ldots, x_{n_j^{(t)}}\}$ is a set of $n_j^{(t)}$ features. We note

$$\mathcal{G} = \left\{(x_i, y_i) \in \mathbb{R}^{128} \times \mathbb{N}\right\}$$

the gallery of accumulated OpenFace features that we update with each new detected face.

When the Euclidean distance between the face $x_j^{(t)}$ of the current tracklet and a face x_i in the gallery is lower than a re-identification threshold τ, then we increment C_{y_i,y_j}, which accumulates the number of matches between identity y_i and y_j. As the tracking goes on, potentially more matches are made, and we consider the tracklet with number y_j to be of identity y_{i^*} as soon as we have $C_{y_{i^*},y_j} > \Lambda$.

Note that the re-identification of a new tracklet can take several frames, depending on how many features (faces) associated with the correct identity in the past have been stored and how many matches each single face of the tracklet is getting. The longer a person has been interacting with the robot, the more features of this person has been stored, therefore the faster the re-identification will be.

C. Sound source localisation and audio-visual fusion

We briefly describe the audio part of the system for completeness although it is not used in the evaluation part V. To detect who is speaking, our system integrates the framework for multiple speaker detection and localization using deep neural networks introduced in [20], [21]. This framework processes audio in time frames of 170 ms and detects sounds sources in the azimuth directions.

The fusion of the audio and visual part is done by doing pairwise comparison of the face directions and the sound directions. A face direction and a speech direction are matched when the angle between them is lower than a tolerance angle, typically 10 degrees.

TABLE II: Detection accuracy

Sequence	TP	FP	FN	Recall	Precision
Easy (Intel)	1836	19	29	98.5	99.0
Easy (Kinect)	2009	53	76	96.4	97.4
Easy (Robot)	907	54	48	95.1	94.5
Hard (Intel)	15159	739	2769	85.0	95.5
Hard (Kinect)	18936	1274	5412	78.1	93.8
Hard (Robot)	6643	413	1486	82.7	94.5
Overall	446130	16606	52837	89.6	96.5

TP: true positive; FP: false positive; FN: false negative.

V. Evaluation

This section presents a benchmark of our audio-visual perception system on the MuMMER data set introduced in this paper. The MuMMER data set contains 33 clips, but for the sake of clarity and analysis, in addition to the overall results we will also present the result on an "easy" and a "hard" sequences. The "easy" sequence (see figure 4) consists of two protagonists interacting with the robot and nobody is in the background, while the "hard" sequence has three protagonists and up do six persons in the background, simulating complex scenarios in shopping malls.

A. Evaluation of the detection

We first analyze the performance of the face detection on the data set. In our system, we used the OpenPose framework to detect the body joints, and the face is extrapolated based on the locations of the nose, eyes, and ears. Evaluating the face detection alone shows how good the tracking can be later on. We use an intersection-over-union score of 0.3 to match a detection with the ground truth, as suggested in [22] for faces.

Table II shows the detection accuracy for the two selected sequence as well as for the entire data set. As stated earlier, the CPM algorithm has a very high accuracy to detect bodies in our context where upper body is visible and not too small because people are closer than 5 meters. The detection has a precision of 96.5% and a recall of 89.6% overall.

B. Metrics used for tracking evaluation

To evaluate the tracking, we use the usual metrics of the MOT challenge [23] which uses the MOTA score [24] defined as follows:

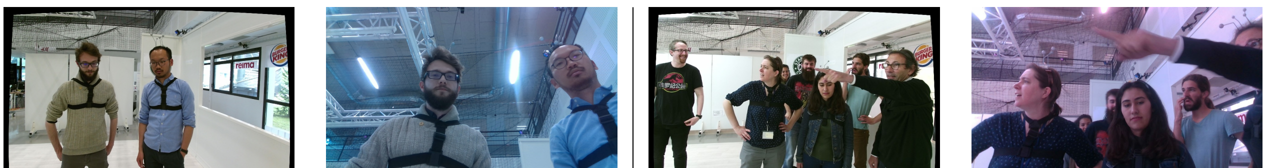
$$MOTA = 1 - \frac{\sum_t (FN_t + FP_t + IDs_t)}{\sum_t GT_t}, \quad (1)$$

where t is the frame index, FN the number of false negatives, FP the number of false positives, and IDs the number of identity switches. The MOTA score combines basics types of errors that are done by a tracker, such as the miss detections (FN), the wrong detections (FP), and the failure to keep a consistent identity across adjacent frames. One criticism usually made about MOTA is that is under-represent identity switches as they are much fewer event of that sort compared to false negatives and false positives.

In our case, we are also interested in evaluating if the tracker was able to re-identify a person when re-entering the field of view (i.e. re-assigning the previous identity), not

Algorithm 1 Re-identification procedure

$\mathcal{G} = \left\{(x_i, y_i), \in \mathbb{R}^d \times \mathbb{N}\right\}$: the gallery features x_i identified by y_i ($\emptyset$ at start).
$\mathcal{Y} = \{y_i\}$: the set of currently valid identities
$C = \left\{C_{y_i,y_j}, y_i < y_j\right\}$: The number of matches between identities y_i and y_j (0 at start).
τ: The re-identification threshold
Λ: The merging threshold
for each new frame at time t **do**
  for each tracked face j represented by $\left(x_j^{(t)}, y_j\right)$ **do** ▷ y_i face identity according to the visual tracker
    $\mathcal{M} = \left\{(x_i, y_i) \in \mathcal{G}, \left\|x_i - x_j^{(t)}\right\| < \tau\right\}$ ▷ Compute matches between $x_j^{(t)}$ and accumulated ones
    $\forall y_i \in \mathcal{Y}, y_i < y_j, C_{y_i,y_j} \leftarrow C_{y_l,y_j} + |\{(x_l, y_l) \in \mathcal{M}, y_l = y_i\}|$ ▷ Increment cumulative matches
    $\mathcal{G} \leftarrow \mathcal{G} \bigcup \left(x_j^{(t)}, y_j\right)$ ▷ Update gallery with current face $\left(x_j^{(t)}, y_j\right)$
    if $\exists\ i^*, C_{y_{i^*},y_j} > \Lambda$ **then** ▷ Tracklet identity y_j has a match with identity y_i^*
      $\forall (x_i, y_i) \in \mathcal{G}, y_i = y_j, (x_i, y_i) \leftarrow (x_i, y_{i^*})$ ▷ Merge identities y_{i^*} and y_j in gallery $\mathcal{G}$
      $y_j \leftarrow y_{i^*}$ ▷ Current tracklet is re-identified as y_{i^*}
    end if
  end for
end for



Easy sequence (Kinect) | Easy sequence (Intel) | Hard sequence (Kinect) | Hard sequence (Intel)

Fig. 4: Illustration of the two sequences used to present the results

only if the new identity was kept consistent after the tracking started again (the MOTA score only takes into account an identity switch, but does not account for the fact that the assignment was correct). We therefore also take into account the precision (IDP), recall (IDR), and F1 (IDF1) scores of the identity assignments. For instance, an IDP of 90% means that for a track of 100 frames produced by the algorithm, 90 frames corresponds to the same person, and 10 to other person IDs.

C. Evaluation of the tracking alone

The tracker presented in section IV-A provides consistent identities across contiguous frames (tracklets) but inconsistent after a person was lost. We first evaluate this tracker alone to have a clear understanding of what the re-identification part brings later on.

Table III (part "Tracker") presents the tracking scores of the tracker alone. As expected, the identity assignment score (IDP, IDR, and IDF1) are low (<30%) because this tracker does not perform identity re-assignment, but has a reasonable MOTA score (>80%) which shows that the tracking is good once a target is tracked.

D. Evaluation of the re-identification alone

We are also interested in evaluating the re-identification part. Since this part tries to reassign a tracklet identity to a previously seen identity, its performance depends a lot on the quality of the detections of the tracker. To remove

TABLE III: Evaluation of the modules (metrics described in V-B)

Sequence	IDF1	IDP	IDR	IDs	MOTA
Tracker (presented in IV-A)					
Easy (Intel)	72.7	72.9	72.5	6	97.1
Easy (Kinect)	96.2	96.7	95.6	3	93.7
Easy (Robot)	37.6	37.5	37.7	26	87.0
Hard (Intel)	10.0	10.6	9.4	516	78.2
Hard (Kinect)	16.5	18.2	15.1	338	71.5
Hard (Robot)	9.0	9.7	8.4	465	72.5
Overall	39.1	40.6	37.7	7746	84.8
Re-identification (presented in IV-B)					
Easy (Intel)	99.3	99.3	99.3	1	99.9
Easy (Kinect)	96.9	96.9	96.9	2	99.9
Easy (Robot)	92.9	92.9	92.9	7	99.3
Hard (Intel)	89.2	89.2	89.2	144	99.2
Hard (Kinect)	89.3	89.3	89.3	117	99.5
Hard (Robot)	87.3	87.3	87.3	105	98.8
Overall	90.4	90.4	90.4	2640	99.5
Full system					
Easy (Intel)	96.7	96.9	96.4	5	97.2
Easy (Kinect)	96.2	96.7	95.6	3	93.7
Easy (Robot)	60.4	60.2	60.6	31	86.4
Hard (Intel)	75.0	79.6	70.9	383	78.9
Hard (Kinect)	77.7	85.5	71.2	172	72.2
Hard (Robot)	46.9	50.2	43.9	454	72.6
Overall	82.8	86.0	79.8	5869	85.1

IDF1: F1 score; IDP: precision; IDR: recall; IDs: identity switches.

this dependency, we have used the ground truth detections and built new tracklets corresponding a perfect detection and association of detection in adjacent frames: if an identity appears in frames 100 to 110, and then from frame 120 to 130, we have one first tracklet with ID 1 (for instance) from frame 100 to 110, and a second tracklet with ID 2 from frame 120 to 130. We want to evaluate if the re-identifier is able to properly re-assignment ID 2 to ID 1.

Table III (part "Re-identification") presents the performance of the re-identifier alone. Since we used the ground truth detections, there are neither false positives nor false negatives, so the IDR, IDP, and IDF1 scores are the same. The re-identifier is properly able to re-identify the faces reaching precision and recall of roughly 90%.

E. Evaluation of the full system

Finally, Table III (part "Full system") presents the overall results of the combination of the tracker and the re-identifier. The re-identification brings a huge improvement over the tracker: the IDP goes from 40.6 to 86.0, and the IDR from 37.7 to 79.8.

VI. Conclusion

We have introduced a new HRI data set in the context of people interacting with a social robot. The data set is available for research purposes[10]. The data set contains color and depth videos of three cameras shooting the interactions from the robot's point of view. The humans are interacting with the robot in entertainment scenarios (quiz, chat, route guidance). The data set contains annotations of the face and identities of the person for a total of 506,713 faces and 28 identities.

We have used this data set to benchmark our audio-visual perception system which consists of a head pose tracker, a face re-identifier, and sound source localizer, and a module performing audio-visual fusion. The system was evaluated on the tracking and re-identification part.

With these results in an unconstrained environment, we have a system which is able to properly perceive who is in front of the robot and to re-identify them correctly 80% of the time, showing that there is still work to be done. Our code is available for research purposes[11].

Acknowledgment

We thank Rachid Alami, Aurélie Clodic and their team from LAAS for welcoming and helping us in the data collection.

References

[1] C. Chen, Y. Yu, and J.-M. Odobez, "Head nod detection from a full 3d model," in *ICCV Workshop,*, 2015.

[2] S. Sheikhi and J. Odobez, "Combining dynamic head pose and gaze mapping with the robot conversational state for attention recognition in human-robot interactions," *Pattern Recognition Letters*, vol. 66, pp. 81–90, Nov. 2015.

[10] `https://www.idiap.ch/dataset/mummer`

[11] `https://www.idiap.ch/software/ihper/`

[3] D. Jayagopi and J.-M. Odobez, "Given that, should i respond? contextual addressee estimation in multi-party human-robot interactions," in *Int. Conf. HRI*, 2013.

[4] D. Klotz, J. Wienke, J. Peltason, B. Wrede, S. Wrede, V. Khalidov, and J.-M. Odobez, "Engagement-based multi-party dialog with a humanoid robot," in *SIGDIAL Conference*, 2011.

[5] D. B. Jayagopi, S. Sheiki, D. Klotz, J. Wienke, J. Odobez, S. Wrede, V. Khalidov, L. Nyugen, B. Wrede, and D. Gatica-Perez, "The vernissage corpus: A conversational human-robot-interaction dataset," in *Int. Conf. on Human-Robot Interaction*, 2013, pp. 149–150.

[6] I. D. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-visual speaker diarization based on spatiotemporal Bayesian fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 5, pp. 1086–1099, 2018.

[7] O. Celiktutan, E. Skordos, and H. Gunes, "Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement," *IEEE Transactions on Affective Computing*, pp. 1–1, 2018.

[8] A. Ben-Youssef, C. Clavel, S. Essid, M. Bilac, M. Chamoux, and A. Lim, "Ue-hri: A new dataset for the study of user engagement in spontaneous human-robot interactions," in *Int. Conf. on Multimodal Interaction*, 2017, pp. 464–472.

[9] M. S. Ryoo and L. Matthies, "First-person activity recognition: What are they doing to me?" in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Portland, OR, June 2013.

[10] M. Najibi, P. Samangouei, R. Chellappa, and L. S. Davis, "Ssh: Single stage headless face detector," in *Int. Conf. on Computer Vision*, 2017, pp. 4875–4884.

[11] J.-M. Odobez and P. Bouthemy, "Robust multiresolution estimation of parametric motion models," *Journal of visual communication and image representation*, vol. 6, no. 4, pp. 348–365, 1995.

[12] B. Amos, B. Ludwiczuk, M. Satyanarayanan, *et al.*, "Openface: A general-purpose face recognition library with mobile applications," *CMU School of Computer Science*, vol. 6, 2016.

[13] Y. Wang, J. Shen, S. Petridis, and M. Pantic, "A real-time and unsupervised face re-identification system for human-robot interaction," *CoRR*, vol. abs/1804.03547, 2018.

[14] K. Koide, E. Menegatti, M. Carraro, M. Munaro, and J. Miura, "People tracking and re-identification by face recognition for rgb-d camera networks," in *Eur. Conf. on Mobile Robots (ECMR)*, 2017.

[15] A. Zaraki, M. Pieroni, D. De Rossi, D. Mazzei, R. Garofalo, L. Cominelli, and M. B. Dehkordi, "Design and evaluation of a unique social perception system for human–robot interaction," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 9, no. 4, pp. 341–355, 2017.

[16] Y. Cao, O. Canévet, and J. Odobez, "Leveraging convolutional pose machines for fast and accurate head pose estimation," in *Int. Conf. on Intelligent Robots and Systems (IROS)*, 2018, pp. 1089–1094.

[17] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, July 2017, pp. 1302–1310.

[18] V. Khalidov and J.-M. Odobez, "Real-time multiple head tracking using texture and colour cues, Tech. Rep. Idiap-RR-02-2017, February 2017.

[19] S. Duffner and J.-M. Odobez, "A track creation and deletion framework for long-term online multiface tracking," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 272–285, Jan. 2013.

[20] W. He, P. Motlicek, and J. Odobez, "Deep neural networks for multiple speaker detection and localization," in *Int. Conf. on Robotics and Automation (ICRA)*, May 2018, pp. 74–79.

[21] W. He, P. Motlicek, and J.-M. Odobez, "Joint localization and classification of multiple sound sources using a multi-task neural network," in *Interspeech*, 2018, pp. 312–316.

[22] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, "Face detection without bells and whistles," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 720–735.

[23] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "MOT16: A benchmark for multi-object tracking," *arXiv:1603.00831 [cs]*, Mar. 2016, arXiv: 1603.00831. [Online]. Available: http://arxiv.org/abs/1603.00831

[24] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa, and P. Soundararajan, "The clear 2006 evaluation," in *Multimodal Technologies for Perception of Humans*, R. Stiefelhagen and J. Garofolo, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 1–44.