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. 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[4] (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, addressee [3] of a speech utterance, 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).
https://www.idiap.ch/software/ihper/
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 MHRR 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.

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, passersby 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 passersby (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.

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

TABLE I: Main figures of the data set
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, passersby 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 [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 tracklets.

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 tracklets 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 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 pre-detected 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 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.

The code 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 person.

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

---

[3] https://www.youtube.com/watch?v=CfsCQzXAMVU
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 presented 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 overall track) with id $y_j$ are re-identified 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 $T_j = \{x_{j}^{1}, \ldots, x_{j}^{(n)}\}$ is a set of $n_j^{(t)}$ features. We note $\mathcal{G} = \{(x_i, y_i) \in \mathbb{R}^{128} \times \mathbb{N}\}$ 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 $\tau$, 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 $y_j$ as soon as we have $C_{y_i, y_j} \geq \Lambda$.

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.

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:

$$MOTA = 1 - \frac{\sum_{t} (FN_t + FP_t + IDS_t)}{\sum_{t} GT_t},$$

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

<table>
<thead>
<tr>
<th>Sequence</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy (Intel)</td>
<td>1386</td>
<td>19</td>
<td>29</td>
<td>98.5</td>
<td>99.0</td>
</tr>
<tr>
<td>Easy (Kinect)</td>
<td>2009</td>
<td>53</td>
<td>76</td>
<td>96.4</td>
<td>97.4</td>
</tr>
<tr>
<td>Easy (Robot)</td>
<td>907</td>
<td>54</td>
<td>48</td>
<td>95.1</td>
<td>94.5</td>
</tr>
<tr>
<td>Hard (Intel)</td>
<td>15159</td>
<td>739</td>
<td>2769</td>
<td>85.0</td>
<td>93.5</td>
</tr>
<tr>
<td>Hard (Kinect)</td>
<td>18936</td>
<td>1274</td>
<td>5412</td>
<td>78.1</td>
<td>93.8</td>
</tr>
<tr>
<td>Hard (Robot)</td>
<td>6643</td>
<td>413</td>
<td>1486</td>
<td>82.7</td>
<td>94.5</td>
</tr>
<tr>
<td>Overall</td>
<td>446130</td>
<td>16606</td>
<td>52837</td>
<td>89.6</td>
<td>96.5</td>
</tr>
</tbody>
</table>

TP: true positive; FP: false positive; FN: false negative.
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 correspond 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
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. 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 finding). 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 purpose.

ACKNOWLEDGMENT

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

REFERENCES


https://www.idiap.ch/dataset/nummer
https://www.idiap.ch/software/ihper/