2022 IEEE International Conference on Robotics and Automation (ICRA)
May 23-27, 2022. Philadelphia, PA, USA

Robot Learning Physical Object Properties from Human Visual Cues: A novel approach to infer the fullness level in containers

Nuno Ferreira Duarte[1], Mirko Raković[1,2] José Santos-Victor[1]

***Abstract*— For collaborative tasks, involving handovers, humans are able to exploit visual, non-verbal cues, to infer physical object properties, like mass, to modulate their actions. In this paper, we investigate how the different levels of liquid inside a cup can be inferred from the observation of the movement of the person handling the cup. We model this mechanism from human experiments and incorporate it in an online human-to-robot handover. Finally, we provide a new dataset with human eye+head+hand motion data for human-to-human handovers and human pick-and-place of a cup with three levels of liquid: empty, half-full, and full of water. Our results show that it is possible to model (non-verbal) signals exchanged by humans during interaction and classify the level of water inside the cup being handed over.**

I. Introduction

Perceiving the physical characteristics (e.g. mass) of objects manipulated by others is often important to prepare our own actions, as during the handover of heavy objects. Humans can infer such properties, even when they are not visually observable, through the analysis of the motor behaviour of another human handling an object [1]. Understanding the existence of water, or any liquid, in a cup has been a challenging problem in computer vision and robotics. There have been attempts at training large neural networks to classify the level of liquid from a single RGB image [2], [3], or RGB-depth cameras [4], [5], [6]. Alternatively, other approaches required pouring liquid in a cup to detect the liquid level [7], [6], [5], [8]. However, some challenges are extremely difficult to handle, such as occlusions, transparency of the liquid, different types of cups, colors, and the most important case, opaque cups. Most existing approaches struggle with opaque cups as you might not get the chance to view the cup from an advantageous angle that allows to view the liquid inside [2], [4], or get the robot to manipulate the cup prior [5], [6], [7], [8]. In most cases the cup or object is handed to the robot without any prior knowledge. The problem we are tackling is different from before. Our aim is not through direct visualization of the cup but through human observation and human motion features. We propose a novel perspective that takes into account the human side of the equation. Experiments have shown that human subjects are capable of estimating the weight of an object, through the observation of other people lifting objects with different weights [9]. Wei et al. [10] used human gaze direction to infer the action being performed. We argue that it is possible to understand the fullness level of a cup, to a certain extent, by observing others manipulating it. Previous approaches from Lastrico et al. [11] and Duarte et al. [12] observed the human kinematic motion during manipulation of cups with or without water to propose models for classifying carefulness motion strategies. Our proposal aims at studying the human-to-human handovers of cups with different water levels [13] and explore the non-verbal gaze cues shared by humans during manipulation. Humans tend to fixate the gaze direction in the regions that are most relevant concerning the executed action [14]. Therefore, exploring the eye-gaze cues should provide us with the relevant information for classifying water levels in cups. Eye-gaze direction is analysed rapidly, automatically and can trigger reflexive shifts of an observer's visual attention. However, understanding another individual focus of attention involves more than simply analysing their gaze direction [15]. This paper addresses the importance of cues from gaze direction. The authors also suggest that head orientation makes for a better approach in capturing others' direction of attention. Although we agree with the author's comments, as it is more challenging to extract eye-gaze information, previous works have accomplished it [16] and there are advantages in doing so. The work presented in this paper explores just that. Castelhano et al. [17] expand on the idea that during real-world scenarios human eye-gaze cues are influenced by others during interactions. They showed that during object manipulation humans fixate the actor's face as well as the object, which is in line with the experiments presented in this work. There are previous works that have shown the importance of robot-to-human eye-gaze movements [18], [19], [20], while for human-to-robot eye-gaze movements informs whether the human is engaged [21], [22], [23]. Huang et al. [24] use eye-gaze information using a head-mounted eye-tracker to predict the ingredients chosen for making a sandwich. Duarte et al. [25] integrated eye-gaze cues for both humans and humanoids for the robot to correctly adapt to the human, by decoding human action from eye-gaze cues and updating the humanoid's eye-gaze cues to express action understanding.

The contributions of this paper are fourfold: (i) a publicly available dataset of human head and wrist body motion plus head-mounted eye-tracker data for human-to-human handovers of a cup in three conditions (empty, half-full, and full); (ii) the analysis of human-to-human handovers,

[1]N.F. Duarte, M. Raković, and J. Santos-Victor are with Vislab, Institute for Systems and Robotics, Instituto Superior Técnico, Universidade de Lisboa, Portugal {nferreiraduarte, rakovicm, jasv}@isr.tecnico.ulisboa.pt

[3]M Raković is with Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia rakovicm@uns.ac.rs

N. F. Duarte is supported by FCT-IST fellowship grant PD/BD/135116/2017. FCT Project UID/50009/2020, the Lisbon ELLIS unit, LUMLIS, and the Research Infrasructure RBCOG-Lab.

2022 IEEE International Conference on Robotics and Automation (ICRA) | 978-1-7281-9681-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICRA46639.2022.9811725

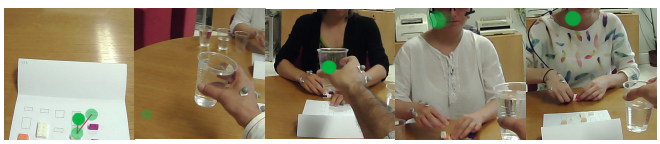
with one cup with 3 different levels of water, to extract the relevant eye-gaze cues and understand the differences in manipulation; (iii) a model capable of learning how to classify three levels of water in a cup from human eye-gaze cues; (iv) the integration of this model in a robot controller for online classification of online human handovers.

II. Human Eye-Gaze Cues in Handover Actions

This section presents the human-human experimental scenario created to extract human eye-gaze information during handovers of cups with different water levels. This includes the scenario and data description as well as a formal presentation of our publicly available dataset. The section continues with an analysis of the human eye-gaze sensor data during handovers of cups and concludes with an overall discussion of the findings from human eye-gaze movements.

A. Human-Human Experiment Description

The experiment consists of two people sitting on opposite sites of a table and completing a set of instructions hidden in a puzzle set provided to each one of the participants. This puzzle, which can be seen in Figure 1 (a) on the bottom region from the point of view (pov) perspective, has a set of LEGO® pieces that are to be picked up, one by one, and beneath specific pieces, there are instructions to manipulate the available cups. On each side of the table there are 3 identical cups but with three possible levels of water inside: (i) empty, (ii) 50% full, and (iii) 90% full; which for simplicity we refer to as empty, half, and full cup. The action instruction involves manipulating one of the 3 different cups as follows: (1) to grasp it and move it from the initial position (the right of the puzzle) to the left side of the puzzle (final position), or (2) grasp it and hand it over to the other participant in front of the table. Figure 1 (b) shows an example of action (i) and Figures 1 (c)-(e) of action (ii). The instruction indicates the type of action, (1) for pick & place and (2) for handover, and which cup to manipulate. The experiment is finished when both participants pick all the puzzle pieces, building a structure in the process, and all the actions are fulfilled. The pair of participants repeat the experiment a total of 5 consecutive times however the location of the action instructions changes in each repetition. The participants did not know, beforehand, which of the pieces contained an action instruction and the number of action changes for each trial. This is to prevent any anticipatory behaviour by the participants.



(a) (b) (c) (d) (e)

Fig. 1. Human-human interaction experiment: video frames from the head-mounted eye tracker field of view camera and corresponding eye-gaze fixation marked in each frame with a green dot. (a) the subject is working on its individual task, (b) moving a cup from the right side of the table to the left, (c)-(e) subject handing over a cup to the other participant.

B. Data Collection Description

The purpose of the experiments is to record the visuomotor movements of both participants. To collect the sensory information of the human eyes-head-arm movements the Pupil Labs head-mounted glasses, and OptiTrack motion capture systems are used. The Pupil Labs glasses are worn by each participant providing the eye-gaze movements. The OptiTrack motion capture (MoCap) system consists of 12 cameras all around the environment and OptiTrack infrared markers are placed on the Pupil Labs glasses and human's right wrist making up rigid bodies for capturing and 3D tracking of head-gaze and arm movements. The MoCap system provides position and orientation data recorded at 120 Hz whilst Pupil Labs Capture [26] system recorded at 60 Hz. PupilLabs Capture also provides additional information related to pupil detection, the two eye cameras frames, the external (world) camera frames, among other information, which is available on the dataset[1]. The dataset includes all the raw data provided by the PupilLabs Capture System (raw videos, gaze information, eyes information, etc). For more information on the specifications, please consult the GitHub repository[2]. It also includes the Cartesian and Quaternion rotations of the rigid bodies present in the study.

A total of 6 participants aged from 22 to 30 years old, 5 females and 1 male, all right-handed, none were members of the lab or the department, and all were naive regarding the purpose of the experiments. A total of 209 cup manipulations are performed: (i) 105 trials are of moving the cups from the right to the left side of the puzzle, (ii) 52 trials are of handing over the cup to the other participant, and (iii) 52 trials of receiving the cup from the other participant (mirror action of (ii)). Present in the dataset are 17, 19, and 16 handovers for empty, half, and full cups, respectively.

C. Human Eye-Gaze Data in Handovers

The eye-gaze fixations provided by the Pupil Labs Capture system are estimations of the focus of attention of the human projected onto the world camera of the glasses giving a 2d pixel location. This location is a representation, in the world camera video reference, of where the participant is looking. Since this is a free-moving reference frame and head-mounted on the participant, the 2d pixel vector points are not useful to understand the non-verbal gaze movements in human-to-human handovers. As a result, it was necessary to process the data acquired from Pupil Labs into meaningful gaze fixations, i.e. eye-gaze cues relevant to the experiments. Henceforth the data was labelled by a master student who followed the sole instruction of identifying the most prevalent gaze fixations in the whole experiment. The student did not participate in the makings of this paper nor was it aware of the purpose of this work. Each video frame from the Pupil Labs recordings was marked with a label corresponding to an eye-gaze cue. The most frequent eye-gaze cues are shown

[1]The human *pick-and-place* and *handover* actions dataset with video, gaze fixations from Pupil eye tracker, head and wrist movements from OptiTrack motion capture is available to anyone on the institution's website.

[2]https://docs.pupil-labs.com/core/software/pupil-player

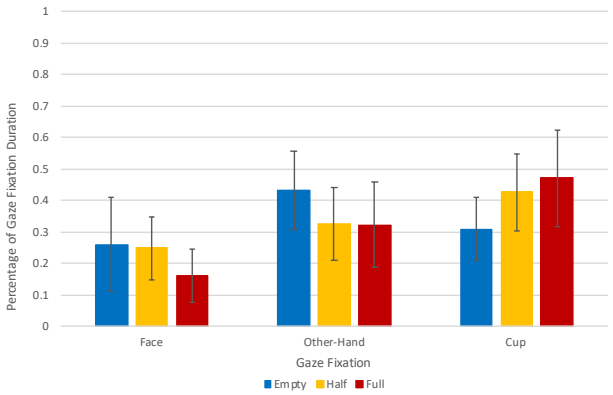


Fig. 2. The average and standard deviation percentage of eye-gaze cues duration during handover actions

Fixation	% of Frames
Cup	30 – 40
Own Hand	< 5
Face	10 – 30
Other's Hand	30 – 40
Other's Cup	< 1
Puzzle	NP
Final Position	NP
Outlier	< 1
No Gaze	< 1

TABLE I

TOTAL PERCENTAGES ON AVERAGE FOR ALL GAZE CUES DURING HANDOVER ACTIONS. NP - NOT PRESENT.

in Table I and correspond to: looking at the cup (Cup), at own hand (Own hand), at the other person's face (Face), at the other person's hand (Other's Hand), at the cup the other person is manipulating (Other's Cup), when picking LEGO® pieces (Puzzle), looking ahead to where the cup will be placed (Final Position), none of the above and with no particular meaning (Outlier), and a frame with no gaze fixation (No Gaze). Figure 1 has video frames with the projected fixation and the labels are: (a) Puzzle, (b) Final Position, (c) Cup, (d) Face, and (e) Face.

The segmentation of the handover actions is initiated when the participant fixates the cup for grasping and concludes when the handover is completed, which can be identified from the video recordings. These segments were collected for all the participants and the three cups. Table I shows the percentages of frames present in all the handovers collected for each of the aforementioned eye-gaze cues. It is reasonable to comprehend the reason that some of the cues are not present in the handover situation, e.g. the puzzle refers to moments where the participant is not performing an action, and the Final Position is related to the other action. The Own Hand is uncommon to occur in handovers and when it does happen it is usually during grasping or manipulation of the cup, hence we consider those cases as fixating the Cup. As a result, we can evaluate the 3 most relevant eye-gaze cues and compare them against the three possible cups. Figure 2 shows the time spent fixating the eye-gaze cues for the three types of the cup during handovers, ignoring the Outliers and No Gaze frames. A Shapiro-Wilk test was performed and demonstrated that the fixations distribution departed significantly from normality (W=0.9121, p=0.0008). As such, non-parametric tests were used in the analysis.

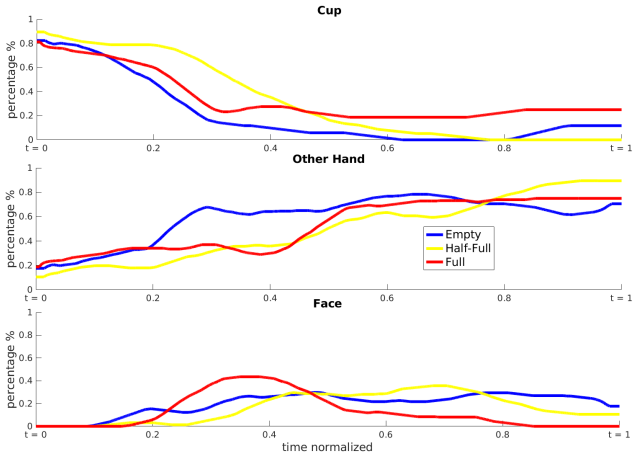


Fig. 3. The eye-gaze cues evolution over the handover action as a percentage in entire dataset.

As expected the analysis revealed that the fuller the cup, the longer the handover duration. To fairly compare the three cup conditions, Figure 2 is representing the percentage of time spent during the handover and not the absolute time spent for each eye-gaze cue. Despite this, there is still a correlation between the time spent fixating the cup and the level of water inside. In the Full condition, more time is spent fixating the Cup than the Face and Hand and the difference is statistically significant (Wilcoxon test p=5.6430e-05 to Face, and Wilcoxon test p=0.0059 to Other-Hand). This must be related to a general focus on not spilling and stabilizing the cup during handover. The eye-gaze cues analysis demonstrates that eye-gaze movements have two purposes: visuomotor control and visual-communication control. The former are the visual cues to guide the motor movement, and the latter are the visual cues to communicate intent to others. The former eye-gaze movements happen most often at the beginning of the action to ensure the object is safely stored in the hand and only after the grasp and manipulation are guaranteed to respect the object conditions then the human moves towards a gaze oriented toward expressing intent. From Figure 2 it can be concluded that the emptier the cup, the more time you can spend communicating your intent. The communication intent is expressed by more time spent looking at the subject's face and hand. Friedman's test confirms that the cup condition changes significantly the fixation percentage of the Face, Hand, and Cup (p=6.4767e-04). For a full cup, the visuomotor control of the action is more important than the visual-communication.

To understand the sequence of events that happen during

a handover, we process the eye-gaze movements as seen in Figure 3, where the eye-gaze cues are plotted over the handover sequence. From analysing the eye-gaze cues during handovers we can first conclude that, as previously seen in Raković et al. [20], the focus, in the beginning, is on fixating the cup. The initial 20% of the duration the Cup is fixated thrice as much as the other two. This is the functional gaze performing the visuomotor control of the arm grasping the cup for safe transportation. Secondly, the visual-communication only occurs during the transportation and after the visuomotor control check. Additionally, fixating the face does not occur in the visuomotor part, indicating that this is a gaze cue for communicating intent and not for visuomotor guidance. Thirdly, the emptier the cup was the sooner participants started communicating intent and for longer. Figure 3 shows that the Face is more likely to be fixated sooner and continue being present throughout the handover for not-full conditions. In the full cup condition, there is one evident discrepancy to the other cases. The fixation of someone's face becomes dominant in a small interval of time during the handover. In comparison to the other two conditions, the face continues to be fixated until the end of the action. From this, we can imagine the face cue as a bell-curve signal that as the level of water increases the amplitude increases while the width shrinks. This can be translated into an increased difficulty in manipulation so more time has to be spent in performing visuomotor control. Resulting in less time to communicate intent so the visual-communication is quicker but more pronounced.

III. Echo State Network

A. Data Selection

This dataset allows us to make a comparative analysis between two non-verbal cues: our approach of using eye-gaze cues, against the most common approach in robotics, head-gaze orientation [27], [28], [29]. From the markers placed on the head-mounted eye-tracker, we can track the head orientation during the handover actions. The head orientation is computed as the absolute rotation starting from the moment of cup pickup (beginning of handover sequence). Figure 4 shows the two non-verbal cues over the handover sequence for the three types of cups. The first major difference is the reaction time where the eyes switch fixation sooner than the head moves. This is in line with the human visuomotor coordination where gaze shows an anticipatory behaviour preceding motor movement [30], [31]. Secondly, from the eye-gaze data we were able to categorize three important cues (cup, other's hand, face), however from solely the head orientation that distinction is not available. This is simply a limitation on what we can extract from head movements. A counter-point can be made on the importance of those eye-gaze cues, and our hypothesis follows the ideas from [20] that there are two properties of gaze movements: (i) visual-communication and (ii) visuomotor control. From the analysis of the handovers we can conclude that: fixating the Cup aims at guiding the grasp and observing the exerted lifting force for potential spilling [32]; fixating the Face aims at expressing handover intent; as for fixating the Other's Hand we hypothesize that this fixation is an intermediate step between the other two, i.e. endows the two properties, visuomotor control for meeting one's hand with the other's, while at the same time, expressing the intent of the action. From this, we can conclude that these eye-gaze cues provide valuable information to classify cup manipulations of different levels of water (difficulty) than only head-gaze cues would not be possible.

B. Model Formulation

This section contains the formalism of a simple Echo State Network (ESN) and the included modifications applied in this work for better performance. ESN is an effective Recurrent Neural Network that has attracted substantial interest due to its performance in time-series [33]. For more details, the reader is referred to [34].

Let's consider a classification problem for a discrete univariate time-series with $\mathcal{T}$ time, and for each t there is an observation $\mathbf{u}(t) :=$ {Cup; Other's Hand; Face}, which in an ESN is the input unit, $x(t) \in \mathbb{R}^{N\times 1}$ denotes the state of the reservoir, and $y(t) :=$ {Empty; Half-full; Full} denotes the output unit. The time-series is represented in compact form as $\mathbf{U}^{\mathcal{T}} = [\mathbf{u}(1), \mathbf{u}(2), ..., \mathbf{u}(\mathcal{T})]^T$. $\mathbf{W_{in}} \in \mathbb{R}^{\mathcal{T}\times N}$ represents the connection weights between the input and hidden layer, $\mathbf{W_{res}} \in \mathbb{R}^{N\times N}$ denotes the connection weights inside the hidden layer. The encoder and decoder functions are formulated as:

$$\begin{aligned} \mathbf{x}(t) &= f(\mathbf{W_{in}}\mathbf{u}(t) + \mathbf{W_{res}}\mathbf{x}(t-1)) \\ \mathbf{y}(t) &= \mathbf{W_{out}}\mathbf{x}(t) \end{aligned} \tag{1}$$

where f is a nonlinear function, in this case the $\tanh$ was applied. This basic idea was first clearly spelled out in a neuroscientific model of the corticostriatal processing loop [35]. To avoid the costly operation of backpropagation through time, the ESN approach takes a different approach. While it continues to implement the encoding function, however the encoder parameters are randomly generated and left untrained. Only $\mathbf{W_{out}}$, the connection weights between the hidden and output layer, are subject to training using fast algorithmic closed form solutions like ridge regression

$$\min_{\mathbf{W_{out}}} \parallel \mathbf{W_{out}}\mathbf{X} - \mathbf{Y} \parallel_2^2 \tag{2}$$

where $\mathbf{W_{out}} = \mathbf{Y} \cdot \mathbf{X}^{-1}$ commonly referred as the readout weights. To compensate for untrained parameters, a large recurrent layer, the reservoir generates a rich pool of heterogeneous dynamics. The reservoir has three main hyperparameters: (i) the spectral radius, i.e. largest eigenvalue, of $\mathbf{W_{res}}$, (ii) the sparsity parameter, i.e. nonzero connections, of $\mathbf{W_{res}}$, and (iii) input scaling of $\mathbf{W_{in}}$. Gaussian noise with standard deviation is also applied in the state update function of Equation 1. Due to the high dimensionality of the reservoir $\mathbb{R}^{N\times N}$, the number of parameters for predicting the next reservoir state would be intractable, which could lead to overfitting, and the ridge regression evaluation is computationally very resourceful. Applying PCA for dimensionality reduction has shown to improve performance and provide

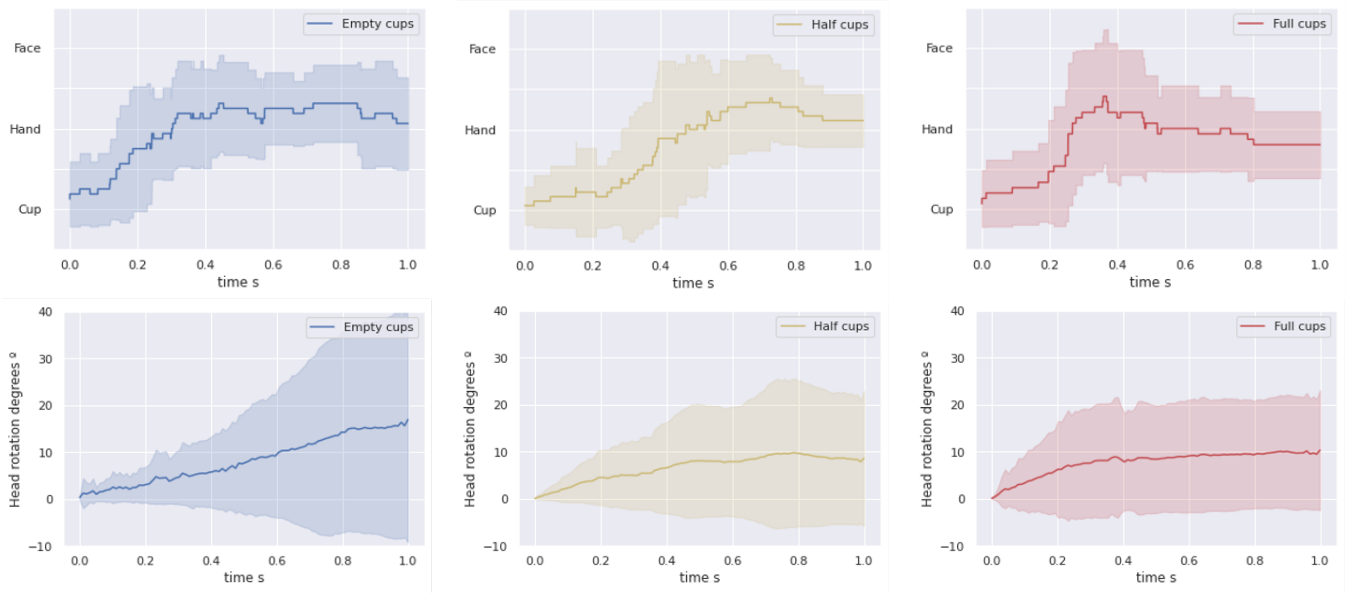


Fig. 4. Eye movements (top) vs head movements (bottom) for the three cases of water levels.

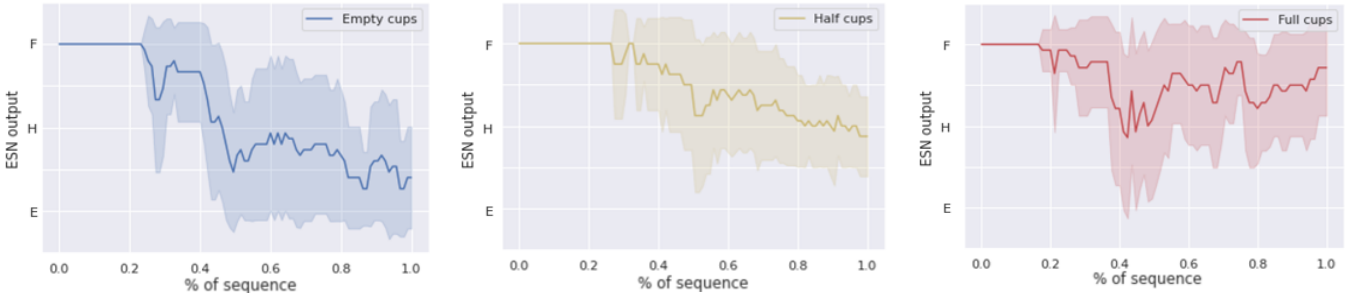


Fig. 5. Classification results for each cup level (E - empty; H - half-full; F - full) over time by the ESN.

good generalization when combined with ESNs [34]. As a result, we perform PCA projection on the data to extract the first D-eigenvectors of the covariance matrix. Additionally, since Recurrent Neural Networks (RNNs) with bidirectional architectures can extract features over a long period, ESNs with a bidirectional reservoir has been shown to improve the classification accuracy.

C. Training and Testing

The full list of hyper-parameters is the following: D-eigenvectors for PCA dimensionality reduction, N neurons, the spectral radius ρ, the sparsity β, input scaling ω, regularization value λ of ridge regression, and Gaussian noise ϵ. The dataset gaze cues sequences are normalized to 100 samples, and the output layer has $M = 3$ for the three types of cups. The hyper-parameter space is explored using grid search and performing 3-fold cross-validation on the whole dataset we achieved $95\% \pm 2\%$ and $72\% \pm 8.5\%$ accuracy in training and testing, respectively. In Figure 5 we get the prediction results of the ESN, at each time step, for the whole dataset. Since ESN requires a series of observations (compared to other methods which require a single sample) to regulate the internal state of its reservoir, it is provided around 20% of the sequence at the beginning. The classification result is given

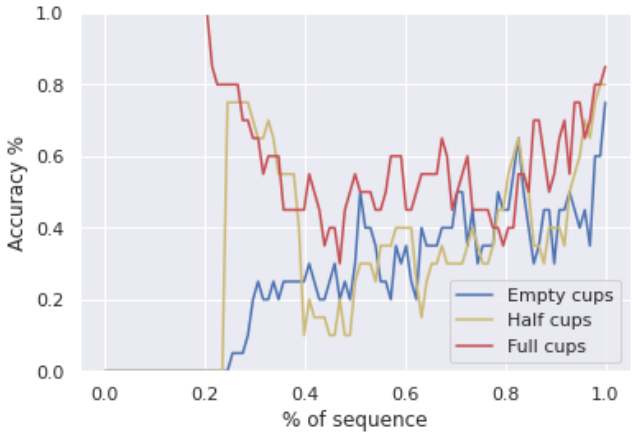


Fig. 6. ESN output accuracy for the three levels of liquid.

by the ESN highest probability output at each time step.

The results in Figure 5 show that the ESN classifies all the actions, in the beginning, as full cups. This makes sense since the cup is mostly fixated at the beginning which, without more knowledge, indicates that the handover is challenging

(full water level) as only the visuomotor control is present. However, during training, we noticed that the default action would change depending on the random initialization of $\mathbf{W_{res}}$ weights, so it might just be a random coincidence. Although, more often the best accuracy networks would output as initial default action the full cup option. Figure 6 illustrates the model's accuracy along the handover sequence for the three water cup levels. As stated before, the ESN is provided with around 20% of the sequence at the beginning before prediction and given that most fixate the cup, the accuracy is falsely indicating a full cup classification for all handovers. The model's accuracy achieves good results, of 60% or more for the three cup conditions, at around 90% completion of the handover. This reflects not only the high variance of gaze cues between humans but also the importance of the visual-communication part, which occurs last and goes on until the completion of the handover. Detecting full cups seems to be the easiest, as the accuracy increases sooner and reaches 80%, which could be impacted by the Face cue as it fades the fastest in those conditions (Figure 3). In the next section, we will incorporate the model in an online handover architecture.

IV. Online Human-to-Robot Handovers

The ESN is capable of classifying unseen handover actions with accuracy more than twice times higher than the chance level (1/3). However, these handover actions are sequences of eye-gaze cues from the human-to-human dataset of Section II. In this section, the model is applied to a human-robot interaction (HRI) scenario where the system is running online with a humanoid robot. The scenario consists of a human handing over cups with different levels of water to a robot that is reading, in real-time, the human eye-gaze cues from the head-mounted Pupil Labs sensor worn by the human. The purpose of this section is a proof of concept to demonstrate the compatibility of the proposed approach to real-robot experiments with an online classification of cups with different levels of water.

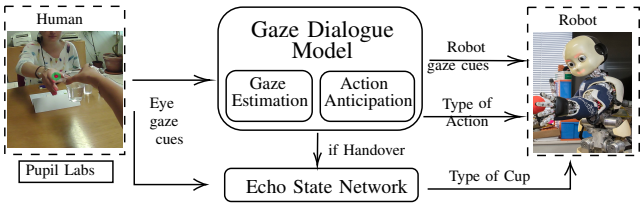


Fig. 7. Schematic of the robotic controller architecture for classification of type of action and cups.

The HRI controller architecture is represented in Figure 7. This system is composed of two important blocks which handle the communication between humans and robots: (i) the Gaze Dialogue Model [20], and (i) the Echo State Network. The Gaze Dialogue Model is an inter-personal gaze coordination system for human-humanoid interactions. It anticipates the human action by reading human eye-gaze cues and estimating the future human gaze cues, while at the same time, generating appropriate robot gaze cues and motor control response for the anticipated action [20]. The Gaze Dialogue Model is also responsible for classifying the type of action, extending the architecture with the ESN model classifying the type of cup during handovers. The Gaze Dialogue model instructs the robot to collaborate with the human (when handover) or merely observe (when pick & place). In a scenario where a handover is recognized, the Gaze Dialogue Model instructs the robot to reach towards the object and grasp it, and from the same eye-gaze cues, the Echo State Network block classifies the level of water in a cup (type of cup). To note that both the ESN and the Gaze Dialogue model process the same eye-gaze cues with the inclusion of other eye-gaze cues that are ignored by the ESN model (e.g. pick & place relevant fixations). An important difference between the two blocks is the processing of the cup fixation. The Gaze Dialogue Model identifies it simply as an Object while the ESN block treats it as a cup. This is crucial since, during HRI scenarios, the Gaze Dialogue Model is interested in the object the human is interacting with, not only cups. The supplementary video illustrates the architecture running online and exemplifies HRI examples where the Gaze Dialogue model classifies the type of action and the ESN classifies the type of cup.

V. Conclusion

From the human-to-human handover analysis, we were able to prove that human behaviour adapts to changing properties on cups. Previous works have shown that human motion strategy changes when the object weight is different [9], furthermore when humans handle liquid containers, such as cups, they constrain their motion behaviour to take into account the risk of spilling, revealing a contrasting strategy when comparing to empty containers [12]. Those findings are corroborated with our analysis of the human eye-gaze movements during handovers of a cup in three conditions: (i) empty, (ii) half-full, and (iii) filled with water. The eye-gaze movement's strategy to perform a handover is altered by an increased level of water inside the cup. As the level of water increases, so thus the risk of spilling, hence the gaze behaviour is spent more time on the visuomotor control role, than in the visual-communication role, as seen in Figure 2. As the visuomotor control focuses on ensuring a safe grasp and safe transportation of the cup (prevent spilling), and the visual-communication focuses on expressing to others the intent of handing over. The human-to-robot handover scenario has proven that the learned human-to-human handover of cups with different spilling risk levels is capable of classifying similar cups in a human-in-the-loop online interaction system with a humanoid robot. In future work, we intend to extend the robotic architecture to include a robot-cup manipulation controller that, according to the classified cup, it adapts the motor control strategy of the robot arm to prevent spilling.

REFERENCES

[1] H. Kjellström, J. Romero, and D. Kragić, "Visual object-action recognition: Inferring object affordances from human demonstration," *Computer Vision and Image Understanding*, vol. 115, pp. 81–90, Jan. 2011.

[2] R. Mottaghi, C. Schenck, D. Fox, and A. Farhadi, "See the Glass Half Full: Reasoning About Liquid Containers, Their Volume and Content," in *2017 IEEE International Conference on Computer Vision (ICCV)*, (Venice, Italy), pp. 1889–1898, IEEE, Oct. 2017.

[3] A. Modas, A. Xompero, R. Sanchez-Matilla, P. Frossard, and A. Cavallaro, "Improving filling level classification with adversarial training," *arXiv:2102.04057 [cs]*, Feb. 2021. arXiv: 2102.04057.

[4] C. Do, T. Schubert, and W. Burgard, "A probabilistic approach to liquid level detection in cups using an RGB-D camera," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, (Daejeon, South Korea), pp. 2075–2080, IEEE, Oct. 2016.

[5] C. Do and W. Burgard, "Accurate pouring with an autonomous robot using an RGB-D camera," *Advances in Intelligent Systems and Computing*, vol. 867, pp. 210–221, 2019. arXiv: 1810.03303 ISBN: 9783030013691.

[6] C. Schenck and D. Fox, "Visual closed-loop control for pouring liquids," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, (Singapore, Singapore), pp. 2629–2636, IEEE, May 2017.

[7] C. Schenck and D. Fox, "Reasoning About Liquids via Closed-Loop Simulation," *arXiv:1703.01656 [cs]*, June 2017. arXiv: 1703.01656.

[8] L.-F. Yu, N. Duncan, and S.-K. Yeung, "Fill and Transfer: A Simple Physics-Based Approach for Containability Reasoning," in *2015 IEEE International Conference on Computer Vision (ICCV)*, (Santiago, Chile), pp. 711–719, IEEE, Dec. 2015.

[9] K. Alaerts, P. Senot, S. P. Swinnen, L. Craighero, N. Wenderoth, and L. Fadiga, "Force requirements of observed object lifting are encoded by the observer's motor system: A TMS study," *European Journal of Neuroscience*, vol. 31, no. 6, pp. 1144–1153, 2010.

[10] P. Wei, Y. Liu, T. Shu, N. Zheng, and S.-C. Zhu, "Where and Why are They Looking? Jointly Inferring Human Attention and Intentions in Complex Tasks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (Salt Lake City, UT), pp. 6801–6809, IEEE, June 2018.

[11] L. Lastrico, A. Carfì, F. Rea, A. Sciutti, and F. Mastrogiovanni, "From Movement Kinematics to Object Properties: Online Recognition of Human Carefulness," in *International Conference on Social Robotics*, vol. 13086, pp. 61–72, 2021. arXiv: 2109.00460.

[12] N. F. Duarte, K. Chatzilygeroudis, J. Santos-Victor, and A. Billard, "From human action understanding to robot action execution: how the physical properties of handled objects modulate non-verbal cues," in *2020 Joint IEEE 10th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, (Valparaiso, Chile), pp. 1–6, IEEE, Oct. 2020.

[13] R. Sanchez-Matilla, K. Chatzilygeroudis, A. Modas, N. F. Duarte, A. Xompero, P. Frossard, A. Billard, and A. Cavallaro, "Benchmark for Human-to-Robot Handovers of Unseen Containers With Unknown Filling," *IEEE Robotics and Automation Letters*, vol. 5, pp. 1642–1649, Apr. 2020.

[14] J. R. Flanagan, G. Rotman, A. F. Reichelt, and R. S. Johansson, "The role of observers' gaze behaviour when watching object manipulation tasks: predicting and evaluating the consequences of action," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 368, no. 1628, pp. 20130063–20130063, 2013. ISBN: 1471-2970 (Electronic)\r0962-8436 (Linking).

[15] S. R. H. Langton, R. J. Watt, and V. Bruce, "Do the eyes have it? Cues to the direction of social attention," *Trends in Cognitive Sciences*, vol. 4, no. 2, p. 10, 2000.

[16] O. Palinko, F. Rea, G. Sandini, and A. Sciutti, "Robot reading human gaze: Why eye tracking is better than head tracking for human-robot collaboration," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5048–5054, Oct. 2016. ISSN: 2153-0866.

[17] M. S. Castelhano, M. Wieth, and J. M. Henderson, "I See What You See: Eye Movements in Real-World Scenes Are Affected by Perceived Direction of Gaze," *Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint*, vol. 4840, pp. 251–262, 2007. ISBN: 978-3-540-77342-9.

[18] K. Kompatsiari, F. Ciardo, V. Tikhanoff, G. Metta, and A. Wykowska, "On the role of eye contact in gaze cueing," *Scientific Reports*, vol. 8, p. 17842, Dec. 2018.

[19] S. M. Anzalone, S. Boucenna, S. Ivaldi, and M. Chetouani, "Evaluating the Engagement with Social Robots," *International Journal of Social Robotics*, vol. 7, pp. 465–478, Aug. 2015.

[20] M. Raković, N. F. Duarte, J. Marques, A. Billard, and J. Santos-Victor, "The Gaze Dialogue Model: Non-verbal communication in Human-Human and Human-Robot Interaction," *Paper under revision for IEEE Transactions on Cybernetics*, p. 14, 2021.

[21] J. Fan, D. Bian, Z. Zheng, L. Beuscher, P. A. Newhouse, L. C. Mion, and N. Sarkar, "A Robotic Coach Architecture for Elder Care (ROCARE) Based on Multi-User Engagement Models," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, pp. 1153–1163, Aug. 2017.

[22] M. Khoramshahi, A. Shukla, S. Raffard, B. G. Bardy, and A. Billard, "Role of gaze cues in interpersonal motor coordination: Towards higher affiliation in human-robot interaction," *PLoS ONE*, vol. 11, no. 6, pp. 1–21, 2016.

[23] K.-S. Tseng and B. Mettler, "Analysis of Coordination Patterns between Gaze and Control in Human Spatial Search," *IFAC-PapersOnLine*, vol. 51, no. 34, pp. 264–271, 2019.

[24] C.-M. Huang, S. Andrist, A. Sauppé, and B. Mutlu, "Using gaze patterns to predict task intent in collaboration.," *Frontiers in psychology*, vol. 6, no. July, p. 1049, 2015.

[25] N. F. Duarte, M. Rakovic, J. Tasevski, M. I. Coco, A. Billard, and J. Santos-Victor, "Action Anticipation: Reading the Intentions of Humans and Robots," *IEEE Robotics and Automation Letters*, vol. 3, pp. 4132–4139, Oct. 2018.

[26] M. Kassner, W. Patera, and A. Bulling, "Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction," *arXiv:1405.0006 [cs]*, Apr. 2014. arXiv: 1405.0006.

[27] P. Claudia and J. A. Shah, "Fast target prediction of human reaching motion for cooperative human-robot manipulation tasks using time series classification.," in *International Conference on Robotics and Automation*, 2015. ISBN: 9781479969234.

[28] M. Zheng, A. J. Moon, E. A. Croft, and M. Q. Meng, "Impacts of Robot Head Gaze on Robot-to-Human Handovers," *International Journal of Social Robotics*, vol. 7, no. 5, pp. 783–798, 2015. Publisher: Springer Netherlands.

[29] A. Kshirsagar, M. Lim, S. Christian, and G. Hoffman, "Robot Gaze Behaviors in Human-to-Robot Handovers," *IEEE Robotics and Automation Letters*, vol. 5, pp. 6552–6558, Oct. 2020.

[30] L. Lukic, J. Santos-Victor, and A. Billard, "Learning robotic eye–arm–hand coordination from human demonstration: a coupled dynamical systems approach," *Biological Cybernetics*, vol. 108, pp. 223–248, Apr. 2014.

[31] R. S. Johansson, G. Westling, A. Bäckström, and J. R. Flanagan, "Eye-hand coordination in object manipulation.," *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 21, no. 17, pp. 6917–6932, 2001. ISBN: 1529-2401 (Electronic)\n0270-6474 (Linking).

[32] H. C. Mayer and R. Krechetnikov, "Walking with coffee: Why does it spill?," *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 85, no. 4, pp. 1–7, 2012.

[33] C. Sun, M. Song, S. Hong, and H. Li, "A Review of Designs and Applications of Echo State Networks," *arXiv:2012.02974 [cs]*, Dec. 2020. arXiv: 2012.02974.

[34] F. M. Bianchi, S. Scardapane, S. Lokse, and R. Jenssen, "Reservoir Computing Approaches for Representation and Classification of Multivariate Time Series," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, pp. 2169–2179, May 2021.

[35] P. F. Dominey and F. Ramus, "Neural network processing of natural language: I. Sensitivity to serial, temporal and abstract structure of language in the infant," *Language and Cognitive Processes*, vol. 15, pp. 87–127, Feb. 2000.