Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

The Role of Social Cues for Goal Disambiguation in Human-Robot Cooperation

Samuele Vinanzi, Angelo Cangelosi, Christian Goerick

2020

Preprint:

This is an accepted article published in RO-MAN 2020. The final authenticated version is available online at: https://doi.org/[DOI not available]

The Role of Social Cues for Goal Disambiguation in Human-Robot Cooperation

Samuele Vinanzi[1]*, Angelo Cangelosi[1,2] and Christian Goerick[3]
[1]Cognitive Robotics Lab, The University of Manchester, Manchester, United Kingdom
[2]AIST-AIRC, Tokyo, Italy
[3]Honda Research Institute Europe GmbH, Offenbach am Main, Germany

*Correspondence: *samuele.vinanzi@manchester.ac.uk*

***Abstract*—Social interaction is the new frontier in contemporary robotics: we want to build robots that blend with ease into our daily social environments, following their norms and rules. The cognitive skill that bootstraps social awareness in humans is known as "intention reading" and it allows us to interpret other agents' actions and assign them meaning. Given its centrality for humans, it is likely that intention reading will foster the development of robotic social understanding. In this paper, we present an artificial cognitive architecture for intention reading in human-robot interaction (HRI) that makes use of social cues to disambiguate goals. This is accomplished by performing a low-level action encoding paired with a high-level probabilistic goal inference. We introduce a new clustering algorithm that has been developed to differentiate multi-sensory human social cues by performing several levels of clustering on different feature-spaces, paired with a Bayesian network that infers the underlying intention. The model has been validated through an interactive HRI experiment involving a joint manipulation game performed by a human and a robotic arm in a toy block scenario. The results show that the artificial agent was capable of reading the intention of its partner and cooperate in mutual interaction, thus validating the novel methodology and the use of social cues to disambiguate goals, other than demonstrating the advantages of intention reading in social HRI.**

***Keywords*—intention reading, social cues, clustering, cognitive robotics, human-robot interaction**

I. Introduction

Current trends in social robotics seem to indicate a future in which the adoption of autonomous robots in human environments, such as houses and working places, will be common. These artificial agents are expected to interact with people in a human-like fashion, as doing so will make them blend with ease into our daily lives. For this to happen, modern robots need to adopt cognitive models that take inspiration from the biological human brain and mimic our mental skills [1]. One of the latter, in particular, seems crucial to achieve social awareness: that is "intention reading", the ability to understand the underlying and unspecified goal that is driving the actions of another observed agent. A natural or artificial brain imbued with this skill is able to understand the behavior of other agents present in the environment and react to them accordingly by performing informed decision making. Additionally, intention reading is the base requirement for cooperation in scenarios where two or more agents must engage in joint action to achieve a shared goal: for them to act collaboratively, they must be able to understand each other's intention and dynamically adjust their behavior.

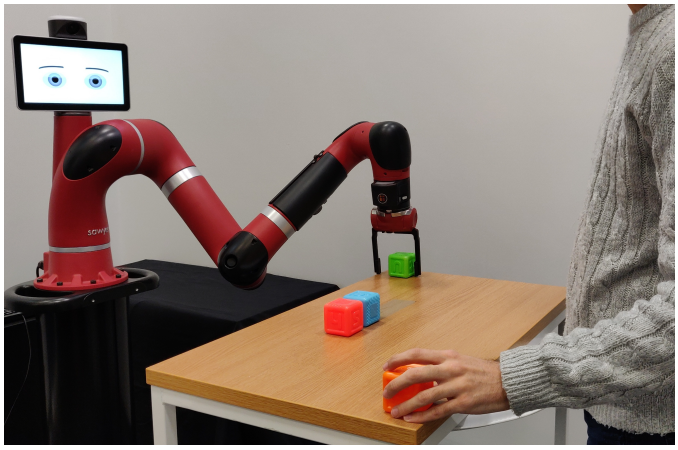

Fig. 1. The Sawyer robot used in the intention reading and goal disambiguation experiment involving an interactive block placing game.

In this paper, we present an artificial cognitive architecture that uses the developmental robotics approach [2] to perform intention reading on a human partner, using social cues to differentiate goals. A collaborative robot will learn by experience to derive intentions from its companion observing their physical clues and will use this knowledge to provide assistance in a cooperative manipulation task. To achieve this result, the architecture will make use of a novel multi-level clustering algorithm that employs social cues to disambiguate observed actions and goals, combined with probabilistic models to associate low-level observations to high-level intentions. We have validated this computational cognitive model through a block placing experiment, showing that the robot was able to correctly predict the partner's intentions using fast and lightweight machine learning methods.

II. Previous Work

Both adults and children possess the innate capability of understanding the behavior of others not as a series of unre-

lated motions through space, but rather as sets of goal-directed actions [3]. In fact, humans are able to see observed actions as a continuous stream of movement that is divided in intervals which are then segmented and interpreted [4]. The cognitive ability that makes us decode the purpose of someone's actions is called intention reading: a mental skill that lays the foundations of our social awareness [5] and enables us to reason about other agents in our environment to perform appropriate decision making. Once perceived, intentions can be shared and used to coordinate joint actions with common goals [6]. Overall, this cognitive ability enables us to be social creatures.

A neuroscientific explanation as to why humans are so good at intention reading compared to other animals is to be found in the brain's mirror neuron system, which maps the visual perception of observed actions with the organism's motor representation [7]. On a higher level, human infants learn how to associate social cues, as biological motion and eye gaze [6], to goals through experience [5], maturing this skill across various stages of development: what is understood by children is originally biological motion, then the pursuit of goals and finally the choice of plans. This process carries on until the fifth year of age [6], the same age at which children fully develop Theory of Mind, the ability to understand that others possess different mental states (as knowledge and goals) from one's own [8]. This makes sense, as the two mental skills overlap: intentions cannot be read if the beliefs and desires that drive the other are not acknowledged.

Several computational methods have been developed to address artificial intention reading. The use of probabilistic methods, such as Bayesian networks [9] or hidden Markov models [10] seem to have gained a large consensus, but researchers in the field have also adopted neural networks [11], support vector machines [12] and other machine learning models. Granada [13] proposes an approach where the intention reading task is divided in two: a convolutional neural network (CNN) extracts low-level features from camera images which are then used in a probabilistic plan recognizer. The use of robots for the exploration of computational intention reading capabilities is promoted by Sciutti [14], which argues that these agents share the same action space as the humans and can exhibit contingent behaviors. In fact, robots have widely been used to investigate shared intentionality with promising results [15, 16]. Many of these works gain information from motion and posture data collected from the human participants, but others also make use of gaze direction [17].

These approaches often assume the pre-existence of plan libraries that limit the scope of the goal search, other than the use of data-intensive methods. In an attempt to overcome these limitations, we have looked into unsupervised learning methods, in particular clustering. Other than the standard, known techniques, more advanced algorithms have been developed over the years. For example, adaptive clustering [18] aims at re-clustering objects when the features that describe them increase, whilst incremental (or sequential) clustering [19] processes data one element at a time. Our methods takes inspiration from a technique known as subspace clustering [20],

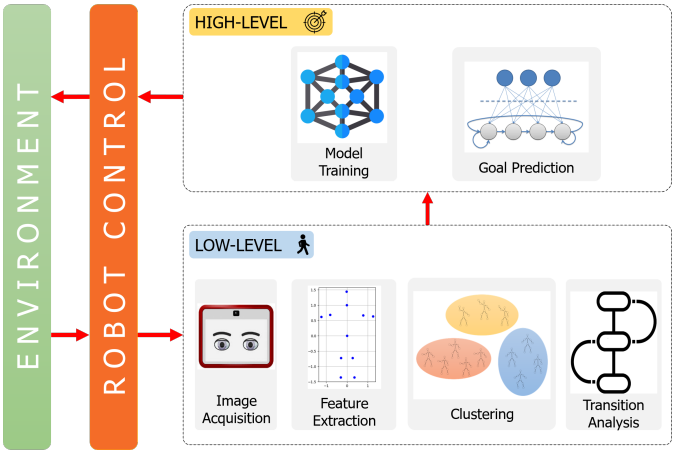

Fig. 2. Overview of the architecture. The low-level extracts data from the optical stream, forms clusters and uses them to represent actions as transitions through clusters. The high-level uses this encoding to probabilistically infer the pursued goal. The robot control manages sensors and actuators.

which consists in finding clusters that are defined by subsets of the available features, computed automatically, and by using these low-dimensional clusters to define higher-dimensional ones. As described in Section III, our method splits the features based on the sensory perception that generated them and works directly with the low-dimensional clusters.

III. Proposed Method

The aim of this research it to design an artificial cognitive architecture that will enable a social robot to perform intention reading in HRI. To do this, we built on the foundations of our previous model [21], scaling it to operate in more complex scenarios by considering a multi-modal perceptual input. This architecture is designed to be used in a collaborative context where the robot and a human partner are working together to reach a shared goal, the nature of which is not directly specified by the human and must be inferred by the artificial agent using non-verbal communication. The robot will use this model to observe its partner's actions, intended as sequences of physical configurations, and perform probabilistic decision-making with the aim to provide appropriate assistance. The design of this architecture follows the developmental robotics principles, defined by Cangelosi and Schlesinger [2] as "the approach to the design of behavioral and cognitive capabilities in artificial agents that takes direct inspiration from the developmental principles and mechanisms observed in the natural cognitive systems of children". In particular, the robot will not be provided with an initial, fixed goal library but it will learn by experience as newborns do [3]. Additionally, in accordance to the developmental theories of Tomasello [6], the architecture is organized in a low- and high-level subdivision of actions and goals using biological motion and gaze direction as a source of clues. Figure 2 shows an overview of the cognitive architecture. The latter will be initially trained by demonstration to associate a series of actions to a set of goals

and will subsequently be able to predict said intentions when observing the human agent's social cues.

A. Low-Level Posture Clustering and Action Representation

The purpose of the low-level is to make use of the partner's multi-modal social cues to encode pattern of movements which will be recognized during their real-time execution. This is done by using the robot's eye cameras to acquire images of the human, from which skeleton and gaze data will be extracted and used in an unsupervised clustering process.

1) Social Data Generation: Our data is sourced from the eye cameras of the robot. We decided to use 2D images because RGB sensors are present in the vast majority of social robots, whilst depth (RGBD) cameras are less common.

We generate skeleton data points using an open source and pre-trained deep CNN architecture named OpenPose [22], specialized in real-time 2D pose estimation. After providing an RGB image as input, we obtain a 18x2 feature vector representing the detected skeleton keypoints as 18 joints expressed in 2D spatial coordinates. We then operate a reduction on the number of keypoints to reduce the size of the vector and normalize the coordinates to gain independence from the position and size of the subject with respect to the robot. More details on this stage can be found in our earlier paper [21].

Another social cue that we collect from the robot's partner is gaze direction. We use Deepgaze [23], a CNN specialized in head pose estimation, to retrieve a 3D vector representing estimated roll, pitch and yaw for each image acquired by the robot. We chose to approximate gaze direction with head orientation to avoid some computational overheads that would impair the real-time computation of several frames per second. This has been proved to be an acceptable approximation [24].

2) Feature-Space Split Clustering (FSSC): We create action representations from the perceptual data through an unsupervised clustering procedure, using a novel algorithm that combines multiple sets of features in several increasingly refined stages. This strategy is adopted because complex and potentially ambiguous actions can be distinguished by increasing the granularity of the clustering operation which keeps into consideration a multitude of social cues. The main idea behind FSSC is the one of a multi-level clustering process that uses only a subset of the features at each level.

Consider a set of M training samples:

$$X = \left\{x^{(1)}, x^{(2)}, \ldots, x^{(M)}\right\} \tag{1}$$

Each sample can be seen as defined by N groups of features:

$$x^{(i)} = \left\{f_1^{(i)}, f_2^{(i)}, \ldots, f_N^{(i)}\right\} \tag{2}$$

Each group defines the feature-space f_n with $n \in [1, N]$ and contains data extracted from a different perceptual input. In our scenario we use $N = 2$ and for each image i we have that $f_1^{(i)}$ is a 22D vector containing the skeleton keypoints configuration and $f_2^{(i)}$ is a 3D vector that specifies the gaze direction.

Algorithm 1: Feature-Space Split Clustering (FSSC)

Input: training samples X; number of feature sets N
Output: A tree of clusters
$tree \leftarrow \{\}$
Initialize the root node with all the samples X
for $\ell \leftarrow 0$ **to** N **do**
  foreach *cluster of level* ℓ **do**
    $x \leftarrow$ samples belonging to $cluster$
    $f \leftarrow f_{\ell+1}^{(x)}$
    $f' \leftarrow$ Dimensionality reduction on f
    $newClusters \leftarrow CLUSTERING(f')$
    Set $newClusters$ to level $\ell + 1$
    $tree \leftarrow tree \cup newClusters$
  end
end
return *tree*

Algorithm 2: Cluster classification

Input: cluster tree T, testing sample s
Output: cluster to which s belongs
$parentNode \leftarrow$ root node of T
$\ell \leftarrow 1$
Loop
  $C \leftarrow$ descendants of $parentNode$ in T
  $cluster \leftarrow \min distance(f_\ell^{(s)}, c \in C)$
  if $cluster$ has descendants **then**
    $parentNode \leftarrow cluster$
    $\ell \leftarrow \ell + 1$
  else
    return $cluster$
  end
EndLoop

FSSC works by implicitly computing a tree of depth $L = N$ whose nodes contain the refined clusters. The root node ($\ell = 0$) contains all the data samples and is considered as a single cluster, while nodes of each subsequent level $\ell > 0$ are the clusters obtained by clustering the samples belonging to the parent node in the feature-space f_ℓ. At each level, we perform Principal Component Analysis (PCA) dimensionality reduction to project the data in a 2D space to avoid the curse of dimensionality [25]. We chose X-Means [26] as the internal clustering method, which is a variation of the traditional K-Means able to automatically find the optimal number of clusters. The motivation behind the choice of both PCA and X-Means is described more in-depth in our previous work [21]. Algorithm 1 describes this computation.

Given the hierarchical and nonlinear nature of this algorithm, we can't perform classification (intended as the association of a new data sample to one of the existing clusters) through a simple Euclidean distance search for the closest centroid. Instead, the procedure described in Algorithm 2 must be adopted. The latter searches through the cluster

tree, comparing the centroids in their respective feature-space coordinates until a leaf node is found.

Despite this approach splits the feature-spaces, in fact it does not decouple the multimodal features extracted from the human, as they share a temporal dependence on the frame from which they were generated. In other words, the features never lose their alignment.

3) Transition Analysis: During its initial training, the agent will observe the actions of its human partner, record the image frames and extract all the relevant features. This assembled dataset will be used to create the clusters using the method described in Algorithm 1: each of them represents a group of similar but not identical postures. These will be used in the next stage, Transition Analysis, to create the low-level encoding: each action will be represented by the sequence of the cluster ids encountered during its performance. To obtain temporal invariance, we include in the encoding only transitions through different clusters, discarding the persistence in the same group: in this way, the representation of an action won't depend on the speed of its execution. The results of this analysis, plus a set of unique names for each goal, are forwarded to the high-level module to train it as described in Section III-B.

After being trained, the system will be able to perform intention reading: the agent will observe its human partner during the execution of the action, each of their physical configurations will be classified to one of the known clusters using the procedure described in Algorithm 2 and the discovered id will be forwarded to the high-level module for probabilistic inference.

B. High-Level Goal Probabilistic Inference

The high-level module is in charge of goal probabilistic inference from the observed actions. What we are trying to achieve is not action recognition but rather prediction, this means that only one part of the action will be known and observable. Our objective is to determine the intention based on as few observations as possible, so that the robot will be able to contribute to the task before it is over.

To achieve this, we have employed the Bayesian network shown in Figure 3. The top node denoted as I represents the intention of the human partner and its probability distribution is equal across all the possible goals. The bottom nodes marked as O_k with $k \in [1, K]$, where K is the maximum length of the encoded actions, represent the observations. The values of these nodes span in the range of the possible cluster ids identified by the low-level module. The conditional probability tables of the observation nodes are fitted from the training data provided by the low-level Transition Analysis (i.e. the action encoding associated to each goal name) using Maximum Likelihood Estimation [27]. We assume that the probability of each observation depends on the driving intention and by the precedent symbol encountered:

$$P(O_1 \mid I) \tag{3}$$

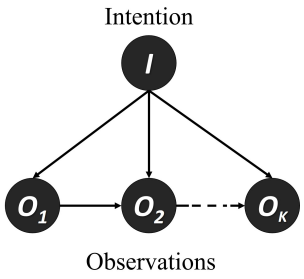

Fig. 3. The Bayesian network used for high-level probabilistic goal prediction. The top node represents the intention of the observed partner, whilst the bottom node symbolize the observations (the action encoding symbols produced by the low-level module).

$$P(O_k \mid O_{k-1}, I) \colon k \in [2, K] \tag{4}$$

Once the probabilistic model is trained, it can be used for inference. During the execution phase, the robot will be observing the human and recording each cluster transition in real-time. The low-level module will forward these symbols to the high-level, which will treat them as sequential observations. Each time a new piece of evidence is added to the model, we use Pearl's Message-Passing algorithm [28] to calculate the marginal probability distribution for node I given the evidence. As soon as one of the goals is predicted with a probability greater than 0.5, it is sent forward in the processing chain to instantiate appropriate collaborative behavior. A higher threshold would lead to a more confident prediction but would also slow down the robot's reaction time, potentially impairing the collaborative effort.

The adoption of a probabilistic model instead of a deterministic one, such as a lookup table, is justified by the potential noisiness of the observations. The classification performed by the low-level will not always be accurate and the agent is required to be robust against invalid input sequences.

C. Robot Control

The robot control module deals with the direct interface between the cognitive architecture and the robotic platform, in this case a Sawyer: an industrial collaborative robot designed for object manipulation, equipped with a 7-DOF arm (Figure 1). In particular, it provides interaction with the ROS middleware to control its sensors and actuators and perform vision, movement and grasping for the shared goal task.

IV. Experiments

The aim of this experiment is to validate the intention reading capabilities of the cognitive architecture on a shared goal scenario. It involves a collaborative block placing game where the robot will help the human to line up different colored toy blocks following a set of rules.

A. Experimental setup and procedure

The experimental setup is shown in Figure 4a. The robot and the human are facing each other on the two sides of a

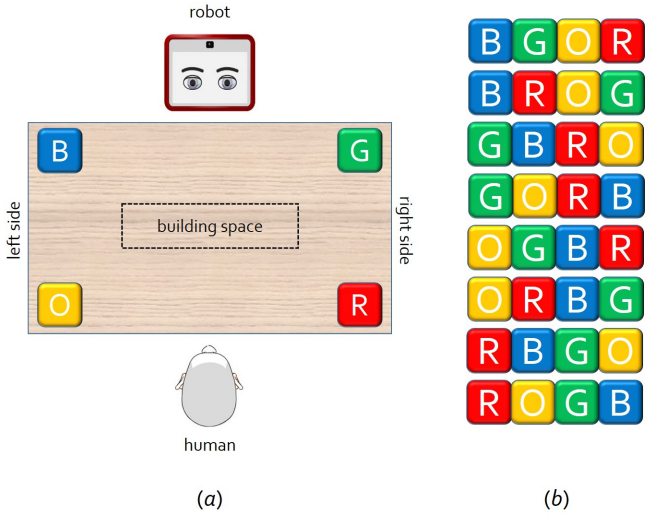

Fig. 4. Experimental setup for the block building game. (a) Schematic of the playing table, depicting the position of the 4 colored blocks: blue (B), orange (O), red (R) and green (G). (b) The 8 admissible block sequences obtained by picking blocks alternatively from each side. These sequences are the goals for this scenario.

table. 4 different colored blocks are positioned on the corners of the playing area; anticlockwise from the top left they are: blue (B), orange (O), red (R) and green (G). The central area of the table is denoted as the building space.

The aim of the game is to use the 4 available blocks to form a line, following a simple rule only known by the demonstrator: the blocks must be chosen one by one from a different side of the table (left or right). The 8 legal combinations of blocks are reported in Figure 4b and each of them forms a goal for our intention reading purposes. During the training phase, the human will demonstrate each goal once and the robot will learn to associate the demonstrator's movements to their intentions, as described in Section III.

The system is sensitive to noise because it is learning each intention and its sequence of actions from a single training example, which is in turn obtained from a non-deterministic and unsupervised process. To reduce this effect, we implement a post-processing computation that aims to capture the regularities of the data, assigning to each end position of our actions (i.e. the grasping position for each of the blocks) one of the cluster ids based on its statistical mode in the computed training dataset. Training actions are then eventually corrected by merging the two representations to avoid any errors in the training set.

During the execution phase, the human will choose one of the goals and start placing blocks accordingly. The robot will have to predict as quickly as possible which of the learned goals is being pursued and act collaboratively by collecting one at a time the remaining blocks and placing them in the building space to complete the construction.

TABLE I
EXPERIMENTAL RESULTS. SEE SECTION V FOR THE FULL EXPLANATION.

Goal	Accuracy	Time	Intercept	Precision	Recall	F1
BGOR	0.80	5.65	51%	0.80	0.64	0.71
BROG	0.70	6.07	53%	0.70	0.74	0.72
GBRO	0.90	6.53	51%	0.90	0.86	0.88
GORB	0.90	5.98	50%	0.90	0.95	0.92
OGBR	0.75	6.36	50%	0.75	0.88	0.81
ORBG	0.85	4.96	50%	0.85	0.89	0.87
RBGO	0.80	6.02	50%	0.80	0.70	0.74
ROGB	0.70	6.66	50%	0.70	0.82	0.76
Mean	0.80	6.03	51%	0.80	0.81	0.80

V. RESULTS AND DISCUSSION

After being trained to recognize the 8 goals referenced in Figure 4b, the cluster tree generated by FSSC is reported in Figure 5. The axes of each graph are the two Principal Components in which the feature-spaces have been projected during the dimensionality reduction process.

During the first step of computation (Figure 5a), all the training samples have been clustered in the f_1 feature-space, which contains the skeleton keypoints of the partner's body. This operation generated 3 clusters: a neutral one (id 1), where the demonstrator was approximately standing straight, and the other two (ids 2 and 3) containing skeletons leaning respectively left and right. The average Silhouette evaluation metric [29] for this clustering is $S = 0.71$, which indicates that the number and placement of the centroids is adequate.

In the second stage of processing (Figures 5b and 5c), clusters 2 and 3 are independently clustered: the samples contained in their domain are clustered in the f_2 space, expressing the gaze direction. Both these clusters are split in other two groups that capture the interaction with the far block (B or G) or the near one (O or R). The Silhouette scores for these two splittings are $S' = 0.86$ and $S'' = 0.69$. The samples contained in cluster 1 are not re-clustered in this stage because the computation in the f_2 feature-space does not distinguish multiple, separate sub-clusters.

It is now possible to see that the two stages of progressive clustering have divided the space of physical configurations of the human partner in 5 partitions: left/right and near/far, which correspond to the position of the 4 colored blocks, plus the neutral position. The adoption of multiple social cues has led to a granularity that would have not been possible by using only one set of features, say the skeleton keypoints. The adoption of a multi-modal perception has resulted in a finer division of space which was able to correctly disambiguate the actions and therefore the goals.

During the execution phase of the experiment, we run 20 trials for each goal, totaling 160 games where the robot had to guess the current intention in progress. The results of this experimentation are reported in Table I. For each goal, the latter displays: the average accuracy across 20 trials, the observation time measured as number of seconds needed to perform a prediction, the percentage of completion of the action by the partner when the intention was predicted, average precision, recall and F1 score. The last line of the table reports

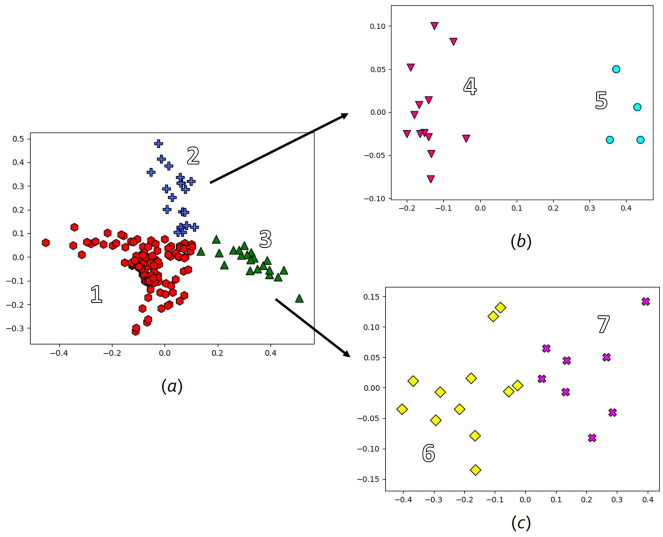

Fig. 5. Clustering results. (a) First level of FSSC in f_1 feature-space (skeleton keypoints) of all the training samples. (b) Second level of FSSC in f_2 feature-space (gaze direction) on cluster 2 and (c) cluster 3. Refer to Section V for the full details.

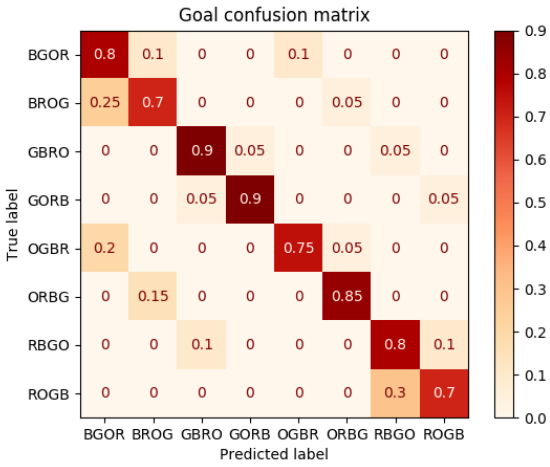

Fig. 6. Confusion matrix for the 8 goals of the block placing experiment.

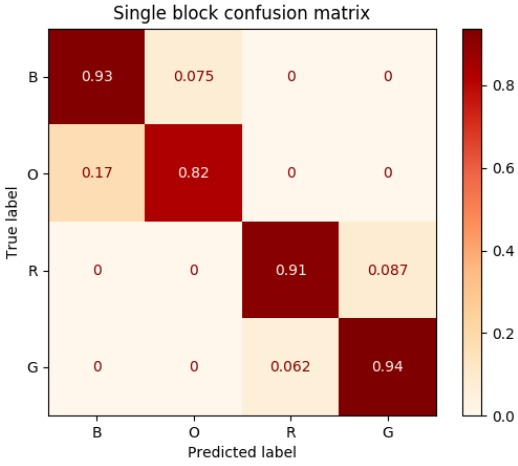

Fig. 7. Confusion matrix for each individual block across the whole 160 experimental trials.

the mean of these scores across all goals. The robot was able to perform correct intention reading approximately 80% of the time, identifying the pursued goal when its partner was halfway along its completion, giving it enough time to act cooperatively.

Finally, we report two confusion matrices: one for the overall goals (Figure 6) and one for the single blocks (Figure 7). Analyzing the latter, it is interesting to note that the system never confused two blocks belonging to different sides of the table, rather it tends to sometimes misclassify the near and far blocks of each side. This was an expected outcome, as those are the actions that present the most commonalities.

In comparison to our previous work [21], we have obtained less overall accuracy but we were successful in performing intention reading in a more complex scenario: with respect to our past setting, we have doubled the number of goals and partially overlapped some of the actions, creating a more error-prone environment that our previous methodology would have struggled to grasp.

VI. Conclusion and Future Work

In this paper, we have presented a robotic cognitive architecture that makes use of unsupervised machine learning and probabilistic models to predict the intention of a human partner engaged in a joint activity. We have explored the social and cognitive importance of intention reading and how it can benefit future robotic agents in proportion to the advantages that it brings to human beings: social understanding, environmental awareness and the development of Theory of Mind. Following the developmental robotics principles, our model takes inspiration from scientific findings in human cognition and is designed according to some psychological theories. In particular, it is based on the agent's experience, it divides the understanding in a twofold process and makes use of some basic sensory perceptions at which infants are particularly susceptible: gaze and motion. We introduced a new multi-level and multi-modal clustering algorithm to combine several sensory inputs and we use its outcome to define a low-level encoding of the observed actions, which are then translated to high-level probabilities for each of the learned goals. We have tested this model on a block placing game where a Sawyer robot must try and help its human partner to complete a construction, demonstrating that the adoption of multiple social cues leads to a better goal disambiguation.

This piece of work represents an incremental step from of our previous experiment [21]. In the latter, the general structure of the cognitive architecture was the same but the operations performed at each stage were simpler, as the experimental setup presented actions easier to represent and quite distinct from each other. In our current work, we have

introduced a more complex scenario that presents a degree of ambiguity between several actions and a wider goal space: this has led us to improve our methods with the addition of the FSSC algorithm and a finer processing of the data at each stage of execution. The results we have obtained are in line with our expectations: the ambiguity of the actions has decreased the overall performance of the robot, but the new methodology made it possible to still correctly read the intention on average 80% of the time.

We plan to further extend this architecture to make it more general and applicable to even more complex scenarios. One possible research question involves the addition of hierarchical goals, i.e. goals composed by multiple sub-goals (for example, uncapping a bottle might be one step to achieve the "drink" goal). Another possible study could explore the use of more social clues and the investigation of their order of application within FSSC. In addition to this, we believe that true collaborative intelligence can emerge from the integration with a model of trust and Theory of Mind. After reading the intention of the human partner, the robot would be able to use its trusting model to determine how much help does the latter need in order to achieve the shared goal. For this reason, we have planned an integration with our previous work on robot trust modeling [30].

REFERENCES

[1] Cynthia Breazeal. Social interactions in HRI: the robot view. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 34(2):181–186, 2004.

[2] Angelo Cangelosi, Matthew Schlesinger, and Linda B Smith. *Developmental robotics: From babies to robots*. MIT Press, 2015.

[3] Bertram F Malle, Louis J Moses, and Dare A Baldwin. *Intentions and intentionality: Foundations of social cognition*. MIT press, 2001.

[4] Dare A Baldwin and Jodie A Baird. Discerning intentions in dynamic human action. *Trends in cognitive sciences*, 5(4):171–178, 2001.

[5] Amanda L Woodward, Jessica A Sommerville, Sarah Gerson, Annette ME Henderson, and Jennifer Buresh. The emergence of intention attribution in infancy. *Psychology of learning and motivation*, 51:187–222, 2009.

[6] Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne, and Henrike Moll. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28(5):675–691, 2005.

[7] Vittorio Gallese and Alvin Goldman. Mirror neurons and the simulation theory of mind-reading. *Trends in cognitive sciences*, 2(12):493–501, 1998.

[8] Kimberly E Vanderbilt, David Liu, and Gail D Heyman. The development of distrust. *Child development*, 82(5):1372–1380, 2011.

[9] Haris Dindo, Francesco Donnarumma, Fabian Chersi, and Giovanni Pezzulo. The intentional stance as structure learning: a computational perspective on mindreading. *Biological cybernetics*, 109(4-5):453–467, 2015.

[10] Richard Kelley, Christopher King, Alireza Tavakkoli, Mircea Nicolescu, Monica Nicolescu, and George Bebis. An architecture for understanding intent using a novel hidden markov formulation. *International Journal of Humanoid Robotics*, 5(02):203–224, 2008.

[11] Bharat Singh, Tim K Marks, Michael Jones, Oncel Tuzel, and Ming Shao. A multi-stream bi-directional recurrent neural network for fine-grained action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1961–1970, 2016.

[12] Alessandro Manzi, Paolo Dario, and Filippo Cavallo. A human activity recognition system based on dynamic clustering of skeleton data. *Sensors*, 17(5):1100, 2017.

[13] Roger Granada, Ramon Fraga Pereira, Juarez Monteiro, Rodrigo Barros, Duncan Ruiz, and Felipe Meneguzzi. Hybrid activity and plan recognition for video streams. In *The 2017 AAAI Workshop on Plan, Activity, and Intent Recognition. Greer, JE and Koehn, GM*, pages 54–59, 1995.

[14] Alessandra Sciutti, Caterina Ansuini, Cristina Becchio, and Giulio Sandini. Investigating the ability to read others' intentions using humanoid robots. *Frontiers in psychology*, 6, 2015.

[15] Peter Ford Dominey and Felix Warneken. The basis of shared intentions in human and robot cognition. *New Ideas in Psychology*, 29(3):260–274, 2011.

[16] Bart Jansen and Tony Belpaeme. A computational model of intention reading in imitation. *Robotics and Autonomous Systems*, 54(5):394–402, 2006.

[17] Nuno Ferreira Duarte, Mirko Raković, Jovica Tasevski, Moreno Ignazio Coco, Aude Billard, and José Santos-Victor. Action anticipation: Reading the intentions of humans and robots. *IEEE Robotics and Automation Letters*, 3(4):4132–4139, 2018.

[18] Alina Câmpan and Gabriela Şerban. Adaptive clustering algorithms. In Luc Lamontagne and Mario Marchand, editors, *Advances in Artificial Intelligence*, pages 407–418, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[19] Steven Young, Itamar Arel, Thomas P Karnowski, and Derek Rose. A fast and stable incremental clustering algorithm. In *2010 Seventh International Conference on Information Technology: New Generations*, pages 204–209. IEEE, 2010.

[20] René Vidal. Subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68, 2011.

[21] S. Vinanzi, C. Goerick, and A. Cangelosi. Mindreading for robots: Predicting intentions via dynamical clustering of human postures. In *2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 272–277, Aug 2019.

[22] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1611.08050*, 2016.

[23] Massimiliano Patacchiola and Angelo Cangelosi. Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods. *Pattern Recognition*, 71:132 – 143, 2017.

[24] S. Jha and C. Busso. Probabilistic estimation of the driver's gaze from head orientation and position. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–6, Oct 2017.

[25] Richard Bellman. *Dynamic programming*. Courier Corporation, 2013.

[26] Dan Pelleg, Andrew W Moore, et al. X-means: Extending k-means with efficient estimation of the number of clusters. In *ICML*, volume 1, pages 727–734, 2000.

[27] John Aldrich et al. Ra fisher and the making of maximum likelihood 1912-1922. *Statistical Science*, 12(3):162–176, 1997.

[28] Steffen L Lauritzen and David J Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 157–224, 1988.

[29] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[30] Samuele Vinanzi, Massimiliano Patacchiola, Antonio Chella, and Angelo Cangelosi. Would a robot trust you? Developmental robotics model of trust and theory of mind. *Philosophical Transactions of the Royal Society of London B*, 2019.