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

Mindreading for Robots: Predicting Intentions via Dynamical Clustering of Human Postures

Samuele Vinanzi, Christian Goerick, Angelo Cangelosi

2019

Preprint:

This is an accepted article published in 9th Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics. The final authenticated version is available online at: https://doi.org/[DOI not available]

Mindreading for Robots: Predicting Intentions via Dynamical Clustering of Human Postures

Samuele Vinanzi*†, Christian Goerick§ and Angelo Cangelosi†‡
†Cognitive Robotics Lab, The University of Manchester, Manchester, United Kingdom
‡ICAR-CNR, Palermo, Italy
§Honda Research Institute Europe GmbH, Offenbach am Main, Germany

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

***Abstract*—Recent advancements in robotics suggest a future where social robots will be deeply integrated in our society. In order to understand humans and engage in finer interactions, robots would greatly benefit from the ability of intention reading: the capacity to discern the high-level goal that is driving the low-level actions of an observed agent. This is particularly useful in joint action scenarios, where human and robot must collaborate to reach a shared goal: if the latter can predict the actions of the former, it will be able to use this information for decision making in order to improve the quality of the cooperation. This research proposes a novel artificial cognitive architecture, based on the developmental robotics paradigm, that can estimate the goals of a human partner engaged in a joint task to modulate synergistic behavior. This is accomplished using unsupervised dynamical clustering of human skeletal data and a hidden semi-Markov chain. The effectiveness of this architecture has been tested through an interactive cooperative experiment involving a block building game, the iCub robot and a human. The results show that the former is able to adopt a collaborative behavior by performing intention reading based on the partner's physical clues.**

***Keywords*—intention reading, clustering, hidden semi-markov model, cognitive robotics, developmental robotics, human-robot interaction**

I. Introduction

Humans are social creatures: they tend to live in communities and societies, which provide fulfillment for many needs that range from basic survival to intellectual and emotional expression. This intrinsic predisposition towards aggregation is rooted in their natural attitude to social abilities: in fact, even newborn babies possess the instinct and capability to perform social behavior [1]. Human interpersonal competencies are bootstrapped by a fundamental cognitive skill named "intention reading", which is the ability to correctly understand the goals that drive the observable actions performed by others. This information can be used for decision making and to gain a social advantage.

As robotics and artificial intelligence progress, we expect that robots will become important social actors in our future societies. For these artificial agents to perform optimally in a human world, it is essential for them to possess the same social skills, including intention reading, as humans: this is considered the key to creating truly social machines [2]. In fact, collaboration between human and machines, such as with robot companions or intelligent cars, requires a shared understanding of common goals and intentions. For example, if a humanoid robot and a person are engaged in joint action, both must be able to understand the goal and intentions of the partner, dynamically adjust their behavior to the shared goal and communicate the intended objective to the other agent. This requires embodied and situated interaction between humans and robotic agents [3].

Following the developmental robotics approach [1], this study addresses an artificial robotic cognitive architecture able to learn by experience to derive goals from observed body movement. This is accomplished using computer vision techniques, unsupervised clustering algorithms and probabilistic Markovian models to associate low-level sequences of postures to high-level intentions. The presented novel architecture has been validated by designing a human-robot interaction experiment regarding a cooperative game and the results show that is was possible for the robot to learn to read intentions in a lightweight, unsupervised manner and to use this information to drive decision making.

II. Previous Work

Human action is driven by intentions: both adults and preschoolers understand human behavior not as an unrelated series of motions through space but rather as actions directed towards goals [4]. A preliminary distinction must be made between goal and intention: whereas the former represents a final desired state, the latter incorporates both a goal and an action plan to achieve it.

Human action can be seen as a continuous stream that is subdivided in intervals which are then decoded [5]. There are two ways in which an observed action can be interpreted: it can be teleological if it is understood as goal-oriented, or referential if its aim is to ascribe attentional states or communicative messages (for example, pointing gestures) [6]. More generally, intentional action can be viewed as a closed loop process in which a goal is formulated (for example, to open a bottle) and, by integrating perception of the environment, knowledge and skills, an intention is generated (grasp the bottle and remove the cap). Subsequently, the plan will be performed and the effects of the actions will be perceived and used as feedback (if the cap is hard to remove by hand,

retrieve a tool). The previous example also demonstrates the hierarchical organization of intentions: some primitive actions may complete goals that serve the purpose of a higher goal (to drink the content of the bottle) [5].

Humans' ability to read others' intentions is argued to have matured through evolution: the capacity to predict other's behavior, cooperate and group has probably been a key factor in human survival. It forms the basis of may cognitive skills [7] and of social interactions: if intentions can be generated and understood, they can also be shared and used as a coordination tool in joint actions with common goals [6, 8]. This skill develops from childhood, evolving through three different stages: understanding autonomous action (i.e. self-produced movement that distinguishes animate from inanimate entities), understanding the pursuit of goals and, finally, understanding the choice of plans [8]. The developmental origin of this skill is still under debate: on rich interpretations the ability to understand intentionality is encoded in the genes, while lean interpretations assume it develops across experience. Although there is no definitive answer yet, many researchers think that preschoolers learn to associate actions to goals via a direct and progressively refined experience of the world [7]. One thing that seems clear is that intention reading fully develops around the fifth year of age [8], coinciding with the maturation of theory of mind: the ability to attribute mental states to others, for instance beliefs, intentions and knowledge, that may differ from one's owns [9]. This makes sense, as a prerequisite to understand someone's intentions is to acknowledge in them a set of desires and beliefs that drive their actions. Children, whose intention reading is typically tested using the visual habituation paradigm [4], at this stage are also able to understand intentions behind failed actions and re-enact a successful version of them [5].

The scientific community has been researching intention reading capabilities in robots for many years. Dominey [10] explored shared intentionality in a turn-taking game between a human and a robotic arm. Bien [11] developed a system that analyzes posture and movements in elder people and tries to decode the inner intentions they are driven by. Sciutti [3] affirms that humans are sensible to subtle kinematic differences that can distinguish similar actions directed to different goals and proposes a robotics approach to investigate this. Other relevant researches involve the use of variations of dynamic Bayesian networks [12], self-organizing maps [13] and first order logic [14].

Granada [15] proposes an approach where the intention recognition task is divided in two levels: the low-level is focused on action recognition and classification, whilst the high-level derives the goal from the observed actions. Action recognition is a topic that has received great attention from computer vision researchers [16]. The most consolidate state-of-the-art techniques involve use of a modern deep learning architecture composed of a convolutional neural network (CNN) connected to a recurrent neural network that takes as input an image frame from a video sequence and outputs a label representing the recognized action [17]. Other similar approaches work not

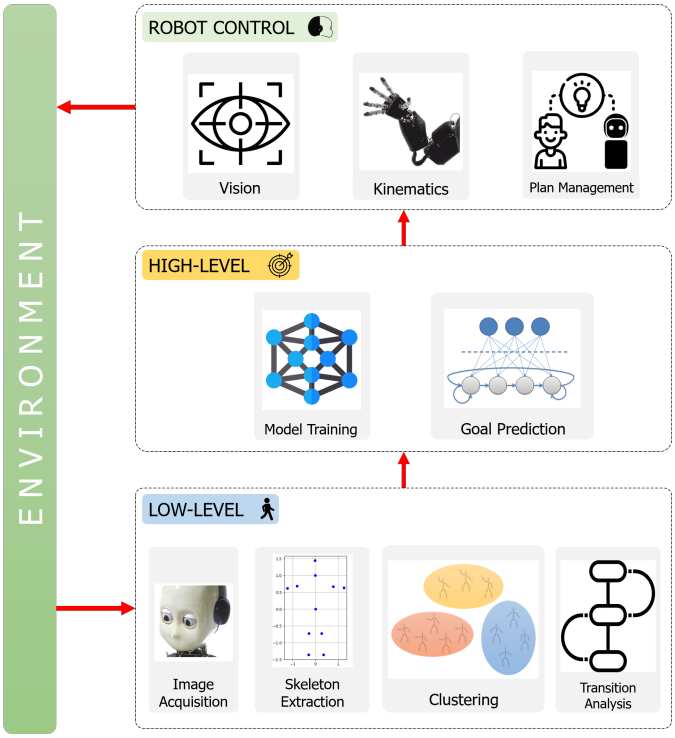

Fig. 1. Overview of the proposed system. The low-level extracts skeleton data from the optical stream, forms clusters and uses them to represent actions as transitions through clusters. The high-level uses this encoding to probabilistically infer the goal pursued by the partner. The robot control is in charge of performing the collaborative behavior.

by operating on an RGB image but on skeletal data extracted by means of computer vision algorithms as OpenPose [18] or depth sensors as the Microsoft Kinect [19]. Plan recognition, in contrast, is usually performed by means of probabilistic modeling such as static or dynamic Bayesian networks [12], first order logic [20] or hybrid architectures [21]. These models often assume the availability of a plan library capturing the system's knowledge about goals and means of accomplishing them [22].

III. Proposed Method

The objective of our research is to develop an artificial cognitive architecture able to perform intention reading on a partner. This model will be used in a joint action context where a human and a robot need to cooperate to reach a shared goal. In such a scenario, the role of the robot will be to observe its human companion initiate a task, infer the underlying intention and offer to collaborate, intervening in the ongoing plan. To do so, we adopted the developmental robotics paradigm, defined by Cangelosi [1] 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 the pursuit of this approach, the system will not be provided with a pre-existing plan library, conversely it will be designed following the psychological theories which state that intention

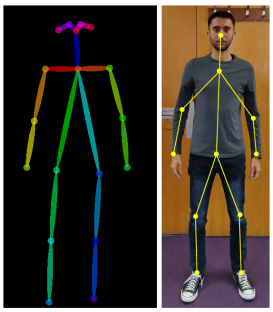

Fig. 2. A comparison between the skeletal keypoints extracted from the camera image (a) and the reduced keypoint set computed by the system (b).

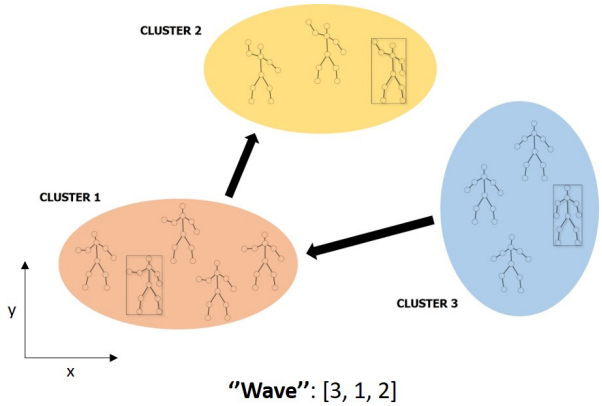

Fig. 3. Skeleton clustering. The observed skeletons are clustered into groups that represent similar but not identical postures. An action is defined as a particular transition through clusters. In this example, there are three clusters and the goal "wave" is performed by a sequence of body postures that transition from cluster 3 to cluster 1 to cluster 2, hence the encoding [3, 1, 2].

reading is developed through experience in the very first months of life [4]. It will also follow the low- and high-level subdivision of actions and goals theorized by Tomasello [8], performing a low-level body movement analysis that will provide inputs for a high-level goal determination mechanism. Our approach is supported by evidence that infants recognize biological motion and use it as the main clue to read others' intentions [23].

An overview of the architecture is shown in Figure 1. The system is designed to undergo a training phase in which a demonstrating agent will teach it an open set of goals and the actions needed to achieve them and an execution phase in which it will use this knowledge to infer as soon as possible the partner's intentions.

A. Low Level Skeleton Clustering and Action Representation

1) Skeleton Generation: The low-level deals with human skeletal data generation and manipulation. The former is achieved through the use of a publicly available and pre-trained deep CNN architecture named OpenPose [18], specialized in real-time multi-person 2D pose estimation. This neural computation receives in input 320x240 images from the robot's eye cameras and outputs a 18x2 feature vector representing the detected skeleton keypoints as 18 joints expressed in 2D spatial coordinates, as reported in Figure 2a. Because it has been proven that classification tasks achieve better results with a reduced set of joints [19] and to optimize memory and speed requirements, we operate a keypoint reduction to diminish the volume of data required for each skeleton. To do so, we discard the keypoints corresponding to the eyes, the ears and the shoulders, whilst calculating a new torso keypoint as a median between the two hips: doing so, we obtain a more compact 11x2 representation shown in Figure 2b.

The skeletons generated by this procedure cannot be used directly for classification purposes, as they are dependent on the position and size of the subject. To overcome this problem and gain spacial invariance, we apply a normalization process first introduced by Cippitelli [24]. For a skeleton with n joints, the feature vector f is defined as:

$$f = [J_1, J_2, ..., J_n] \tag{1}$$

where J_i is a vector containing the normalized 2D coordinates of the ith keypoint:

$$J_i = \frac{J_i - J_0}{\|J_1 - J_0\|} \tag{2}$$

where J_0 and J_1 are, respectively, the neck and torso joint. The latter will be located on the origin of the cartesian space, so its components will all be zero. For this reason, it is removed from the feature vector, which at this point will have a dimension of 10x2: this corresponds to a size reduction of approximately 44.5% compared to the original representation.

2) Dynamical Clustering: Skeleton manipulation for action representation is done through an unsupervised clustering process. During the training phase, a teacher will demonstrate a set of goals and the actions required to achieve them. The robot will observe and detect, process and memorize skeletons from the video stream with an empirically chosen frequency of 2 fps. When the demonstration ends, the robot will perform a clustering operation on the set of acquired skeletons. Each of the latter is represented in a high, 20-dimensional space which will subject the dataset to the curse of dimensionality [25]. In addition, clustering relies on euclidean distance as a metric, but in high dimensional spaces the concept of distance becomes less precise, since it tends to converge. To overcome these problems, we apply Principal Component Analysis [26] to perform a dimensionality reduction by projecting the 20D feature vectors on a 2D space. At this point, clustering can be performed. We employ X-Means [27], a variation on the traditional K-Means algorithm that overcomes its principal limitation: the need to manually specify the parameter K that defines the desired number of clusters. The model selects the optimal one by performing model selection among a finite set of models through the optimization of the Bayesian Information Criterion. At the end of this process, the system will possess the 2D coordinates of the n_c centroids and each

skeleton will have been grouped to the closest one. Each cluster will represent a group of similar but not identical postures.

Once the clusters are obtained, transition analysis is performed: each action to be learned is encoded as a sequence of integers that represent the temporally ordered ids of the clusters through which the demonstrator transitioned during the performance of the action, as shown in Figure 3. We impose an important restriction: we include in the representation only the transitions through different clusters, discarding the persistence in the same group. This is done to make the final representation independent from the speed at which the action is performed, in fact gaining temporal invariance. Once computed, these action encodings are sent to the high-level module to train it as described in Section III-B.

The execution phase relies on the data structures built during training: it receives a stream of camera images and for each of them it tries to detect a skeleton, process it as described in Section III-A1, project it on the reduced 2D space and associate it to one of the clusters. Once the cluster id is obtained, it is immediately forwarded to the high-level module.

B. High Level Goal Probabilistic Inference

This module deals with probabilistic intention inference using the results of the low-level computation on the partner's body posture: we need to predict the other agent's goals based on its actions. In other words, we are trying to infer a hidden, unobservable state from a series of observable events: this is a classic hidden Markov process. However, the application of a Hidden Markov Model is not possible because in this scenario each state can emit a sequence of observations and not just a single one (a goal will be formed from a series of low-level cluster transitions). To take this into account, we adopt a Hidden Semi-Markov Model (HSMM) [28] as the high-level computational model: in it, the probability of a state change depends not only on the observable events but also on the amount of time elapsed since the entry in the current state. During the training of the system, the action representations learned from the low-level module are used to tune the parameter matrices of the HSMM through a supervised Maximum Likelihood Estimation process. During the execution phase, the high-level receives in real time from the low-level a stream of single transitions observed in real time from the partner's posture and applies Viterbi's algorithm to predict which hidden state is generating the observed events.

This is, however, not sufficient: as we want the robot to act collaboratively before the action itself has ended, we are pursuing action prediction, not simply recognition. This means that we are trying to work with incomplete data, because we want to guess the hidden state before the sequence of observations is fully complete. To solve this problem, we support HSMM with the Anticipator: a module that tries to reconstruct in advance the missing observations, artificially generating enough data for the HSMM to look ahead of time and decode the goal correctly. Once the HSMM is able to formulate a prediction, the Anticipator tries to match the sequence of incremental input observations with the ones learned during the training phase: if an unambiguous match is possible, it will emit the remaining symbols of that sequence for the HSMM to perform a second inference, otherwise the previous one will be maintained.

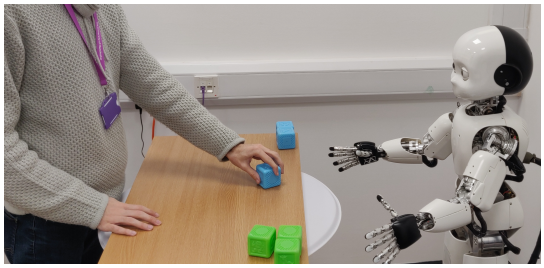

Fig. 4. Experimental setup for the block building game.

As soon as one of the goals is confidently predicted, it is sent forward in the processing chain to instantiate appropriate collaborative behavior.

C. Robot Control

The robot control module deals with the direct interface between the cognitive architecture and the iCub robot: a humanoid research platform that resembles a three and half year-old child with 53 degrees of freedom [29]. In particular, it provides interaction with its YARP middleware to control the sensors and actuators and perform vision, movement and grasping tasks for the shared goal task.

IV. Experiments

The goal of this experiment is to validate the intention reading capabilities of the cognitive architecture on a shared goal scenario. In particular, it involves a collaborative block building game where the robot will help the human to assemble constructions using different colored toy blocks.

A. Experimental Setup

Figure 4 shows the experimental setup. The iCub and the experimenter are facing each other with a table in between. Some colored toy cubes are situated next to the robot: greens to its right and blues to its left: these are denominated the green and blue block areas. A playing area, where the buildings will be assembled, is situated in between.

B. Procedure

The toy cubes can be used to build four different kind of structures: towers, walls, castles and stables, each of one consisting of three blocks. The first two may only be made of one single color each, c_1 and c_2, whilst the castle and the stable must be built using blocks from both groups. An example of these configurations can be seen in Figure 5. At the beginning of the experiment, iCub will not know the rules of the game and the experimenter will demonstrate all the intentions by performing them one at a time and giving them a name. In the second phase, the human experimenter will start

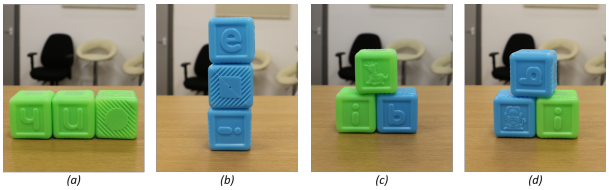

Fig. 5. An example of the four structures taken into account in the experiment: *(a)* wall, *(b)* tower, *(c)* castle and *(d)* stable. In this case, $c_1 = green \wedge c_2 = blue$

pursuing one of the goals and the robot will try to predict the intention as soon as possible and act in a helpful manner, that means fetching the remaining blocks needed to finalize the construction. It is important to note that the experimenter can choose to demonstrate either $c_1 = blue \wedge c_2 = green$ or $c_1 = green \wedge c_2 = blue$.

A video demonstration of the experiment is available online [30].

V. Results and Discussion

Figure 6 shows the resulting knowledge base of the robot after the training phase: the learning data was autonomously divided in three clusters, each of which represents a particular pose of the human partner. The average Silhouette evaluation metric [31] for this clustering is $S = 0.858$, which indicates that the number and placement of the centroids is adequate.

Each intention was then encoded as a series of transitions between these clusters: details are reported in Table I. To better understand the latter, Figure 7 visualizes the recorded skeletons that, during the teaching of the goal "wall", have triggered cluster transitions. As noticeable from the skeleton ids, which don't progress linearly, the transition analysis method was successful in filtering out uninteresting data and provide temporal invariance to the encoding.

Finally, the performance of the high-level inference during the execution or playing phase, consisting of 80 games with the robot, is illustrated in Table II. The latter shows, in order: the goal which was tested, number of trials performed, accuracy of the prediction, average number of cluster transitions encountered before formulating an hypothesis, average time (in seconds) necessary for the inference and average percentage of completion of the action by the partner when the intention was predicted. Note how the synergic combination of the HSMM and Anticipator has led to an intention prediction that was both perfectly accurate (these results compare well with the ones obtained from Manzi [19]) and fast enough to make the robot engage collaboratively when the partner had completed more or less half of the action (higher percentages of completion would diminish the usefulness of the collaborative robot, as the partner would already have completed most of the action without assistance).

VI. Conclusion and Future Work

In this paper, intention reading has been explored by a psychological, developmental and computational point of view.

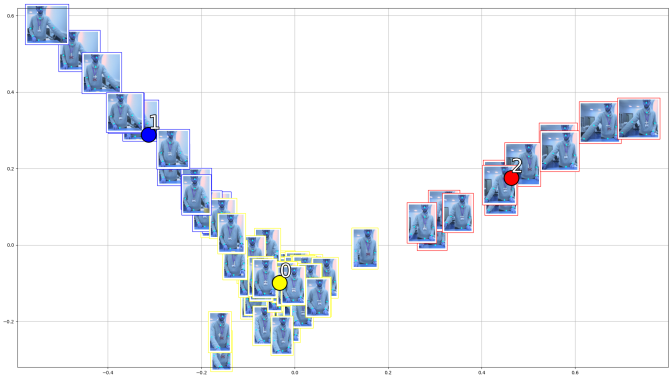

Fig. 6. The 2D projection of the resulting clustering obtained during the training phase of the block building game. Skeletons have been divided into three groups of similar postures and the numbered dots represent the centroids.

TABLE I
Trained experimental intentions: each goal is associated to a sequence of cluster transitions.

Goal	Action representation
wall	[0, 2, 0, 2, 0, 2, 0]
tower	[0, 1, 0, 1, 0, 1, 0]
castle	[0, 2, 0, 1, 0, 2, 0]
stable	[0, 1, 0, 2, 0, 1, 0]

The evolutionary and cognitive importance of this skill has been universally recognized by researchers in various fields, so it is almost natural to think that even robots would benefit from it. Using the developmental robotics approach, this work aims to propose a new method inspired by scientific findings in human cognition.

We presented a novel artificial cognitive architecture able to perform intention reading on a humanoid partner for collaborative behavior towards the achievement of a shared goal that uses experience, as human children do. This model works on two levels: a low-level skeleton extraction and action

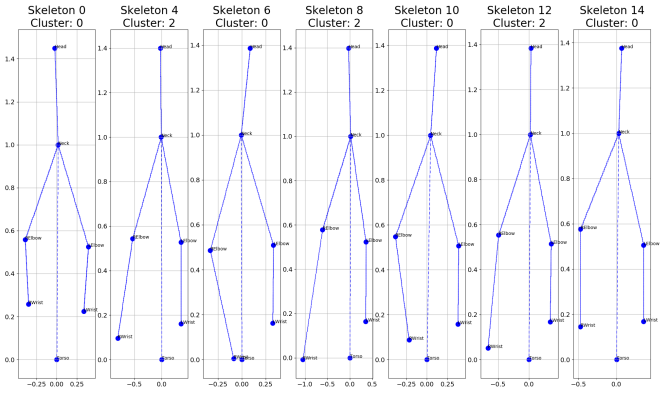

Fig. 7. Decomposition of one of the intentions that are trained to the robot during the block building game, corresponding to the goal "wall". The skeletons showed are the ones that triggered a cluster transition and show the visible upper body of the demonstrator manipulating blocks on the table.

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

Goal	Trials	Accuracy	Transitions	Time	Interception
Wall	20	100 %	4,00	4,41 s	57,14 %
Tower	20	100 %	4,05	3,69 s	57,86 %
Castle	20	100 %	4,00	4,63 s	57,14 %
Stable	20	100 %	4,05	5,22 s	57,86 %
Total	80	100 %	4,03	4,49 s	57,5 %

representation and a high-level goal probabilistic inference. The system has been validated on a human-robot interaction experiment consisting in a block building game where the iCub humanoid robot was initially instructed on the rules and, after that, was able to understand which goal was being pursued by its partner and act helpfully towards its achievement.

Many models in the current literature share the need for handcrafted plan libraries, limiting their capacities and flexibility. By using the developmental robotics approach, we managed to overcome this limitation: the set of goals learnable by the robot is open-ended. In addition to this, our approach proves to be robust, meaning that the low-level is both space and time invariant. Finally, unsupervised learning makes the robot learn on the fly with no need for big datasets or long traning processes, making this architecture extremely lightweight on a computational point of view. Despite this model being designed and tested for human-robot interaction, the way in which skeletons are detected from the visual field makes possible the use of this architecture even for humanoid-robot interaction scenarios, opening opportunities for new and different kind of studies.

We plan on extending this architecture by including new sources of information and by implementing hierarchical high-levels. In addition to this, we believe that true collaborative intelligence can emerge from the integration with a model of trust and theory of mind. For example, if the robot understands that its human partner is not able to fulfill the required steps to achieve a specific goal, it can try to act in a way that will ensure the successful completion of the task. For this reason, we have planned an integration with our previous work [32].

REFERENCES

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

[2] 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.

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

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

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

[6] Gergely Csibra. Teleological and referential understanding of action in infancy. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 358(1431):447–458, 2003.

[7] 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.

[8] 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.

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

[10] 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.

[11] Z Zenn Bien, Kwang-Hyun Park, Jin-Woo Jung, and Jun-Hyeong Do. Intention reading is essential in human-friendly interfaces for the elderly and the handicapped. *IEEE Transactions on Industrial Electronics*, 52(6):1500–1505, 2005.

[12] Karim A Tahboub. Intelligent human-machine interaction based on dynamic bayesian networks probabilistic intention recognition. *Journal of Intelligent & Robotic Systems*, 45(1):31–52, 2006.

[13] Miriam Buonamente, Haris Dindo, and Magnus Johnsson. Discriminating and simulating actions with the associative self-organising map. *Connection Science*, 27(2):118–136, 2015.

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

[15] 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.

[16] Volker Krüger, Danica Kragic, Aleš Ude, and Christopher Geib. The meaning of action: A review on action recognition and mapping. *Advanced robotics*, 21(13):1473–1501, 2007.

[17] 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.

[18] 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.

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

[20] Henry A Kautz and James F Allen. Generalized plan recognition. In *AAAI*, volume 86, page 5, 1986.

[21] Haris Dindo and Antonio Chella. What will you do next? a cognitive model for understanding others' intentions based on shared representations. In *International Conference on Virtual, Augmented and Mixed Reality*, pages 253–266. Springer, 2013.

[22] Sandra Carberry. Techniques for plan recognition. *User Modeling and User-Adapted Interaction*, 11(1):31–48, 2001.

[23] Bennett I Bertenthal. Origins and early development of perception, action, and representation. *Annual review of psychology*, 47(1):431–459, 1996.

[24] Enea Cippitelli, Samuele Gasparrini, Ennio Gambi, and Susanna Spinsante. A human activity recognition system using skeleton data from RGBD sensors. *Computational intelligence and neuroscience*, 2016:21, 2016.

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

[26] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[27] 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.

[28] Shun-Zheng Yu. Hidden semi-markov models. *Artificial intelligence*, 174(2):215–243, 2010.

[29] Giorgio Metta, Giulio Sandini, David Vernon, Lorenzo Natale, and Francesco Nori. The iCub humanoid robot: an open platform for research in embodied cognition. In *Proceedings of the 8th workshop on performance metrics for intelligent systems*, pages 50–56. ACM, 2008.

[30] Samuele Vinanzi. Experiment video. http://youtu.be/BV41zoTtMtw.

[31] 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.

[32] 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.