Towards Anchoring Self-Learned Representations to Those of Other Agents

Martina Zambelli Tobias Fischer Maxime Petit Hyung Jin Chang Antoine Cully Yiannis Demiris

Personal Robotics Laboratory, Department of Electrical and Electronic Engineering
Imperial College London, United Kingdom
{m.zambelli13, t.fischer, m.petit, hj.chang, a.cully, y.demiris}@imperial.ac.uk

***Abstract*—In the future, robots will support humans in their every day activities. One particular challenge that robots will face is understanding and reasoning about the actions of other agents in order to cooperate effectively with humans. We propose to tackle this using a developmental framework, where the robot incrementally acquires knowledge, and in particular 1) self-learns a mapping between motor commands and sensory consequences, 2) rapidly acquires primitives and complex actions by verbal descriptions and instructions from a human partner, 3) discovers correspondences between the robots body and other articulated objects and agents, and 4) employs these correspondences to transfer the knowledge acquired from the robots point of view to the viewpoint of the other agent. We show that our approach requires very little *a-priori* knowledge to achieve imitation learning, to find correspondent body parts of humans, and allows taking the perspective of another agent. This represents a step towards the emergence of a mirror neuron like system based on self-learned representations.**

I. Introduction

Daily-life home environments are typical examples where robots are expected to provide a tremendous amount of support in our day-to-day duties. However, in spite of the recent advances in robotics [1, 2], the currently closest form of such an assistant is only capable to achieve a limited number of pre-programmed tasks, such as dusting the floor, or preparing meals. Before seeing robots able to assist humans in their daily chores, many scientific challenges need to be addressed. For example, robots need to be able to learn how to deal with new situations on their own (*e.g.* performing new tasks or using new objects) [3] without requiring the intervention of an engineer, as every tool and every house is different. Moreover, they need to understand the point of views and the different abilities of other people for better interaction and collaboration. Such perspective taking abilities, for instance, reduce the ambiguity in the interactions [4] and allow robots to adjust their actions according to the abilities of its users [5].

In this paper, we present a robotics architecture for anchoring representations that are autonomously learned by the robot into the perspective of other agents (see Fig. 1). This architecture is composed of five components:

1) Multimodal Sensorimotor Representation component [6],
2) Symbolic Representation component [7],
3) Kinematic Structure Correspondence component [8],
4) Perspective Taking component [9], and
5) Autobiographical Memory component [10].

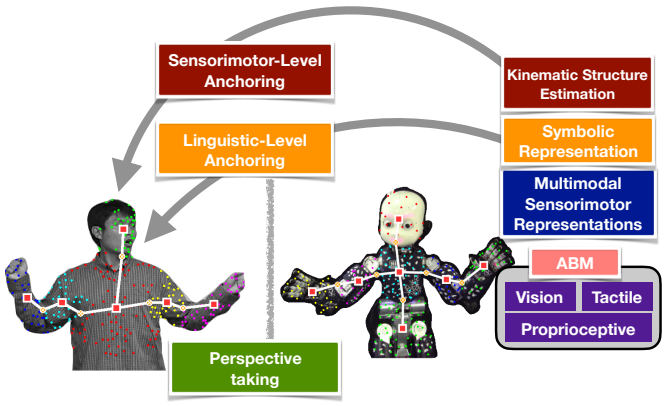


Fig. 1. Overview of the proposed architecture. The robot learns multi-modal sensorimotor representations, and employs this knowledge to anchor self-knowledge to that of other agents on the linguistic and sensorimotor levels. This can then be used by the perspective taking component to reason about the other agents' perspective. The common interface is provided by an autobiographical memory (ABM).

The first four components are tied together with a long-term **Autobiographical Memory** [10], which is used as a common interface to exchange a) streaming data originating from the robots' sensors, b) their internal representations, and c) augmented versions of the data. In other words, it is designed to store annotated multi-modal data, and allows the augmentation of these data when new knowledge concepts emerge from reasoning modules.

The **Multimodal Sensorimotor Representation** component [6] allows the robot to self-learn sensorimotor representations from visual, proprioceptive and tactile stimuli. This knowledge can be employed by two other components which anchor the learned self-representations to the representations of others. The **Symbolic Representation** component [7] performs anchoring on the linguistic level (*e.g.* joint number to body part name); whereas the **Kinematic Structure Correspondences** component [8] anchors the representations on the sensorimotor level (*e.g.* body part correspondences between two agents).

The combination of these components allows our robot to not only rapidly perform imitation learning , but also to reason about the observed actions based on representations acquired from self-exploration. This follows the developmental principle shown by Baraglia *et al.* [11], where action production alters action perception. Moreover, using the **Perspective Taking**

component [9], we are able to use spatial reasoning algorithms which were learned from the robot's perspective, to also reason from the human's perspective.

By taking advantage of the synergies between its different components, our architecture is a step towards an implementation of the "others like me" paradigm [12, 13]. This paradigm is based on self-exploration and self-others mapping. It allows the robot to expand its understanding of the actions performed by others, and their underlying intentions. The paradigm is of particular interest for developmental robotics, due to the potential reuse of learned models in bootstrapping the learning of further knowledge from other agents.

The remainder of the paper is organised as follows. We first briefly review some related works in Section II, and then introduce each of the individual components and their contributions to the overall architecture in Section III. In Section IV, we discuss the advantages and limitations of our architecture. We propose experiments to evaluate our architecture in Section V along with a conclusion.

II. Related Work

To the best of our knowledge, our proposed architecture is the first architecture that is able to simultaneously discover sensorimotor contingencies, to extract higher level representation, to detect kinematic correspondences with other agents, and to project these self-learned representations into the point of view of these agents. However, several works can be compared to sub-parts or functionalities of our architecture.

For instance, Tani *et al.* [14] created a connectionist model that allows robots to both generate and recognise behaviour patterns based on recurrent neural networks. It has been used for imitative interactions, action learning, and linguistic behaviour bindings. While these applications represent important abilities for a robot, this work is targeted towards self-learning and generalisation, and thus does not provide insight into the behaviour or capabilities of others.

Conversely, several works consider the generation of a mirror neuron system [15] used to distinguish and recognise actions performed by others from those executed by the robot itself. Among those, Nagai *et al.* [16] investigated the emergence of such a mirror neuron system from an immature visual system. However, several open questions remain: firstly, the integration of other sensory modalities (*e.g.* tactile and auditory perceptions), and secondly the scalability of the system toward more complex actions and representations. Similarly, Rebrova *et al.* [17] proposed a mirror system for a simulated iCub robotics platform that links the observed actions with the respective motor commands. Such a mapping facilitates the perspective taking mechanisms when observing an agent. In a first phase, the robot learns to associate its own actions with an egocentric observation, and in a second phase, it uses bidirectional associations to estimate the observation of the self-movement from another perspective. However, the perspective taking does not take the specific view of a human into account, but rather consists of a geometric change in the frame using the full knowledge of a simulated world, making its application in real world conditions challenging.

An overview of experimental data which supports the mechanisms of a mirror-neuron system can be found in the work of Demiris *et al.* [18]. The authors also highlight that there are only very few neural models of the mirror system, whereas most models which are implemented on a robot are based on internal action models (and so is our model). Interestingly, they have shown that decomposing models of the mirror system into brain operating principles can be used to compare these models, and link them to neuroimaging data in order to find common predictions among other things.

The architecture presented in this paper attempts to overcome these different challenges by combining the abilities of its different modules. This results in a physical robotic platform with a large set of abilities ranging from low-level sensorimotor contingencies learning to high-level perspective taking.

III. Material and Methods

In this section we introduce the different components of the architecture. In order to provide a general overview of the architecture, we limit the description of technical details in this paper. More information can be found in the publications associated with each component.

A. Long-term Autobiographical Memory

In a real-life condition, the training-data acquired and used by the robot presents specific particularities. First, the data represents a multi-modal stream, in our case containing proprioception, visual and tactile information. Second, the data is scattered in time, as the robot has to infer relationships between events that happened at different moments. For example, the robot may need to associate actions performed previously with labels provided by the user at any moment. To fulfil these requirements, the framework is based on the implementation of a long term autobiographical memory that is able to store raw data along with augmented memories from different reasoning modules [10].

For instance, the sensorimotor contingency exploration creates memories that are later augmented by the kinematic structure estimator. Then, the estimated kinematic structures of self and others, which are anchored in the autobiographical memory, can be compared and correspondences found. As the robot can also acquire self-knowledge in a symbolic form (*e.g.* naming of its body parts), this self-knowledge can later be transferred to other agents using the previously found correspondences.

B. Multimodal Sensorimotor Representations for Imitation Learning

Humans perceive and interact with their environment using a large variety of senses, even when they accomplish simple tasks, like drawing a circle. For example, when finger-painting, they engage proprioception to perform smooth movements, vision to check and adjust their motion, and touch to know when their fingers are in contact with the surface. However, in

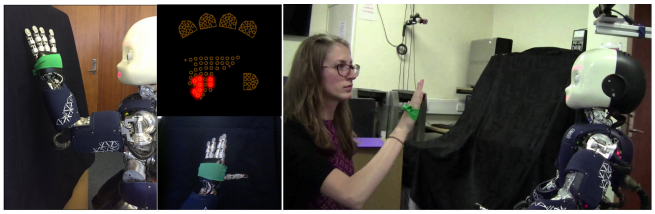


Fig. 2. (Left) iCub faces a surface and performs movements in order to touch it, with the palm touch sensor being activated during touch events (centre top) and the view from the robot's left eye camera (centre bottom). (Right) iCub learns to imitate a circular trajectory demonstrated by a human.

spite of the wide sensing abilities of modern robots, which are provided with a variety of different sensors, the vast majority of approaches using imitation learning relies only on data coming from a single modality (*i.e.* vision) [19, 20].

The architecture presented in this paper relies on a new component that learns the multimodal sensorimotor contingencies that the robot encounters from visual, proprioceptive and tactile stimuli [6]. This component builds data matrices from the multimodal perceptions in order to encode the contingencies. Thanks to these matrices, a variation in the proprioception is associated to variations in the other "senses" of the robot. Based on this representation, the robot can determine the motor commands that are likely to produce the desired changes in its perception in order to achieve multi-modal imitation tasks.

The construction of these matrices of data is achieved through a motor babbling sequence in which the robot performs random motion and observes the consequences of its movements. The main advantage of this approach is that it does not require explicit model formulation or the inversion of complex kinematic problems. Conversely, it only relies on observed data, which makes the application of this method possible on a large variety of systems.

This approach has been evaluated in two situations where a humanoid iCub robot 1) has to learn how to draw a circle on a board, and 2) learns how to press two keys on a piano. In these two tasks, the robot aims to reproduce a visual trajectory demonstrated by the user, while fulfilling requirements on the other sensory spaces, *e.g.* touching a surface or the keyboard (see Fig. 2). Successful imitation behaviours have been obtained even with a limited number of samples (respectively 70 and 135 samples were recorded during the motor babbling sequences).

These experimental results demonstrate that this approach allows robots to achieve multi-modal imitation tasks with no *a-priori* knowledge about themselves or about the task. Moreover, this approach puts almost no constraint on the type of sensory information used, making it particularly scalable in terms of modalities that can be combined. For example, the experiments presented previously will be extended by incorporating audio perceptions.

C. Symbolic Knowledge Acquisition from Human-Robot Interactions

In this section, we focus on the anchoring of self-representations at the linguistic level, in order to create concepts about body parts or actions. The human partner acts as a tutor, who provides a common ground for future communication. Specifically, we made progress toward sequence learning of body movements when the robot cannot rely on prior knowledge of body parts or motor skills. Rather, the robot engages in a social interaction with a nearby human to acquire this information, in line with the works of Heinrich *et al.* for object learning [21] and Petit *et al.* for shared plan learning [22].

The framework [7] is composed of three hierarchically organised components to 1) learn body part names from human labelling after a robot motor babbling activity, 2) discover proto-actions (*i.e.* a single position command of a single joint) with on-the-fly descriptions by the human of these motor babbling activities, and 3) learning by instructions with the human as a teacher to scaffold the newly acquired skills into motor primitives or more complex actions.

The framework relies on the direct grounding and transfer grounding mechanisms defined by Cangelosi and Riga [23], which are used to ground language into autonomous cognitive systems. Direct grounding is the capacity to link internal and perceptual representations to symbols with a supervised feedback, and is used to learn body part names or proto-actions concepts. Transfer grounding creates new symbols by combining already known symbols, and is used to acquire new motor primitives or complex actions using learning by instructions.

An important aspect of our framework is that the learning by instruction does not require *a-priori* knowledge about available actions because the proto-actions are learned from interactions with the human.

Another advantage of the framework is that only a small amount of data is required to learn new concepts. For instance, we used only 3 motor babbling sequences of 18 seconds each (with a velocity change every 3 seconds) to create the fold and unfold proto-actions with the thumb, index, and ring fingers. The use of a linear model to estimate the desired angle that a joint should have in order to produce a specific proto-action is providing 1) the effect of the proto-action *per se*, and 2) the potential correction for the body part used. A generalisation of the proto-action is thus feasible. For example, to generate proto-actions of body parts that have never been used, the body part correction effect is set to 0. This knowledge allows a human to subsequently teach new primitive capabilities by scaffolding the available proto-actions using a learning by instructions method. Similarly, more complex actions can be taught using not only the proto-actions but also the primitives, which reduces the number of instructions needed.

D. Anchoring Kinematic Structure Representation to Others

We find correspondences within the sensorimotor level based on the estimated kinematic structures of agents. A kinematic structure represents articulated objects in a topological manner, and describes how rigid body parts are connected by kinematic joints. We tackled the problem of complex kinematic structure estimation, demonstrating that combining motion information

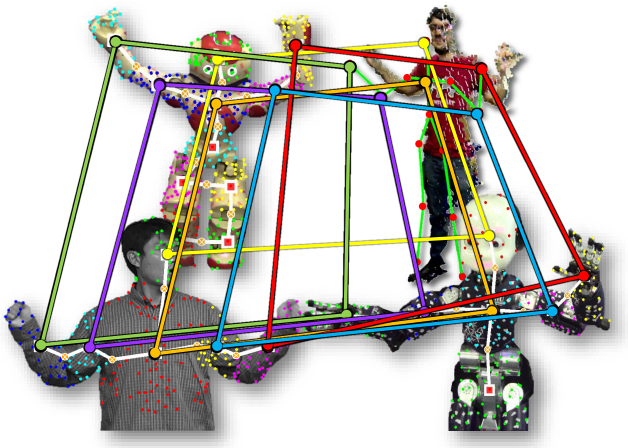
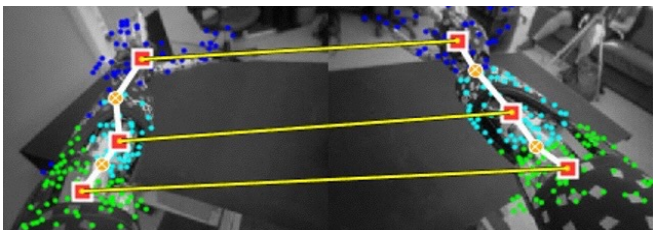
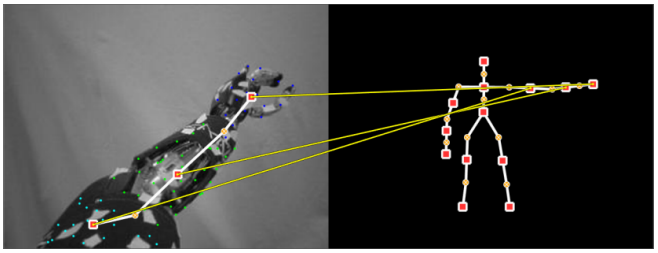


Fig. 3. Various kinematic structure correspondence matching results using the proposed method. The iCub humanoid robot (bottom right) can find correspondences to human partners, either sensed through the iCub eyes (bottom left) or a RGB-D camera (top right). Also, correspondences to other humanoid robots like the NAO robot (top left) can be found. Figure from [8].



(a) Correspondence iCub self-body to self



(b) Correspondence iCub self-body to others

Fig. 4. The proposed method can be used (a) for an iCub robot to find correspondences between iCub's partial arm structure captured using the iCub's RGB camera to its own other body part, and even (b) to full body human structure captured using RGB-D camera.

and skeletal topology can be used reliably to estimate the kinematic structure of the self (*i.e.* the robot after self-observation) and that of bodies of other agents [24].

Based on this work, we developed a method that allows a robot to anchor two objects' kinematic structure joints by observing object movements [8]. We formulated the problem of finding kinematic joint matches between two articulated kinematic structures via hypergraph matching. The method was shown to be accurate under appearance and motion variations.

As shown in Fig. 3, our method is able to anchor similar kinematic structure joints even between visually very different appearances, and in the presence of strong motion variations. Furthermore, we consider the kinematic structure as a mid-level representation, so the proposed method can be applied to any kind of input device as long as the kinematic structure can be produced. For example, the kinematic structure of a human extracted by a RGB-D camera can be matched to the observation of the same human by the robot's eye cameras, and similarly the observation can also be matched to a self-observation of *e.g.* the robots arms (see Fig. 4).

E. Perspective Taking and the Mirror Neuron System

The Perspective Taking component of our architecture takes inspiration from the simulation theory of mind, which proposes that humans use their own mental processes applied to the perspectives of other agents to understand their internal states. However, the actions resulting from these mental processes are only simulated rather than executed [13]. Various works have shown the importance of perspective taking abilities in the context of human-robot interactions (see [9] for an overview). For example, it was shown that, when using perspective taking, a robot can learn from ambiguous demonstrations [4].

In the previous sections, we have demonstrated that our robot can find correspondences to other agents in both, the linguistic as well as the sensorimotor levels. Here, we present how these correspondences can be used to understand others through a table top scenario as shown in Fig. 5 and by using our Perspective Taking component [9].

This component is based on three aspects to perceive the environment. Firstly, as no prior knowledge about the environment is assumed, we map the environment using random exploration. Secondly, while interacting with the human, the head pose of the human is continuously estimated. Thirdly, objects in front of the robot are recognised and tracked in real time.

Based on these three sources of information, it is possible to perform mental rotations of the robot's perception such that the coordinates origin coincides with the head frame of the human rather than the robot. This new, self-learned representation allows the robot to reason from the user's perspective without any changes to the underlying algorithms. For example, our experimental results have demonstrated that the spatial error is sufficiently small when observing where humans are looking so that it allows the robot to estimate whether an object is to the left or to the right of the human and if this objective is visible to the human.

Interestingly, it was suggested that not all perspective taking tasks might rely on mental rotation [25]. Rather, simple tasks such as finding whether another object is visible from the others perspective (level 1 perspective taking) might be solved using a line of sight tracing. Only more complex tasks such as the earlier mentioned left/right judgements or imagining how the world visually looks from another perspective (level 2 perspective taking) might rely on a mental rotation. Thus our Perspective Taking component differentiates these two pathways and implements both of them separately, as shown in Fig. 6.

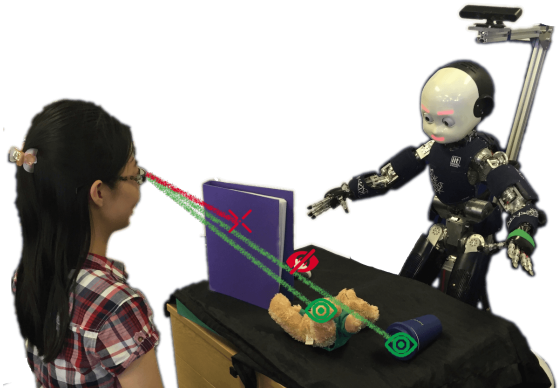


Fig. 5. Typical set-up in a perspective taking scenario. The world perceived by the human and that of the robot differ in various aspects. For example, in this figure, one object is occluded to the human but visible by the robot. Furthermore, the blue cup is to the left of the robot, but to the right of the human. Figure best viewed in colour.

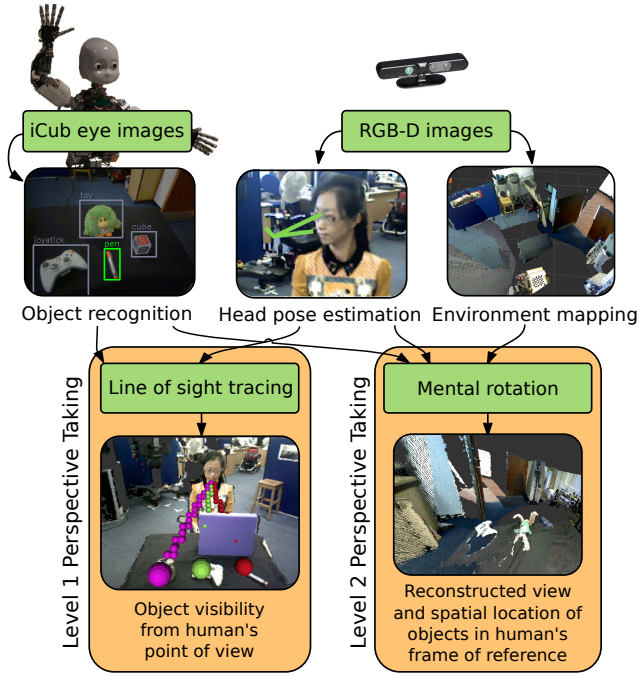


Fig. 6. Overall flow of the Perspective Taking method, with images acquired by RGB-D camera or iCub eyes as input. In the first step, the robot recognizes objects, estimates the head pose of surrounding humans, and maps the environment. Two separate processes are employed for level 1 and level 2 perspective taking. Figure from [9].

IV. Discussion

In this paper, we presented a robotic architecture for anchoring representations autonomously learned by the robot into the perspective of other agents. The robot progressively builds these high-level representations by discovering the sensorimotor contingencies and by interacting with the users.

While the current architecture increases the interaction abilities of the robot by augmenting its understanding of the user's actions, we forecast that the need of scalability will be the next challenge that the architecture will have to face. For example, the complexity of human-robot interactions, particularly in home environments, requires the robot to rely on its entire range of motion and reasoning capabilities. The robot does not only need to identify the users' abilities, but also how they change over time or with the use of tools, and to adapt accordingly to the humans' behaviour. Each of these points represent open scientific questions. However, the hierarchical structure of the proposed architecture offers a singular perspective to decompose these scientific questions into smaller problems and to make progresses in this direction.

Our ambition with this architecture is to design and implement a developmental approach to generate a mirror neuron-like system by bootstrapping the understanding of the actions of others based on the learned model of the self. This development is thought to take place by progressively augmenting each layer of the architecture. For example, a progressive increase in either the number of modalities or in the number of degrees of freedoms used in the Multimodal Sensorimotor Contingency component will result in an augmentation of the scope of possibilities for the emergence of symbolic representations (*e.g.* auditory sensations and dual-handed actions). The augmentation of the motion abilities of the robot is also expected to have an impact on the range of actions performed by other agents that the robot can understand or reproduce, via the dual kinematic structures correspondences. Finally, we also anticipate that this enlarged set of skills will allow the robots to extend its perspective taking abilities through a better understanding of the complex actions and interactions carried out by other agents. In other words, we will experiment with this architecture if the scalability of the entire framework can be improved by augmenting the scalability of each of its components.

V. Conclusion and Future Work

Our framework is a step towards the emergence of a mirror neuron like system. With this framework, the robot is currently able to predict the consequences of movements performed by other agents, based on the learned consequences of its own movements, and assist humans effectively in cooperative tasks. The key property of this system is that does not use any prior knowledge about the body schema of the human. Therefore, the robot is able to indifferently adapt to the physical limitations of the users, which is of great importance for various robotics applications, such as assistive robots for health-care.

We want to investigate applications of our architecture with *in-situ* experiments. For example, we are going to demonstrate the capabilities of our architecture in a home scenario, where a robot is helping a human in daily life tasks such as cleaning a table or assisting in cooking. Our robot will be able to rapidly learn fulfilling new tasks, as they can be acquired through a combination of learning by imitation and instructions. The acquired representations can then be applied to a human in a personalised manner (*i.e.* depending on the humans capabilities and the current situation), which allows anticipating human behaviour.

Specifically, we consider the following scenario: at the beginning, the robot does random exploration to acquire a self-representation, *e.g.* its own kinematic structure and its body part names. Then, a human teacher is instructing the robot with an action sequence. The robot starts acquiring knowledge about others while being taught. For example, the robot extracts the kinematic structure of the human. In another step, the knowledge about the human is augmented by finding correspondences to its own body. Thus, without any further guidance the robot knows the body part names of the human based on its self-representation and the found correspondences.

The correspondences can also be used to refine the learnt action sequences, and the perspective taking abilities can be used to detect potential flaws during the action execution. For example, the robot can inform the human about objects which are hidden from the humans perspective. Similarly, the robot will be able to predict potential problems in the task execution for users with limited mobility; without explicitly being informed about the limitation but rather by observing the movements of the user.

In the longer term, we also want to explore the social aspects of the mirror system. We are going to use the perspective taking capabilities to estimate the intentions and goals of humans, and are going to model the underlying desires that drive their behaviours. This will allow the robot to improve its capacity to coordinate with the human by not only detecting the current action among a sequence that the human is executing, but also by predicting the goal of such a plan. Then, the robot can provide appropriate support in real-time, or it can follow a contingency plan when upcoming problems are detected.

ACKNOWLEDGEMENT

This work was supported by the EU FP7 project WYSIWYD (612139). The authors gratefully acknowledge the support from the members of the Personal Robotics Lab.

REFERENCES

[1] L. Iocchi, D. Holz, J. Ruiz-Del-Solar, K. Sugiura, and T. Van Der Zant, "RoboCup@Home: Analysis and results of evolving competitions for domestic and service robots," *Artificial Intelligence*, vol. 229, pp. 258–281, 2015.

[2] H. Robinson, B. MacDonald, and E. Broadbent, "The Role of Healthcare Robots for Older People at Home: A Review," *International Journal of Social Robotics*, vol. 6, no. 4, pp. 575–591, 2014.

[3] A. Cully, J. Clune, D. Tarapore, and J.-B. Mouret, "Robots that can adapt like animals," *Nature*, vol. 521, no. 7553, pp. 503–507, 2015.

[4] C. Breazeal, M. Berlin, A. Brooks, J. Gray, and A. L. Thomaz, "Using perspective taking to learn from ambiguous demonstrations," *Robotics and Autonomous Systems*, vol. 54, no. 5, pp. 385–393, 2006.

[5] Y. Gao, H. J. Chang, and Y. Demiris, "User Modelling for Personalised Dressing Assistance by Humanoid Robots," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2015, pp. 1840–1845.

[6] M. Zambelli and Y. Demiris, "Multimodal Imitation Using Self-Learned Sensorimotor Representations," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2016.

[7] M. Petit and Y. Demiris, "Hierarchical Action Learning by Instruction Through Interactive Body Part and Proto-Action Grounding," in *IEEE International Conference on Robotics and Automation*, 2016, pp. 3375–3382.

[8] H. J. Chang, T. Fischer, M. Petit, M. Zambelli, and Y. Demiris, "Kinematic Structure Correspondences via Hypergraph Matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4216–4425.

[9] T. Fischer and Y. Demiris, "Markerless Perspective Taking for Humanoid Robots in Unconstrained Environments," in *IEEE International Conference on Robotics and Automation*, 2016, pp. 3309–3316.

[10] M. Petit, T. Fischer, and Y. Demiris, "Lifelong Augmentation of Multi-Modal Streaming Autobiographical Memories," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 8, no. 3, pp. 201–213, 2016.

[11] J. Baraglia, J. L. Copete, Y. Nagai, and M. Asada, "Motor Experience Alters Action Perception Through Predictive Learning of Sensorimotor Information," in *International Conference on Developmental Learning and Epigenetic Robotics*, 2015, pp. 63–69.

[12] A. N. Meltzoff, "'Like me': a foundation for social cognition," *Developmental Science*, vol. 10, no. 1, pp. 126–134, 2007.

[13] W. G. Kennedy, M. D. Bugajska, A. M. Harrison, and J. G. Trafton, "'Like-Me' Simulation as an Effective and Cognitively Plausible Basis for Social Robotics," *International Journal of Social Robotics*, vol. 1, no. 2, pp. 181–194, 2009.

[14] J. Tani, M. Ito, and Y. Sugita, "Self-Organization of Distributedly Represented Multiple Behavior Schemata in a Mirror System: Reviews of Robot Experiments Using RNNPB," *Neural Networks*, vol. 17, no. 8, pp. 1273–1289, 2004.

[15] G. Rizzolatti and C. Sinigaglia, *Mirrors in the brain: How our minds share actions and emotions*. Oxford University Press, 2008.

[16] Y. Nagai, Y. Kawai, and M. Asada, "Emergence of Mirror Neuron System: Immature vision leads to self-other correspondence," in *IEEE International Conference on Development and Learning*, 2011.

[17] K. Rebrova, M. Pecháč, and I. Farkaš, "Towards a robotic model of the mirror neuron system," in *IEEE International Conference on Development and Learning and Epigenetic Robotics*, 2013.

[18] Y. Demiris, L. Aziz-Zadeh, and J. Bonaiuto, "Information Processing in the Mirror Neuron System in Primates and Machines," *Neuroinformatics*, vol. 12, no. 1, pp. 63–91, 2014.

[19] J. Bandera, J. Rodriguez, L. Molina-Tanco, and A. Bandera, "A Survey of Vision-based Architectures for Robot Learning by Imitation," *International Journal of Humanoid Robotics*, vol. 9, no. 1, pp. 1–40, 2012.

[20] M. Hoffmann, H. Marques, A. Arieta, H. Sumioka, M. Lungarella, and R. Pfeifer, "Body Schema in Robotics: A Review," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 4, pp. 304–324, 2010.

[21] S. Heinrich, P. Folleher, P. Springst, E. Strahl, J. Twiefel, C. Weber, and S. Wermter, "Object Learning with Natural Language in a Distributed Intelligent System A Case Study of Human-Robot Interaction," in *IEEE International Conference on Cognitive Systems and Information Processing*, 2012, pp. 811–819.

[22] M. Petit, S. Lallée, J.-D. Boucher, G. Pointeau, P. Cheminade, D. Ognibene, E. Chinellato, U. Pattacini, I. Gori, U. Martinez-Hernandez *et al.*, "The coordinating role of language in real-time multimodal learning of cooperative tasks," *IEEE Transactions on Autonomous Mental Development*, vol. 5, no. 1, pp. 3–17, 2013.

[23] A. Cangelosi and T. Riga, "An Embodied Model for Sensorimotor Grounding and Grounding Transfer: Experiments With Epigenetic Robots," *Cognitive Science*, vol. 30, no. 4, pp. 673–689, 2006.

[24] H. J. Chang and Y. Demiris, "Unsupervised Learning of Complex Articulated Kinematic Structures combining Motion and Skeleton Information," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3138–3146.

[25] P. Michelon and J. M. Zacks, "Two kinds of visual perspective taking," *Perception & Psychophysics*, vol. 68, no. 2, pp. 327–337, 2006.