A Computational Model for Embodied Visual Perspective Taking: From Physical Movements to Mental Simulation

Tobias Fischer and Yiannis Demiris

Personal Robotics Laboratory, Electrical and Electronic Engineering Department, Imperial College London, UK

{t.fischer, y.demiris}@imperial.ac.uk

Abstract

To understand people and their intentions, humans have developed the ability to imagine their surroundings from another visual point of view. This cognitive ability is called perspective taking and has been shown to be essential in child development and social interactions. However, the precise cognitive mechanisms underlying perspective taking remain to be fully understood. Here we present a computational model that implements perspective taking as a mental simulation of the physical movements required to step into the other point of view. The visual percept after each mental simulation step is estimated using a set of forward models. Based on our experimental results, we propose that a visual attention mechanism explains the response times reported in human visual perspective taking experiments. The model is also able to generate several testable predictions to be explored in further neurophysiological studies.

1. Introduction

The socio-cognitive skills of the human brain are the product of prolonged childhood development, where fundamental skills such as a *theory of mind* [2, 15] and a capacity for *perspective taking* [9–11] are developed. Possessing a theory of mind implies being aware that other people's visual and mental states differ from one's own. This requires an understanding of how the physical space is perceived from the viewpoint of another person, which is referred to as visual perspective taking [11]. Together these skills are used to analyze and infer intentions of others [2].

One hypothesis, known as the *embodied transformation account* [9], suggests that perspective taking is the mental simulation of the physical rotation or translation necessary to acquire another perspective. Moreover, the body representations used for the mental simulation are identical to the ones used for physical movement. While this hypothesis is supported by psychological [9, 15] and neurophysiological [16] data, it has not yet been investigated using a computational model. Here we introduce such a model and implement it on a simulated robot platform (setup shown in Figure 1) in order to systematically test this hypothesis computationally.

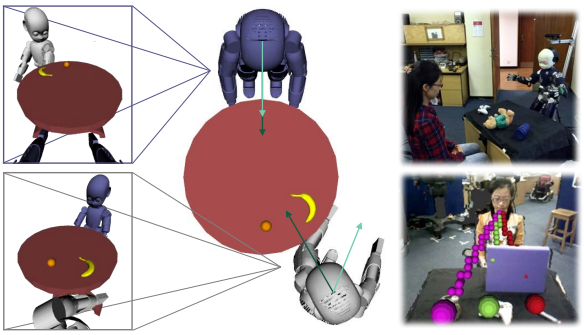

Figure 1. Experimental setup. The task of the blue robot is to take the perspective of the gray robot, and to decide whether one of the objects (*e.g.* the orange) is to the left or right as perceived by the gray robot. The model can be integrated in an architecture for human-robot interaction (right, see our previous work in [7]).

We advocate that perspective taking is governed by a competition process for visual attention between multiple forward models, where execution of recent models is preferred. The forward models predict the agent's state given the current state and a (mentally simulated) motor input [4]. As the forward models are recurrently executed, an internal (mental) representation of the simulated state is required. The model explains the response times of humans, which are often used as an evaluation metric in experiments where the body posture of the self and target agents are varied [9]. We believe our model provides further explanation of the precise mechanisms of embodied simulation and generates predictions to be tested in further neurophysiological studies.

2. Related Works

On the utility of perspective taking: Visual perspective taking offers many benefits to social capabilities and therefore has received a lot of interest within the fields of human-human and human-robot interaction. For example, it allows the reduction of the search space in ambiguous situations where multiple objects are visible, but some only to one of the agents [11]. Similarly, it can be used to understand spatial language [14]. It has also been suggested that perspective taking minimizes the collective task load between involved agents [12].

On the mechanisms of perspective taking: There are two levels of visual perspective taking with different underlying mechanisms [11]. The first level emerges around two years of age and comprises the abilities to identify objects that are occluded from one perspective but not another and determine whether an object is in front or behind another agent. Level two perspective taking develops later (between three and five years) and is an understanding of *how* the object is perceived from another perspective. Here, we investigate the more advanced level two perspective taking.

There is extensive debate about the mechanisms behind level two perspective taking [9, 10, 15]. The sensorimotor interference account suggests that mental rotations are difficult due to conflicts of spatial information, which emerge when a simulated perspective is taken [15]. In contrast, the embodied transformation account proposes that the representations used for perspective taking are analogue to those employed by the motor system, and that perspective taking is the simulation of the body rotation required to physically align perspectives [9].

Computational models of perspective taking: Perspective taking was used to anticipate motions of other agents in a perspective invariant manner [13] and to distinguish self-motion from other-motion [3]. A computational model for understanding ambiguous spatial language by integrating several cues, including the perspectives of the speaker and listener, was proposed in [5]. Imitating a demonstrator whose perspective differs from the imitator has been implemented by remapping the frame of reference of the demonstrator so that it matches that of the imitator [8]. In a similar fashion, an embodied model to understand another person's bodily motions and mapping these motions to the egocentric frame of the observer was presented in [6].

3. Methodology

In this section, we formalize visual perspective taking as an embodied simulation of physical movements using a set of forward models. We also introduce a visual attentional mechanism that improves computational efficiency by favoring previously activated forward models which we will show in Section 5 is essential to explain response times of humans in perspective taking tasks.

Formalization: Our architecture makes extensive use of forward models $\{f_i\big(\mathbf{z}(t), \mathbf{u}_i(t)\big) \mid f_i \in \mathbf{F}\}$, which provide the (simulated) predicted state $\mathbf{z}'(t+1)$ given the current state $\mathbf{z}(t)$ of the self-agent and a motor input $\mathbf{u}_i(t)$ at time t. The $\mathbf{u}_i(t)$ are action primitives (move forward, move left, rotate torso left, *etc*.) of 0.1 units of translation or 10 degrees of rotation. We instantiate one forward model per action primitive and follow [4] with their proposal that these operate in parallel. The goal of the self-agent is to mentally adopt the visual perspective, $\hat{\mathbf{z}}$, of the target agent, which we assume is static. We inhibit the motor inputs from being sent to the motor system, which results in a feed-forward control system. Hence, we suggest that there is a visuo-spatial memory representation of the mentally transformed self, which is updated over time in a *simulation loop*.

We define the distance metric $d\big(\mathbf{z}(t), \hat{\mathbf{z}}\big)$ so that:

$$d\big(\mathbf{z}(t), \hat{\mathbf{z}}\big) = d_S\big(\mathbf{z}(t), \hat{\mathbf{z}}\big) + d_\theta\big(\mathbf{z}(t), \hat{\mathbf{z}}\big),$$

where d_S measures the Euclidian distance between the translational components of $\mathbf{z}(t)$ and $\hat{\mathbf{z}}$, and d_θ is a measure for the angular disparity between the agents. The aim is to find a control policy $\mathbf{u}^*(t) = \boldsymbol{\pi}\big(\mathbf{z}(t), \mathbf{F}\big)$ that minimizes d, such that $d\big(\mathbf{z}(t), \hat{\mathbf{z}}\big) < d\big(\mathbf{z}(t-1), \hat{\mathbf{z}}\big), \forall t < t_{\text{goal}}$ and $d\big(\mathbf{z}(t_{\text{goal}}), \hat{\mathbf{z}}\big) < \epsilon$, where ϵ is a distance threshold and t_{goal} is the point in time where the estimated distance falls below this threshold. We introduce a cost function $c\big(\mathbf{u}_i(t)\big)$ and the accumulated cost C that corresponds to the response time of the model. An agent state $\mathbf{z} = \{\mathbf{z}_1, ..., \mathbf{z}_N\}$ is composed of N joint states, with each $\mathbf{z}_n(t)$ containing two translational components $x_n(t)$ and $y_n(t)$, and one rotational component $\theta_n(t)$. All components are relative to the parent joint $n-1$ and we define joints for the torso, head and eyes.

Thus far it remains unclear to which frame of reference the self-agent aligns the perspective to, and whether each joint is individually matched or a combination of the reference frames is matched [1]. Hence, we introduce a weighted average with mixing parameter $0 \leq w \leq 1$ such that:

$$d_\theta\big(\mathbf{z}(t), \hat{\mathbf{z}}\big) = (1-w)\, d_I\big(\mathbf{z}(t), \hat{\mathbf{z}}\big) + w\, d_\Sigma\big(\mathbf{z}(t), \hat{\mathbf{z}}\big), \text{with}$$

$$d_I\big(\mathbf{z}(t), \hat{\mathbf{z}}\big) = \sum_{n=0}^{N} \Big|\theta_n(t) - \hat{\theta}_n(t)\Big|, \text{and}$$

$$d_\Sigma\big(\mathbf{z}(t), \hat{\mathbf{z}}\big) = \left|\sum_{n=0}^{N} \theta_n(t) - \sum_{n=0}^{N} \hat{\theta}_n(t)\right|$$

while ensuring that all angle differences are in $[-\pi, \pi]$.

We choose the control policy $\boldsymbol{\pi}$ such that the optimal action primitive $\mathbf{u}^*(t)$ is found by executing all available forward models f_i and choosing the corresponding action primitive $\mathbf{u}_i$ that results in the lowest expected distance:

$$\mathbf{u}^*(t) = \arg\min_{\mathbf{u}_i} d\Big(f_i\big(\mathbf{z}(t), \mathbf{u}_i(t)\big), \hat{\mathbf{z}}\Big),$$

and stop the process once $d\big(\mathbf{z}(t_{\text{goal}}), \hat{\mathbf{z}}\big) < \epsilon$.

Attentional component: To reduce the computational complexity, we extend this model with an attentional component that selects a subset $\mathbf{A}(t) \subsetneq \mathbf{F}$ of forward models to be executed at time t (rather than executing all forward models $f_i \in \mathbf{F}$). The selection is governed such that the forward model $f^*(t-1)$ that was executed at the previous time step becomes one element of $\mathbf{A}(t)$. The other elements are selected based on the similarity of the associated motor inputs $\mathbf{u}_i$. For example, the motor inputs 'rotate torso left' and 'rotate head left' have a high similarity, while 'move

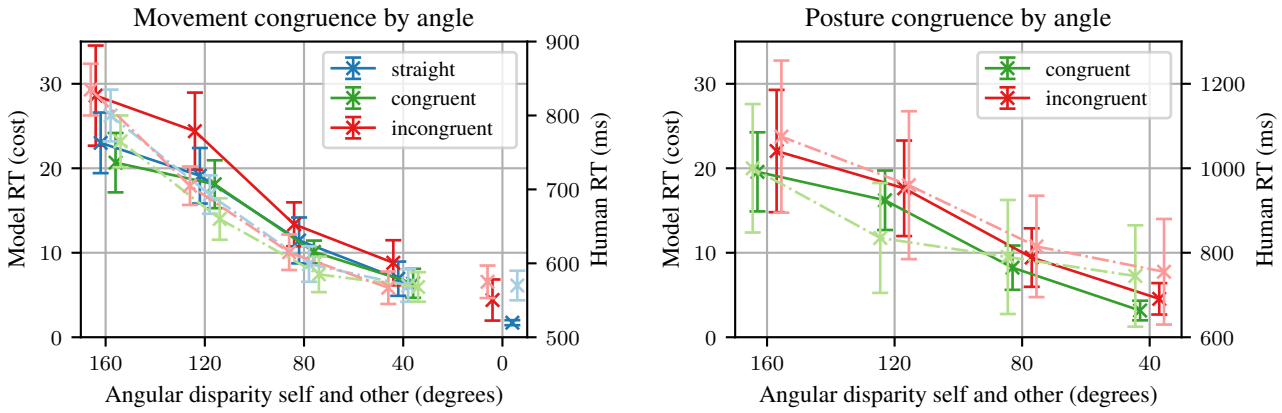

Figure 2. The response times with respect to *movement congruence* (left) and *posture congruence* (right) of our computational model (dark, solid lines) are compared to experimental data in humans (bright, dashed lines, data from [9]). Left: In both the model and human data the movements where the self agent's posture is aligned with the movement direction lead to faster response times than movements where the self agent starts with a straight posture, and incongruent movements are the slowest. These differences are more pronounced for larger angular disparities in both human and model data. Right: There is a good match of the model and human data when varying the posture congruence. The model's mixing parameter w is set to 0.8, which indicates a heavy bias towards d_Σ as further discussed in the main text.

forward' and 'move backward' have a low similarity. The forward models that are dissimilar $\left(f_j \notin \mathbf{A}(t)\right)$ are only executed if no suitable forward model is found within $\mathbf{A}(t)$ (*i.e.* none of the initially executed models reduced d). This reduces the cost c for the forward models that are selected by the attentional component (while increasing the cost for all other forward models).

4. Experiments

In this section, we investigate the properties of our computational model applied to two simulated iCub humanoid robots in a visual perspective taking task. Simulated robots are used to allow large-scale systematic evaluation of our model. The model is compatible with our previous work on perspective taking implemented on a real iCub robot [7].

Experimental setup: As shown in Figure 1, the two robots are placed in a scenario where two objects are positioned on a table top. One of the robots (self agent) is tasked with mentally adopting the perspective of the other static robot (target agent) and decide whether an object is to the left or right of the target agent. We try to replicate a typical setup from experiments with humans [9] as closely as possible, so that comparisons can be drawn easily. That is, we introduce variations in the perspective difference between the two robots, and variations of the body postures of the robots. The perspective difference is set between 0 and 160 degrees in 40 degree steps. The posture of the robot executing the mental rotation can be congruent or incongruent with the required movement direction, or straight (movement congruence). Furthermore, the body posture of both robots is varied so that the torso is either rotated to the left or the right, while the head is always facing the center of the table. Therefore, the postures can be either congruent (torsos are facing the same way), or incongruent (torsos facing opposite ways). As evaluation metrics, we employ the response time and accuracy.

Computational efficiency: In the first experiment, we examined the computational efficiency of our model. The model's response time generally grows linearly with the angular perspective difference. The attentional mechanism impacts the response time in the following way. The response time for congruent movements is lowered as the forward models corresponding to the physical movement required to take the initial pose become elements of $\mathbf{A}(t=0)$, and these forward models are the same as the ones required for the mental simulation, which in turn decreases the cost $c(t)$. Similarly, an incongruent posture leads to forward models being elements of $\mathbf{A}(t=0)$ that are not required for the mental simulation, which leads to an increased cost. These effects are shown in Figure 2 (left). Without the attentional mechanism, the impact of the movement congruence diminishes. In Section 5, we will show that the response times are only then qualitatively similar to those of humans if the attentional mechanism is employed.

Perspective matching mechanism: In the second experiment, we vary w to obtain the model properties if each joint is individually matched ($w=0$), only the overall perspective difference is matched ($w=1$), or the two terms are blended ($0<w<1$). Instantiating a model with $w=1$ leads to a response time that is irrespective of the body posture of the target agent, so that the posture congruence does not impact the response time. Analogously, for smaller w, the response time difference between incongruent and congruent body postures increases; for $w=0$ the response time for incongruent postures is nearly flat and very high. In the following section, we argue that w is close to 1 (*i.e.* favoring

d_Σ over d_I) in humans. In Figure 2 (right), we show a good qualitative match of human and model data for $w = 0.8$.

5. Discussion

Psychophysical validation data: As in our computational model, the response times of humans in perspective taking tasks grow approximately linearly with the angular disparity between the self and the target agent [9, 11, 15]. Several experimental variations have shown to impact the response times [9]. A body posture that is congruent with the required movement direction lowers the response time (movement congruence), while an incongruent posture increases the response time. We advocate that the initial posture variation acts as a prior to the attended forward models; *e.g.* physically rotating the torso to the left results in internal forward models responsible for the prediction of movements towards the left being elements of the attended set $\mathbf{A}(t = 0)$.

Another experimental variation investigates the impact of the target agent's body posture (rather than the self agent's), which in humans is much smaller compared to the movement congruence discussed above. In our computational model, the mixing parameter w controls the impact of the target agent's posture on the response times. The smaller w, the higher the impact as each joint state is individually matched. On the other hand, large w leads to a small impact as the combination of all states is matched. Therefore, by comparison with human response times, our model suggests that the cost function for angular disparity is heavily biased towards d_Σ in humans, *i.e.* w is close to 1.

Another model property we would like to discuss is the speed-accuracy trade-off governed by the distance threshold ϵ. The larger ϵ, the larger the remaining distance between the two agents and the earlier the mental simulation is stopped, but the higher the chance of an incorrect response. Investigating this property in more detail leads to a model prediction as described below.

Model predictions: Our model offers the following testable predictions. Firstly, when a response is forced early, our model predicts that the response is biased towards the egocentric perspective. Notably, due to the visual attention mechanism, the model hypothesizes that the self agent is less impacted in situations where the movement is congruent with the initial pose. Secondly, our model suggests that there should be a significant difference in response times depending on the congruence between the required movement direction of the current and previous mental simulations. We suggest that this habituation effect is also due to the attentional component and the faster processing of forward models that have been employed in the previous time step.

6. Conclusions

In this work, we presented a computational model for visual perspective taking that contains a set of forward models as building blocks. To reduce the computational complexity, we proposed an attentional component that introduces a competition between multiple forward models. We have shown that our model's response time is similar to those of humans only if this attentional component is employed, and therefore argued that humans implement an attentional mechanism similar to that of our model.

In future work, we will discuss the developmental aspects of perspective taking by learning the forward models and will show that accurate forward models are needed for perspective taking. Ongoing work continues the implementation of the computational model on a real iCub humanoid robot. This requires an architecture with perceptual components for the state of the human, the positions of objects, and the physical layout of the surroundings, as outlined in [7].

References

[1] A. J. Alsmith, E. R. Ferrè, and M. R. Longo, "Dissociating contributions of head and torso to spatial reference frames: The misalignment paradigm," *Conscious. Cogn.*, 2017. 2

[2] S.-J. Blakemore and J. Decety, "From the Perception of Action to the Understanding of Intention," *Nat. Rev. Neurosci.*, 2001. 1

[3] J. L. Copete, Y. Nagai, and M. Asada, "Motor development facilitates the prediction of others' actions through sensorimotor predictive learning," in *ICDL-EPIROB*, 2016. 2

[4] Y. Demiris and G. Hayes, "Imitation as a dual-route process featuring predictive and learning components: a biologically-plausible computational model," in *Imitation Anim. Artifacts*, 2002. 1, 2

[5] N. Duran and R. Dale, "Toward Integrative Dynamic Models for Adaptive Perspective Taking," *Top. Cogn. Sci.*, 2016. 2

[6] S. Ehrenfeld and M. V. Butz, "An embodied kinematic model for perspective taking," in *Biannu. Conf. Ger. Cogn. Sci. Soc.*, 2014. 2

[7] T. Fischer and Y. Demiris, "Markerless Perspective Taking for Humanoid Robots in Unconstrained Environments," in *ICRA*, 2016. 1, 3, 4

[8] R. J. Gentili *et al.*, "A Neural Architecture for Performing Actual and Mentally Simulated Movements During Self-Intended and Observed Bimanual Arm Reaching Movements," *Int. J. Soc. Robot.*, 2015. 2

[9] K. Kessler and L. A. Thomson, "The embodied nature of spatial perspective taking: Embodied transformation versus sensorimotor interference," *Cognition*, 2010. 1, 2, 3, 4

[10] M. May, "Imaginal perspective switches in remembered environments: Transformation versus interference accounts," *Cogn. Psychol.*, 2004. 1, 2

[11] P. Michelon and J. M. Zacks, "Two kinds of visual perspective taking," *Percept. Psychophys.*, 2006. 1, 2, 4

[12] A. K. Pandey and R. Alami, "Affordance graph: A Framework to Encode Perspective Taking and Effort based Affordances for day-to-day Human-Robot Interaction," in *IROS*, 2013. 1

[13] F. Schrodt, G. Layher, H. Neumann, and M. V. Butz, "Embodied learning of a generative neural model for biological motion perception and inference," *Front. Comput. Neurosci.*, 2015. 2

[14] L. Steels and M. Loetzsch, "Perspective Alignment in Spatial Language," in *Spat. Lang. Dialogue*, 2009. 1

[15] A. Surtees, I. A. Apperly, and D. Samson, "The use of embodied self-rotation for visual and spatial perspective-taking," *Front. Hum. Neurosci.*, 2013. 1, 2, 4

[16] H. Wang *et al.*, "Rhythm makes the world go round: An MEG-TMS study on the role of right TPJ theta oscillations in embodied perspective taking," *Cortex*, 2016. 1