Coupling Dynamical and Connectionist Models: Representation of Spatial Attention via Learned Deictic Gestures in Human-Robot Interaction

Barış Serhan (baris.serhan@plymouth.ac.uk)
Centre for Robotics and Neural Systems (CRNS), Plymouth University
B110, PSQ, Drake Circus, Plymouth, PL4 8AA, UK

John Spencer (j.spencer@uea.ac.uk)
School of Psychology, University of East Anglia
Norwich Research Park, Norwich, Norfolk, NR4 7TJ, UK

Angelo Cangelosi (a.cangelosi@plymouth.ac.uk)
Centre for Robotics and Neural Systems (CRNS), Plymouth University
A316, PSQ, Drake Circus, Plymouth, Devon, PL4 8AA, UK

Abstract

A proper representation of space and a joint attention mechanism are indispensable for an effective deictic communication with embodied agents. Taking inspiration from developmental psychology may help us to tackle computational challenges for robots. Although some developmental joint attention models for robots have already been proposed, to the best of our knowledge, there is no such model that can stand for the effects of pointing gestures on covert attention in infants. Thus we have designed and implemented a developmental robotics model for joint spatial attention combining connectionist and dynamical approaches. The hybrid architecture was structured over two existing computational models: a connectionist model of gesture comprehension and a Dynamic Field (DF) model of spatial attention in infants. These models were extended with various perceptual modules and dynamical neural fields, and implemented on the state-of-art iCub humanoid robot. In this paper, the computational architecture is introduced with some preliminary results that show the model's capability of representing deixis and perceived objects, and their effects on attention over space and time.

Keywords: cognitive modelling; cognitive robotics; artificial neural networks; dynamic field theory; joint attention; pointing gestures; spatial attention; deixis; grounded cognition

Introduction

Inherently simple tasks, such as jointly attending to a particular object or an event in the scene, might be very challenging for artificial cognitive agents. Human infants acquire joint attentional skills very early in infancy, starting from gaze following, later comprehension and production of deictic gestures such as pointing (Tomasello, Carpenter, & Liszkowski, 2007). These abilities are essential for human communication. On the other hand, the ability to accomplish proper deictic communication with humanoid robots, key processing mechanisms, such as attention synchronisation, object recognition and object indication, are needed (Sugiyama, Kanda, Imai, Ishiguro, & Hagita, 2007).

Pointing gestures are observed prior to verbal communication and they are universal social tools to direct attention (Bates, Camaioni, & Volterra, 1975). In attentional cueing tasks, it has been shown that comprehension of pointing gestures occurs several months before their production (Gredebäck, Melinder, & Daum, 2010) and if the pointing hand provides also motion information, then the attentional sensivity in congruent cases can be observed even in 4.5-month-old infants (Rohlfing, Longo, & Bertenthal, 2012).

Developmental (or epigenetic) Robotics is a multidisciplinary field where insights from developmental psychology guides the implemention of adaptive intelligent embodied agents (Cangelosi & Schlesinger, 2015). In developmental robotics, Artificial Neural Networks (ANNs) have been used to map some of the developmental changes of joint attention (Nagai, Hosoda, Morita, & Asada, 2003). ANNs were also used to model pointing gesture comprehension using edge features in robot-robot interaction (Hafner & Kaplan, 2005), as well as using the motion and edge information in human-robot interaction (Nagai, 2005). However, time is not an inherent property of these feed-forward networks. Even if the number of learning steps can be used to denote the time, once the network is trained, the network's output reaction for a given input is immediate in terms of computation steps. On the other hand, time is a built-in feature in dynamical systems approaches, as well as in certain hybrid approaches such as Nengo, which integrates time dynamics into its framework using LIF spiking neurons (Eliasmith, 2013).

The Dynamic Field Theory (DFT) is a dynamical approach to model cognition at the neural population level (Schöner, Spencer, & the DFT Research Group, 2015). The theory has been used to model wide variety of cognitive processes, as well as the developmental aspects of cognition. The DFT is notably robust to simulate reaction times of the underlying processes of spatial cognition (Spencer, Simmering, Schutte, & Schöner, 2007). The DF model of the proposed architecture of this paper was constructed over the existing IOWA model that can capture the developmental changes in spatial attention and saccade planning (Ross-Sheehy, Schneegans, & Spencer, 2015).

It has been proposed that, on the theoretical basis, connectionism and dynamical systems accounts do not have competing positions in child development (Thelen & Bates, 2003). In this paper, we introduced a computational architecture that integrates and extends the conceptual link between connec-

tionism and dynamical modelling approaches. The proposed developmental robotics model is able to represent the overt and the covert spatial attention over time and space by extending a connectionist model of deictic gesture comprehension (Nagai, 2005) with the DFT approach.

In the next section, the architecture of the model is explained in detail considering, respectively, the robot and its related modules, the connectionist parts of the model, and the integration of the DFT modules. The model is then evaluated in three case scenarios in the results section. Finally, the paper is concluded with the discussions and the future directions.

The Computational Architecture

The cognitive robotics model was designed to interpret the contribution of the low level features of pointing gestures such as movement and edge information, to a higher understanding of deixis. To learn the association between the pointing hands and the indicated locations in the space through the low level features, a variation of a gesture comprehension neural network model was implemented (Nagai, 2005). This model can learn autonomously the intended direction of a pointing gesture. However, once the network is trained, it immediately produces outputs for any given input sequences. This makes it impossible to model developmental psychology studies based on reaction time. We extended this model using the Dynamic Field Theory so that the new architecture is able to represent some parts of the high level cognition such as spatial attention and spatial working memory. The overall structure of the model can be seen in Figure 1. Its constitutive modules and their functions are explained in detail in the following subsections.

The Embodied Agent

The iCub robot was used as an embodied agent of the computational model in this study. This section briefly explains the robot and associated modules.

iCub Humanoid Robotic Platform The iCub humanoid robotic platform is an open source robotic platform developed at the Italian Institute of Technology with the contribution of more than 20 laboratories. The iCub robot was designed as a 3 to 4 year-old child (Figure 1) and equipped with binocular vision, binaural audition, haptic and inertial sensors to perceive the surrounding environment, as well as its bodily states. It has also 53 degrees of freedom that enables it to perform wide variety of actions to interact with its environment. The main purpose of the iCub is to provide an interdisciplinary platform for cognitive development research via HRI and autonomous learning studies (Metta et al., 2010).

YARP Modules Yet Another Robot Platform (YARP) is an open source robotics middleware that is designed to ease communication between different hardware and software systems (Metta, Fitzpatrick, & Natale, 2006). In our study, YARP was used to ensure the link between all physical machines such as server and client PCs, and the iCub's boards. The communication of the modules that work on different platforms (e.g. Matlab for DFT modules and Python for neural network modules) were also accomplished by implementing modules using YARP's functionalities.

Motor Control Module This module was implemented to create an actual saccade on the particular location where the saccade motor field of the DFT module was indicated. It takes one dimensional input, turns into a point on a semicircle on 2D plane and sets the motor decoders of the robot's eyes to fixate that point.

Learning Pointing Direction via Low Level Features

The modules that are responsible to get raw data from the robot's camera and process them to understand pointed location, are illustrated in the upper half of Figure 1. Pointing gestures are first captured by the iCub's left camera and images are transferred to the main computer using YARP protocols. Then, the images are passed separately through preprocessing steps in three modules before they are fed into neural networks and dynamical neural fields.

Perception Module This module detects the object in the current scene according to its colour using basic openCV masking methods and forms the object location into an angle value. This angle value is later sent to the DFT module and represented as the location of the object on the perception dynamic neural field.

Feature Detection Modules The Edge Detection (ED) and the Optic Flow Detection (OFD) modules were implemented very similarly to Nagai's model (Nagai, 2005). The centre of the image (168x168) was considered as foveal area. The ED detects the edges of the hand image in this area with an orientation selective filter so that each edge pixel has a 4 dimensional vector for 4 different orientations ($\leftrightarrow, \nearrow, \updownarrow, \nwarrow$). The foveal region was split into 49 small regions (24x24) called receptive fields. A cumulative orientation vector was calculated for each receptive fields.

The displacement of the receptive fields between consecutive frames was calculated by the OFD module using a template matching algorithm. For each field, an 8 dimensional vector is constructed to keep the displacement amounts on 8 different directions ($\leftarrow, \swarrow, \downarrow, \searrow, \rightarrow, \nearrow, \uparrow, \nwarrow$).

Learning Modules Two separate feed-forward neural networks were implemented in PyBrain (Schaul et al., 2010) to learn the association between pointing gestures and the pointed location. The edge neural network (edge-NN) has three layers. The input layer has 196 neurons which receive the orientation vectors of the receptive fields as inputs (49x4) from the ED. The edge-NN also has a fully connected hidden layer consisting of 49 neurons and an output layer of 8 neurons. The output neurons represent the magnitude of the 8 different direction vectors of the indicated location.

The optic flow neural network (flow-NN) has one input (392 neurons) and one output layer (8 neurons). It receives displacement information of the receptive fields in 8 direc-

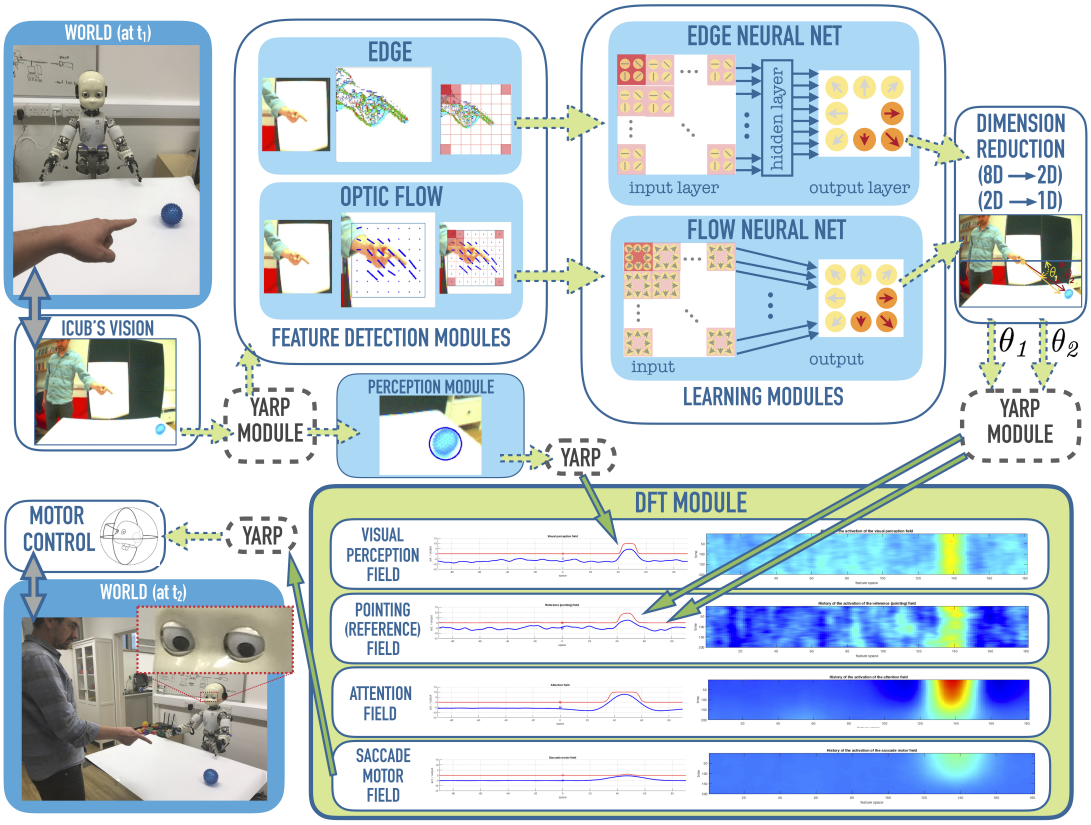

Figure 1: Architecture of the model

tions (49x8) and outputs an 8 dimensional vector as edge-NN.

The learning of pointing gestures with these networks was done offline by using 104 labelled videos from the same participant. These videos had 37 different object locations that were congruent to the pointing hand. Around 2500 image frames were extracted from these videos to train the networks for 10000 epochs with a learning rate $\sigma = 0.05$.

Dimension Reduction Once the pointed location is determined by neural networks with 8 dimensional vectors, these vectors are first projected onto 2D plane by taking the means of horizontal and vertical components of the 8 directions as in (Nagai, 2005). After that, their dimension is again reduced by using a multi-valued inverse tangent function (i.e. numpy.arctan2) to an angle in radian $(\,]-\pi;0[\,)$. The two angle predictions of two networks are sent separately to the DFT module through YARP.

High Level Representation of Pointed Location

To extend the sensorimotor understanding of our connectionist network with spatial attention and spatial working memory mechanisms, we took advantage of the dynamic field theory (Schöner et al., 2015). In the DFT, the activation of the neural populations are represented with dynamic neural fields that are defined by the following differential equation:

$$\tau \dot{u} = -u(x,t) + h(x) + s(x,t) + \int g(u(x',t))k(x-x')dx \quad (1)$$

where $u(x,t)$ is the activation level of the related neural field over the predefined metric dimension x such as space, τ is the time constant, $h(x)$ is the resting level of the activation, $s(x,t)$ is the external input to the field, and the integral term stands for the convolution of the interaction kernel k (e.g. a Gaussian function) with the output sigmoid function $g(u)$.

The DFT is a powerful approach to model neural dynamics of high level cognitive processes at the population level. The DF model of this study has been built by extending the existing IOWA model (Ross-Sheehy et al., 2015) using the COSIVINA framework in Matlab. The IOWA model is an infant saccade planning model that can capture developmental changes in spatial attention and memory. The IOWA model was extended by adding two other dynamic neural fields, namely perception field and reference (pointing) field (the DFT module in Figure 1).

The continuous metric dimension (x) of all the neural fields is defined as the angles on a semicircle $(\,]-\pi;0[\,)$ that is centred at the mid horizontal line of the image plane (the dimension reduction in Figure 1).

Visual Perception Field This neural field is used to represent the perceived object location as a peak of activation of a

neural population. This can be thought as a sort of retinotopic representation of the object in the current scene by a population of neurons in early visual areas such as V1. The visual perception field has excitatory projections onto the attention field so that if the visually salient object remains a sufficient amount of time on a particular location, that might cause a peak of activation in the spatial attention for that location. The field has also lateral interactions so that when a detected object gives rise to a peak of activation, this peak may became an attractor state with the local excitation around that location and the global inhibition over the space dimension.

Pointing (Reference) Field The purpose of the neural population of the pointing field is to stand for the referential relation between the observed space and the pointed location. The angle values received from two neural nets are projected onto the pointing field as the summation of two Gaussian stimuli (i.e. $s(x,t)$ in eq. 2) where their centres were characterised by these angles on the field's metric dimension (x). The reference field has self excitatory-inhibitory lateral interactions and an excitatory projection onto the attention field (just as the visual perception field). The field equation of the pointing field is the following:

$$\begin{aligned}\tau \dot{u_r} = &- u_r(x,t) + h_{u_r}(x) + s(x,t) \\ &+ \int k_{u_r u_r}(x-x')g(u_r(x',t))dx + q\xi(x,t)\end{aligned} \tag{2}$$

where u_r is the activation variable of the pointing (reference) field, τ is the time coefficient, g is the output sigmoid function, ξ is the random noise function and $k_{u_r u_r}$ is a Mexican hat shaped function for the lateral interaction in the form of:

$$\begin{aligned}k_{u_r u_r}(x-x') = &\frac{c_{exc}}{\sqrt{2\pi}\sigma_{exc}} \exp\left(-\frac{(x-x')^2}{2\sigma_{exc}^2}\right) \\ &- \frac{c_{inhib}}{\sqrt{2\pi}\sigma_{inhib}} \exp\left(-\frac{(x-x')^2}{2\sigma_{inhib}^2}\right) - c_{glob}\end{aligned} \tag{3}$$

where c is the strength of the interaction (c_{exc} for excitatory, c_{inhib} for inhibitory interactions), σ parameters are the width of the Gaussians and c_{glob} is used for the global inhibition over the metric feature dimension of the pointing field.

Basically, when the distance between the referential locations related to motion and edge information is small, then the amplitude of the attractor that appears on that location becomes higher, which then causes an increase in the attention on that location with its excitatory projection onto the spatial attention field.

Spatial Attention Field The spatial attention and the saccade motor fields are constructed based on the existing IOWA model (Ross-Sheehy et al., 2015). The strength of a localised activation on the spatial attention field represents the amount of attention on that particular location. The lateral interactions permits the emergence of a rivalry mechanism between different localised activations over its feature dimension which is the same metric dimension as the others. Our contributions to its dynamical equation were two excitatory projections coming from the visual perception (bottom-up) and the pointing field (top-down) which are defined by the convolutions of those fields' outputs with Gaussian kernels.

Once the activation passes the resting level on a particular location, the strong global inhibition suppresses all the other rival spots over the feature dimension. Together with the local excitation, one single attractor state that represents the spotlight of the attention emerges on this field.

If there is a conflicted situation, for example, if a hand points in the opposite direction while there is a visually salient object on the other side, the competition increases the time needed to create an attractor on an attentional locus. This process is also non-deterministic in terms of who the winner will be.

Saccade Motor Field This field represents motor areas responsible of saccadic eye movements. It receives excitatory input from the spatial attention field and this interaction is done by convolving the sigmoid output function of the attention field with an inverse Gaussian kernel so that foveal areas are more inhibited while approaching to the center. With this mechanism, if the spotlight of attention is already at the fovea, the motor responses are suppressed, on the other hand, if the attention is on another locus, an attractor state emerges at that location to saccade and fixate there. Once the resting level is exceeded, attention field is inhibited to reset the system by discrete nodes (Ross-Sheehy et al., 2015).

Results

The model was tested with three different videos that were not in the training dataset for the learning experiments. Each video had a different scenario. A blue plastic ball was located in the same place on the table in all cases. No pointing gesture appears in the scene in the first scenario while a congruent or an incongruent pointing gesture appears, respectively, in case II and case III.

The behaviour of the model was illustrated in these scenarios in three columns in Figure 2. Each panel of a column simultaneously represents the activation patterns on the respective neural field during the associated experimental scenario. In each panel, x-axis stands for the continuous metric feature dimension that is linked to locations on a spatial map of the environment. Y-axis represents the computational time steps whereas z-axis shows the activation levels of the field over time and space. Except for the saccade motor field, the surface areas were covered by connected red lines where the output of the sigmoid activation functions were higher than 0.5. In the saccade motor fields, only the first attractor was covered in red, since once threshold passed, the location is sent immediately to the motor control module to direct the iCub's eye gaze to that fixation point.

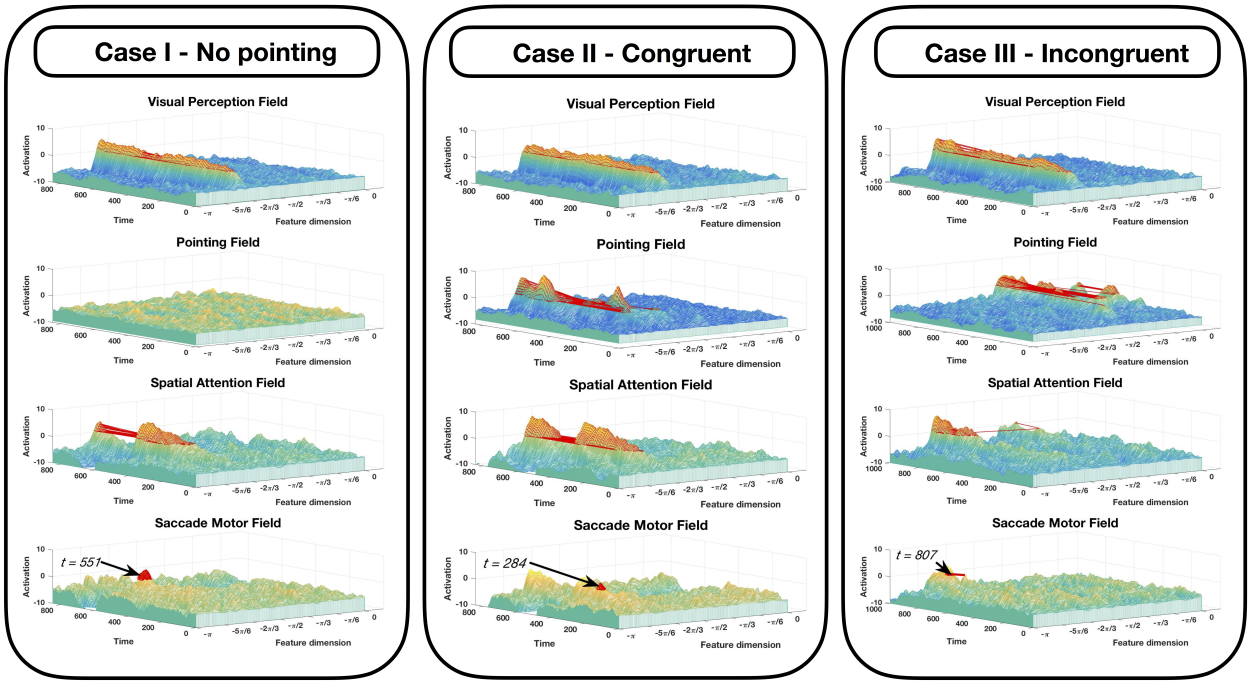

Figure 2: Behaviours of the dynamical neural fields in 3 different conditions

Case Scenario I - Object Perception

In the first case, the stationary object located at $-2\pi/3$ can be seen on the visual perception field as a continuous peak of activation over time. Random fluctuations at the resting level ($h = -5$) on the pointing field demonstrate that there was no pointing gesture observed during this scenario. It can be seen in the attention field that the persistent activation of the perception field triggered an increase in the attention at the same location and the field's output function passed its threshold (i.e. $g(u_a(x',t)) > 0.5$) after around 300 computation cycles. The projection of this activation then caused a rise on the saccade motor field and created a saccade signal at $t = 551$.

Case Scenario II - Congruent Pointing

In this case, the stationary object was again located in the same spot, however this time a moving pointing gesture directed at the object was also included in the scene. Thus, the peaks of activation can be observed around the object location ($-2\pi/3$) on the attention field in this occasion, whilst the behaviour of the perception field was very similar to that of the case I. Since the combination of both field activations left a trace together onto the spatial attention field, in this scenario the attractor states in the attention field had more strength than the previous case and more importantly, attractors were formed faster starting after around 200 time steps. This effect was then reflected also to the motor field and the saccade reaction time decreased to $t = 284$.

Case Scenario III - Incongruent Pointing

In the final scenario, the object was positioned in the same location, and a pointing gesture was presented. However this time the pointing hand was indicating another spatial region ($-\pi/6$) which was incongruent to the object location ($-2\pi/3$) (see the first and the second panels from the top). In this example, when the activations of the perception and the pointing fields were forwarded onto the spatial attention field, the incongruity of the locations initiated a rivalry between two conflicting regions. The localised peaks on the two sides of the attention field were trying to suppress the other because of the strong global inhibition defined through the lateral interaction kernel. Similarly, the strong local excitation was helping these peaks to self-sustain whilst being under inhibition of the other. Therefore, this attentional rivalry mechanism elicited an increase in reaction time of the motor field ($t = 807$).

The pattern of results described here reflect the typical dynamics of the system with the three visual stimulus configurations. Future explorations of the control of the object location and the pointing hand will help to clarify the full dynamics of the model.

Conclusion and Future Directions

The proposed cognitive architecture is able to learn and represent the joint attentional intention underlying pointing gestures while also taking into consideration its time dynamics and its deictic nature. The implementation of our computational model is consistent with the formal account of grounded cognition (Barsalou, 2008). Connectionism was taken into our hybrid model as a bottom-up understanding mechanism of pointing gestures so that its learning can be seen similar to 'sensorimotor toil' method in the Sym-

bolic Theft Hypothesis (Cangelosi, Greco, & Harnad, 2002). Grounding symbols over the pointing neural field might be possible in future, as, for example, demonstratives are accompanied by pointing gestures in early infancy and the exophoric use of demonstratives is to create a joint attentional frame (Diessel, 1999). In addition, the model may provide practical advantages for the design of robots that have more natural interaction and communication skills.

In this paper, we introduced a cognitive robotics model and validated its key mechanisms. The next step is to design experiments to compare the iCub's behaviours with infants' reactions to the pointing gestures in attentional cueing tasks (e.g. Rohlfing et al. (2012)). Moreover, replicating developmental psychology studies with robots may improve our understanding of the underlying mechanisms of cognitive processes. Furthermore, this line of research might also provide new directions of investigations for developmental psychologists (Cangelosi & Schlesinger, 2015).

Since speech is an essential modality for human communication, it can be an obvious extension for the model in future. The DFT is again an option as it has been already used for modelling spatial language (Richter, Lins, Schneegans, Sandamirskaya, & Schoner, 2014), as well as in word learning tasks (Samuelson, Spencer, & Jenkins, 2013). Another option might be using a deep neural network (DNN) to classify the speech and gesture couplings.

Acknowledgements

This study was conducted as a part of Deictic Communication (DComm) project and has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skodowska-Curie Actions (Grant agreement No 676063). We would also like to thank Prof. Dr. Gregor Schöner and the DFT Research Group at Ruhr University Bochum, as well as the researchers at the iCub facility at the Italian Institute of Technology (IIT, Genoa).

References

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.*, *59*, 617–645.

Bates, E., Camaioni, L., & Volterra, V. (1975). The acquisition of performatives prior to speech. *Merrill-Palmer Quarterly of Behavior and Development*, *21*(3), 205–226.

Cangelosi, A., Greco, A., & Harnad, S. (2002). Symbol grounding and the symbolic theft hypothesis. *Simulating the evolution of language*, 191–210.

Cangelosi, A., & Schlesinger, M. (2015). *Developmental robotics: From babies to robots*.

Diessel, H. (1999). *Demonstratives: Form, function and grammaticalization* (Vol. 42). John Benjamins Publishing.

Eliasmith, C. (2013). *How to build a brain: A neural architecture for biological cognition*. Oxford University Press.

Gredebäck, G., Melinder, A., & Daum, M. (2010). The development and neural basis of pointing comprehension. *Social Neuroscience*, *5*(5-6), 441–450.

Hafner, V. V., & Kaplan, F. (2005). Learning to interpret pointing gestures: experiments with four-legged autonomous robots. In *Biomimetic neural learning for intelligent robots* (pp. 225–234). Springer.

Metta, G., Fitzpatrick, P., & Natale, L. (2006). Yarp: yet another robot platform. *International Journal of Advanced Robotic Systems*, *3*(1), 8.

Metta, G., Natale, L., Nori, F., Sandini, G., Vernon, D., Fadiga, L., ... others (2010). The icub humanoid robot: An open-systems platform for research in cognitive development. *Neural Networks*, *23*(8), 1125–1134.

Nagai, Y. (2005). Learning to comprehend deictic gestures in robots and human infants. In *Robot and human interactive communication, 2005. roman 2005. ieee international workshop on* (pp. 217–222).

Nagai, Y., Hosoda, K., Morita, A., & Asada, M. (2003). A constructive model for the development of joint attention. *Connection Science*, *15*(4), 211–229.

Richter, M., Lins, J., Schneegans, S., Sandamirskaya, Y., & Schoner, G. (2014). Autonomous neural dynamics to test hypotheses in a model of spatial language. In *Proceedings of the 36th annual meeting of the cognitive science society.*

Rohlfing, K. J., Longo, M. R., & Bertenthal, B. I. (2012). Dynamic pointing triggers shifts of visual attention in young infants. *Developmental Science*, *15*(3), 426–435.

Ross-Sheehy, S., Schneegans, S., & Spencer, J. P. (2015). The infant orienting with attention task: Assessing the neural basis of spatial attention in infancy. *Infancy*, *20*(5), 467–506.

Rumelhart, D. E., & McClelland, J. L. (1986). Parallel distributed processing: explorations in the microstructure of cognition.

Samuelson, L. K., Spencer, J. P., & Jenkins, G. W. (2013). A dynamic neural field model of word learning. *Theoretical and computational models of word learning: Trends in psychology and artificial intelligence*, 1–27.

Schaul, T., Bayer, J., Wierstra, D., Sun, Y., Felder, M., Sehnke, F., ... Schmidhuber, J. (2010). Pybrain. *Journal of Machine Learning Research*, *11*(Feb), 743–746.

Schöner, G., Spencer, J. P., & the DFT Research Group. (2015). *Dynamic thinking: A primer on dynamic field theory*. Oxford University Press.

Spencer, J. P., Simmering, V. R., Schutte, A. R., & Schöner, G. (2007). Insights from a dynamic field theory of spatial cognition. *The emerging spatial mind*, 320.

Sugiyama, O., Kanda, T., Imai, M., Ishiguro, H., & Hagita, N. (2007). Natural deictic communication with humanoid robots. In *Intelligent robots and systems, 2007. iros 2007. ieee/rsj international conference on* (pp. 1441–1448).

Thelen, E., & Bates, E. (2003). Connectionism and dynamic systems: Are they really different? *Developmental Science*, *6*(4), 378–391.

Tomasello, M., Carpenter, M., & Liszkowski, U. (2007). A new look at infant pointing. *Child development*, *78*(3), 705–722.