# Combining Deep Learning for Visuomotor Coordination with Object Identification to Realize a High-level Interface for Robot Object-picking

Manfred Eppe[1] and Matthias Kerzel[1] and Sascha Griffiths[1] and Hwei Geok Ng[1] and Stefan Wermter[1]

***Abstract*— We present a proof of concept to show how a deep network for end-to-end visuomotor learning to grasp is coupled with an attention focus mechanism for state-of-the-art object detection with convolutional neural networks. The cognitively motivated integration of both methods in a single robotic system allows us to realize a high-level interface to use the visuomotor network in environments with several objects, which otherwise would only be usable in environments with a single object. The resulting system is deployed on a humanoid robot, and we perform several real-world grasping experiments that demonstrate the feasibility of our approach.**

## I. INTRODUCTION

Deep Learning has proven to be a successful method to realize robot control and Inverse Kinematics (IK). Specifically, it is well suited for end-to-end approaches for learning to solve tasks like grasping and other visuomotor problems. For example, Levine et al. [16] and Kerzel et al. [12] show that for the specific case of picking up objects, a deep neural network can successfully be trained to approximate a function that maps camera images of single objects to grasping trajectories. The models have proven to be very robust, in that they can be used to grasp a variety of objects under different lighting conditions at different locations within the robot's arm range. However, a problem with this approach is that it only works if there is a single kinematic goal trajectory admitted for the specific input data. This means that there can only be one graspable object within the robot's field of view; otherwise, the task to solve would be ambiguous. However, real-world scenarios usually involve several objects, and humans that interact with robots require an interface that allows one to indicate the object that the robot should grasp. For example, if there are several balls in the environment, the robot should be capable of understanding high-level instructions like *"Pick up the yellow ball!"* vs *"Pick up the green ball!"*.

In this paper, we present a cognitively inspired approach to solve this problem. Humans execute such high-level instructions in three steps: They first identify the object of interest, second they focus on it, and finally, they execute the motion trajectory. This sequence of task planning and execution is also described by Land and Hayhoe [14]. However, while the authors report that humans monitor task execution extensively, studies on children's hand-shapes during reach-to-grasp tasks show that motion planning seems to occur mostly *before* the execution of the motion plan [25]. That is, children first focus visually on the target object, and then perform the grasp.

Furthermore, just like other aspects of cognition, human perception is often situated in the sense that people focus on a task more than on the details of a given stimulus. An example from linguistics is the *good enough parsing* hypothesis [9]. Asking, "How many animals of each kind did Moses take on the Ark?" [2], will prompt most people to answer "two" – focusing on the task of answering the question – instead of answering "none" which is the correct answer (as it was Noah and not Moses who built the Ark). Similarly, in vision, the task-driven perception can lead to what has been dubbed *inattentional blindness* [20]. In experiments, people were given a counting task such as how often a basketball bounces or how many passes a team scores, in a video. While performing the task participants often fail to notice an otherwise salient feature of the video such as a woman with a bright colored umbrella or a person in a gorilla costume walking through the observed scene [23]. This illustrates that visual information unnecessary for a task can be disregarded by humans as the focus of attention is placed on a particularly relevant set of features in the field of vision. Although visual attention and visual awareness need to be separated conceptually in terms of processing in the brain [13], conscious access [1] to information in a given stimulus can be likened to a spotlight being shown on a soloist among "performers on a stage" [24] while other performers remain "in the shadows". The model presented here in some way reflects that spotlight idea.

In this paper, we realize a similar approach: First, to identify the object of interest, we first query a knowledge model with qualitative information about the objects and their properties in the robot's field of view. This knowledge model is generated using a deep convolutional neural network for object detection. Second, we implement an attention focus mechanism that eliminates all unnecessary information from the camera image, akin to inattentional blindness. Third, we use this focused input data to generate the grasp trajectory, by using a deep neural network for motor control. In doing so, we provide a proof of concept which addresses the problem of realizing a general interface that maps high-level action descriptions to grasping trajectories in environments with multiple objects. We focus on pick-up and grasping tasks, but the approach can in principle also be transferred to other control problems.

[1] Knowledge Technology, Department of Informatics, University of Hamburg, Germany {eppe, kerzel, griffiths, 5ng, wermter} @ informatik.uni-hamburg.de

We realize the object detection with a state-of-the-art method based on convolutional region proposal networks [22]. The system is given a video stream of the robot's camera and it generates bounding boxes around objects of the classes that it is trained with. To solve the problem of detecting object properties like color and shape, we do not only train the detector on object classes but also on object properties. For example, it also generates bounding boxes around all objects of a given color or shape. This way, our interface can refer to individual objects by the object's class and also by the object's properties. To realize the attention focus mechanism, the system produces a modified camera image, where only the object of interest is present, and all the other objects are removed. This allows the deep network for grasping to function in multi-object environments while retaining the advantages of the original approach like fast and mostly autonomous acquisition of visuomotor abilities.

## II. Background and Related Work

### A. *Continuous end-to-end deep learning for visuomotor abilities*

The robotic task of grasping and manipulating objects is usually decomposed: different software modules accomplish subtasks like object localization and computation of inverse kinematics, e.g. [15]. In contrast, recent advances in deep neural architectures allow end-to-end learning setups, where a robot acquires hand-eye coordination through the interaction with its environment. These approaches follow the developmental robotics paradigm [3], where increasingly sophisticated sensorimotor skills are developed through learning. With the recent success of deep reinforcement learning in discrete [19] and continuous [17] virtual environments a suitable way to train robots has become available. The large number of training iterations required and the resulting strain on robot hardware from hundreds of hours of training [21] render this approach non-applicable for non-industrial robot platforms.

To overcome this problem, endeavors have been made to transform the reinforcement learning task into a supervised learning task [16], [12], where the neural architecture is trained with sufficiently annotated and correct training samples instead of having to find these through trial-and-error. The work of Levine et al. [16] utilizes a model of the robot and a forward kinematics computation to create annotated training samples. The approach by Kerzel et al. [12] uses repeated random placing of an object to link visual stimuli to joint configurations by exploiting the fact that the same joint configuration used to place an object can be used to grasp an object. Both approaches facilitate end-to-end learning of visuomotor skills.

Both architectures initially process a visual input without depth information through a sequence of convolutional layers and use dense layers for computing motor commands. The method by Levine et al. [16] additionally uses a spatial soft-max layer to transform the pixel-wise information to a spatial coordinate representation for faster computation into motor commands. This can be seen as a type of attention mechanism where image features that do not contribute to the task are filtered out from the computation in later layers.

### B. *Faster R-CNN for object property detection*

To realize the object detection, we employ the *Faster R-CNN* (FRCNN) method by Ren et al. [22]. The approach uses a *Region Proposal Network* (RPN) that informs the object detection network about where to look for an object in the full image. By combining both networks with shared feature maps, the performance is significantly better than the performance of predecessor approaches. This method has been extensively used by several competitors of the MS-COCO object detection challenge, including the first, second and fourth place on the current 2017 MS-COCO leaderboard. The approach was originally intended to be used for labeling the category (e.g. *'car'*, *'cat'*, *'chair'*) of multiple objects in an image. However, the method can also be used to identify other object properties, such as shape and color. We exploit this additional capability to enable a richer identification method for object disambiguation.

## III. Framework

Our framework allows us to provide the NICO robot with high-level instructions for grasping. Specifically, the interface allows us to distinguish objects in the robot's field of view by three properties, namely shape, color and object class. The framework is depicted in Figure 1.

### A. *The NICO platform*

NICO (Neuro-Inspired COmpanion) is a child-sized humanoid robot that is designed to have human-like sensing and motor capabilities [11]. The NICO robot is designed to be a research platform for multimodal human-robot interaction and neurocognitive models. Figure 2 shows NICO in its training setup for grasping. NICO's head is adapted from the iCub [18], together with the body size it was chosen to create a positive, human-like but not uncanny appearance.

NICO's arms have six degrees of freedom to emulate a human-like range of motion: Three motors form a cluster in the shoulder area that emulates a ball joint, one motor each bend the elbow and rotate the wrist, while the sixth motor bends the wrist. The arms end in three-fingered hands. The fingers are operated by a tendon mechanism and can reliably grasp small objects and also serve as haptic sensors for successful grasp attempts. NICO's head can tilt and yaw to allow the two cameras that are embedded into the head to adjust its field of view. Furthermore, NICO features fully articulated legs.

### B. *Object property detection and data annotation*

The *Faster R-CNN* method [22] was originally proposed to detect objects and to predict their object class, as, e.g., required for the MS-COCO dataset with its 80 object classes that include *'car'*, *'cat'*, *'dog'*, etc. In this work, we exploit

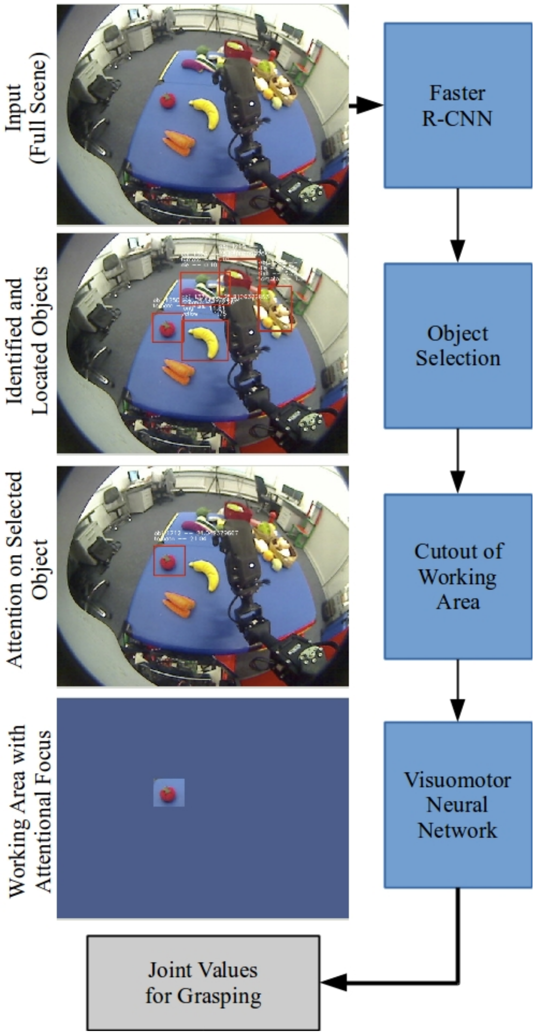

Fig. 1. Overview of our framework. The robot's camera image is first processed by an object detection mechanism that generates one bounding box for each object, labeled with class, shape and color information of the respective object. To identify a specific object, the user can input a set of object properties, and the framework selects the object with the highest cumulative score for the properties that the user entered (e.g. *'red'* and *'round'* may refer to the tomato). The attentional focus on this object is realized by removing all other objects from the visual input. From the resulting image with a focus on a single object, the relevant target area of the robot is extracted and processed by a visuomotor neural network to generate an appropriate joint configuration for grasping.

the capabilities of the network to not only detect object classes, but also additional properties, namely shape, and color. Additional object properties are important for real-world scenarios, where one may want to distinguish between several objects of the same class in the same image, e.g. *"grasp the red ball"* vs *"grasp the green ball"*, or, if one wants to refer to an object of a class that has not been trained, by naming its shape and color, e.g. *"grasp the white square object"*. In our parameterization, the network generates 300 bounding boxes for 21 classes in total, i.e., we use a single network model for shape, object class and color. The network architecture and hyperparameterization is identical to the Zeiler and Fergus (ZF) model [26] that has also been used in the end-to-end non-alternating setting of the Faster-RCNN-approach [22] for object detection.

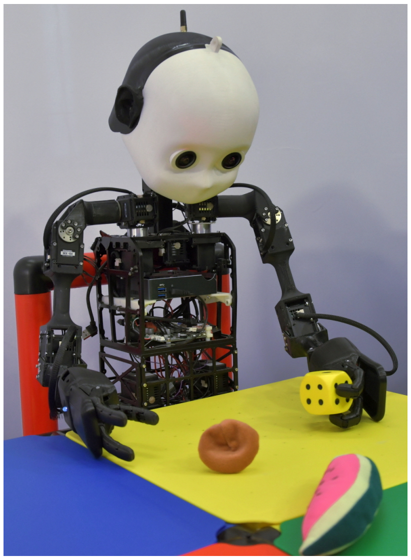

Fig. 2. NICO robot in its training setup with multiple objects.

To generate training data we implement a semi-automated annotation tool that is based on color-shift object tracking [4]. To use the tool, we first record videos with an external camera that we slowly move over a table with a set of obects that we want the system to learn, to display the objects from different angles. We annotate the first frame of the videos by drawing and labeling bounding boxes with the three properties that we consider in this study, i.e., object class, shape, and color. We then jump frame by frame through the video by pressing a hotkey, and, in most cases, the object tracker correctly shifts the initially drawn bounding boxes. When necessary, we can correct bounding boxes by re-drawing and re-annotating them. Assuming an average of 2 seconds for each frame, which shows, say, 6 objects with 3 property values each, the tool makes it possible to generate 540 annotations per minute or 32,400 annotations per hour. There are cases where the class, color or shape is not clearly identifiable, e.g. where an object has a color between yellow and orange. In such cases, it is possible to annotate a bounding box with multiple values for each property. It is also possible to annotate only a subset of all three properties, e.g. one can specify only shape and color of an object, omitting the object class.

### C. *Object identification and attentional focus*

The object detection Faster-RCNN (FRCNN) network that we describe in Section II-B generates a set of bounding boxes, along with scores for the object properties and classes that it is trained on. Hence, there are several bounding boxes for each object in the robot's field of view. To merge these

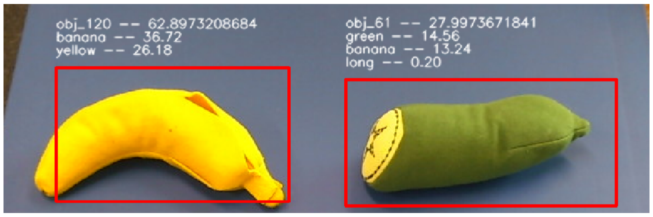

Fig. 3. The object identification mechanism assigns a set of label-score tuples to each object in the robot's field of view. Here, a yellow banana and a cucumber-like green object have been detected. The system is not trained on the cucumber-like object, and there is no dedicated object class for it. Hence, the system classifies it as the closest and most appropriate class, which is also *'banana'*, with a confidence score of 13.24, and as *'green'* with 14.56. This is reasonable, because the object indeed shows visual similarities with a green unripe banana, even though it s note entirely green. In comparison, the actual banana is labeled *'banana'* with a confidence of 36.72, and *'yellow'* with 26.18.

boxes, we perform clustering using an affinity propagation algorithm, which has the advantage that the number of clusters, and therefore also the number of objects, is variable. We do not run into scalability problems because the number of bounding boxes is limited (in our case to 300). The clustering generates one bounding box for each object that is annotated with labels and scores for object classes and object properties (see Figure 3). For example, for the detection of the banana, it generates one bounding box with a high score $v_{banana}$ for an object class *'banana'* and another bounding box with a high score $v_{yellow}$ for the color property *'yellow'*. There could also be a third bounding box for the color *'green'* if the banana is not yet ripe and its color is somewhere between yellow and green.

In allowing to select objects based on multiple labels, the system can not only detect but also *identify* objects based on the labels. Specifically, if the user enters a set $\mathcal{L}$ of labels, for example, $\{banana, yellow\}$. The system now sums up all scores for these labels for each object and selects the object with the highest total score, $v_{total}$, according to Equation 1.

$$v_{total} = \sum_{l \in \mathcal{L}} v_l \tag{1}$$

In the case depicted above, the yellowest banana would be identified as the object that the user is interested in. It is also possible to refer to less stereotypical or unknown objects, such as the green object in Figure 3, e.g., by asking for the greenest banana $\{banana, green\}$, or the greenest long object $\{green, long\}$. After the object is identified, all other objects are removed from the robot's visual input by computing the average background color in the image and by flood-filling everything around the bounding box for the selected objects with that color. The result is an image that only shows the identified object. This image is the input for the deep sensorimotor network to generate the grasping trajectory.

### D. The visuomotor network

The neural architecture for grasping follows Kerzel and Wermter [12]; it realizes grasping by a supervised end-to-end deep learning approach: The network architecture combines two constitutional layers and two fully connected layers to simultaneously locate an object in the visual input and compute the corresponding motor commands to grasp the object. As the architecture directly outputs the joint values that move the arm into a grasping position, it alleviates the need for an inverse kinematics component.

The network architecture and its hyper-parameters for training were adapted from [12], they were originally informed by successful approaches for learning visuomotor abilities, e.g. [19], and empirically optimized. The input to the network is an 80*60 RGB image, that shows the area directly in front of the robot; the peripheral areas are cut from the input as they do not contribute to the grasping task. The two constitutional layers that process the input have 16 filters each with a size of 3x3 and 4x4 (stride=1). This architecture does not have pooling layers to avoid translational invariance which would be obstructive to the localization functionality [19]. The two fully connected layers with 900 neurons follow the convolution layers. The output layer of the architecture consists of six neurons that correspond to the six degrees of freedom of the grasping arm. For each of the experimental conditions reported below, the network was trained for 2000 epochs with stochastic gradient descent with Nesterov momentum (learning rate = 0.01, momentum = 0.9). The batch size depended on the size of the training set. The squared error was used as a loss function.

The visuomotor network was trained in a semi-autonomous self-learning cycle ([12]) by letting the robot repeatedly place an object at random positions with minimal human assistance to generate training samples.

## IV. EXPERIMENTS AND EVALUATION

The experiments involve the grasping of seven different objects. The objects are distinguished by seven possible object classes, three possible shapes, and five possible colors. We have trained the network on 2,500 annotated frames with approximately 20 object property annotations each, which results in approximately 50,000 training samples. The visuomotor network has been trained with 535 samples. These samples are distributed differently among the different objects to be grasped as shown in Figure 4; grasping of some objects was trained with 200 samples, while others were trained with less than 100 samples or not trained at all. These values were deliberately chosen to analyze the influence of the number of training samples on the grasping success as an earlier study showed that without the attentional focus mechanism, i.e., in cases where only one object is present in the robot's field of view, the grasping success is positively correlated with the number of training samples [12].

| Object | Class | Shape | Color | Grasp | Grasp with Attention | Grap Training Samples |
|---|---|---|---|---|---|---|
| Die | die | squared | yellow | 88.9% | 16.7% | 200 |
| Tomato | tomato | round | red | 61.1% | 27.8% | 150 |
| Pepper | pepper | round | green | 88.9% | 22.2% | 75 |
| Banana | banana | long | yellow | 61.1% | 77.8% | 60 |
| Green-yellow lemon | lemon | round | yellow, green | 44.4% | 33.3% | 45 |
| Carrots | (no class) | long | orange, yellow | 94.4% | 61.1% | 5 |
| Eggplant | eggplant | long | purple, red | 100% | 83.3% | 0 |

Fig. 4. Success of detecting and grasping the objects we used in our experiments. Columns 2-4 describe object properties, column 5 shows how often the grasping works for single object images without attention focus, column 6 shows how often grasping works if one out of multiple objects is selected using the attention focus, and column 8 shows the number of samples used for grasp training.

### A. *Grasping*

First, we evaluated the performance of the visuomotor network for grasping without the attentional focus. For this, a single object was placed in front of NICO 18 times in a 3 x 6 grid. For each position a single grasping trial was performed, a grasp was counted as successful if NICO was able to grasp the object in such a way that it was able to lift up the object. We performed this experiment for all seven objects shown in Figure 4. Overall, objects were grasped successfully in 76.4% of all trials, with significant variations between the different objects: Easy to grasp objects that could be enclosed well in the robotic hands, like the eggplant or the carrots, were grasped successfully in over 90% of all problem instances. Objects that are hard to grasp tend to slip out of the fingers if not grasped at their exact center. These were only grasped with less than 50% accuracy (see Figure 4).

The result strongly indicates that the neural architecture is not only able to generalize from a relatively small set of training samples to a large number of spatial configurations from which the objects are picked up, but it also generalizes to different objects as long as these objects are visually similar. The exact number of training samples per object seems to be less relevant to the grasping success rate than its shape.

We repeated the experiments with the attentional focus mechanism in order to evaluate how well the FRCNN and the visuomotor network work together. To evaluate the FRCNN we placed two distractor objects as well as the target object in front of the robot and used the FRCNN to select the target object for grasping.

The FRCNN was 100% successful in visually selecting the desired object in all of the trials. The accuracy of grasping with an attentional mechanism diverged significantly from the evaluation of the visuomotor network: While easy to grasp objects could be grasped as well as before, the change to the network input that was caused by the attention focus significantly decreased the success rate of hard to grasp objects like the die or the pepper. With such objects, even a slightly different positioning of the robotic fingers can cause the object to slip out of the hand during grasping.

Overall, this leads to a combined accuracy of 46.0%. The results show that an attention mechanism that is based on manipulating the visual input to the visuomotor network can facilitate grasping without retraining the network. The changes to the input, however, seem to cause slight deviations in the resulting motor commands that lead to unsuccessful grasping attempts for hard to grasp objects. This could be caused by the strong color contrast between the object cutout and the homogeneous blue background in the input image (see bottom left image with the tomato cutout in Figure 1). A possible explanation for the performance decrease is that the grasping network considers the squared cutout to be the object to grasp, and not the object within the cutout. A more smooth contrast between the bounding box and the background may resolve this problem.

Further analyzing the results in relation to the number of training samples, it becomes apparent that the visuomotor network benefits from a large amount of training data for hard to grasp objects and can also generalize fairly well to little or never before seen objects. When applying the attention mechanism, however, the grasping precision gained from a large number of training samples is lowered significantly while the less precise grasping learned from fewer training samples is not affected as strongly. This could indicate overfitting on case of a larger sample number.

## V. CONCLUSION

We have presented a proof of concept that shows how a simple attention focus mechanism can realize a high-level interface for an end-to-end deep network to map camera images to grasping trajectories. This makes it possible to give symbolic grasping instructions where the object to be grasped is described by its properties, as in *"Grasp the red tomato!"* or *"Grasp the green long object!"*. We currently consider three object properties, namely object class, color, and shape, but the framework can also be trained to consider more object properties. We have shown that two different tasks, i.e., object identification and grasping, which require a different type and amount of training data and which are solved by different neural network architectures, can be integrated into a joint architecture by using an attention mechanism where the object detection architecture manipulates the input to our visuo-motor control architecture.

The results indicate that manipulation of the visual input to the visuomotor network can facilitate grasping of a selected object within a complex scene. The applied image manipulation by the FRCNN-based attention focus, however, causes less precise motor behavior when compared to unmodified input for single item grasping. In future work, we will address this problem by developing an architecture that provides a closer integration of the two networks and allows to fine tune the visuomotor network to the modified input.

Also, we will integrate other sensory modalities, like haptic perception, into the neural architecture to facilitate object identification in case of visually ambiguous objects [10] and improve the overall grasping accuracy.

We also want to use the framework to provide a high-level abstraction layer and interface for symbolic reasoning and action planning methods [5], [6] and for Natural Language Processing tools [7], [8]. Both applications are a step towards a tighter and more integrated interaction between humans and robots, which we will investigate in our future research.

## ACKNOWLEDGEMENTS

The authors gratefully acknowledge partial support from the German Research Foundation DFG under project CML (TRR 169) and the Hamburg Landesforschungsförderungsprojekt CROSS. Manfred Eppe acknowledges funding by the German Academic Exchange Service (DAAD) through the FITweltweit programme.

## REFERENCES

[1] B. J. Baars, "The conscious access hypothesis: origins and recent evidence," *Trends in cognitive sciences*, vol. 6, no. 1, pp. 47–52, 2002.

[2] H. C. Bottoms, A. N. Eslick, and E. J. Marsh, "Memory and the Moses illusion: Failures to detect contradictions with stored knowledge yield negative memorial consequences," *Memory*, vol. 18, no. 6, pp. 670–678, 2010.

[3] A. Cangelosi and M. Schlesinger, *Developmental Robotics: From Babies to Robots*. Cambridge, MA: MIT Press, 2015.

[4] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Accurate Scale Estimation for Robust Visual Tracking," in *British Machine Vision Conference BMVC*, 2014.

[5] M. Eppe and M. Bhatt, "Narrative based Postdictive Reasoning for Cognitive Robotics," in *International Symposium on Logical Formalizations of Commonsense Reasoning (CR)*, 2013.

[6] M. Eppe, M. Bhatt, and F. Dylla, "Approximate Epistemic Planning with Postdiction as Answer-Set Programming," in *International Conference on Logic Programming and Nonmonotonic Reasoning (LPNMR)*, 2013, pp. 290–303.

[7] M. Eppe, S. Trott, and J. Feldman, "Exploiting Deep Semantics and Compositionality of Natural Language for Human-Robot-Interaction," in *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 731–738.

[8] M. Eppe, S. Trott, V. Raghuram, J. Feldman, and A. Janin, "Application-Independent and Integration-Friendly Natural Language Understanding," in *Global Conference on Artificial Intelligence (GCAI)*, 2016, pp. 340–352.

[9] F. Ferreira and N. D. Patson, "The good enoughapproach to language comprehension," *Language and Linguistics Compass*, vol. 1, no. 1-2, pp. 71–83, 2007.

[10] M. Kerzel, M. M. M. Ali, H. G. Ng, and S. Wermter, "Haptic material classification with a multi-channel neural network," in *International Joint Conference on Neural Networks (IJCNN)*. Anchorage, Alaska: IEEE, May 2017, pp. 439–446.

[11] M. Kerzel, E. Strahl, S. Magg, N. Navarro-Guerrero, S. Heinrich, and S. Wermter, "NICO – Neuro-Inspired COmpanion: A developmental humanoid robot platform for multimodal interaction," in *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2017 accepted.

[12] M. Kerzel and S. Wermter, "Neural end-to-end self-learning of visuomotor skills by environment interaction," in *International Conference on Artificial Neural Networks (ICANN)*, 2017 accepted.

[13] V. A. Lamme, "Why visual attention and awareness are different," *Trends in Cognitive Sciences*, vol. 7, no. 1, pp. 12–18, 2003.

[14] M. F. Land and M. Hayhoe, "In what ways do eye movements contribute to everyday activities?" *Vision Research*, vol. 41, no. 25, pp. 3559–3565, 2001.

[15] J. Leitner, S. Harding, A. Förster, and P. Corke, "a modular software framework for eye–hand coordination in humanoid robots," *Frontiers in Robotics and AI*, vol. 3, p. 26, 2016.

[16] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-End Training of Deep Visuomotor Policies," 2015. [Online]. Available: http://arxiv.org/abs/1504.00702

[17] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[18] G. Metta, L. Natale, F. Nori, G. Sandini, D. Vernon, L. Fadiga, C. Von Hofsten, K. Rosander, M. Lopes, J. Santos-Victor, *et al.*, "The iCub humanoid robot: An open-systems platform for research in cognitive development," *Neural Networks*, vol. 23, no. 8, pp. 1125–1134, 2010.

[19] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[20] U. Neisser and R. Becklen, "Selective looking: Attending to visually specified events," *Cognitive Psychology*, vol. 7, no. 4, pp. 480–494, 1975.

[21] L. Pinto and A. Gupta, "Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours," in *Proceedings of the 2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 3406–3413.

[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Advances in Neural Information Processing (NIPS)*, 2015.

[23] D. J. Simons and C. F. Chabris, "Gorillas in our midst: Sustained inattentional blindness for dynamic events," *Perception*, vol. 28, no. 9, pp. 1059–1074, 1999.

[24] G. A. Wiggins, "The minds chorus: creativity before consciousness," *Cognitive Computation*, vol. 4, no. 3, pp. 306–319, 2012.

[25] S. A. Winges, D. J. Weber, and M. Santello, "The role of vision on hand preshaping during reach to grasp," *Experimental Brain Research*, vol. 152, no. 4, pp. 489–498, 2003.

[26] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in *European Conference on Compter Vision (ECCV)*, 2014, pp. 818–833.