

Neurocognitive Shared Visuomotor Network for End-to-end Learning of Object Identification, Localization and Grasping on a Humanoid

Matthias Kerzel, Manfred Epe, Stefan Heinrich, Fares Abawi, and Stefan Wermter
Knowledge Technology, Department of Informatics, University of Hamburg, Germany
kerzel / epe / heinrich / gabawi / wermter @informatik.uni-hamburg.de

Abstract—We present a unified visuomotor neural architecture for the robotic task of identifying, localizing, and grasping a goal object in a cluttered scene. The RetinaNet-based neural architecture enables end-to-end training of visuomotor abilities in a biological-inspired developmental approach. We demonstrate a successful development and evaluation of the method on a humanoid robot platform. The proposed architecture outperforms previous work on single object grasping as well as a modular architecture for object picking. An analysis of grasp errors suggests similarities to infant grasp learning: While the end-to-end architecture successfully learns grasp configurations, sometimes object confusions occur: when multiple objects are presented, salient objects are picked instead of the intended object.

Index Terms—Developmental robotics, bio-inspired visuomotor learning, cognitive robotics

I. INTRODUCTION

Recent advances in neural networks for robotic vision tasks like object classification and object localization [12] as well as for visuomotor abilities like reaching and grasping tasks [8] have shown the ability of fully learning mappings from visual perception to categories, locations, and motor actions. An open problem for integrating object localization and grasping tasks into a single framework is the importance of accurate spatial information, the complexity of visual features, and the varying availability of suitable training data. Contemporary architectures that integrate these capabilities usually use separate neural networks for each of these tasks. For example, our preliminary work [3] for integrating vision and visuomotor coordination is based on separate neural networks for object localization and grasping that are integrated by means of a non-neural attentional focus mechanism. Using separate neural approaches that are integrated via non-neural mechanisms seems to work to some extent, however, in biological systems, such as mammals, the integration occurs within a single modular neural architecture. Moreover, in the human brain, visual processing *shares* initial steps before branching off into the more specialized dorsal “where” and ventral “what” pathways [9]. Specifically, in low-level visual processing a high degree of hierarchical processing includes feedback,

The authors gratefully acknowledge partial support from the German Research Foundation DFG under project CML (TRR 169).

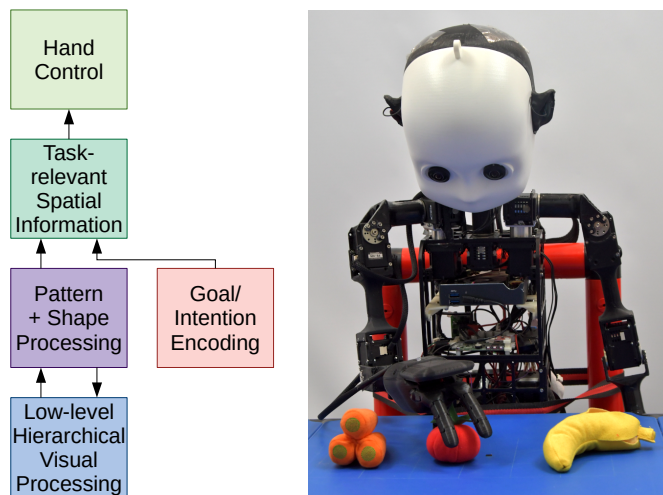


Fig. 1. Left: End-to-end architecture for object picking. Right: Experimental setup with NICO humanoid robot picking a selected object (red tomato) from a desk with two distractor objects.

and in the dorsal pathway the pattern, shape, and motion information gets filtered by task driven goal cues.

We suggest that a neurocognitively more plausible computational model, where different capabilities are integrated on the neural level, and thus the processing of visual information and computation of motor actions is shared – should even be advantageous over non-neural integration approaches (compare Figure 1 for an overview). Specifically, we address this issue by posing the following **research question**: *Can a neurocognitively inspired, unified neural architecture, embodied in a developmental robot, learn visuomotor abilities for distinguishing and for grasping objects?*

We pursue the question by studying a neural architecture that coherently integrates three different robotic capabilities. Specifically, we test combining convolution neural network (CNN) layers for object classification and localization [12], neurally encoded linguistic labels, and on a higher level, feed-forward layers for robotic grasping [8]. Our main contribution is a better insight, how such an integrated architecture interlinks spatially unspecified object locations with labeled object identifications and with correct grasping movements in an end-to-end real robotic system (illustrated in Figure 1). A potential

implications for developing human-level robotic agents is to benefit from cross-training effects between different tasks and to transfer and share learned architectures between them.

II. RELATED WORK

A. Object Detection with Neural Vision Networks

For image classification, pretrained vision architectures with various depths are available, e.g., the VGG-16 and VGG-19 models [16] with 16 and 19 layers, and the ResNet with up to 98 layers [5]. The architectures follow a conic scheme of interleaving convolutional and pooling layers, thus going from low-level features in high resolution to high-level features in a low resolution. These deep feature hierarchies perform well in classification tasks. However, object localization, in contrast, requires precise spatial information. While the architectures perform well at determining *what* is shown in a picture, the low spatial resolution in the deeper layers make it difficult to precisely determine *where* an object is located in an image. To address this issue, two-stage architectures have been introduced for object detection. These architectures have two networks, one for proposing a set of regions of the input image where an object might be located and a second stage that classifies these proposed regions [4]. High classification confidence indicates that the region is determined well, thus providing an accurate localization. This issue of repeated computation due to the two stages was in part addressed by *Faster R-CNN* [15] by reusing full-image convolutional features for classification.

As an extension, one-stage approaches achieve lowered processing time by classifying over a regular sampling of possible object localizations. RetinaNet [12] is the first single-stage approach that surpassed the performance of two-stage approaches by introducing a novel loss-function that addresses the class imbalance between the large set of background images versus the small set of actual objects. Another essential feature of RetinaNet is the use of a Feature Pyramid Network (FPN) [11]. An FPN complements a traditional CNN architecture with a parallel top-down pathway with additional lateral connections. It provides a multi-scale feature pyramid, where each layer can be used to detect objects at different scales. FPNs have been shown to improve the performance of detection architectures [12]. In summary, RetinaNet offers a singular architecture that can localize and classify objects in a cluttered scene. Due to the FPN architecture, both categorical “what” and spatial “where” features are available.

B. Neural Approaches for Grasping

Neural end-to-end learning of visuomotor abilities is explored in developmental robotics, where complex abilities are learned through the interaction of an embodied agent with its environment [2]. Kerzel and Wermter [8] suggest an approach that create fully annotated training samples of successful visuomotor actions by interacting with the environment to train neural architectures. Trial-and-error learning is avoided, but time-consuming environment interaction is still required. While it is easily possible to scale up collected training data in

the visual domain by means of data augmentation, it remains far more challenging for motor abilities, motivating coupling of visuo and motor tasks in a unified neural architecture that can benefit from cross-training effects.

C. Visuomotor-ability-learning in Humans

The parallel development of visual (and multimodal) object representations and motor abilities is highly plausible [6]. Once an agent has acquired the ability to grasp and manipulate an object, richer visual impressions of the object are generated by, e.g., rotating the object. This should be reflected in the underlying neural architectures. According to Oztop et al. [14], human infants perform open-loop grasping. I.e., they look at a scene and decide on a trajectory for grasping without correcting their motor action during execution. Oztop argues that visual processing abilities to relate hand and object pose are not developed in infants. In contrast, adults perform closed-loop grasping: the spatial relation between hand and object is monitored, and the grasping action is updated accordingly. Though closed-loop grasping is more robust, open-loop grasping is an important developmental step towards complex visuomotor abilities. This open-loop grasping is realized by Kerzel and Wermter [8] in a neurobotic model that is extended by the presented approach for picking objects in cluttered scenes. For such tasks, Libertus et al. [10] report that infants show an initial grasp preference for visually salient objects. We will analyze if similar effects occur in our approach.

III. METHODOLOGY AND NEURAL ARCHITECTURE

Visuomotor abilities, like object picking, inherently combine the processing of visual information, including object classification and object localization with motor control policies for the specific object identities. In our approach, we aim to realize these characteristics in a neurocognitively plausible unified neural network architecture as presented in Figure 2. First, we base our architecture on the RetinaNet [12], since it is capable of reliably identifying as well as locating objects in a scene and at the same time mimics hierarchical processing steps as found for processing in the brain. This component does not explicitly make a distinction between the “where” and “what” pathways in the brain, but rather reflects the preprocessing and abstraction steps. Second, we concatenate one-hot encoded object labels with the output of the Feature Pyramid Net, to introduce information for specifying a certain object, similar as an intention or goal can actively steer the perception in the brain. And third, we integrate the layers (two dense and one output) of our recent visuomotor-grasping network [8], as it is capable of learning joint configurations for grasping single objects end-to-end. Finally, the training is performed end-to-end on cluttered scenes, e.g., providing a raw pixel image of a scene with multiple objects as well as two one-hot-encoded vectors describing the category, color, and shape of the desired object and expecting the selected object getting picked up with a correct joint configuration.

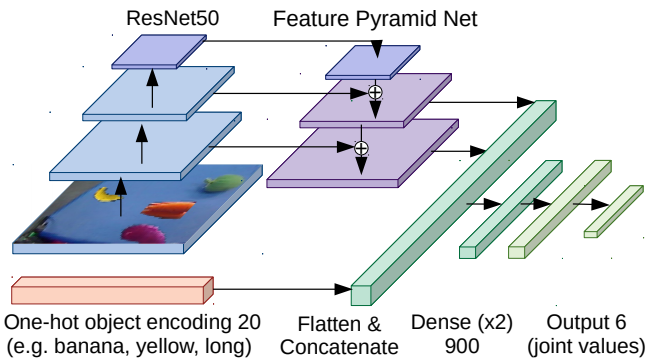


Fig. 2. RetinaNet-based architecture for neuro-inspired end-to-end learning of object picking. Figure adapted from [5], [8].

In particular, the convolutional part of the architecture has the purpose to differentiate and locate objects in an image, i.e., it extracts features that are rich enough to enable object selection but still provide a high spatial resolution to enable motor action with regard to this object. We evaluate both the use of relatively shallow fully convolutional architectures without pooling [8], [12], [13], as well as the Feature Pyramid Net introduced in RetinaNet for visual processing. The input to the convolutional network consists of a 60×80 pixel RGB image. The output of the network is flattened to 22563 units and concatenated with a vector of length 20 that encodes the object to be grasped and its visual properties; e.g., green, round pepper. The concatenation layer is followed by dense layers and directly outputs six joint angles to control the robotic arm. The architecture is trained supervised with back propagation, the training procedure is detailed in section V.

IV. EXPERIMENTAL SETUP

A. NICO Robot

Our experiments are realized on NICO, a robotic research platform for embodied neuro-cognitive models and crossmodal human-robot collaboration [7] (see Figure 1)¹. When standing up, the anthropomorphic robot NICO has a size of about one meter and child-like proportions. Similar to a child of age four to five, it is large enough to interact with normal domestic environments but requires smaller furniture for proper seating. NICO’s arms have a human-like range of motion with six degrees of freedom (DoF): three DoFs form a ball joint in the shoulder area, one DoF bends the elbow, and two DoF rotate and flex its robot hand. The hand is a tendon-operated, three-fingered Seed Robotics SR-DH4D² hand, that can grip objects well that would fit into the hand of a child. For visual perception, NICO is equipped with two Logitech C9056 cameras.

B. Dataset

In our experimental setup, NICO is seated at a table and presented with various toy objects. Figure 1 shows the experimental setup with three objects: carrots, tomato, and



Fig. 3. Top row: unmodified training examples gained from the self-learning cycle. Each sample is annotated with a joint configuration to grasp the object. Bottom rows: augmented images with one, two and three distractor objects.

banana. The robot’s task is to pick one selected object. The training dataset is created in two phases, self-sampling and data augmentation. During self-sampling, the self-learning paradigm by Kerzel and Wermter [8] is used to create fully annotated samples for grasping objects in single-object scenes. This approach requires no information about the kinematic model of the robot. We extended the dataset collected in [3] and during the data augmentation used image manipulation to add distractor objects for end-to-end learning of grasping in cluttered scenes.









1) *Self-learning through interaction with the environment:* For learning to grasp a single object we utilize a neuro-cognitively inspired self-learning approach where the robot generates fully annotated training samples by repeatedly placing and re-grasping an object. We utilize the fact that grasping an object can be transferred into the much simpler task of placing an object at a random position.

The robot’s hand is first manually moved over a table surface for a few seconds to collect data for the subsequent phase. Thereafter, the robot enters a self-learning cycle where the robot’s hand begins in a home position. Then, the experimenter puts one of the objects for grasp learning into the robot’s hand. The robot then moves to a random joint configuration from the initial training; the robot memorizes the joint configuration and places the object on the table. Next, the robot moves the hand away to take an unoccluded picture of the scene. The robot moves back to the memorized joint configuration and grasps the object again. The robot can automatically detect whether the grasp is successful using the proprioceptive haptic information from the hand motors. If the re-grasping attempt was successful, the recorded picture is stored together with

¹ Further information and videos: <http://nico.knowledge-technology.info>

² <http://www.seedrobotics.com/>

TABLE I
TOP ROW: OBJECTS FOR PHYSICAL GRASP TRAINING:
BOTTOM ROW: OBJECTS FOR AUGMENTATION AND EVALUATION.

Type	apple	pepper	tomato	die	banana
Color	green	green	red	yellow	yellow
Shape	round	round	round	cube	long
Image					
Type	eggplant	kiwi	die 2	carrot	orange
Color	black	black	blue	orange	orange
Shape	long	round	cube	long	round
Image					

the memorized joint values as a training sample and the self-learning cycle is repeated. If the re-grasping attempt fails, picture and memorized joint values are discarded, the hand is moved back into the home position and human assistance is requested. The self-learning was performed with five different objects: apple, tomato, pepper, banana, and yellow die. In total, 232 equally distributed samples were recorded. Figure 3 shows examples of training images. In postprocessing, each training image was cropped to the relevant 80×60 pixel table area. According to Kerzel and Wermter [8] this small number of samples is suitable to train a neural grasping model.

2) *Dataset Augmentation*: To make the sample suitable for learning to pick an object from a cluttered scene, we augment them by pasting distractor objects into the images using image manipulation. From the collected samples and additional pictures, images of a single graspable object without background are manually extracted (10 images per object). The additional objects are from the following categories: eggplant, kiwi, blue die, carrot, and orange. Each of these distractor objects is assigned with a color and shape descriptor (see Table I). To augment a fully annotated sample, a number of random cutouts are selected, and empty areas are found in the original sample using edge detection. The pasted objects do not overlap with an object in the scene or each other based on bounding boxes. It is also ensured that the pasted objects also do not overlap with regard to categories from the original image (color, shape, class). To maximize the augmentation benefit, two affine transformations are applied to the pasted objects: rotation (random 360 degrees), scaling (random, minimum size = 20 pixel, maximum = 40 pixels). This way, any required number of augmented training images can be created, each with two annotations: the semantic category of the object in the original sample and the corresponding joint values for grasping this object. Figure 3 shows examples of augmented images with one, two and three added distractor objects.

V. EXPERIMENTS AND RESULTS

We first perform baseline experiments with a single object to evaluate the ability of the network to learn visuomotor abilities and to investigate the influence of distractor objects. In the main experiment, we train the architecture to grasp five different objects with the above described augmented training set. Finally, we embody the best-trained model in the

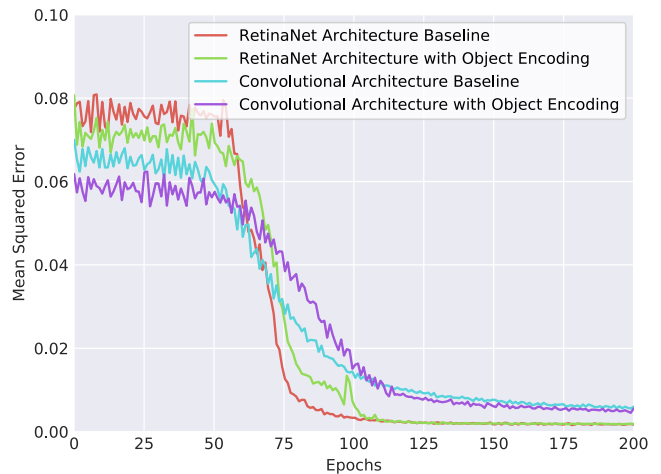


Fig. 4. MSE for different architectures, single object grasp learning.

NICO robot for physical grasping. For all models, we use 323 fully annotated sample of five different objects. The object is represented by an n-hot encoded vector, resulting from the super-imposition of one-hot vectors signifying the shape, color, and category. The images were reshaped to 80×60 pixels. The output of all models is the joint configuration for the robot arm, represented as a vector of six floating-point values, normalized on a scale between 0 and 1.

A. Baseline Experiments: Single Object Grasping

We train the architecture without semantic information or distractor objects. Both the baseline system with a convolutional architecture [8] consisting of two convolutional layers (16 4×4 filters), followed by two dense layers with (900 units) and the modified architecture that replaces the convolutional layers with the ResNet and Feature Pyramid Net are evaluated. Mean squared error (MSE) is used as loss function. Hyperparameters were informed by models from literature [5], [8]: learning rate = 0.01, with a momentum of 0.9 using Nesterov accelerated stochastic gradient descent on batches of 20 images for 200 epochs. Models trained with RetinaNet models had a ResNet-50 model as a backbone, initialized with ImageNet pre-trained model weights.

Results and discussion: All experiments are repeated 10 times with randomly initialized weights. The MSE is computed over the validation set (10% of the dataset). The results depicted in Figure 4 show that both architectures can learn joint values for object grasping with an MSE of 0.006 for the convolutional architecture and 0.002 for the RetinaNet architecture. We repeated the experiment while adding semantic information about the object to be grasped and did not detect a significant difference in MSE. These results add evidence to the hypothesis that a pyramidal network performs similarly good for the given spatial visual task and that adding semantic information does not influence the performance.

B. Main Experiment: Object Picking in Cluttered Scene

Each image is augmented with one to three distractor objects as described in subsection IV-B2. An n-hot encoding of

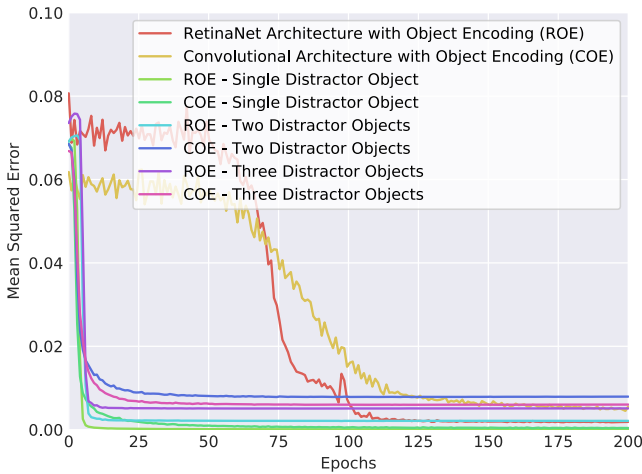


Fig. 5. MSE for different architectures, object picking with zero to three distractor objects.

the desired object’s attributes is fed into the network by concatenating the information with the flattened output of the Feature Pyramid Net. All hyperparameters were taken from previous baseline experiments. The model is trained with 10 trials of randomly initialized weights for each architecture.

Results and discussion: The mean squared error is computed over the validation set (composed of 10% of the entire NICO dataset). From the results shown in Figure 5 it becomes apparent that the object augmentation increases the learning speed but it does not lead to decreased MSE for the converged models. The averaged mean squared joint error increases with the number of distractor objects: For the RetinaNet architecture, the MSE increases from 0.0006 (one distractor object) to 0.0021 (two distractor objects) and 0.0051 (three distractor objects). A similar increase was found for the convolutional architecture. This result has two competing explanations. First, the overall ability of the network slightly deteriorates, i.e., the output joint angles are less suited to grasp the intended object in all cases. Second, the network can not always focus on the intended object, which results in the network producing joint values for a non-goal object and being less prone to overfitting. To address these hypotheses, we employ the model on the physical robot to perform object picking.

Hyperparameter Optimization: To further improve the grasp accuracy, we performed hyperparameter optimization using a tree-structured Parzen estimator [1]. During optimization, images with one to three distractor objects were used. We explored: learning rate: $1e-4$ and 0.9, with the option to reduce the learning rate on plateau; number and type of layers after the pyramidal network: 1 to 4 dense layers with 32 to 1800 units or convolutional layers with 32 to 2048 filters, and proportional 2D kernel sizes between (3,3) and (6,6) optionally followed by a max. or avg. pooling layer. A dropout layer can follow any layer of the two variants, with a dropout rate between 0 and 0.5. An activation function for each layer was selected from the available set (ReLU, Tanh, and sigmoid). After 110 trials, we found the 900 unit dense layer with sigmoid activation,

TABLE II
RESULTS OF SINGLE OBJECT GRASPING TRIALS ON A PHYSICAL ROBOT. DIFFERENT OBJECT WERE PLACED IN 3×6 GRID.

Experiment Type	Object	Grasp accuracy	Touch accuracy
In training	Tomato	94.4%	100%
In training	Banana	100%	100%
In augmented set	Eggplant	88.8%	100%
In augmented set	Kiwi	94.4%	100%
Unseen	Red Sponge	100%	100%
Unseen	Cut Apple	100%	100%
Average	all objects	96.3%	100%

followed by a 2D convolutional layer with ReLU activation and 349 filters of a kernel size (3,3) as well as a max. pooling layer with size (2,2) followed by a dropout of 0.12, achieving the lowest MSE of 0.001 given a learning rate of 0.14 with learning rate reduction on plateau enabled.

C. Robotic Experiments

In the robotic experiments we evaluate grasp accuracy under realistic conditions and investigate whether the increased MSE on the joint values is caused by distractor objects stems from an overall decrease in accuracy or from confusing objects. We embed our optimized model the physical robot platform. We first perform baseline experiments on grasping a single object and then add distractor objects. We also test the grasping performance on never-before-seen objects based on semantic object descriptions.

1) *Single Object Grasping:* In the experimental setup, one object is placed in a 3×6 grid in the 30×60 cm workspace of the robot. The grasped objects differ in term of being present in the recorded training set (tomato, banana), present in the augmented training set (eggplant, kiwi), and never-seen-before objects (red sponge, apple slice). For all grasping trials, the number of successful grasps and is also the non-successful grasps during which the robot finger at least touched the object were recorded. The results are summarized in Table II.

Results and discussion: We reach an average grasp accuracy of 96.3%. This exceeds the accuracy reported by Kerzel and Wermter [8] by 10%. We attribute this increase in grasp accuracy to the pyramidal vision architecture that can preserve spatial information while offering a more complex visual feature extraction. Little difference in the overall high grasp accuracy for objects with regard to the object present in the training sets can be attributes to random fluctuations. Overall, this result can be interpreted as the network being able to learn grasp never-seen-before objects based on a description of their shape and color.

2) *Robot Object Picking: Known and Unknown Objects:* We evaluated object picking by placing two and three objects, into the workspace of the robot. The target object was placed at 3×6 grid positions while the distractor object(s) shifted its position accordingly to be non-connected and non-overlapping with the target object. Trial I used two objects from the training set (tomato, banana). Trial II and III used three objects, two from the training set (pepper, carrots), and one novel object (red grapes). In trial II, the target object was from

TABLE III

RESULTS OF OBJECT PICKING EXPERIMENTS TRIALS ON REAL PHYSICAL ROBOT. TARGET OBJECTS WERE PLACED ON 3×6 POSITIONS, TWO TRIALS PER CATEGORY RESULTED IN 36 GRASP ATTEMPTS.

Experiment	I	II	III
No. of objects	2	3	3
Training set	yes	yes	no
Grasp correct obj.	66.6%	50.0%	83.3%
Touch correct obj.	6.6%	27.8%	16.7%
Grasp incorrect obj.	12.8 %	22.2%	0%
Touch incorrect obj.	8.3 %	0%	0%
None	5.6 %	0%	0%

the training set, in trial III, it was novel. The target objects were fully specified in the semantic encoding (e.g., red, round, tomato). Finally, to evaluate picking never-seen-before objects, we place three sponges, two red, one green and vice versa, to be selected by color.

Results and discussion: Table III shows the results of the two and three object picking tasks. We achieve a grasp accuracy of 66.6% and 50.0% for two and three objects on average. The architecture outperforms the accuracy of 46.0% achieved in previous work using decoupled vision and motor networks [3]. Usually grasps fail due to small deviations from an optimal grasp configuration. However, the system shows a tendency to grasp the incorrect object (12.8% and 22.2%). This result can be interpreted as still purposefully directed grasp movements towards an object, but not the specified one. Therefore the error in the multi-object picking can be attributed to difficulties in correctly identifying objects. For grasping never-seen-before objects, this accuracy increases to 83.3%. We observe these results consistently for visually more distinct objects. In this trial the objects had clearly different color characteristics, e.g. green versus red grasping objects, whereas in trials with lower accuracy these characteristics were more similar, e.g. yellow versus yellowish-orange objects. On the one hand, this means that the correct differentiation is naturally bound to the differentiability of objects that are supposed to differ in shape and color given the noise camera input. On the other hand, the result shows that the visuomotor network can generalize well to abstract semantic object descriptions, in this case, color.

VI. CONCLUSION

We present a developmental robotics approach for neural end-to-end learning for object picking in a scene with multiple objects. This is achieved by encoding object properties and fusing them with preprocessed visual input in a neural architecture. Our end-to-end training allows shared and parallel learning of visuo and motor abilities. This is a neurocognitively more plausible approach than existing architectures, where the individual tasks are decoupled. We propose a sample augmentation procedure that minimizes the need for physical robot interaction while still allowing learning of complex visuomotor tasks with a large variety of objects. Using our architecture, we successfully realize object picking tasks in a physical robot setup. Both for single-object grasping and for picking known objects, our modified architecture exceeds

the accuracy of previous work [3], [8]. By using general descriptors of shape and color, the architecture can generalize abilities for object picking to never-seen-before objects. In line with findings from human grasp learning, grasp errors often stem from grasping (or trying to grasp) incorrect objects. To further improve the performance, in future work, the set of training objects will be extended to increase the robustness. Also, we will extend the approach to simultaneous learning of multiple auxiliary tasks, such as learning to detect objects and object labels while learning the motor joint values.

REFERENCES

- [1] J. Bergstra, D. Yamins, and D. Cox, "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures," in *Proc. of the 30th International Conference on Machine Learning (ICML 2013)*, 2013, pp. 115–123.
- [2] A. Cangelosi, M. Schlesinger, and L. B. Smith, *Developmental robotics: From babies to robots*. MIT Press, 2015.
- [3] M. Eppe, M. Kerzel, S. Griffiths, H. G. Ng, and S. Wermter, "Combining Deep Learning for Visuo-motor Coordination with Object Detection and Tracking to Realize a High-level Interface for Robot Object-picking," in *IEEE RAS International Conference on Humanoid Robots (Humanoids)*, 2017, pp. 612–617.
- [4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.
- [5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [6] S. Heinrich, M. Kerzel, E. Strahl, and S. Wermter, "Embodied multi-modal interaction in language learning: the emil data collection," in *Proceedings of the ICDL-EpiRob Workshop on Active Vision, Attention, and Learning (ICDL-EpiRob AVAL)*, 2018.
- [7] 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, pp. 113–120.
- [8] M. Kerzel and S. Wermter, "Neural end-to-end self-learning of visuo-motor skills by environment interaction," in *International Conference on Artificial Neural Networks (ICANN)*, 2017, pp. 27–34.
- [9] N. Kruger, P. Janssen, S. Kalkan, M. Lappe, A. Leonardi, J. Piater, A. J. Rodriguez-Sanchez, and L. Wiskott, "Deep hierarchies in the primate visual cortex: What can we learn for computer vision?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1847–1871, 2013.
- [10] K. Libertus, J. Gibson, N. Z. Hidayatallah, J. Hirtle, R. A. Adcock, and A. Needham, "Size matters: how age and reaching experiences shape infants preferences for different sized objects," *Infant Behavior and Development*, vol. 36, no. 2, pp. 189–198, 2013.
- [11] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.
- [12] T.-Y. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2999–3007, 2017.
- [13] 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.
- [14] E. Oztop, N. S. Bradley, and M. A. Arbib, "Infant grasp learning: a computational model," *Experimental Brain Research*, vol. 158, no. 4, pp. 480–503, 2004.
- [15] 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 Systems*, 2015, pp. 91–99.
- [16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.