Exploring Deep Models for Comprehension of Deictic Gesture-Word Combinations in Cognitive Robotics

DOI:
10.1109/IJCNN.2019.8852425

Document Version
Accepted author manuscript

Exploring Deep Models for Comprehension of Deictic Gesture-Word Combinations in Cognitive Robotics

Gabriella Pizzuto
School of Computer Science
University of Manchester
Manchester, UK
gabriella.pizzuto@manchester.ac.uk

Angelo Cangelosi
School of Computer Science
University of Manchester
Manchester, UK
angelo.cangelosi@manchester.ac.uk

***Abstract*—In the early stages of infant development, gestures and speech are integrated during language acquisition. Such a natural combination is therefore a desirable, yet challenging, goal for fluid human-robot interaction. To achieve this, we propose a multimodal deep learning architecture, for comprehension of complementary gesture-word combinations, implemented on an iCub humanoid robot. This enables human-assisted language learning, with interactions like pointing at a cup and labelling it with a vocal utterance. We evaluate various depths of the Mask Regional Convolutional Neural Network (for object and wrist detection) and the Residual Network (for gesture classification). Validation is carried out with two deictic gestures across ten real-world objects on frames recorded directly from the iCub's cameras. Results further strengthen the potential of gesture-word combinations for robot language acquisition.**

***Index Terms*—cognitive developmental robotics, embodied language acquisition**

Fig. 1: The iCub learns new vocabulary through the use of complementary gesture-speech combinations.

I. Introduction

Humans have the inherent ability to acquire language skills during interaction, and gestures have an important role in this task [1]. Research has shown that gestures are the harbinger of language and fulfil a key role in language acquisition, by serving as predictors for verbal and sentence complexity [2]. Additionally, human interactions also benefit greatly from non-verbal cues, as these complement or supplement vocal utterances. However, natural language interaction, comprehension and acquisition in robots is still an open problem [3].

The motivation behind this work is two-fold. First, it opens a window of possibilities in engaging humans to act as teachers when vocally labelling objects referred to by gestures, and secondly, using deictic gestures adds a more natural component to any interaction. Deictic gestures are used for making reference, and comprise the four gestures of requesting, showing, giving, and pointing [2]. A complete integrated system would allow the robot to understand and generate more complex language.

Indeed, numerous methods have been proposed for language acquisition [4], [5] and vision-language integration [6] in robots; yet, none seem to include gestures in this process. Recent Deep Learning (DL) methods have shown promising results in bringing this to fruition. This technique has shown unparalleled achievements in research areas such as object detection and language generation [7]. However, results obtained from such methods rely heavily on the quality and quantity of the training data. Also, most datasets are made up of objects and people, and lack the integration of gestures and objects. When it comes to gesture datasets, most of these are recorded in isolation. This is quite unnatural in human interaction, as humans carry out deictic gestures with respect to a reference. Another reason for the lack of datasets could be that image annotation is a tedious task as it requires object segmentation and labelling. Our approach not only deals with images recorded directly from the robot's cameras, including both gestures and objects, but the dataset is also annotated in the JavaScript Object Notation (JSON) format that is widely used in DL architectures. Hence, other architectures can easily replace the Regional-Convolutional Neural Network (R-CNN) used.

This paper aims at addressing the implementation of R-CNNs, specifically by leveraging Mask R-CNNs, for gesture-

The authors acknowledge the support of the Horizon 2020 Marie Sklodowska-Curie Innovative Training Networks through the project grant number 676063.

speech comprehension. Here, our objective is to ground advancements in machine learning in a real-world application, i.e to enhance iCub's language acquisition skills, which is illustrated in Fig. 1. The main technical contributions of this paper are: (i) the application of Mask R-CNN for localising the hand and subsequently classifying the deictic gesture used (ii) the use of Mask R-CNNs for gesture and object recognition in the same scene and (iii) the use of an integrated system for comprehension of complementary gesture-word combinations on a real robot. Our architecture demonstrates a solution for using deictic gestures to obtain a mask around the object of interest. To the best of our knowledge, no previous system has been implemented in this manner.

The rest of the paper is structured as follows: Section II gives a review of related work on complementary gesture-speech combinations in language acquisition, object detection and gesture recognition, and Section III reports the architecture used together with the integration framework for implementing the latter on the iCub. Section IV details the experimental evaluation carried out, including the results obtained. Finally, Section V provides a discussion on the aforementioned results, draws conclusions and gives direction for future work.

II. Related Work

Our focus here is on related work in developmental language acquisition, where we look into the importance of complementary gesture-speech combinations for language acquisition in infants. Subsequently, we illustrate methods that have been used in the scenarios of object detection and gesture recognition. Here, the body of literature on computer vision techniques is vast; however, we will look primarily at deep learning methods. In all subsections, we relate our approach to other research endeavours and provide the motivation behind our chosen direction.

A. Complementary Gesture-Speech Combinations in Child Language Acquisition

Prior to the onset of speech, infants produce gesture-speech combinations which predict vocabulary spurt during language development [2]. Here, the primary focus has been on two types of gesture-speech combinations: complementary and supplementary. Complementary combinations occur when the gesture does not provide additional knowledge to the vocal utterance, whereas in the supplementary modality, the gesture adds semantic information to the speech [8]. Research in this domain has shown that complementary gesture-speech combinations have a positive correlation with the verbal complexity and vocabulary size at later stages. In another study, researchers also found that using complementary combinations at the age of 18 months predicted vocabulary size and mean length of utterance at 24 months [9]. Additionally, it has been shown that children used deictic gestures in combination with speech during the babbling and one-word stage [10]. Furthermore, studies have illustrated that when complementary combinations are used to modify nouns, for example point + cup resulting in "the cup", they provide insight into the learning process of new constructions in speech [11].

In general, it is believed that the complementary gesture-speech combinations appear at around the age of 12 months, prior to the transition to the two-word stage. Although it is the supplementary combinations which predict this transition, the complementary combination symbolise a number of important milestones in early language development [12]. Moreover, as the complementary modality precede other gesture-speech combinations we believe this is an important starting point for robot language acquisition that is influenced by infant development. Hence, this fuelled our research work in this direction.

B. Methods for Object Detection

Object detection is a challenging problem, consisting of classification and localisation of objects, humans and animals, amongst other classes. The current state-of-the-art in this field is the use of DL architectures [13] - [16]. CNNs have been applied with great success for detection, segmentation and recognition of objects and regions in images [7]. As conventional CNNs are relatively slow, these were succeeded by a new method that applied sliding window detectors, known as R-CNNs [17]. These high-capacity CNNs make use of a selective search algorithm to generate region proposals, compute CNN features and use Support Vector Machines (SVMs) for classification [18]. The bottleneck of the training is leveraged through the introduction of Fast R-CNN [19]. This method applies the CNN to the full image and consecutively uses Region of Interest (RoI) pooling on the feature map. Faster R-CNN followed this method, where the selective search method was replaced with Region Proposal Networks (RPNs) [20].

The method of Faster R-CNN is extended by adding an object mask detector in parallel, that is used for pixel-level image segmentation; this architecture is known as the Mask R-CNN [21]. Here, the parallel branch predicts the object segmentation mask. This network takes advantage of the Feature Pyramid Network (FPN) to improve its feature representation capability and eases the challenge of small object detection [22]. Given the effectiveness of Mask R-CNN, we opted for this network here since it has shown remarkable results at general-purpose object instance segmentation and is the current state-of-the-art in this field.

C. Gesture Recognition

The core of a gesture recognition system is in segmenting body parts (mainly hands), extracting features from the segmented region and labelling the gesture through a recognition algorithm [23]. Traditionally, feature extraction methods relied on the use of histogram of oriented gradients features; nevertheless, similarly to object detection, DL is currently also considered the state-of-the-art in gesture recognition [24]. Another element of such a system is classification, where SVMs and Recurrent Neural Networks (RNNs) are some of the methods that have been used. One of the more recent works

combines the use of CNNs with the Long Short-Term Memory (LSTM) for classifying gesture frame sequences [25] .

As we are currently not focusing on sequences of gestures, but more so in isolating the gesture-object interaction from the rest of the scene, the choice of re-using the Mask R-CNN architecture for hand isolation and combining it with a classical CNN to classify the gesture seemed a good approach.

In the robotics domain, gesture understanding has mostly pivoted around Human-Robot Interaction (HRI) scenarios [26]. The majority of these projects focused on using representational gestures, whilst others used pointing as a form of deictic gesture. Since we are more concerned with using deictic gestures for comprehension of complementary gesture-speech combinations, this will be our primary focus; yet we will not limit our work here to merely pointing.

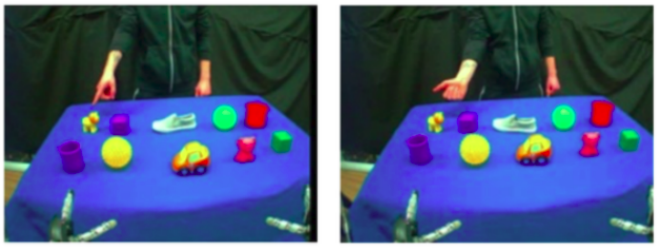
III. Methods

Our system is composed of a two step cascade. The first stage is a Mask R-CNN for wrist keypoint detection; its function is to focus on the area close to the gesture. Our second stage performs classification and consists of two parallel branches. Here, the cropped image is fed into another Mask R-CNN which performs recognition on the object closest to the wrist, whilst a Residual Network (ResNet) labels the deictic hand gesture. This system was trained on a dataset recorded directly from the iCub's cameras described in Section III-A. An in depth overview of the complete architecture is given in Section III-B; however a more comprehensive analysis of the performance of each part is outlined in Section IV. Since the latter is deployed on the iCub humanoid robot, the fully integrated system outlining how this is achieved is described in Section III-C.

A. iCub's Gesture-Object Dataset

To train and evaluate the system, a dataset comprising participants carrying out deictic gestures in front of a table of objects, was recorded through the iCub's cameras. A subset of this dataset was used for this work. This consisted of 1,200 frames for training, 400 frames for validation and 400 frames for testing. These frames were labelled in JSON format using a modified Common Objects in Context (COCO) style dataset generator graphical user interface [27].

The full dataset consists of 20 participants (13 females, 7 males) performing three of the four deictic gestures (showing, requesting and pointing) with respect to each of the ten different objects, that are commonly used when infants are learning new vocabulary, such as cup, ball and car. However, for this work, only the frames concerned with the gestures of requesting and pointing were included as the frames that illustrate the showing gesture did not fully isolate the gesture from the object. The Yet Another Robot Program (YARP) framework [28] was used to record videos at 30 frames per second. The iCub was placed in an empty room with a black background behind the participant. The robot stood in front of the human with a table where the objects were placed in between. A sample from this dataset is illustrated in Fig. 2.



(a) The participant is carrying out the deictic gesture of pointing.

(b) The participant is carrying out the deictic gesture of requesting.

Fig. 2: Frames extracted from the gesture-object dataset recorded by the iCub humanoid robot. The dataset comprises 3 deictic gestures (only 2 are shown here since these were the ones used in this research work) and 10 objects.

B. The Gesture-Object Detection System

The Mask R-CNN was chosen as the core of our system since compared to other methods, it has shown remarkable performance in object and keypoint detection [21]. The strengths of this approach lie in its ability to adopt RPNs resulting in suitable inference times for HRI, and provides an attractive solution to get the centre of the object from the mask as a coordinate that can be given to the robot's motor system for gaze control or motor feedback.

The input to the system is an image captured directly from the iCub's left camera, fed into the first Mask R-CNN network for extracting the wrist keypoint. This is an end-to-end trained keypoint-only Mask R-CNN with initial weights taken from the 2014 COCO dataset [29]. Here, the backbone is the ResNet with a FPN. The output from this network is the main keypoints that represent the skeleton of the body; nonetheless, since we are focusing on gesture recognition we were only interested in the wrist keypoint. Therefore, the image was cropped around this point in preparation for the next stage of the network. The size of the cropped image was chosen in such a manner that for the tested scenarios the object referred to by the gesture was sufficiently captured.

The cropped image is fed into two different networks for object detection and gesture recognition respectively. With regards to object detection, a Mask R-CNN with the ResNet-50 backbone and initial weights trained on the 2017 COCO dataset was used. COCO has 91 different classes and only the ball and cup categories from both datasets are shared. Regarding the rest of the objects, these were not classified using off-the-shelf pre-trained networks. Hence, this network was fine-tuned on the dataset that was collected in Section III-A.

The training process consisted of freezing the shared layers from the COCO pre-trained model and training the other layers using the features from the acquired iCub Gesture-Object dataset. One of the strengths of this finetuning process is that it alters the parameters of the network's layers with respect to the new dataset. For this reason, it adapts better to the low

resolution images obtained directly from the robot when it comes to the inference stage. In the end, the mask obtained as an output from this network is stored.

For the gesture recognition system, the cropped image is fed into another ResNet-50 for classifying the gesture into one of the two deictic gestures of pointing or requesting. A comprehensive block diagram illustrating the proposed architecture is depicted in Fig. 3.

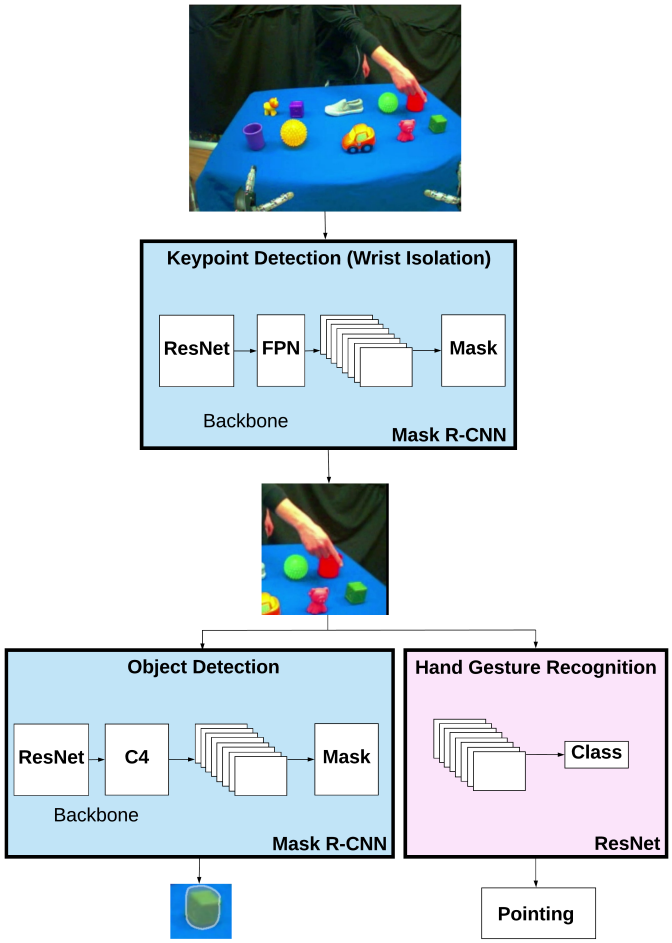


Fig. 3: An overview of the gesture-object detection architecture (a more detailed description can be found in Section III-B).

In this scenario, the Mask R-CNNs used for object detection and wrist key-point detection are both powered by Detectron for PyTorch [30], whilst the ResNet model for gesture recognition is also implemented in PyTorch. Training and inference were carried out using a machine equipped with an AMD Ryzen Threadripper 1950X 16 Core CPU @3.4GHz and a single NVIDIA TITAN Xp Graphical Processing Unit (GPU).

C. Implementation on the iCub

The architecture presented in the previous section was designed with respect to the realistic scenario where a humanoid robot is taught new vocabulary aided by deictic gestures. Naturally, this can only be realised by adding speech understanding and motor control functions to the gesture-object detector. This gesture-language integration is achieved as a result of implementing the system illustrated in Fig. 4. The computer running the trained model communicates with the robot's main control unit through the open source software package of YARP. In this manner, the online inference of the mask is done on this computer and the results are fed to the motor control module of the robot. The human-robot interaction experiments are carried out using the iCub robot, which is an open source platform designed for developmental robotics research. Its frame is based on a child-like morphology and it comprises 53 degrees of freedom [31].

The system runs in two modes of operation: *LEARN* and *RECALL*. The goal of the *LEARN* scenario is for the human to teach a new object to the robot. This can be realised by having the human vocally uttering the name of the object whilst performing a deictic gesture relative to it. Here, the function of the vision system is to isolate the mask of the object and simultaneously classify the gesture correctly. Additionally, the language module is responsible for transcribing the vocal utterance into its textual representation. As a result, the system stores the object, labelled by the vocal utterance and the label of the recognised gesture. On the other hand, in the *RECALL* stage, the human will ask the robot to show the learned object, and the objective of this framework is to understand whether that object has been previously learned and if this is the case, it will look at the correct object, signifying the comprehension task has been fulfilled. In this scenario, the language system once again converts the vocal utterance into its textual format, so that the latter can be used to search the stored database of learned words. In case this becomes true, it signifies that the robot has already been taught that word. Hence, the centre of the mask will be computed and fed to the motor control system so that the robot will look at the object of interest.

IV. Experimental Evaluation

The proposed model is validated in two scenarios. In Section IV-A (Experiment I), the performance of the isolated architecture illustrated in Fig. 3 is evaluated with respect to a subset of the dataset used (test dataset), whereas in Section IV-B (Experiment II) the behaviour of this architecture is studied in a HRI setting. The former experiment was carried out to find the parameters for optimal operation. On the other hand, in the latter, experiments were carried out to substantiate correct performance of the two modalities of operation for teaching a real humanoid robot new words through the use of deictic gestures. With that, our goal is to outline how the proposed system behaves in a language acquisition task.

A. Experiment I

The aim of this experiment is to provide insight into the performance of each part of the model. Since the three networks work in unison to fulfil the task of a gesture-object detector, obtaining the most favourable performance from each part of the network is vital to ensure overall correct functionality.

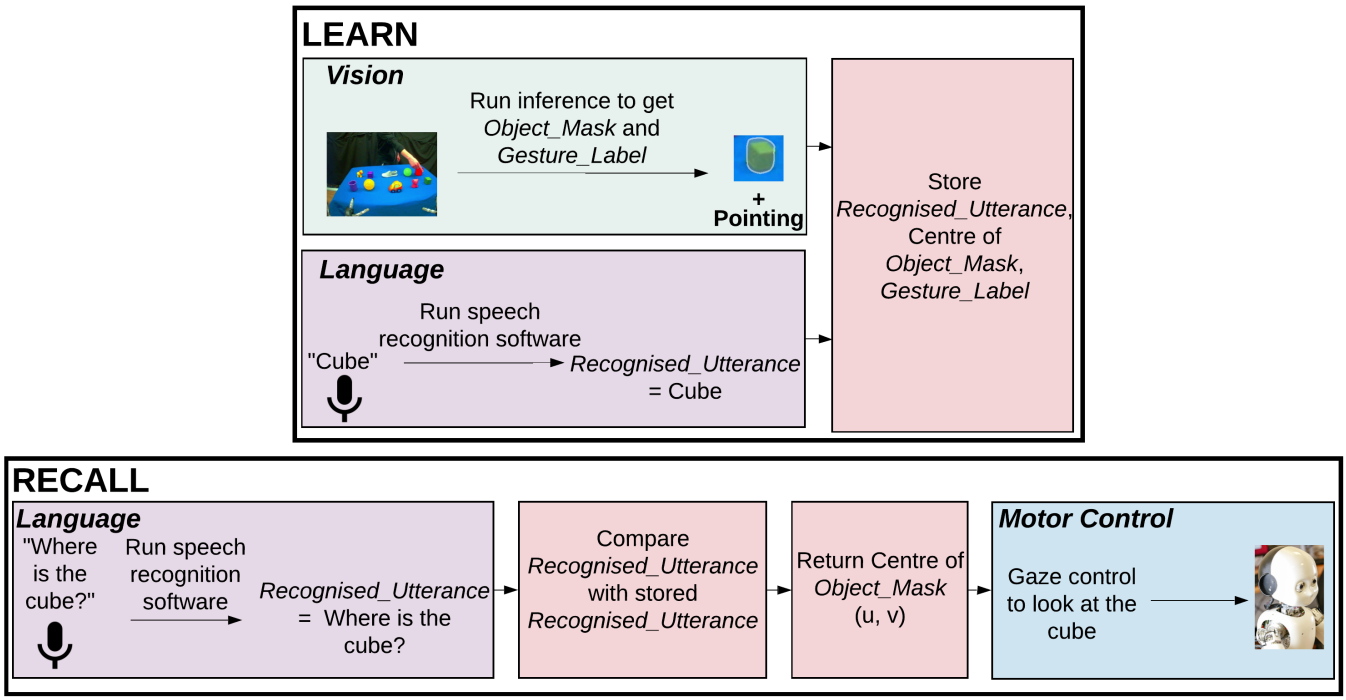


Fig. 4: The overall block diagram for deploying the gesture-language system on the iCub.

Here, we report the results for the keypoint detection network, for gesture classification and for object detection. The effectiveness and robustness of our approach is assessed on a subset of our gesture-object dataset, which has a similar setting to that used for training.

1) Keypoint Detection Network Evaluation: The pre-trained keypoint detection system was evaluated on 300 frames for each of the deictic gestures. Three different backbones, ResNet-50, ResNet-101 and ResNet-X101 were chosen on which the tests were run. The results obtained are illustrated in Table I. In these results, it can be clearly seen that when the participant is performing the deictic gesture of pointing, the wrist is easier localised than when the same participant performs the requesting deictic gesture. This is mainly a result of the wrist being more visible in the pointing scenario. Moreover, most misclassifications occurred due to the other wrist being present in the frame or the participants carrying out the gesture with the other hand as they were free to use whichever hand felt more natural. Overall the results obtained show that the method chosen for wrist detection gives positive results. Compared to keypoint detection results [21], the accuracy obtained here is quite high; however, we need to take into account that in all the images the wrist is highly exposed and we are recording the performance for only one joint. In comparison, in other works using keypoint detection, this method tends to be used for tracking several body joints and hence the overall performance tends to decrease.

TABLE I: The testing accuracy (%) for each model backbone of the end-to-end keypoint-only Mask R-CNN for different deictic gestures reported in Section IV-A1.

Keypoint Wrist Detection	Class	Accuracy
ResNet50-FPN	Pointing	97%
	Requesting	95.3%
ResNet101-FPN	Pointing	98%
	Requesting	92.67%
ResNetX101-FPN	Pointing	98.3%
	Requesting	94.3%

2) Gesture Recognition Network Evaluation: The same testing procedure as for the keypoint detection module was adopted to evaluate the gesture recognition system. Here, we tested using only the 50-layer ResNet. For pointing, we obtained a testing accuracy of 74%, whilst 66% for the deictic gesture of requesting. The results obtained are good considering how similar the two classes are and comparable with the results Tsironi obtained using CNNs [25]. In future work, it is envisaged that the other two deictic gestures of showing and giving will also be included.

3) Object Detection Network Evaluation: For the object detection system, three classes among the ten available were chosen. To this end, the two backbones of R101-FPN and R50-C4 were considered. The results obtained are illustrated in Table II. With regards to the results obtained from the object detector, it can be observed that the networks behave similarly with the two classes, and slightly worse with the other class.

This could be mainly due to the fact that since the first few layers were pre-trained on COCO, two of the classes were more similar to the ones found in this dataset. In addition, comparing this to current state-of-the-art using the Mask R-CNN network, our model is not able to generalise as well as these. Nonetheless, with the quality of the images and the size of our dataset, the performance is as expected.

TABLE II: The testing accuracy (%) for each model backbone of the end-to-end Mask R-CNN for some of the object classes reported in Section IV-A3.

Object Detection Network	Class	Accuracy
Object Detection (R50-C4)	Cup	82%
	Car	54%
	Ball	86%
Object Detection (R101-FPN)	Cup	86%
	Car	56%
	Ball	88%

As the system is deployed on a real robot and it is envisioned that on-the-fly training will be achieved in future work, results regarding the training times were collected and are outlined in Table III. Taking into account the training times illustrated in Table III, the final model chosen was the R50-C4.

TABLE III: The average training time for the Mask R-CNN used for object detection. As previously mentioned, the models were trained on a single NVIDIA TITAN Xp GPU.

		Maximum Iterations	
Network	2000	5000	10,000
R50-C4	$\approx$ 2min	$\approx$ 5min	$\approx$ 10min
R50-FPN	$\approx$ 5min	$\approx$ 10min	$\approx$ 20min
R101-FPN	$\approx$ 7min	$\approx$ 16min	$\approx$ 35min

Indeed, in all scenarios, when taking into consideration the quality of the images captured directly by the iCub's cameras and the size of the dataset, the performance achieved is good and shows promising results for deploying the model on the iCub. This result is a vital first step towards attaining a system that can perform well in a HRI environment, as it illustrates that in a constrained task it succeeds in fulfilling its main goals.

B. Experiment II

The aim of this experiment is to illustrate the implementation of the gesture-object detection system on the iCub. Our focus is on having the iCub understand complementary gesture-word combinations. Hence, this experiment was carried out to analyse how the iCub performs in such an example.

In the evaluation process of Experiment I, training and testing took place in analogous settings; here, we observe how the complete architecture performs in a more realistic HRI setup. This is a critical aspect of every model that will be implemented on a real robotic platform, as we would like our system to generalise to the real world. For this reason, we carried out this experiment to determine to what extent the learning system generalises here.

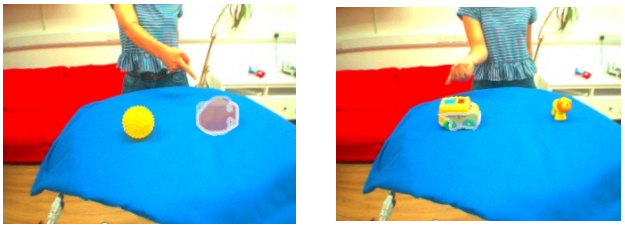


Fig. 5: Illustration of the performance of the system during the HRI experiment. Left image depicts an object taken directly from the training database, whilst the object in the right frame is a similar car, but not the exact same one used in the training database. With the help of the deictic gesture, the system is able to obtain the mask for the object of interest.

As previously mentioned, the trained model communicated with the robot's main control unit through YARP. After mask inference is done on the main computer, the results are sent to the motor control module of the robot in order to have the iCub look at the correct object.

During this experiment, the image acquisition setup is the same as that used to capture the frames for the training dataset, i.e. we capture the images using the RGB camera on the iCub. In this case, the objects used are similar objects to those in the training dataset. A subject stands in front of the table with the objects in front of the robot and performs a deictic gesture with respect to an object, whilst simultaneously uttering the name of the object. An example of this would be that the participant points at a cup and says *"cup"*. Here, the architecture is operating in the *LEARN* scenario. In turn, this action is followed up by providing an instruction to the robot such as *"show me the cup"* and the iCub looks at the cup to illustrate correct comprehension and consequently result in fulfillment of the task. In such a case, the robot is operating in the *RECALL* mode.

Fig. 5 illustrates how the system operates in the HRI scenario. The participant is carrying out deictic gestures in front of familiar objects and unfamiliar, but similar, objects. In the left figure, the mask obtained for the object has a higher confidence score than the mask obtained in the right hand figure, as can be seen by the outline of the mask. This is a result of the car in the left frame being taken directly from the training database, whilst the car in the right frame is a similar car, but not the exact same one used in the training database. Hence, the system was able to generalise well. Fig. 5 shows that the performance in this testing case remains adequate; therefore, the networks performed well in this scenario.

V. Discussion and Conclusion

This paper addresses the challenge of using deictic gestures for equipping humanoid robots with language skills. A successful completion of this task is important for natural language acquisition in robotics as research in infants' developmental stages has shown that gestures play a pivotal role in communication. State-of-the art deep learning-based solutions contribute to achieving this integration as they have shown

groundbreaking performance in both vision and language domains. Here, we proposed a pipeline that uses the method of Mask R-CNNs to detect the wrist keypoint before gesture recognition, together with a Mask R-CNN for object detection. Hence, we showed how together with speech recognition and motor control modules, the aforementioned architecture can fulfill its task of gesture-word comprehension in a HRI scenario. Specifically, we showed how the iCub humanoid robot learns new words with the help of deictic gestures.

Other interesting research directions for this work include focusing more on gestures' role in predicting sentence complexity. In addition, we would also like to look in more depth at how the system can handle occlusions and how deictic gestures can be used to help in such an occasion. Nonetheless, we believe the addressed problem is an initial step towards realising robot platforms that improves its language mechanism by learning from a teacher. In this perspective, the presented contribution is a stride towards robots that can learn language in a similar way to humans.

VI. Acknowledgment

The work is supported through the DCOMM project by the *Horizon 2020 Marie Skodowska-Curie Innovative Training Networks* Grant No. 676063. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU that was used for the initial development stages of this research. We also thank Prof. Gavin Brown and Dr. Daniel Hernández García for supporting this project through constructive comments and advice on the production of the manuscript.

References

[1] J. M. Iverson and S. Goldin-Meadow, "Gesture paves the way for language development," *Psychological Science*, 2005, pp. 367-371.

[2] O. Capirci, J. M. Iverson, E. Pizzuto and V. Volterra, "Gestures and words during the transition to two-word speech," *Journal of Child Language*, 1996, pp. 645-673.

[3] A. Cangelosi et al., "Integration of action and language knowledge: a roadmap for developmental robotics," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 3, pp. 167-195, Sept. 2010.

[4] C. Lyon, C. L. Nehaniv and J. Saunders, "Interactive language learning by robots: the transition from babbling to word forms," *PLOS ONE*, 2015.

[5] P. F. Dominey and J. Boucher, "Developmental stages of perception and language acquisition in a perceptually grounded robot," *Cognitive Systems Research*, 2005, pp. 243-259.

[6] S. Cho, W. H. Lee and J. H. Kim, "Implementation of human-robot VQA interaction system with dynamic memory networks," *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Banff, AB, 2017, pp. 495-500.

[7] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," *Nature*, 2015, pp. 436-444.

[8] M. Capobianco, E. Antinoro Pizzuto and A. Devescovi, "Gesture-speech combinations and early verbal abilities. New longitudinal data during the second year of age," *Interaction Studies*, 2017, 18(1), pp. 5576.

[9] M. Fasolo and L. DOdorico, "Gesture-plus-word combinations, transition forms, and language development Gesture," *Gesture*, 2012, 12 (1), pp. 1-15.

[10] N. Esteve-Gibert and P. Prieto, "Infants temporally coordinate gesture-speech combinations before they produce their first words," *Speech Communication*, 2013, pp. 301-316.

[11] E. Cartmill, D. Hunsicker and S. Goldin-Meadow, "Pointing and naming are not redundant: children use gesture to modify nouns before they modify nouns in speech," *Developmental Psychology*, 2014.

[12] S. Özçalişkan and S. Goldin-Meadow, "When gesture-speech combinations do and do not index linguistic change," *Language and Cognitive Processes*, 2009, 24(2), pp. 190-217.

[13] S. O'Keeffe and R. Villing, "Evaluating pruned object detection networks for real-time robot vision," *IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, 2018, pp. 91-96.

[14] Y. Zhang, H. Wang and F. Xu, "Object detection and recognition of intelligent service robot based on deep learning," *IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM)*, 2017, pp. 171-176.

[15] E. Maiettini, G. Pasquale, L. Rosasco and L. Natale, "Interactive data collection for deep learning object detectors on humanoid robots," *IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, 2017, pp. 862-868.

[16] J. Leitner, A. Förster and J. Schmidhuber, "Improving robot vision models for object detection through interaction," *International Joint Conference on Neural Networks (IJCNN)*, 2014, pp. 3355-3362.

[17] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 580-587.

[18] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580-587.

[19] Ross Girshick, "Fast R-CNN," *Proceedings of the International Conference on Computer Vision (ICCV)*, 2015.

[20] Shaoqing Ren, Kaiming He, Ross Girshick and Jian Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *Neural Information Processing Systems (NIPS)*, 2015.

[21] K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," *IEEE International Conference on Computer Vision (ICCV)*, Venice, 2017, pp. 2980-2988.

[22] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Harihar and S. J. Belongi, "Feature pyramid networks for object detection," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[23] A. Camurri and G. Volpe, "Gesture-based communication in human-computer interaction," *5th International Gesture Workshop*, 2004.

[24] M. Asadi-Aghbolaghi et al., "A survey on deep learning based approaches for action and gesture recognition in image sequences," *12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 2017, pp. 476-483, doi: 10.1109/FG.2017.150.

[25] E. Tsironi, P. Barros, C. Weber and Stefan Wermter, "An analysis of convolutional long short-term memory recurrent neural networks for gesture recognition," *Neurocomputing*, 2017, pp. 76-86.

[26] M. A. Goodrich and A. C. Schultz, "Human-robot interaction: a survey, foundations and trends in human-computer interaction," *Foundations and Trends in Human-Computer Interaction*, 2004, pp. 23-275.

[27] Deep Magic, "Deep-Magic/COCO-Style-Dataset-Generator-GUI," 2018, [online] Available at: https://github.com/Deep-Magic/COCO-Style-Dataset-Generator-GUI [Accessed 11 Dec. 2018].

[28] G. Metta, P. Fitzpatrick and L. Natale, "YARP: yet another robot platform," *International Journal of Advanced Robotic Systems*, 2006.

[29] T. Lin et al., "Microsoft COCO: Common objects in context," *ECCV*, 2014, pp. 740-755.

[30] R. Girshick, I. Radosavovic, G. Gkioxari, P. Dollar and K. He, "Detectron," 2018, *https://github.com/facebookresearch/detectron*.

[31] G. Metta, L. Natale, F. Nori, G. Sandini, D. Vernon, L. Fadiga, C. von Hofsten, K. Rosander, M. Lopes, J. Santos-Victor, A. Bernardino and L. Montesano, "The iCub humanoid robot: an open-systems platform for research in cognitive development," *Neural Networks*, vol. 23, pp. 1125-1134, 2010.