Natural Language Processing for Object Affordance: A Review

Bernard Renardi
Faculty of Science and Engineering
Rijksuniversiteit Groningen
Groningen, The Netherlands
b.jo@student.rug.nl

Duygu Bayram
Faculty of Arts
Rijksuniversiteit Groningen
Groningen, The Netherlands
d.bayram.1@student.rug.nl

***Abstract*—Natural language is an important element of Human-Robot Interaction and improvements in this area can be helpful as the involvement of humans may facilitate learning for the robot. Mi et al. (2020) [6] present an architecture that aims to extract intention from natural language sentences and map the extracted semantic knowledge onto detected object affordances to improve learning. To this end, the authors implement and develop methods for object affordance detection, intention semantic extraction, and target object grounding. As a novel contribution, they extract texture features for the object affordance detection task and build an Attention-Based Multi-Visual Feature Fusion network to fuse texture and visual features.**

***Index Terms*—natural language grounding, object affordance, object affordance detection, intention semantic extraction**

I. Introduction

Vision plays a significant role in human lives and people often rely on object affordance to pragmatically interact with their environment. Object affordance, in this case, refers to the function of an object. As such, an example would be that a "*drink*" or a "*glass*" would have the affordance, or the function, of "*drinking*". By relying on feature information like texture, shape, size, color, and so on, humans are able to infer the affordance of new objects.

In addition to vision, natural language can be an important guidance in this task when it involves other participants. The sentence *"Put the plate on the table."* contains important information regarding objects and their affordance, and other pragmatic sentences such as "*It is too warm, I want to get some air.*" may signal an intention of asking someone to open the windows. This kind of information can be helpful as well as crucial in Human-Robot Interaction.

Mi et al. (2020) [6] present a model in which they detect object affordance, and implement natural language processing and grounding to extract semantic intention from instructions before mapping this semantic knowledge onto target objects to allow the robot to infer affordance. This task can improve Human-Robot Interaction and exploit the human potential of helping robots learn new object affordances.

II. Methodology

The architecture is constructed from three sub-tasks:

1) Object affordance detection from images
2) Intention semantic word extraction
3) Mapping the extracted intention onto detected object affordances

A. Object Affordance Detection

The authors limit their target affordances to 10 categories (*calling, drinking(I)* for drinking appliances such as a "*glass*", *drinking (II)* for drinks, *eating(I)* for utensils, *eating(II)* for foods, *playing, reading, writing, cleaning*, and *cooking*). For this sub-task, they build a model that consists of parts for detection, multi-visual feature extraction (in other words, features consisting of various visual information such as texture and shape), a fusion module with attention, and a Multi-Layer Perceptron (MLP).

After reviewing previous work, they decide to include texture features in addition to visual features, and the extraction task is separated into visual features and textures features. For visual features RetinaNet is used to produce Regions of Interest (RoIs) and the features are processed through L2 normalization. For texture feature extraction, they choose to implement a deep learning model which consists of encoding and convolutional layers [2]. An advantage of this system is its allowance for the transference of object recognition CNNs for texture recognition. The authors use this model to encode texture features on top of the RoIs that were extracted from the previous step, and perform L2 normalization for processing again. Futhermore, they perform a hyperbolic tangent operation on the resulting feature vector as it produces better results, and expand the vector via the tile operation for the following fusion step.

The extracted features are dynamically fused by the use of Factorization Machines (FM) [3]. FMs are developed for large and sparse data and model the interactions between different features. As a result, they are appropriate to use for this paper's authors.

To prevent redundant information, the authors also implement an attention model for the fusion task, which assigns weights according to which feature interactions are more relevant for learning.These attention weights are then exploited by an MLP with a softmax layer to learn object affordances.

B. Intention Extraction

For natural language processing, the authors approach the task with the idea that words have different roles in meaning expression, hence they want to assign different weights to different words in a sentence.

After tokenization, they employ GloVe [4] for word embeddings, obtaining the probabilities on co-occuring words, resulting in a 300-D vector for each word representation. Then, they concatenate the word representation vectors to form a sentence.

They use the resulting sentence representation vector as input for a pre-trained BiLSTM model with an attention mechanism, and derive the weights for the individual words through this model. As a final step, they adjust the word order according to the weights, the word with the highest weight being used to represent the intention, and the verb in the sentence being forwarded to the grounding task to for affordance mapping.

C. Object Grounding

The final task of the study is the mapping of the intention semantic words and the detected affordances of the objects. They propose a method that transforms the intention semantic word and the affordance vectors into the 300-D vectors of GloVe, then they find the semantic similarity between them. If the semantic similarity between the object and the intention semantic word is high, then that object is chosen as the target object. The word is then mapped onto the related affordance.

III. Results

A. Object Affordance Detection

1) Dataset: As the available datasets are insufficient for the affordance set used in the study, the authors decide to develop a new, more appropriate dataset for their purposes. The new dataset consists of indoor MSCOCO [7] and ImageNet [11] scenes, as well as new images obtained by a Kinect V2 sensor. Making up these, there are 12,349 RGB images available with 14,695 corresponding annotations tailored for object affordance detection tasks. Approximately 23% of these annotations are taken from MSCOCO and ImageNet.

The authors randomly select 56.1%, 22.1% and 21.8% regions for the training, validation, and test sets respectively. The distribution of the affordances can be seen in Figure 1. It is important to note that the "*writing*" and "*cleaning*" objects were largely obtained by the Kinect sensor as the available datasets had limited examples of these classes.

2) Results: As can be seen in Figure 2, the "*writing*", "*cleaning*", and "*cooking*" labels have lower accuracy compared to the others. The authors interpret the issue to be related to the shape and texture differences between the objects in these categories, and as such the model demonstrated a lower achievement on detecting similarities and generalizing.

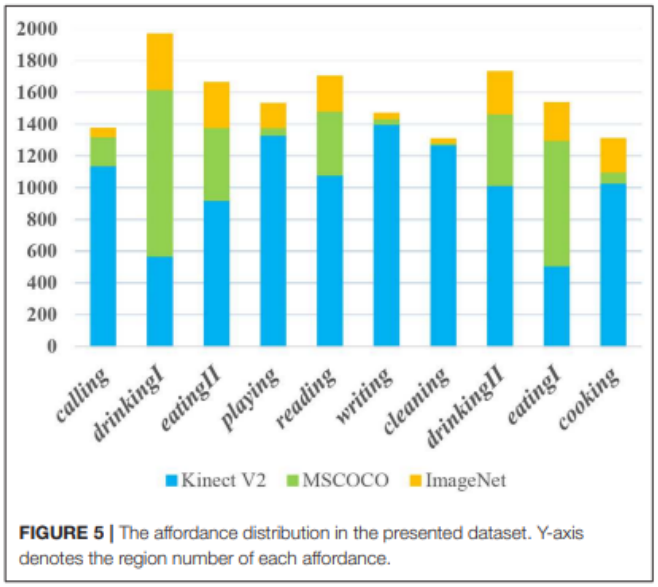

FIGURE 5 | The affordance distribution in the presented dataset. Y-axis denotes the region number of each affordance.

Fig. 1. Distribution of affordances as taken from Mi et al. (2020)

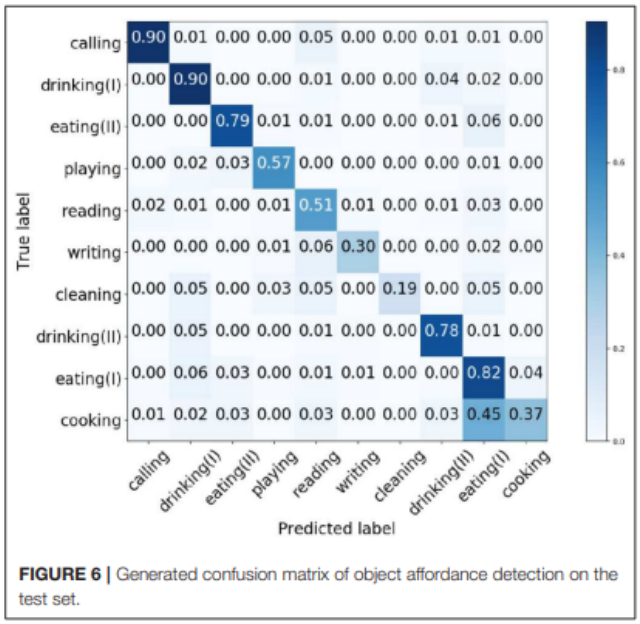

FIGURE 6 | Generated confusion matrix of object affordance detection on the test set.

Fig. 2. Confusion matrix of results as taken from Mi et al. (2020)

3) Ablation: The authors also compare several networks and feature fusion methods to gain insight into which approaches may be preferable. They compare three fusion approaches (Attention multi-visual features fusion, as described, VGG19 deep features, and Naive concatenation), and two networks instead of the attention-based architecture they have presented, RetinaNet [1] and YOLO V3 [8]. The results of these can be viewed in Figure 3, as taken from [6].

In terms of feature fusion approaches, the method described in the initial architecture performs best on most affordance

TABLE 1 | Object affordance detection results acquired by different networks, deep features and feature fusion method.

	Attention multi-visual features fusion	VGG deep features	Naive concatenation	RetinaNet	YOLO V3
calling	0.9036	**0.9096**	0.8723	0.7747	0.5783
drinkingI	**0.8991**	0.7785	0.8195	0.7806	0.4771
eatingII	**0.7943**	0.7658	0.7569	0.6829	0.5696
playing	0.5676	0.4791	0.5305	**0.8305**	0.7871
reading	0.5148	0.4938	0.5297	**0.6424**	0.652
writing	**0.2995**	0.2028	0.286	0.2628	0.2028
cleaning	0.1875	0.1625	0.175	**0.375**	0.3327
drinkingII	**0.7838**	0.7627	0.7248	0.6128	0.5824
eatingI	**0.8162**	0.7103	0.7049	0.6738	0.4837
cooking	0.3719	0.2893	**0.4214**	0.2562	0.2968
Average	**0.6138**	0.5554	0.5821	0.5892	0.4963

The bold value of each row is the acquired best accuracy of each affordance.

Fig. 3. Table of results as taken from Mi et al. (2020)

categories and obtains the best score on average, with "*calling*" performing better on VGG19 deep features, and "*cooking*" performing better with a Naive concatenation method. All are fed into the MLP which is run for 100 epochs.

As for the different networks, RetinaNet reaches an average score of 58.92% after 100 epochs, performing better on the "*playing*", "*reading*", and "*cleaning*" categories. YOLO V3, on the other hand, reaches an average score of 49.63% after 100 epochs, with no improvement on any categories.

Out of the tested feature fusion approaches, the presented attention-based multi-visual features fusion network obtains the best results. Similarly, compared to RetinaNet and YOLO V3, the presented attention-based network performs better. As a result, the use of attention-based multi-visual feature fusion network proves to be preferable in similar object affordance detection tasks.

B. Natural Language Grounding

To test the performance of the grounding model, the authors choose 100 images from the test set and ask 10 human participants to provide one or two sentences for each item. As a result, they obtain 150 natural language instructions. They then use these as part of the data to test the intention semantic word extraction model, which achieves a 90.67% accuracy score.

Moreover, they forward the extracted semantic words to the grounding model to perform a simple error analysis, and report that the model performance is considerably affected by the affordance detection performance.

C. In Robotics

As a final demonstration, the authors run the architecture on a UR5 robotic arm. They employ the Kaldi speech recognition tool [9] to convert spoken language instructions to text, and then use their own grounding model for processing. Once they obtain the target objects, they combine this data with the depth data from the Kinect V2 camera to locate them. For the grasping task, they use their previously presented model [10].

IV. Conclusion

Human language can be helpful for facilitating learning for robots, and incorporating the use of it can produce favorable results for both learning and Human-Robot Interaction. In the presented paper, the authors build a model which extracts intention from natural language instructions and uses this information to learn object affordance. To perform this task, they build architectures for object affordance detection, intention semantic word extraction, and grounding. As a result, they find that an attention-based multi-visual feature extraction network performs best.

V. Author Contributions

Both authors contributed equally to this paper.

References

[1] Lin, T., Goyal, P., Girshick, R., He, K., and Dollár, P. (2020). "Focal loss for dense object detection". IEEE Trans. Pattern Anal. Mach. Intell. 42, 318–327.

[2] Zhang, H., Xue, J., and Dana, K. (2017). "Deep ten: texture encoding network," in Proceedings 1of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2896–2905.

[3] Rendle, S. (2010), "Factorization machines," in IEEE International Conference on Data Mining (ICDM), 995–1000.

[4] Pennington, J., Socher, R., and Manning, C, (2014), "Glove: global vectors for word representation," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Doha), 1532–1543.

[5] R. Kartmann, D. Liu, and T. Asfour, (2021) "Semantic Scene Manipulation Based on 3D Spatial Object Relations and Language Instructions," IEEE. Germany, vol. 20, 306-313.

[6] J. Mi, et.al., "Intention-Related Natural Language Grounding via Object Affordance Detection and Intention Semantic Extraction," Frontiers in Neurorobotics. China, vol. 14, 1-12.

[7] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014), "Microsoft coco: common objects in context," in European Conference on Computer Vision (ECCV), 740–755.

[8] Redmon, J., and Farhadi, A. (2018), "Yolov3: an incremental improvement".

[9] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., et al. (2011), "The kaldi speech recognition toolkit," in IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.

[10] Liang, H., Ma, X., Li, S., Görner, M., Tang, S., Fang, B., et al. (2019), "Pointnetgpd:1 detecting grasp configurations from point sets," in International Conference on Robotics and Automation (ICRA), 3629–3635.

[11] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015), "Imagenet large scale visual recognition challenge," International Journal of Computer Vol. 115, 211–252.