electronics

MDPI

Article

Perceptual and Semantic Processing in Cognitive Robots

Syed Tanweer Shah Bukhari and Wajahat Mahmood Qazi *

Intelligent Machines & Robotics, Department of Computer Science, COMSATS University Islamabad, Lahore 54000, Pakistan; stsbukhari@gmail.com

* Correspondence: wmqazi@cuilahore.edu.pk

Abstract: The challenge in human–robot interaction is to build an agent that can act upon human implicit statements, where the agent is instructed to execute tasks without explicit utterance. Understanding *what to do* under such scenarios requires the agent to have the capability to process object grounding and affordance learning from acquired knowledge. Affordance has been the driving force for agents to construct relationships between objects, their effects, and actions, whereas grounding is effective in the understanding of spatial maps of objects present in the environment. The main contribution of this paper is to propose a methodology for the extension of object affordance and grounding, the Bloom-based cognitive cycle, and the formulation of perceptual semantics for the context-based human–robot interaction. In this study, we implemented YOLOv3 to formulate visual perception and LSTM to identify the level of the cognitive cycle, as cognitive processes synchronized in the cognitive cycle. In addition, we used semantic networks and conceptual graphs as a method to represent knowledge in various dimensions related to the cognitive cycle. The visual perception showed average precision of 0.78, an average recall of 0.87, and an average F1 score of 0.80, indicating an improvement in the generation of semantic networks and conceptual graphs. The similarity index used for the lingual and visual association showed promising results and improves the overall experience of human–robot interaction.

Keywords: cognitive robots; semantic memory; affordance; object grounding

Citation: Bukhari, S.T.S.; Qazi, W.M. Perceptual and Semantic Processing in Cognitive Robots. *Electronics* **2021**, *10*, 2216. https://doi.org/10.3390/electronics10182216

Academic Editors: Janos Botzheim and Maysam Abbod

Received: 3 August 2021
Accepted: 7 September 2021
Published: 10 September 2021

1. Introduction

This paper proposes an affordance- and grounding-based approach for the formation of perceptual semantics in robots for human–robot interaction (HRI). Perceptual semantics play a vital role in ensuring a robot understands its environment and the implication of its actions [1,2]. The challenge is to build robots with the ability to process their acquired knowledge into perception and affordance [1–6]. In this context, the significance of affordance can be rationalized by the following scenario taken from human–human interaction (HHI): if we state the following information to another human "X" that "*I am feeling thirsty*" rather than "*I want to drink beverage using a red bottle*", the human "X" will be able to understand the relationship between "*drink*" and "*thirst*". The link between these two terms is "*thirst causes the desire to drink*". This ability to establish a relationship between "*drink*" and "*thirst*" based on semantic analysis is called affordance. Consequently, the human "X" will offer something to drink. Let us assume we have a robot with the ability to process affordance and a similar situation is present in human–robot interaction; then, it is expected that the robot may perform a similar action as the human "X" in HHI. For robots, this type of interaction is currently a challenge, although there are various contributions in this direction [5,6] with the focus on visual affordance. In the scenario presented above, visual affordance is not sufficient for understanding of the relationship between "*thirst*", and "*drink*". The robot also needs to *ground the objects* placed on the table. Object grounding is an approach that allows the robot to profile objects in the environment [6] i.e., "*How many objects belong to a drinkable category?*" will be answered with the response having the position of the object as "*There is one drinkable object located at the left*

side of the table". This challenge becomes even more complex when it is implemented for cognitive robots, because their design rationale considers factors such as internal regularization, control, and synchronization of autonomous processes through a cognitive cycle (understanding, attending, and acting) [7–11]. A reference cognitive cycle may consist of a variant of the phases of perception, understanding, and action [7,11]. In this study, we adopted an extended version [12] of Bloom's taxonomy as a cognitive cycle. The reason for using the Bloom-based cognitive cycle is that it provides a map between the level of cognitive processes and the type of knowledge domain. The detailed Bloom taxonomy map is provided in a previous paper [12]. In addition, Bloom uses the action verbs to steer the cognitive-knowledge-domain map [12]. The control structure used in this study is shown in Figure 1, which is an extract from our previously reported work [13]. The detailed utilization of the control structure in Figure 1 is discussed in Section 4.

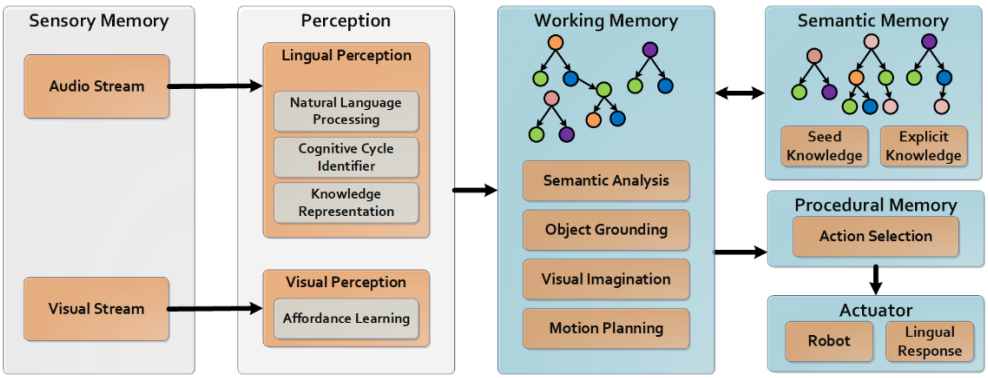


Figure 1. NiHA's minimal architecture for a cognitive robot [13].

In this study, we proposed perceptual semantics based on extended-object grounding and machine perception (visual and lingual). In this regard, we performed a table-top experiment using a Universal Robot (UR5). A dataset for affordance learning comprising 7622 images (Section 4.1) was prepared for the training of a YOLOv3-based perception module (Section 4.1). A Bloom-based cognitive cycle identifier was also implemented for the identification of cognitive levels (see Section 4.2). The semantic memory was constructed from ConceptNet and WordNet having 1.47 million nodes and 3.13 million relationships (see Section 4.3). Our analysis of the experimental data/results (see Section 5) suggests that perceptual learning alone is not sufficient to access the environment; the inclusion of seed knowledge is important to understand the extended affordance features (i.e., the relationship between "*drink*" and "*thirst*"). Moreover, the inclusion of a cognitive cycle identifier helps the robot to choose between "*what to reply*", "*what not to reply*", "*when to reply*", and "*what would be the procedure*". The work reported in this paper is an effort to contribute to the advancement of building robots with a better understanding of the environment.

2. Related Work

There is a growing need for robots and other intelligent agents to have safe interactions with partners, mainly human beings. In this regard, the need for perceptual semantics formulated using affordance learning and object grounding is vital for human–robot interaction (HRI) [14–16].

Affordance is considered to be the catalyst in establishing a relationship between accessible objects, their effects, and actions carried out by robots [17,18]. Affordance capability can be induced in an agent through interaction, demonstration, annotation, heuristics, and trails [19]. Most of the work undertaken in object affordance learning is based on visual perception [3,20–26], whereas lingual cues can also provide additional advantages that can significantly improve affordance [19]. Therefore, in this study, we

focused on both visual and lingual cues. In addition to visual and lingual cues, Breux et al. [16] considered ontologies based on WordNet to extract the action cues and ground the relationships between objects and features (properties). This improved the results and HRI but covered only seven types of relationships (*isA*, *hasA*, *prop*, *usedFor*, *on*, *linked-to*, and *homonym*), which limits the agent's recognition and understanding capabilities to the stated semantic associations. Implementation of semantic memory has also been reported in the literature [3,27,28]. Antunes et al. [27] reported the use of semantic memory for HRI and discussed the scenario of "make a sandwich" having explicit information objects and their actions. This system [27] does not have the capabilities to cater to situations such as "*I am feeling hungry*", in which the robot understands that there is a need to make the sandwich. This suggests that semantic memory is more like a knowledge repository.

Object grounding based on either lingual or visual perception is used to profile the object in the environment [6,29]. The grounding is mapped in terms of exact and relative location(s) of the object(s), i.e., "*left-of*" [16]. Oliveira et al. [3,4] discussed the importance of semantic memory for HRI and interaction-based acquisition of knowledge. The mentioned system uses object grounding without incorporating object affordance; therefore, it is unable to process the feature-based lingual cues. The object grounding techniques used in this paper are similar to that introduced in the Robustness by Autonomous Competence Enhancement (RACE) project [1]. The table-top setup is represented as a grid having nine positions, i.e., $center$, $right$, $left$, top, $bottom$, $bottom-left$, $bottom-right$, $top-right$, $top-left$ (see Section 4.4).

An agent by design has a control structure that can be as simple as a sense act [30] or as complex as a cognitive architecture [8,13,31]. These control structures may be a collection of memories, learnings, and other mental faculties depending on the architectural complexity [8,13]. The systems with semantic memories are those that fall within the domain of cognitive architecture [8,13,31]. These processes in these control structures are regularized, controlled, and synchronized through a cognitive cycle. A reference cognitive cycle consists of perception, understanding, and action [7–10]. A limitation of these cognitive cycles is that they do not provide explicit guidelines to map the degree of processing on various knowledge levels and dimensions [7–10]. This is a challenge in the implementation of cognitive agents in the selection of appropriate cognitive processes and knowledge dimensions from the lingual cues. Bloom's taxonomy provides the method to map lingual cues with the cognitive levels and knowledge dimensions [12], but it has not yet been used as a cognitive cycle.

The analysis of the current work suggests that significant improvement in the state-of-the-art can be made by increasing the number of semantic relationships, combining both supervised and heuristic approaches for acquiring affordance and formulation of object grounding. We proposed a semantic memory consisting of 53 types of relationships having 1.47 million nodes and 3.13 million relationships to enhance the benefits of affordance learning (see Section 4.3).

The control structures and design of existing systems build a strong case for the inclusion of architecture having semantic memory, perception, and other required modalities [3,16,21,22,27,28,32–37]. We used the minimalistic design of the previously reported Nature-inspired Humanoid Cognitive Architecture for Self-awareness and Consciousness (NiHA) (see Figure 1) [13].

We incorporated Bloom's taxonomy in a standard cognitive cycle to identify the cognitive process and knowledge dimensionality based on the identification of action verbs from lingual cues (see Section 4.2). The detailed comparison of the state-of-the-art with the proposed method is symbolized in Table 1.

Table 1. State-of-the-art comparison with the proposed method.

Work	Platform	Task	Perception	Data Source	Control Structure	Grounding	Affordance Dataset	Knowledge Base/Ontology	Evaluation Metric/Method
[20]	N/A	Object Manipulation	Visual	Demonstration	No	No	6 Categories/330 Views	No	Accuracy
[3]	PR2	Object Manipulation	Visual	Interaction	RACE		10 Categories /339 Views	Semantic Memory	Accuracy
[33]	Toe	Object Manipulation	Visual	Labels	Robot Imagination System	Yes	Geometric Shapes	No	Token Test
[38]	Toe	Object Manipulation	Visual	Labels		Yes	10 Classes/30 Sign Symbols	No	Accuracy
[34]	Khepera III	Navigation	Visual	Labels	Multi-robot Control System	No	N/A	No	N/A
[39]	PR2	Action Prediction	Visuo-Spatial	Interaction		N/A	N/A	Graph	N/A
[35]	Atlas	Manipulation /Navigation	Visual	Labels	Yes	No	8 Classes	N/A	N/A
[27]	iCub	Object Manipulation	Visual	Heuristic	Yes	Yes	N/A	Semantic Memory	N/A
[21]	Bioloid	Object Manipulation	Visual	Labels	C5M	No	5 Classes/4 Affordance Classes	N/A	Accuracy
[22]	NAO	Action Prediction/Navigation	Visual	Labels/Trail & Error	Yes	No	8 Action Classes	N/A	Accuracy
[18]	iCub	Object Manipulation	Visual	Heuristic	No	Yes	N/A	N/A	Accuracy
[23]	N/A	Action Prediction	Visual	Heuristic/Labels	No	No	9 Classes/10 Object Categories	N/A	Weighted F Measure
[24]	N/A	Object Manipulation	Visual	Labels	No	No	7 Classes/105 Objects	N/A	Recognition Accuracy
[36]	Fanuc	Object Manipulation	Visual	Labels	Yes	No	13 Classes	N/A	Accuracy
[25]	N/A	Action Prediction	Visual	Labels	No	No	13 Classes/6 Action Affordances	N/A	Accuracy
[37]	Valkyrie	Object Manipulation	Visual	Labels	Yes	No	N/A	N/A	N/A
[28]	PR2	Object Manipulation	Visual	Labels	Yes	No	N/A	KnowRob	N/A
[26]	Walk man	Action Prediction	Visual	Labels	No	No	10 Object Classes/ 9 Affordance Classes	No	Weighted F Measure
[16]	Wheeled Robot	Object Manipulation	Visual/Auditory	Labels	Yes	Yes	N/A	Knowledge Graph	N/A
Ours	**UR5**	**Grasping/Object Manipulation**	**Visual/Auditory**	**Labels/Heuristic**	**NiHA**	**Yes**	**7 Affordance Classes**	**Semantic Memory**	**F1 Measure/Semantic Similarity**

3. Problem Formulation

This study proposed a method for human–robot interaction (HRI). For this purpose, semantic memory S_m for an agent from the atom of knowledge ($Atom$) is an essential first step. The atom of knowledge is generated from visual and auditory sensory stimuli. The human-centered environment consists of household items (objects) present on the table.

Let the items that exist in the workspace be presented as $I = \{i_1, i_2, i_3, \ldots, i_n\}$, and the properties of the item I_p be represented as $I_p = \{name, affordance, location, direction\}$. The affordance of the item is defined as $I_{paffordance} = \{callable, cleanable, drinkable, edible, playable, readable, writable\}$ and the location as $I_{plocation} : item \rightarrow \mathbb{R} \times \mathbb{R}$ gives the parameters concerning the visual frame. The direction of the items is presented according to the center of the visual frame as $I_{pdirection} : Item \times Item \rightarrow [center, right, left, top, bottom, bottom - left, bottom - right, top - right, top - left]$.

Let the auditory stream be based on m number of words as $W = \{N, A_{dj}, V\}$. The word W can be recognized as noun node $N = \{NN, NNS, NNP, NNPS\}$, adjective node $A_{dj} = \{JJ, JJR, JJS\}$ and the verb node can as $V = \{VB, VBD, VBP, VBN, VBG, VBZ\}$. The noun, adjective, and verb are checked with an a priori knowledge base of concepts and features. The concept is defined as $C = \{c_1, c_2, c_3 \ldots . c_k\}$, whereas feature $F = \{f_1, f_2, f_3 \ldots . f_k\}$. The atom of knowledge is represented as $Atom = \{\langle c, c \rangle, \langle c, f \rangle\}$. The semantic memory of the system is the collection of k atoms of knowledge $S_m = \{Atom_1, Atom_2, Atom_3 \ldots . Atom_k\}$. Let the cognitive cycle based on Bloom's revised taxonomy be selected as cognitive level $Cog_{level} = [Perception, Understanding - Comprehensions, Execution - Control, Post - Execution - Analysis, Evaluation, Synthesis]$. The knowledge dimension $Knowledge_{Dimension} = [Factual, Conceptual, Procedural, Meta-$

$Cognition$] can be selected based on action verbs proposed in revised Bloom's taxonomy. The two-dimensional array, i.e., matrix can be represented as $BCog_{matrix} = Cog_{level} \times Knowledge_{Dimension}$. The selected cognitive cycle is an instance of a matrix as $BCog_{cycle} = (Cog_{level_i}, Knowledge_{Dimension_i})$.

4. Methodology

This section explains the methodology for the development of artifacts highlighted in the problem formulation. These artifacts include perception (i.e., visual, and lingual), working memory (i.e., object grounding and semantic similarity analysis), and construction of semantic memory (seed knowledge and explicit knowledge). The lingual perception is further divided into knowledge representation, cognitive cycle identifier, and natural language processing module. The core architecture is depicted in Figure 2.

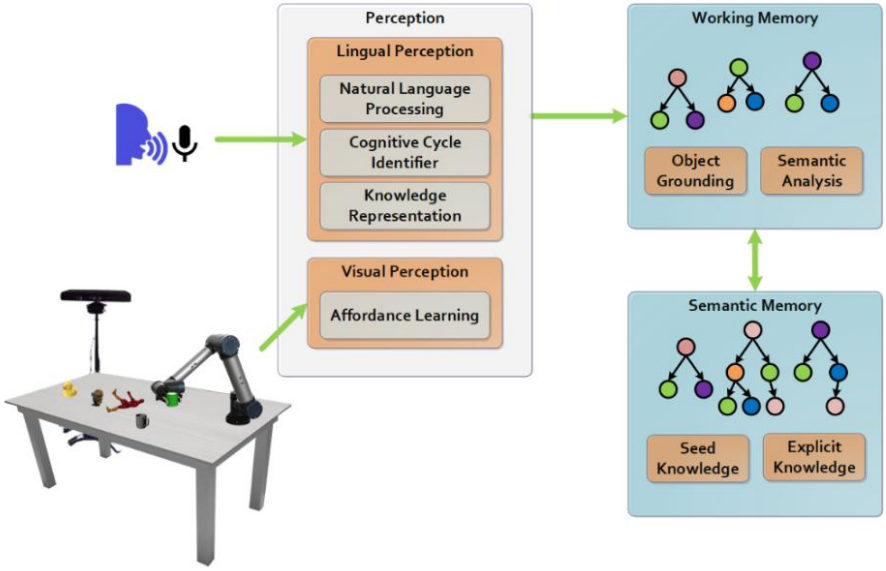


Figure 2. System overview.

4.1. Visual Perception

The visual perception module is based on multiple levels. The first level is based on affordance learning and the next is based on item name identification.

Affordance Learning: The affordance module is trained on a dataset [40] consisting of objects used commonly in the household. The 30 items chosen to date can be on categorized as callable, cleanable, drinkable, edible, playable, readable, writable, and wearable [6]. A total of 8538 images were taken by a Samsung Galaxy 7 camera. The system (see Figure 2) was trained to recognize seven classes, i.e., callable, cleanable, drinkable, edible, playable, readable, and writable. The number of total images used for training purposes was 7622 excluding the wearable category. The system trained on YOLOv3 [41] to identify the items placed on the table-top setup with 18,000 iterations having an average loss of 0.1176 (see Appendix A, Figure A1). The architecture of YOLOv3 with its configuration is shown in Figure 3. The detailed configuration of the training pipeline is presented in Table A1 in Appendix A.

Item Name Identification: The items classified based on affordance learning are further assigned names, i.e., Drinkable as Bottle or Cup. For the said purpose, a pre-trained YOLOv3 classifier [42] was used to identify the name of the commonly used household items. This module uses the position of the image determined by the YOLOv3 classifier to localize the detected object in the table-top setup. The system returns the item sets as $I = \{i_1, i_2, i_3, \ldots, i_n\}$, and the properties of items are classified as $I_p = \{name, affordance, location, direction\}$.

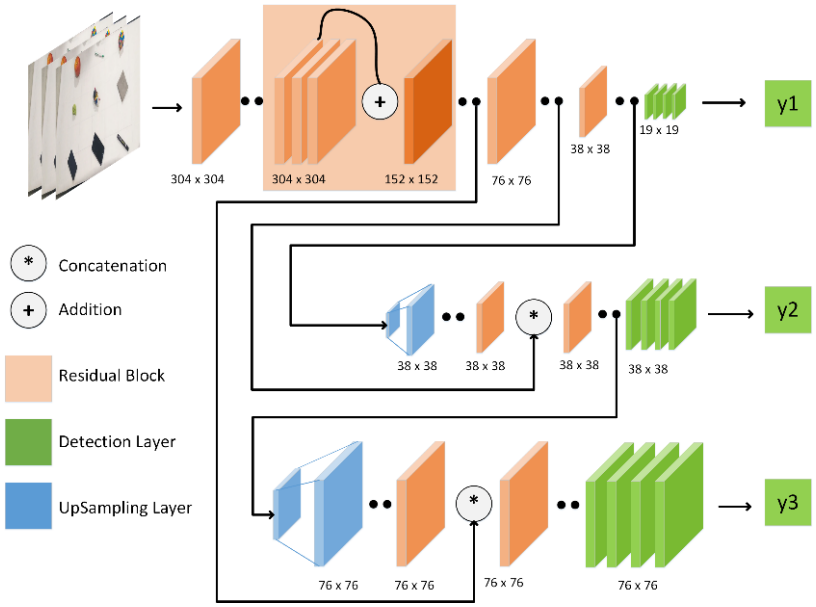


Figure 3. Affordance learning: YOLOv3 architecture.

4.2. Lingual Perception

We developed a rule-based chatbot for the acquisition of perceptual semantics from the auditory stream. The co-worker (i.e., human) communicates with the robot through a speech engine based on Google Speech-to-Text API. The stream is then sent to the *Natural Language Processing* module for tokenization, part-of-speech tagging, name entity tagging, and basic dependency tagging. Further processing is done in *Knowledge Representation* for the formation of the conceptual graph and semantic network, whereas *Cognitive Cycle Identifier* modules are used for the classification of cognitive processes in the cycle.

Natural Language Processing: The *Natural Language Processing* module consists of four submodules: *Tokenization*, *Part of Speech (POS)* tagger, *Name Entity* (NE) tagger and *Basic Dependency* (BD) (see Figure 4). The input stream (sentence) is tokenized in the *Tokenization* module and further tagged using the *Part of Speech (POS)* tagger. The stream is then classified into noun $N = \{NN, NNS, NNP, NNPS\}$, adjective $A_{dj} = \{JJ, JJR, JJS\}$, and verb $V = \{VB, VBD, VBP, VBN, VBG, VBZ\}$ using NLTK (Natural Language Toolkit). Detail about the POS tags can be found at [43]. Furthermore, CoreNLP is used to identify BD and NE tags for the formulation of atom of knowledge elements (concepts, relations, and features).

Figure 4. Natural language processing module.

Knowledge Representation: *Knowledge Representation* module consists of *Triplet Extraction*, Semantic *Association* and *Atom of Knowledge* (see Figure 5). Knowledge is constructed after the NLP module has processed the stream. The *Knowledge Representation*

module extracts the triplets (i.e., predicate, object, and subject) from the processed sentences. The predicate is extracted from the previously processed verb set V, whereas the subject is extracted from the noun set N. The last triplet is assigned based on the last of the adjective set A_{dj} or noun. The association between concepts is created using an a priori knowledge base by searching the concept nodes for similarities based on relationship types such as "InstanceOf", "IsA", "PartOf", "DerivedFrom", "Synonym", "CreatedBy", "HasProperty", "UsedFor", "HasA", "FormOf", and "RelatedTo". Based on these associations, the atom of knowledge is constructed.

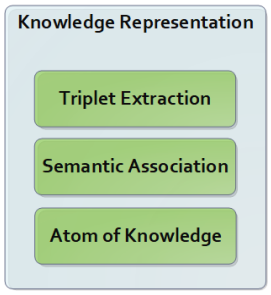


Figure 5. Knowledge representation module.

Cognitive Cycle Identifier: The sensory stimuli based on sentences are evaluated in Bloom's taxonomy-based cognitive module. The module is constructed on a system trained on Long-Short-Term Memory (LSTM) with an improvised dataset based on Yahya's model with 300 epochs and a cost function of 1.903×10^{-6} [44]. The cognitive level is determined as $Cog_{level} = [Perception, Understanding - Comprehensions, Execution - Control, Post-Execution - Analysis, Evaluation, Synthesis]$. The cognitive levels are dataset classes. The stream is then tokenized and parsed using the Natural Language Toolkit (NLTK). The knowledge domain is further classified based on the action verbs of Bloom's revised taxonomy [12] to determine the instance of $BCog_{matrix}$. The instance then initiates the designated cognitive process applicable for potential knowledge dimension and action.

4.3. Semantic Memory

Semantic Memory is constituted in an a priori and an a posteriori manner. The seed knowledge is developed from ConceptNet and WordNet, whereas the posterior knowledge is constructed when the agent interacts with the environment and stored in the *Explicit Knowledge* repository.

Seed Knowledge: Seed knowledge is constituted based on atoms of knowledge from WordNet and ConceptNet. The knowledge base has 1.47 million nodes and 3.13 million relationships (53 relationship types, i.e., *AlsoSee*, *Antonym*, *AtLocation*, *Attribute*, *CapableOf*, *Cause*, *Causes*, etc.). The nodes consist of 117,659 Synsets (WordNet Nodes), 1.33 million Concept (ConceptNet), and 157,300 Lemma nodes. The Lemma nodes are extracted from Concept nodes based on "root words". These nodes are partially or fully matchable with Synset nodes. The semantic memory-based seed (tacit) knowledge is represented as $S_m = \{Atom_1, Atom_2, Atom_3 \ldots . Atom_k\}$ and atoms as $Atom = \{\langle concept, concpet\rangle, \langle concept, feature\rangle\}$. The transformation of ConceptNet and WordNet ontologies to the proposed seed knowledge, i.e., semantic memory can be seen in Tables 2 and 3.

Explicit Knowledge: *Explicit Knowledge* is constructed based on the semantic network [45] and conceptual graph [46] drawn from working memory. These graphs are constructed in the *Knowledge Representation* module.

Table 2. ConceptNet to semantic memory node and edge transformation detail.

ConceptNet to Semantic Memory				
Items	**Original Terms**	**Attached to**	**Adopted Terms**	**Attached to**
Unit of Knowledge	Edge or Assertion	ConceptNet	Concept Node	Graph-Based Ontology
Attributes	Fields	Assertions	Properties	Nodes/Edges
Attribute_1	Uri	Assertion	conceptUri	Node
Attribute_2	rel	Assertion	RelationShip Type	Edge
Attribute_3	start (Concept)	Assertion	Concept Node	Node
Attribute_4	end (Concept)	Assertion	Concept Node	Node
Attribute_5	weight	Assertion	weight	Edge
Attribute_6	sources	Assertion	-	-
Attribute_7	license	Assertion	-	-
Attribute_8	dataset	Assertion	dataset	Edge
Attribute_9	surfaceText	Assertion	Name	Node
	-	-	pos (Extracted frm Uri)	Node
	-	-	Id (Extracted frm Uri)	Node
			<id> (Graph Index)	Node/Edge

Table 3. WordNet to semantic memory nodes and edge transformation detail.

WordNet to Semantic Memory		
Items	**Adopted Terms**	**Attached to**
Hyponym	IsA	Edge
Hypernym	IsA	Edge
Member Homonym	PartOf	Edge
Substance Holonym	PartOf	Edge
Part Holonym	PartOf	Edge
Member Meronym	PartOf	Edge
Substance Meronym	PartOf	Edge
Part Meronym	PartOf	Edge
Topic Domain	Domain	Edge
Region Domain	Domain	Edge
Usage Domain	Domain	Edge
Attribute	Attribute	Edge
Entailment	Entailment	Edge
Causes	Causes	Edge
Also See	AlsoSee	Edge
Verb Group	VerbGroup	Edge
Similar To	SimilarTo	Edge

4.4. Working Memory

Working memory acts as an executive control in the proposed system, whose primary responsibility is to formulate object grounding and semantic similarity analysis.

Object Grounding: The localization is further used to determine the accessibility coordinates of the robotic arm. We started with the simplest approach by dividing the table-top setup into a 3×3 grid as shown in Figure 6.

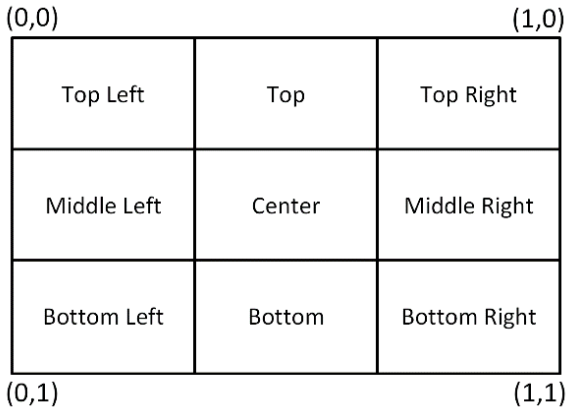


Figure 6. Image grid (3 × 3).

The grid is divided into several directions as defined in $I_{direction} : Item \times Item \rightarrow [center, right, left, top, bottom, bottom-left, bottom-right, top-right, top-left]$. This approach is workable to determine the exact position of the item. However, we want to know the relative positions of the items. The grid is further described based on a reference point, i.e., center in Figure 7.

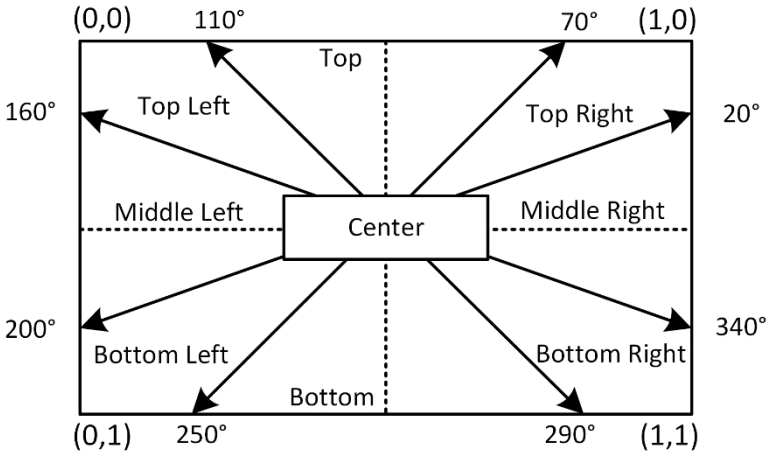


Figure 7. Grid reference point.

This reference point consideration is further extended to position the item relative to others as shown in Figure 8.

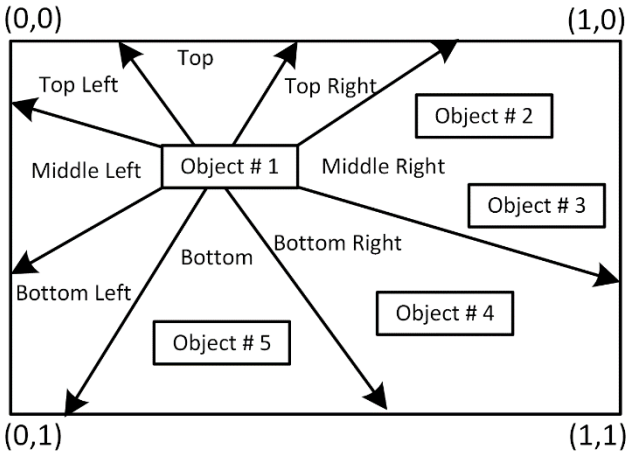


Figure 8. Relative position of objects (items).

Semantic Analysis: The semantic similarity between atoms of knowledge constructed from words W coming from *Lingual Perception* and atoms of knowledge constricted from $I_{p_{affordance}}$ coming from *Visual Perception* is evaluated.

$$\mathfrak{S}\left(Atom_{W},\ Atom_{I_{p_{affordance}}}\right) = \frac{\left|Atom_{W} \cap Atom_{I_{p_{affordance}}}\right|}{\left|Atom_{W} \cup Atom_{I_{p_{affordance}}}\right|} \quad (1)$$

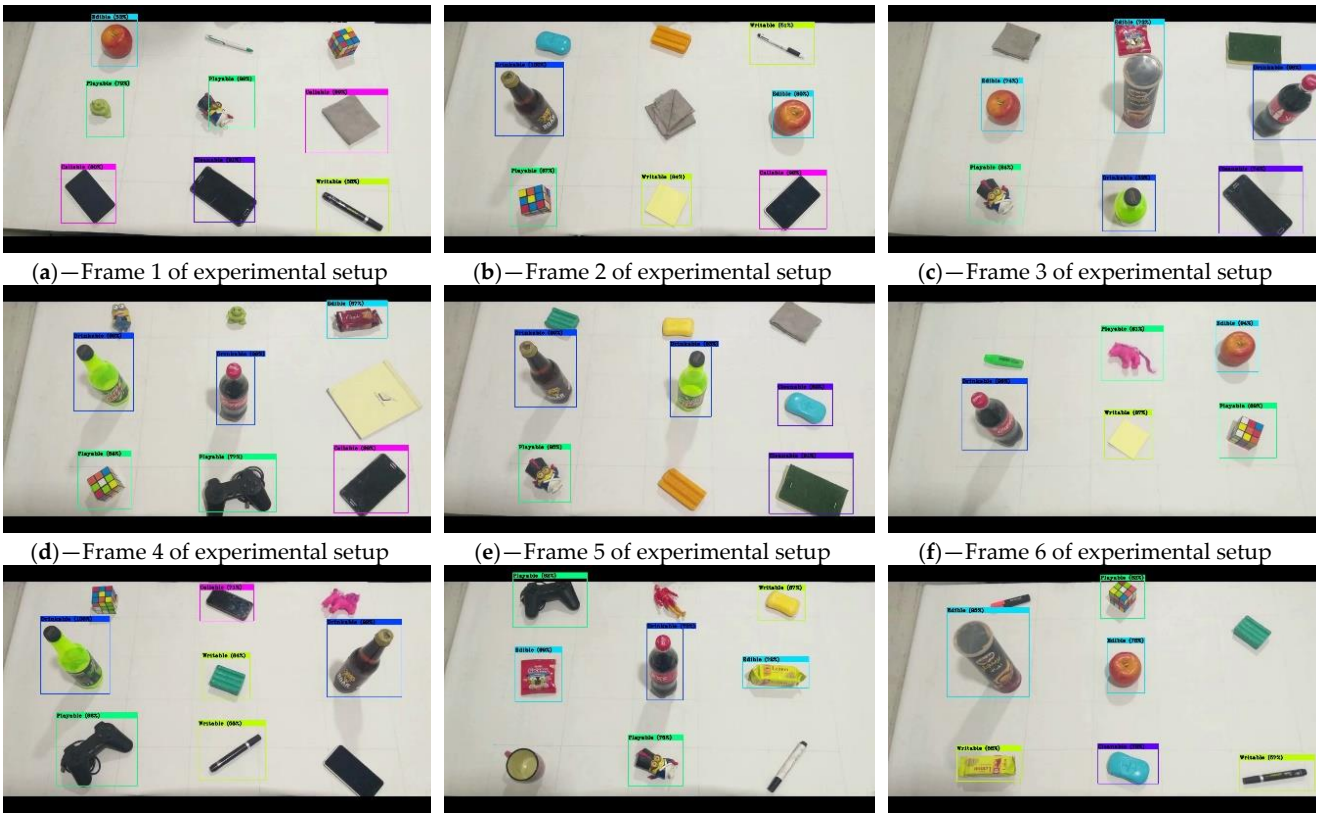
The $\mathfrak{S}$ maximum scores indicate the similarity between $Atom_{W}$ and $Atom_{I_{p_{affordance}}}$.

5. Results

To validate the proposed methodology, we conducted multiple experiments. The experimental results are based on human collaborator interaction with the agent. In the first phase, visual perception analysis is discussed, and the subsequent phases are based on object grounding, cognitive, and semantic analysis.

5.1. Visual Perception

We trained the agent on YOLOv3 and tested it to validate the proposed methods on 160 video frames comprising a collection of 783 different household objects (see Figure 9 for a subset of video frames). The categorization of objects placed on the table-top scenarios is based on callable, cleanable, drinkable, edible, playable, readable, and writable affordance classes. Each frame contains an average of nine objects placed on various areas of the table to identify and relate the location with spoken commands.



(**a**)—Frame 1 of experimental setup (**b**)—Frame 2 of experimental setup (**c**)—Frame 3 of experimental setup

(**d**)—Frame 4 of experimental setup (**e**)—Frame 5 of experimental setup (**f**)—Frame 6 of experimental setup

(**g**)—Frame 7 of experimental setup (**h**)—Frame 8 of experimental setup (**i**)—Frame 9 of experimental setup

Figure 9. Affordance recognition results of 9 frames out 160 in total.

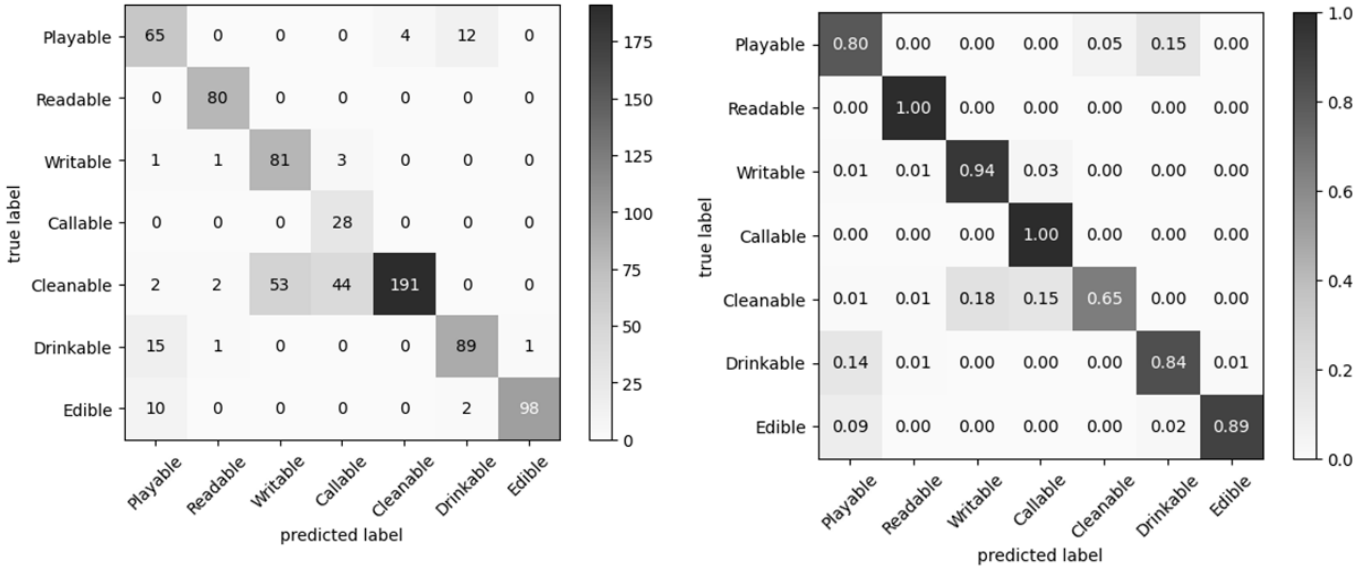
The results are shown in confusion matrices in Figure 10. The results indicate that for cleanable affordance the negative predictions were mostly callable and writeable. This happens in the case of *duster* (*cleanable*) and *spunch* (*cleanable*) because their shape is similar

to that of a *cellphone* (*callable*). In some cases, the yellow *spunch* was misclassified as *sticky-note* (*writable*). Furthermore, the *toys* (*playable*) were misclassified as drinkable objects in 12 instances due to geometric similarities. Moreover, performance metrics were calculated for affordance recognition and can be seen in Table 4.

$$Precision = \frac{True\ Positives}{True\ Positives + false\ positives} \tag{2}$$

$$Recall = \frac{True\ Positives}{True\ Positives + false\ negatives} \tag{3}$$

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}. \tag{4}$$



(**a**) Confusion Matrix (**b**) Normalized Confusion Matrix

Figure 10. Affordance recognition–confusion matrices.

Table 4. Performance metrics: precision, recall, F1 score.

	True Positive	False Positive	False Negative	True Negative	Precision	Recall	F1 Score
Playable	65	28	16	674	0.699	0.802	0.747
Readable	80	4	0	699	0.952	1.000	0.976
Writeable	81	53	5	644	0.604	0.942	0.736
Callable	28	47	0	708	0.373	1.000	0.544
Cleanable	191	4	101	487	0.979	0.654	0.784
Drinkable	89	14	17	663	0.864	0.840	0.852
Edible	98	1	12	672	0.990	0.891	0.938
Average					**0.78**	**0.87**	**0.80**

Table 4 contains false positive, false negative, true positive, and true negative parameters. Based on these parameters' precision, recall, and f1 score are calculated for all seven affordance classes. The results indicate that the precision is good in the case of the cleanable object but the recall has a low value, whereas callable has good recall but has low precision.

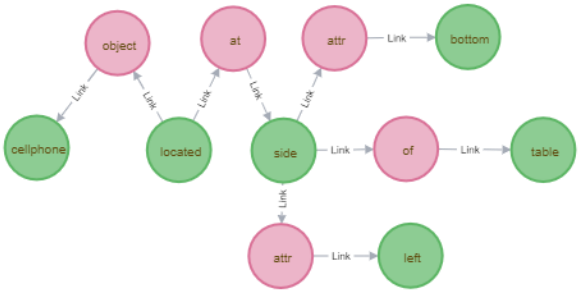
Moreover, the f1 score of callable is lowest amongst the remainder of the classes. The affordance learning is compared with the current state-of-the-art in Table 5. Furthermore, the objects were classified using a pre-trained COCO model for object grounding and knowledge representation. The knowledge is represented using both a conceptual graph and a semantic network. The conceptual graph is used for further action selection and the semantic network becomes part of the semantic memory.

Table 5. State-of-the-art comparison with proposed affordance learning.

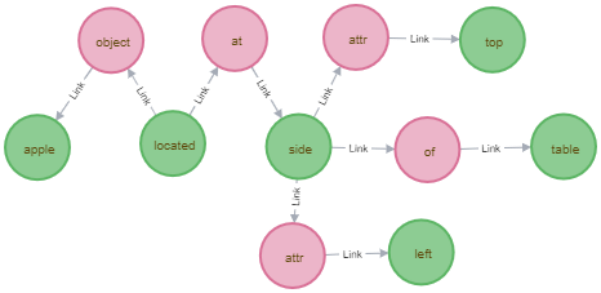
Work	Affordance/Objects	Robotic Task	Size of Dataset	Evaluation Metrics
[47]	9 Classes/10 Object Categories	Action Prediction	8835 RGB Images	Weighted F Score = 73.35
[48]	7 Classes/105 Objects	Object Manipulation	30,000 RGB-D Image Pairs	Recognition Accuracy = 95.0%
[6]	7 Classes/42 Objects	Object Grasp	8960 RGB Images	Recognition Accuracy = 100%
[49]	28 Homes/24 Offices/17 Classes	Action Prediction	550 RGB-D Views	Max Precision = 88.40
[50]	17 Classes	Action Prediction	250 RGB-D Videos	Time Saving Accuracy
[51]	9 Objects	Object Manipulation	RGB-D Images	Confidence Level
[52]	7 Classes/17 Categories/105 Objects	Object Manipulation	28k+ RGB-D Images	Weighted F Measure
Ours	**7 Classes/26 Objects (Originally 8 Classes/30 Objects)**	**Object Grasp**	**7622 (Originally 8538) RGB Images**	**Average F Score = 0.80**

5.2. Lingual Perception and Object Grounding

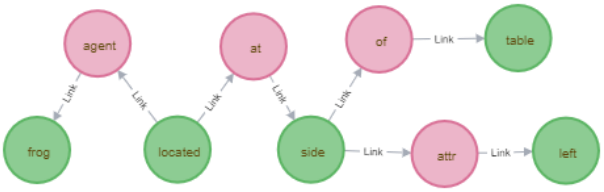
This section is based on the object grounding results, formation of the conceptual graph, and semantic network. To display the formulation of the conceptual graph, one of the previously discussed video frames is used (Figure 9a). In this phase, the information extracted from video frames is used to address the affordance of an object and position with respect to the center of the frame and the position of other objects. This information is further transformed using the COCO model as "*The cellphone is located at the bottom left side of the table*" (see Figure 11). For this purpose, two types of graphs were generated, i.e., a conceptual graph (CG) and a semantic network (SN). CG is generated separately for each instance in the frame. CG is composed of two nodes, i.e., *conceptNode* (*cellphone*, *located*, *side*, *table*, *bottom*, *left*) and *relationNode* (*object*, *at*, *attr*, *of*), whereas the empty relation is represented as "*Link*" (see Figure 11a). This type of graph helps the agent to check the dependency factor. In the case of "*The frog is located on the left side of the table*" the nodes are slightly different, i.e., *conceptNode* (*frog*, *located*, *side*, *table*, *left*) and *relationNode* (*agent*, *at*, *attr*, *of*) (see Figure 11c). Both examples have two *relationNodes* distinct nodes, i.e., "*cellphone*" is represented as "*object*" whereas "*frog*" is represented as an "*agent*". This information helps understand the nature of the item, its role, and placement.



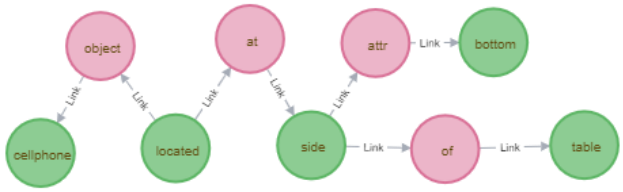
(**a**) The cell phone is located at the bottom left side of the table.



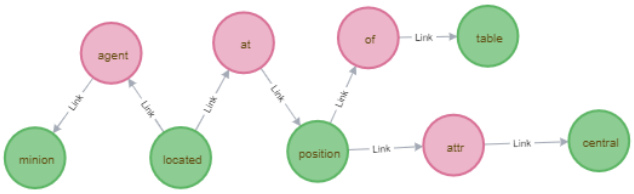
(**b**) The apple is located at the top left side of the table.



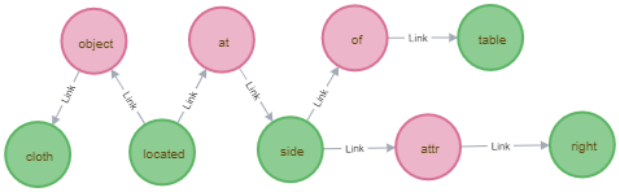
(**c**) The frog is located on the left side of the table.



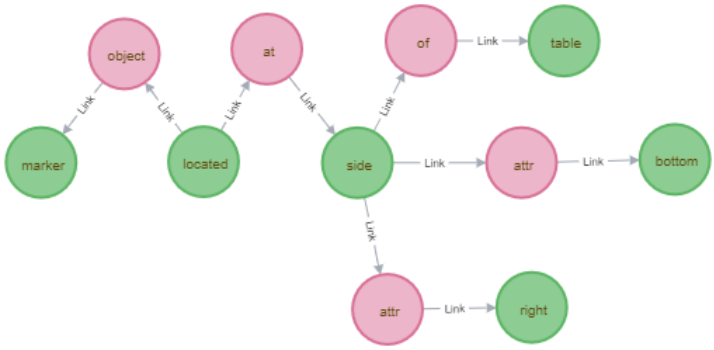
(**d**) The cellphone is located at the bottom side of the table.



(**e**) The minion is located at the central position of the table.



(**f**) The cloth is located on the right side of the table.



(**g**) The marker is located at the bottom-right side of the table.

Figure 11. Conceptual graphs.

The semantic network (SN) for a video frame (Figure 9a) is constructed to be stored in the semantic memory for future processing (see Figure 12). SN is composed of the ConceptNode and relationship in the form of edges. The edges for "*LocatedAt*" and "*LocatedOn*" indicate the path towards the position of the item, whereas "*NEXT*" is an empty relationship that points towards the succeeding node(s).

If the item in the frame is not recognized based on affordance (see Figure 9a), i.e., "*Rubik's cube*" (*as playable*) and "*pen*" (*as writeable*), then the agent will not be able to ground the position and direction of an item. The grounding is formulated after the affordance recognition in the form of sentences and then as a conceptual graph (see Figure 11) and a semantic network (see Figure 12).

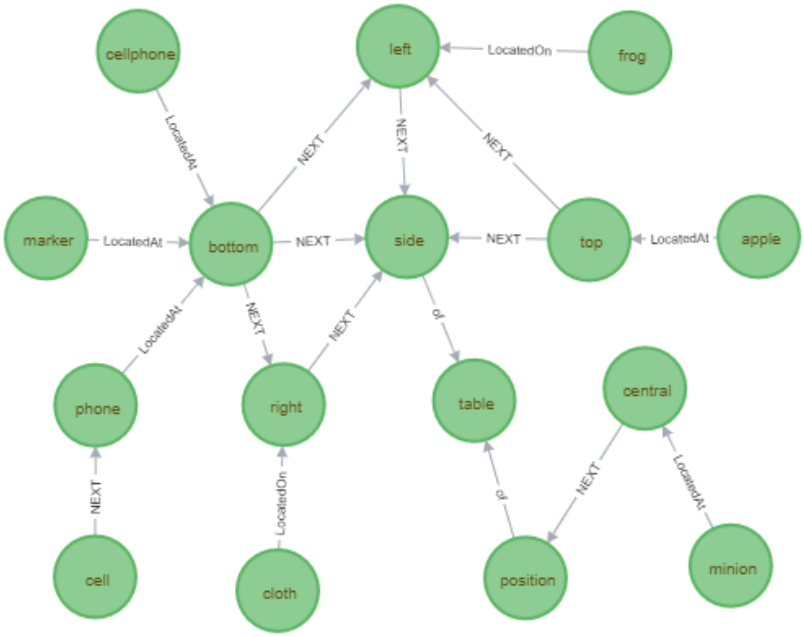
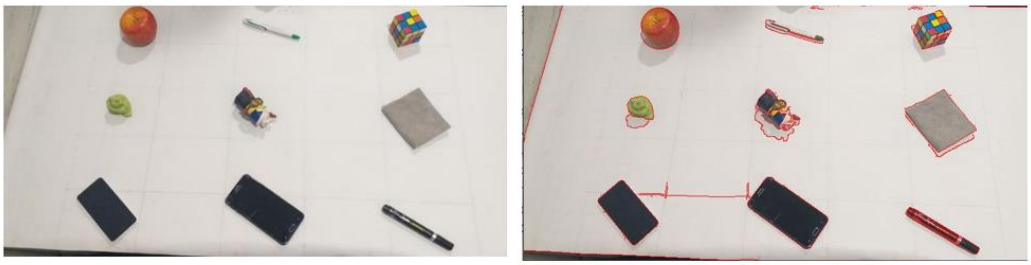


Figure 12. Semantic network.

5.3. Cognitive Cycle Identifier

This section is based on the results from the Bloom-based Cognitive Cycle Identifier. In this phase, the verbal sensory stimuli are analyzed for the action selection, i.e., "*How many items are present on the table?*", "*How many objects belong to a drinkable category?*", "*Which object is used to reduce the hunger?*", and "*Which item is used to reduce the intensity of thirst?*". The action verbs are further accessed for the identification of the "*cognitive domain*", as described in Bloom's revised taxonomy [12]. After the identification of the "*cognitive domain*," the agent chooses its actions as "*Blob Detection and Counting*", "*Affordance Recognition*", and "*Jaccard Semantic Similarity*" (see Figures 13–18). The results shown here are encouraging and represent an important step towards an advancement in perceptual semantics in cognitive robots.

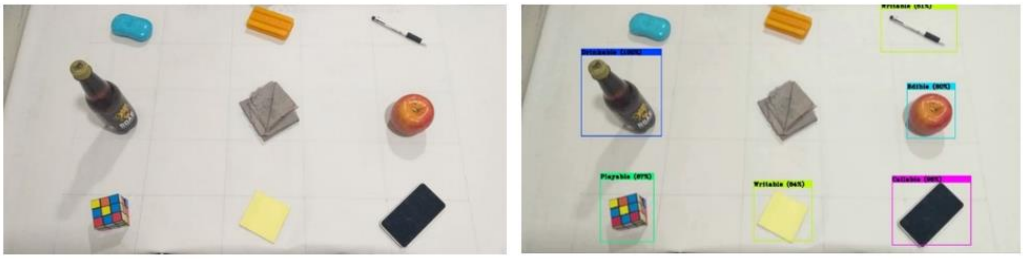


(**a**) (**b**)

Query1:	*How many items are present on the table?*
Cognitive Domain:	*Perception*
Action:	*Blob Detection and Counting*
Output Sentence:	*There are 9 items present on the table.*

(**c**)

Figure 13. (**a**) Original frame, (**b**) blob detection and counting for query, (**c**) query.



(**a**) (**b**)

Query2: *How many objects belong to a drinkable category?*
Cognitive Domain: *Understanding-Comprehensions*
Action: *Affordance Recognition*
Output Sentence: *There is 1 drinkable object located at the left side of table.*

(**c**)

Figure 14. (**a**) Original frame, (**b**) object affordance results for query, (**c**) query.

Query	Object Recognized	Similarity
hunger	**Edible**	**0.000108**
hunger	Drinkable	0.000032
hunger	Playable	0.000008
hunger	Writable	0.000007
hunger	Callable	0.000005

(**a**) (**b**)

Query3: *Which object is used to reduce the hunger?*
Cognitive Domain: *Understanding-Comprehensions*
Action: *Affordance Recognition & Jaccard Semantic Similarity*
Output Sentence: *Affordance Recognition*
There is one edible object located at the right side of table.

(**c**)

Figure 15. (**a**) Similarity score, (**b**) semantic network for query, (**c**) query.

Query	Object Recognized	Similarity
Thirst	**Drinkable**	**6.3607 × 10^{-5}**
Thirst	Edible	3.2921 × 10^{-5}
Thirst	Writable	1.3092 × 10^{-5}
Thirst	Playable	5.2537 × 10^{-6}
Thirst	Callable	0.0000

(**a**) (**b**)

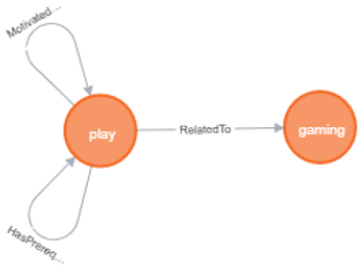
Query4: *Which item is used to reduce the intensity of thirst?*
Cognitive Domain: *Understanding-Comprehensions*
Action: *Affordance Recognition & Jaccard Semantic Similarity*
Output Sentence: *Affordance Recognition*
There is one drinkable object located at the left side of table.

(**c**)

Figure 16. (**a**) Similarity score, (**b**) semantic network for query, (**c**) query.

Query	Object Recognized	Similarity
gaming	**Playable**	**1.7×10^{-5}**
gaming	Drinkable	0.0
gaming	Writable	0.0
gaming	Edible	0.0
gaming	callable	0.0

(**a**)



(**b**)

Query5: *Which object is used for gaming?*
Cognitive Domain: *Understanding-Comprehensions*
Action: *Affordance Recognition & Jaccard Semantic Similarity*
Output Sentence: *Affordance Recognition*
There is 1 playable object located at the bottom left side of table.

(**c**)

Figure 17. (**a**) Similarity score, (**b**) semantic network for query, (**c**) query.

Query	Object Recognized	Similarity
note	**Writable**	**3.4×10^{-5}**
note	Playable	1.3×10^{-5}
note	Callable	1.2×10^{-5}
note	Edible	5.4×10^{-6}
note	Drinkable	4.8×10^{-6}

(**a**)

(**b**)

Query6: *Which item is used to take notes?*
Cognitive Domain: *Understanding-Comprehensions*
Action: *Affordance Recognition & Jaccard Semantic Similarity*
Output Sentence: *Affordance Recognition*
There are 2 writable objects present at the table.

(**c**)

Figure 18. (**a**) Similarity score, (**b**) semantic network for query, (**c**) query.

The Universal Robot (UR5)-based demonstrations can be accessed through Table 6.

Table 6. Links to demonstration videos.

Human Cues	Video
I am feeling thirsty	https://youtu.be/A16Q0Od7vg4 (Accessed on 8 September 2021)
I am hungry, I need something to eat.	https://youtu.be/YJe9CCo1z-M (Accessed on 8 September 2021)
Give me anything to play a video game.	https://youtu.be/R46WCwMzryc (Accessed on 8 September 2021)
I am hungry (unsuccessful)	https://youtu.be/f2vJswBkpZs (Accessed on 8 September 2021)

6. Conclusions

In this work, we proposed perceptual and semantic processing for human–robot interaction in the agent. The contribution of the proposed work is the extension of affordance

learning, Bloom's taxonomy as a cognitive cycle, object grounding, and perceptual semantics. The experiments were conducted on the agent using 160 video frames with household objects in a table-top scenario and human cues that contained implicit instructions. The results suggest that the overall HRI experience was improved due to the proposed method and the agent was able to address implicit lingual cues (see Table 6).

Author Contributions: Conceptualization, W.M.Q.; Methodology, S.T.S.B. and W.M.Q.; Supervision, W.M.Q.; Validation, S.T.S.B.; Writing—original draft, S.T.S.B.; Writing—review & editing, W.M.Q. Both authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Services Syndicate Private Limited for providing access to Universal Robot (UR5) for experimentations.

Acknowledgments: Authors acknowledge the support of Services Syndicate Private Limited for providing access to Universal Robot (UR5) for experimentations.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Layer configurations.

		Layer Type	Layer	Filters	Concatenation	Size/Strd(dil)	Output
0			conv	32		3 × 3/ 1	608 × 608 × 32
1		Convolutional	conv	64		3 × 3/ 2	304 × 304 × 64
2			conv	32		1 × 1/ 1	304 × 304 × 32
3			conv	64		3 × 3/ 1	304 × 304 × 64
4		Residual	Shortcut Layer				304 × 304 × 64
5			conv	128		3 × 3/ 2	152 × 152 × 128
		Convolutional	conv	64		1 × 1/ 1	152 × 152 × 64
	2 ×		conv	128		3 × 3/ 1	152 × 152 × 128
11		Residual	Shortcut Layer				152 × 152 × 128
12			conv	256		3 × 3/ 2	76 × 76 × 256
		Convolutional	conv	128		1 × 1/ 1	76 × 76 × 128
	8 ×		conv	256		3 × 3/ 1	76 × 76 × 256
36		Residual	Shortcut Layer				76 × 76 × 256
37			conv	512		3 × 3/ 2	38 × 38 × 512
		Convolutional	conv	256		1 × 1/ 1	38 × 38 × 256
	8 ×		conv	512		3 × 3/ 1	38 × 38 × 512
61		Residual	Shortcut Layer				38 × 38 × 512
62			conv	1024		3 × 3/ 2	19 × 19 × 1024
		Convolutional	conv	512		1 × 1/ 1	19 × 19 × 512
	4 ×		conv	1024		3 × 3/ 1	19 × 19 × 1024
74		Residual	Shortcut Layer				19 × 19 × 1024
	3 ×		conv	512		1 × 1/ 1	19 × 19 × 512
80		Convolutional	conv	1024		3 × 3/ 1	19 × 19 × 1024
81			conv	39		1 × 1/ 1	19 × 19 × 39
82		Detection	yolo				
83			route	79		->	
84		Convolutional	conv	256		1 × 1/ 1	19 × 19 × 256
85		Upsampling	upsample			2x	38 × 38 × 256
86			route: 85 -> 61	85	61		38 × 38 × 768
			conv	256		1 × 1/ 1	38 × 38 × 256
	3 ×						
92		Convolutional	conv	512		3 × 3/ 1	38 × 38 × 512
93			conv	39		1 × 1/ 1	38 × 38 × 39

Table A1. *Cont.*

		Layer Type	Layer	Filters	Concatenation	Size/Strd(dil)	Output
94		Detection	yolo				
95			route	91		->	
96		Convolutional	conv	128		1 × 1/ 1	38 × 38 × 128
97		Upsampling	upsample			2 x	76 × 76 × 128
98			route: 97 -> 36	97	36		76 × 76 × 384
	3 ×		conv	128		1 × 1/ 1	76 × 76 × 128
104		Convolutional	conv	256		3 × 3/ 1	76 × 76 × 256
105			conv	39		1 × 1/ 1	76 × 76 × 39
106		Detection	yolo				

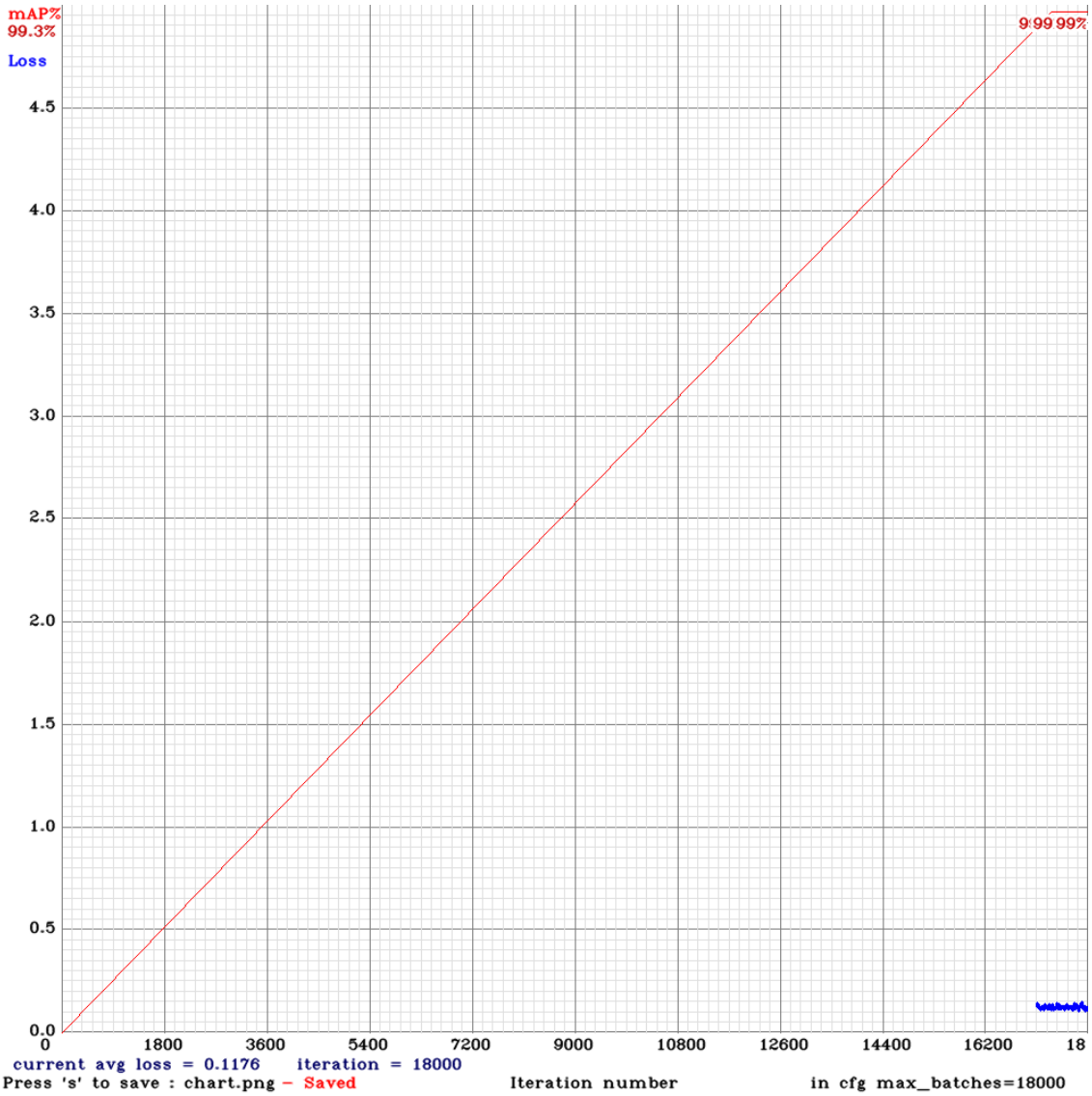


Figure A1. Affordance training iterations graph.

References

1. Dubba, K.S.R.; Oliveira, M.R.d.; Lim, G.H.; Kasaei, H.; Lopes, L.S.; Tome, A.; Cohn, A.G. Grounding Language in Perception for Scene Conceptualization in Autonomous Robots. In Proceedings of the AAAI 2014 Spring Symposium, Palo Alto, CA, USA, 24–26 March 2014.
2. Kotseruba, I.; Tsotsos, J.K. 40 years of cognitive architectures: Core cognitive abilities and practical applications. *Artif. Intell. Rev.* **2020**, *53*, 17–94. [CrossRef]
3. Oliveira, M.; Lopes, L.S.; Lim, G.H.; Kasaei, S.H.; Tomé, A.M.; Chauhan, A. 3D object perception and perceptual learning in the RACE project. *Robot. Auton. Syst.* **2016**, *75*, 614–626. [CrossRef]
4. Oliveira, M.; Lim, G.H.; Lopes, L.S.; Kasaei, S.H.; Tomé, A.M.; Chauhan, A. A perceptual memory system for grounding semantic representations in intelligent service robots. In Proceedings of the 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, Chicago, IL, USA, 14–18 September 2014; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2014; pp. 2216–2223.
5. Lopes, M.; Melo, F.S.; Montesano, L. Affordance-based imitation learning in robots. In Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, San Diego, CA, USA, 30 November 2006; IEEE: New York, NY, USA, 2007; pp. 1015–1021.
6. Mi, J.; Tang, S.; Deng, Z.; Goerner, M.; Zhang, J. Object affordance based multimodal fusion for natural Human-Robot interaction. *Cogn. Syst. Res.* **2019**, *54*, 128–137. [CrossRef]
7. Sowa, J.F. The Cognitive Cycle. In Proceedings of the 2015 Federated Conference on Computer Science and Information Systems (FedCSIS), Lodz, Poland, 13–16 September 2015; IEEE: New York, NY, USA, 2015; Volume 5, pp. 11–16.
8. McCall, R.J. Fundamental Motivation and Perception for a Systems-Level Cognitive Architecture. Ph.D. Thesis, The University of Memphis, Memphis, TN, USA, 2014.
9. Paraense, A.L.; Raizer, K.; de Paula, S.M.; Rohmer, E.; Gudwin, R.R. The cognitive systems toolkit and the CST reference cognitive architecture. *Biol. Inspired Cogn. Archit.* **2016**, *17*, 32–48. [CrossRef]
10. Blanco, B.; Fajardo, J.O.; Liberal, F. Design of Cognitive Cycles in 5G Networks. In *Collaboration in A Hyperconnected World*; Springer Science and Business Media LLC: London, UK, 2016; pp. 697–708.
11. Madl, T.; Baars, B.J.; Franklin, S. The Timing of the Cognitive Cycle. *PLoS ONE* **2011**, *6*, e14803. [CrossRef] [PubMed]
12. Krathwoh, D. A Revision of Bloom's Taxonomy: An Overview. *Theory Pract.* **2002**, *41*, 213–264.
13. Qazi, W.M.; Bukhari, S.T.S.; Ware, J.A.; Athar, A. NiHA: A Conscious Agent. In Proceedings of the COGNITIVE 2018, The Tenth International Conference on Advanced Cognitive Technologies and Applications, Barcelona, Spain, 18–22 February 2018; pp. 78–87.
14. Marques, H.G. Architectures for Embodied Imagination. *Neurocomputing* **2009**, *72*, 743–759. [CrossRef]
15. Samsonovich, A.V. On a roadmap for the BICA Challenge. *Biol. Inspired Cogn. Archit.* **2012**, *1*, 100–107. [CrossRef]
16. Breux, Y.; Druon, S.; Zapata, R. From Perception to Semantics: An Environment Representation Model Based on Human-Robot Interactions. In Proceedings of the 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), Nanjing and Tai'an, China, 27–31 August 2018; IEEE: New York, NY, USA, 2018; pp. 672–677. [CrossRef]
17. Bornstein, M.H.; Gibson, J.J. The Ecological Approach to Visual Perception. *J. Aesthet. Art Crit.* **1980**, *39*, 203. [CrossRef]
18. Cruz, F.; Magg, S.; Weber, C.; Wermter, S. Training Agents With Interactive Reinforcement Learning and Contextual Affordances. *IEEE Trans. Cogn. Dev. Syst.* **2016**, *8*, 271–284. [CrossRef]
19. Min, H.; Yi, C.; Luo, R.; Zhu, J.; Bi, S. Affordance Research in Developmental Robotics: A Survey. *IEEE Trans. Cogn. Dev. Syst.* **2016**, *8*, 237–255. [CrossRef]
20. Kjellström, H.; Romero, J.; Kragić, D. Visual object-action recognition: Inferring object affordances from human demonstration. *Comput. Vis. Image Underst.* **2011**, *115*, 81–90. [CrossRef]
21. Thomaz, A.L.; Cakmak, M. Learning about objects with human teachers. In Proceedings of the 2009 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI), San Diego, CA, USA, 11–13 March 2009; IEEE: New York, NY, USA, 2009; pp. 15–22.
22. Wang, C.; Hindriks, K.V.; Babuška, R. Robot learning and use of affordances in goal-directed tasks. In Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, Japan, 3–7 November 2013; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2013; pp. 2288–2294.
23. Nguyen, A.; Kanoulas, D.; Muratore, L.; Caldwell, D.G.; Tsagarakis, N.G. *Translating Videos to Commands for Robotic Manipulation with Deep Recurrent Neural Networks.* 2017. Available online: https://www.researchgate.net/publication/320180040_Translating_Videos_to_Commands_for_Robotic_Manipulation_with_Deep_Recurrent_Neural_Networks (accessed on 17 September 2018).
24. Myers, A.; Teo, C.L.; Fermuller, C.; Aloimonos, Y. Affordance detection of tool parts from geometric features. In Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA), Seattle, WA, USA, 26–30 May 2015; IEEE: New York, NY, USA, 2015; pp. 1374–1381.
25. Moldovan, B.; Raedt, L.D. Occluded object search by relational affordances. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, 7 June 2014; IEEE: New York, NY, USA, 2014; pp. 169–174.
26. Nguyen, A.; Kanoulas, D.; Caldwell, D.G.; Tsagarakis, N.G. Object-based affordances detection with Convolutional Neural Networks and dense Conditional Random Fields. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; IEEE: New York, NY, USA, 2017; pp. 5908–5915.

27. Antunes, A.; Jamone, L.; Saponaro, G.; Bernardino, A.; Ventura, R. From human instructions to robot actions: Formulation of goals, affordances and probabilistic planning. In Proceedings of the 2016 IEEE International Conference on Robotics and Automation (ICRA); Institute of Electrical and Electronics Engineers (IEEE), Stockholm, Sweden, 16–21 May 2016; IEEE: New York, NY, USA, 2016; pp. 5449–5454.
28. Tenorth, M.; Beetz, M. Representations for robot knowledge in the KnowRob framework. *Artif. Intell.* **2017**, *247*, 151–169. [CrossRef]
29. Roy, D.; Hsiao, K.-Y.; Mavridis, N. Mental Imagery for a Conversational Robot. *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* **2004**, *34*, 1374–1383. [CrossRef] [PubMed]
30. Russell, S.; Norvig, P. *Artificial Intelligence: A Modern Approach*; Pearson Education, Inc.: Upper Saddle River, NJ, USA, 1994.
31. Madl, T.; Franklin, S.; Chen, K.; Trappl, R. A computational cognitive framework of spatial memory in brains and robots. *Cogn. Syst. Res.* **2018**, *47*, 147–172. [CrossRef]
32. Shaw, D.B. Robots as Art and Automation. *Sci. Cult.* **2018**, *27*, 283–295. [CrossRef]
33. Victores, J.G. *Robot Imagination System*; Universidad Carlos III de Madrid: Madrid, Spain, 2014.
34. Diana, M.; De La Croix, J.-P.; Egerstedt, M. Deformable-medium affordances for interacting with multi-robot systems. In Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, Japan, 3–7 November 2013; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2013; pp. 5252–5257.
35. Fallon, M.; Kuindersma, S.; Karumanchi, S.; Antone, M.; Schneider, T.; Dai, H.; D'Arpino, C.P.; Deits, R.; DiCicco, M.; Fourie, D.; et al. An Architecture for Online Affordance-based Perception and Whole-body Planning. *J. Field Robot.* **2014**, *32*, 229–254. [CrossRef]
36. Sun, Y.; Ren, S.; Lin, Y. Object–object interaction affordance learning. *Robot. Auton. Syst.* **2014**, *62*, 487–496. [CrossRef]
37. Hart, S.; Dinh, P.; Hambuchen, K. The Affordance Template ROS package for robot task programming. In Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA), Seattle, WA, USA, 26–30 May 26 2015; IEEE: New York, NY, USA, 2015; pp. 6227–6234.
38. Gago, J.J.; Victores, J.G.; Balaguer, C. Sign Language Representation by TEO Humanoid Robot: End-User Interest, Comprehension and Satisfaction. *Electronics* **2019**, *8*, 57. [CrossRef]
39. Pandey, A.K.; Alami, R. Affordance graph: A framework to encode perspective taking and effort based affordances for day-to-day human-robot interaction. In Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems; Institute of Electrical and Electronics Engineers (IEEE), Tokyo, Japan, 3–7 November 2013; IEEE: New York, NY, USA, 2013; pp. 2180–2187.
40. Bukhari, S.T.S.; Qazi, W.M.; Intelligent Machines & Robotics Group, COMSATS University Islamabad, Lahore Campus. Affordance Dataset. 2019. Available online: https://github.com/stsbukhari/Dataset-Affordance (accessed on 8 September 2021).
41. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 18–20 June 1996*; IEEE: New York, NY, USA, 2016; pp. 779–788.
42. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
43. Taylor, A.; Marcus, M.; Santorini, B. The Penn Treebank: An Overview. *Treebanks* **2003**, *20*, 5–22.
44. Yahya, A.A.; Osman, A.; Taleb, A.; Alattab, A.A. Analyzing the Cognitive Level of Classroom Questions Using Machine Learning Techniques. *Procedia-Soc. Behav. Sci.* **2013**, *97*, 587–595. [CrossRef]
45. Sowa, J.F. Semantic Networks. In *Encyclopedia of Cognitive Science*; American Cancer Society: Chicago, IL, USA, 2006.
46. Sowa, J.F. Conceptual graphs as a universal knowledge representation. *Comput. Math. Appl.* **1992**, *23*, 75–93. [CrossRef]
47. Do, T.-T.; Nguyen, A.; Reid, I. AffordanceNet: An End-to-End Deep Learning Approach for Object Affordance Detection. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018; IEEE: New York, NY, USA, 2018; pp. 1–5.
48. Myers, A. From Form to Function: Detecting the Affordance of Tool Parts using Geometric Features and Material Cues. Ph.D. Thesis, University of Maryland, College Park, MD, USA, 2016.
49. Jiang, Y.; Koppula, H.; Saxena, A.; Saxena, A. Hallucinated Humans as the Hidden Context for Labeling 3D Scenes. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 18–20 June 1996; IEEE: New York, NY, USA, 2013; pp. 2993–3000.
50. Koppula, H.S.; Jain, A.; Saxena, A. Anticipatory Planning for Human-Robot Teams. In *Experimental Robotics*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 453–470.
51. Baleia, J.; Santana, P.; Barata, J. On Exploiting Haptic Cues for Self-Supervised Learning of Depth-Based Robot Navigation Affordances. *J. Intell. Robot. Syst.* **2015**, *80*, 455–474. [CrossRef]
52. Chu, F.-J.; Xu, R.; Vela, P.A. Learning Affordance Segmentation for Real-World Robotic Manipulation via Synthetic Images. *IEEE Robot. Autom. Lett.* **2019**, *4*, 1140–1147. [CrossRef]