A COGNITIVE INTEGRATED MULTI-MODAL PERCEPTION MECHANISM AND DYNAMIC WORLD MODELING FOR SOCIAL ROBOT ASSISTANTS

E. Daglarli

Abstract**—** In this study, it is investigated that robots may acquire environmental awareness based on perceptual attention. The purpose of this study is to model a human-like perception system architecture for a humanoid robot to establish robust and efficient communication with humans and its environment. The modeling of the robot's environment and deficiencies in the coordination of multi-modal perceptual stimuli are the main challenges for achieving this purpose. Previous works do not fulfill some of these requirements. We present a novel solution which covers a cognitive multi-modal integrated perception system. The computational framework contains features of basic feature extraction, recognition tasks, and spatial-temporal inference. In addition, it provides to help modeling perceptual attention and awareness. It is convenient that implementation works are carried out on the developed open-source software involving this architecture for social robots. The model's performance can be evaluated by various interaction scenarios. In the future, it is considered that the framework presented in this study will guide to develop next-generation cognitive models for social robots.

Keywords**—** *Cognitive perception, attention modeling, perceptual awareness, human-robot interaction.*

1. INTRODUCTION

THE perceptual ability to interact with their social environment is very important not only for humans but also for social robots. These skills depend on their spatial-temporal world model representation, perceptual awareness, and attention (focus) abilities [1]. In nature, these functions and features are biologically achieved by sensory and perceptual regions of the neocortex in the human brain. The anatomical construct of the neocortex contains two major cerebral structures such as frontal and posterior parts [2]. Cognitive skills related to perception functions are involved in the posterior part of the cerebral cortex [2, 3]. This part of the cortex is divided into three sub-regions such as occipital, parietal, and temporal lobes. The occipital lobe which hosts regions of the primary visual cortex realizes post feature extraction on visual stimuli. The temporal lobe involves pattern recognition on visual and auditory stimuli. The parietal lobe which accepts visual and somatosensory stimuli is responsible for spatial perception [3]. However, smart digital assistants or social robots have confronted severe difficulties in accomplishing these abilities during human-machine interaction experiments [4]. To establish better interaction between the social robot and the human, cognitive perception systems are currently very critical issues for human-robot interaction (HRI) studies and social robotics [5].

In daily life, as personal assistants, social robots which contain cognitive-perceptual functions can be utilized to support individuals attempting to interact with their social environment [6, 7]. Therefore, social robots should be equipped with a human-like perception system that incorporates spatial-temporal cognitive perception skills to interpret world model representation and evaluates the human-machine interaction via joint attention in a shared workspace [1].

The spatial cognition deals with environmental (situation) awareness which involves spatial perception (e.g. locations, orientations, distances, and movements) of the objects [1, 8]. The temporal cognition deals with mid-level abstraction processes that involve temporal or non-spatial encoding (e.g. color, shape) of the objects, recognition of the patterns (e.g. objects, faces, spoken words). It is a very hard problem that high-level cognitive skills that make it possible to respond to multi-modal perceptual stimuli can provide some properties having environmental awareness as well as capabilities of pattern recognition and modeling of attention [1, 9]. The modeling of the robot's social environment (e.g. the spatial world model representation and the interaction of physical behavior models) is one of the biggest problems [10]. Another major issue is the temporal perception that involves event (or situation) based world model representation [11-13]. Deficiencies in the representation of the world modeling or the coordination of multi-modal perceptual stimuli may cause interaction failures.

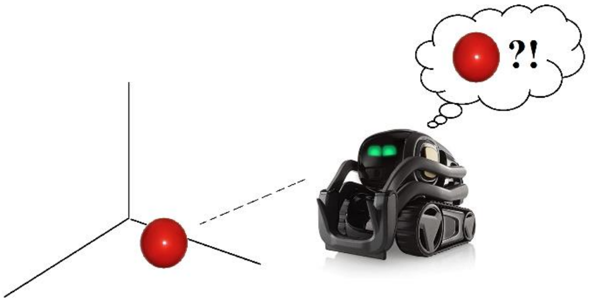


Fig.1. Cognitive perception in social robots.

The purpose of this study is to sketch a cognitive integrated multi-modal perception system for mobile robots that can be used as a social assistant. The computational approximation of the posterior neocortex may guide in which the cognitive perception system is developed in a software framework. It is expected that this solution may succeed in the case of world model representation having dynamic environments with uncertainties. This approach provides several contributions. For example, cognitive perception functions accepting multi-modal stimuli (e.g. visual, auditory, and somatosensory)

Evren DAGLARLI. is with Computer and Informatics Engineering Faculty Istanbul Technical University, Istanbul, Turkey,
(e-mail: evren.daglarli@itu.edu.tr, evrendaglarli@ieee.org)

Manuscript received Sep 15, 2020; accepted Oct 18, 2020.

perform several tasks including feature extraction, pattern recognition, and spatial perception. Perceptual coordination property includes several skills such as perceptual association (or sensory fusion) and competition between stimuli. This property plays an important role in the modeling of perceptual attention. To achieve them, supervised and unsupervised learning progress are performed on different modules of the cognitive perception system.

The paper follows with section 2 that express related works. In section 3, design principles of the computational framework of cognitive integrated multi-modal perception system realizing the world model representation model and spatial-temporal situation awareness under dynamic environments with uncertainties are presented. After all, discussion, concluding remarks, and future works are expressed in section 4.

2. RELATED WORKS

Computational cognitive architectures have been developed to solve perceptual and environmental modeling problems for social robots. The number of projects are rapidly increasing, and they promise to increase as a developing subject in the future. Certain quite interesting instances of computational architectures based on cognitive perception ensuring world model representation and attention model have been presented in recent years.

Inceoglu et al. present a visual scene representation framework for service robots to produce and keep up exact models of their workplaces for object manipulation [14]. Their framework is intended to ensure a conventional system to both humanoid and manipulator utilizing various sorts of algorithms and sources of vision data streams. Different perception algorithms processing visual data are implemented to develop and persistently refreshing a world model representation.

Kim et al. [15] investigated a curiosity-driven Dynamic World Model Learning (AWML) framework. To realize that, as visually exploring a 3D physical workplace plenty with the refinement of representative real-world agents, a curious agent building world models was built [16]. They aimed an AWML framework guided by -Progress: efficient and adaptive learning progress-based curiosity indicator and introduce that -Progress inherently provides ascend to an exploration policy. Thus, their - Progress-driven controller accomplished altogether higher AWML performance than controllers embodied with cutting edge exploration methodologies like random network distillation and model disagreement.

Riedelbauch and Henrich introduced a highly adaptable method to the human-robot collaboration framework in which a robot dynamically chooses actions contributing to a common objective from a given behavior model [17]. According to this, a world model built from eye-in-hand camera images derived knowledge on the task progress. They utilized a human-aware world model sustaining an observation for trust in stored items concerning the ongoing human presence and past assignment progress, since data created by fractional workspace perceptions is not suitable over time, as humans may interact with resources. Their contribution was a decision-making mechanism utilizing this confidence indication score to interleave task operations with an active vision to replenish the world model. Comprehensive system tests spread different kinds of human interest in various benchmark assignments through the recreation of streamlined, fractionally randomized human models. The results of their system displayed scores related to functions for various parametrizations of their human-robot joining structure.

Rosinol et al. offered an integrated model for dynamic spatial perception: 3D Dynamic environment networks. The term environment networks expressed in their framework composed of a directed network including nodes as entities in the scene (e.g., objects, walls, rooms), and edges as relations (e.g., inclusion, adjacency) between nodes [18]. This notion is extended by Dynamic environment networks (DENs) to show dynamic scenes with moving agents (e.g., humans, robots), and to incorporate dynamic data that assists planning and decision-making (e.g., spatiotemporal relations, topology at different levels of abstraction). Another novelty is to realize an automatic Spatial Perception eNgine (SPIN) to develop a DEN from visual-inertial data. They studied cutting edge strategies for human and object recognition and posture computation, and depicted how to detect objects, robot, and human nodes in crowded places. The visual-inertial SLAM and the dense human mesh tracking are incorporated by their work. Besides, they also offered algorithms to get hierarchical models of indoor areas (e.g., corridors, warehouses, hall, lobby, rooms) and their relations. Their last innovation is to show a spatial perception engine in a photo-realistic Unity-based simulator. 3D Dynamic environment networks technique seems a deep effect on action selection, task planning, human-robot interaction, long-term autonomy, and environment modeling.

Venkataraman et al. dealt with the issue of creating overall 3D models for genuine items utilizing a robot, that can expel items from the mess for better order [19]. They realized models of grasped objects using simultaneous manipulation and monitoring. Then their model processed visual data utilizing a kinematic representation of the robot to incorporate observations from various scenes and cancel background noise. For evaluation of their model, they employed a robot composed of a mobile platform with a manipulator and mounted with an RGBD camera to assemble voxelized representations of undefined items and then classify them into new categories.

Persson et al. dealt with the issue of semantic world representation by merging statistical training and item linking. Their paradigm employs a top-down item binding approach based on full permanent property scores observed from perceptual sensor data [20]. According to their study, a binding pairing model trains to sustain item entities and is verified utilizing a big set of trained manually labeled ground truth data of real-world items. To add more complicated instances, a high-level probabilistic item tracker was organized with the binding architecture and handled the tracking of occluded items with reasoning about the state of unobserved items. They displayed the performance of their system with scenarios including the shell game scenario. In this scenario, it is explained how binder items are stored by maintaining relations through probabilistic reasoning.

Martires et al. aimed a semantic scene representation paradigm based on top-down item linking utilizing an item-originated model of the world [21]. Perceptual linking processes continuous perceptual sensor data and sustains a correlation to a symbolic model. They continued the descriptions of linking to conduct multi-modal probability distributions and symbol linking model to a probabilistic logic reasoning for performing inference. In addition, they utilized stochastic correlation learning to allow the linking system to train symbolic information in the form of a set of probabilistic relationships of the world model from noisy and sub-symbolic sensor input. As exploiting the significance of relationships to reason about the state of items which are not directly detected by sensory input data, their system that incorporates perceptual linking and statistical relational learning, may sustain a semantic world model of all the items that have been perceived over time. They indicated the performance of the framework to verify their system, to execute probabilistic reasoning over multi-modal likelihood, the learning of probabilistic logical rules from linked items generated by perceptual observations.

3. COGNITIVE PERCEPTION

The main responsibilities of the cognitive integrated multi-modal perception system are to build world model representation for dynamic and uncertain environments and to support to constitute joint attention with robot's partners. In this study, the proposed framework providing initialize and establish social interaction evaluates its operations according to three main grounds such as human, robot's entity, and environment. The proposed novel structure will represent spatio-temporal relations or features of the dynamic interaction between the robot and the world model. To improve the world model of a robot, the attention model is extracted from high-level perceptual processing. This is a key element for measuring or detecting the spatio-temporal situation awareness level of the robot during interaction with its environment.

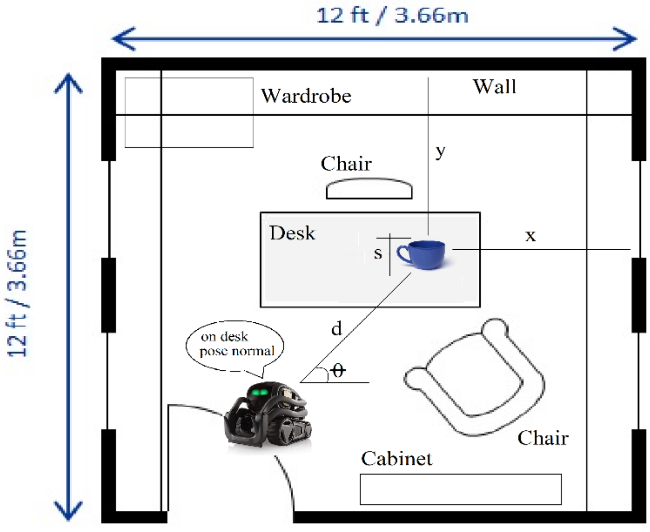


Fig.2. Spatial perception.

In order to constitute the robot's world model representation, the perceptual cognition need to deal with multi-modal perceptual data fusion and anchoring of perceptual concepts [22]. In addition, perceptual cognition of the robot is organized by its world model representation and joint attention ability involving to establish a communicational link with the other (human or robot) during the interaction. The realizing situation awareness that is the ability to resolve and discover all perceptual relations built by the robot's attention and world model representation model may be a big problem for the efficient social interaction between the human and the social robot. On the other hand, problems in manipulation between multi-modal perceptual stimuli make it difficult for the building of the robot's attention model and world model representation during interaction with the social robot. Another issue in human-robot interaction might exhibit difficulties in modeling dynamic environments and associative learning of perceptual response with multi-modal stimuli. Perceptual corruption in processes of recognition and joint attention drastically restrain human-robot interaction (HRI) in socially interactive and shared workspaces with uncertainties.

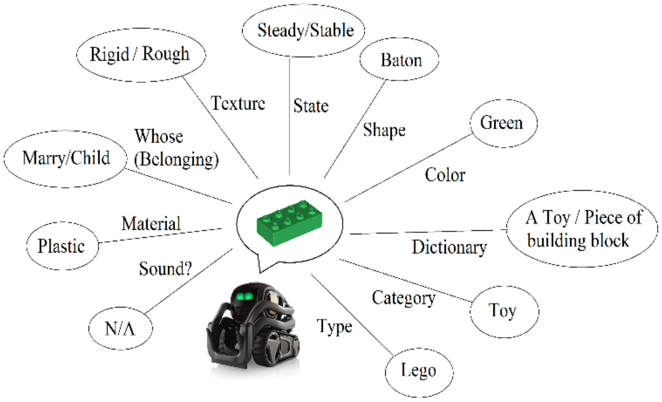
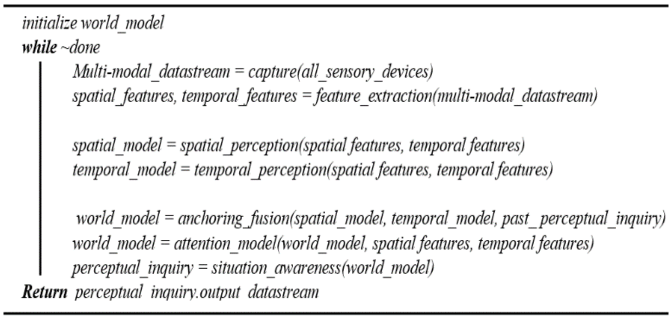


Fig.3. Temporal or non-spatial perception.

The proposed framework has multiple modalities for visual perception, auditory perception, and somatosensory (physical) perception. According to the order of data flow, the pipeline of the generic algorithm is proceeded by several main tasks such as pre-processing, feature extraction, basic perceptual operations, and processing (perceptual cognition) respectively. In our cognitive framework, the most basic functions are feature extraction and grabbers capturing multi-modal input data streams (e.g. visual, auditory, and somatosensory). It is expected that the integration of visual data sources beside of the other data streams coming from non-visual sensor modalities (e.g., microphone array, tactile sensor, laser range finder, etc.) are considered to help to achieve human-like perceptual cognition [23]. In this section, the computational framework of the cognitive perception system is described so that it constitutes the perceptual model of the robot interacting with the environment and humans. Before cognitive perception processes involving world model representation and situation awareness, some feature extraction tasks realizing segmentation, edge detection, and filtering need to be performed as data pre-processing activities. After these preliminary processes are completed, higher-level cognitive modeling is realized by developing situation awareness and attention models for a social robot so that social interaction skills such as human-like communication between robots and humans are established. Our proposed cognitive perception system introduces a generic template for world modeling which is based on anchoring knowledge between semantic concepts and perceptual relationships for social robots. It has to be analyzed about how knowledge representation can be used together with an anchoring mechanism. As a computational framework, the proposed cognitive integrated multi-modal perception system for social robot assistants annotates descriptions of recognized objects and makes anchoring these annotated representations to build a semantic network of perceptual relationships. It is considered this integrated solution in the context of social robots having perceptual cognition. In particular, we study how to constitute semantic relationships between percepts and concepts referring to features of items for social robots.



```
initialize world_model
while ~done
    Multi-modal_datastream = capture(all_sensory_devices)
    spatial_features, temporal_features = feature_extraction(multi-modal_datastream)

    spatial_model = spatial_perception(spatial features, temporal features)
    temporal_model = temporal_perception(spatial features, temporal features)

    world_model = anchoring_fusion(spatial_model, temporal_model, past_ perceptual_inquiry)
    world_model = attention_model(world_model, spatial features, temporal features)
    perceptual_inquiry = situation_awareness(world_model)
Return perceptual  inquiry.output  datastream
```

Fig.4. Pseudo algorithm related to the framework.

In this system architecture, for multi-modal perceptual data fusion and knowledge anchoring in world model representation, various machine learning methods may be utilized such as Bayesian networks, deep belief networks, support vector machine (SVM), Helmholtz machines and convolutional neural network (CNN or ConvNet). Recognition tasks include supervised learning methodologies. Generalization, classification, and clustering tasks are usually driven by unsupervised learning procedures. The hybrid formulation of methods depicts the effectiveness of the computational cognitive perception system, which realizes related perceptual-cognitive skills for a social robot.

This framework which utilizes the knowledge anchoring model and determines semantic relations via machine learning tools offers human-like perception abilities enabling social robots to continuously learn object categories and affordances for achieving augmented world model representation and situation awareness with attention model.

4. CONCLUSIONS

Social robots need to interact with users to establish joint attention enabling behavioral, emotional, and intentional synchronization. To achieve these robots to do that, they should be equipped with a human-like cognitive perception system. In the present study, it is investigated design principles of a novel computational framework of cognitive perception system for social robots as the digital smart assistant. The cognitive perception system is a framework incorporating multi-modal sensor fusion to build a more realistic world model [24].

3D physical environment and its dynamics of the world model representation are evaluated by the spatial perception model. The temporal perception model of this framework is responsible for the recognition of non-spatial and event-based features. Also, this framework has an attention mechanism assisting to handle situation awareness under the supervision of semantic memory.

It is expected that this sketched cognitive framework ensuring a human-like multi-modal perception mechanism will have robust performance against dynamic environments having uncertainties. In addition, a social robot with this proposed cognitive framework that allows constituting joint attention will be successful in social environments such as interactive areas existing humans.

The proposed architecture represents a prospective model for the perceptual cognition system of social robots. Hence, the framework can be employed by digital assistants, smart devices, or social robots. The presented structure can be further developed in the future, by incorporating approximate models of the other cortical regions related to cognitive perception.

REFERENCES

[1] Yan, Z., Schreiberhuber, S., Halmetschlager, G., Duckett, T., Vincze, M., & Bellotto, N. (2020). Robot Perception of Static and Dynamic Objects with an Autonomous Floor Scrubber. arXiv preprint arXiv:2002.10158.

[2] Freud, E., Behrmann, M., & Snow, J. C. (2020). What Does Dorsal Cortex Contribute to Perception?. Open Mind, 1-18.

[3] Bear, M., Connors, B., & Paradiso, M. A. (2020). Neuroscience: Exploring the brain. Jones & Bartlett Learning, LLC.

[4] Taylor, A., Chan, D. M., & Riek, L. D. (2020). Robot-centric perception of human groups. ACM Transactions on Human-Robot Interaction (THRI), 9(3), 1-21.

[5] Ronchi, M. R. (2020). Vision for Social Robots: Human Perception and Pose Estimation (Doctoral dissertation, California Institute of Technology).

[6] Müller, S., Wengefeld, T., Trinh, T. Q., Aganian, D., Eisenbach, M., & Gross, H. M. (2020). A Multi-Modal Person Perception Framework for Socially Interactive Mobile Service Robots. Sensors, 20(3), 722.

[7] Russo, C., Madani, K., & Rinaldi, A. M. (2020). Knowledge Acquisition and Design Using Semantics and Perception: A Case Study for Autonomous Robots. Neural Processing Letters, 1-16.

[8] Lee, C. Y., Lee, H., Hwang, I., & Zhang, B. T. (2020, June). Visual Perception Framework for an Intelligent Mobile Robot. In 2020 17th International Conference on Ubiquitous Robots (UR) (pp. 612-616). IEEE.

[9] Mazzola, C., Aroyo, A. M., Rea, F., & Sciutti, A. (2020, March). Interacting with a Social Robot Affects Visual Perception of Space. In Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (pp. 549-557).

[10] Mariacarla, B. Special Issue on Behavior Adaptation, Interaction, and Artificial Perception for Assistive Robotics.

[11] Sanneman, L., & Shah, J. A. (2020, May). A Situation Awareness-Based Framework for Design and Evaluation of Explainable AI. In International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems (pp. 94-110). Springer, Cham.

[12] Kridalukmana, R., Lu, H. Y., & Naderpour, M. (2020). A supportive situation awareness model for human-autonomy teaming in collaborative driving. Theoretical Issues in Ergonomics Science, 1-26.

[13] Tropmann-Frick, M., & Clemen, T. (2020). Towards Enhancing of Situational Awareness for Cognitive Software Agents. In Modellierung (Companion) (pp. 178-184).

[14] Inceoglu, A., Koc, C., Kanat, B. O., Ersen, M., & Sariel, S. (2018). Continuous visual world modeling for autonomous robot manipulation. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 49(1), 192-205.

[15] Kim, K., Sano, M., De Freitas, J., Haber, N., & Yamins, D. (2020). Active World Model Learning in Agent-rich Environments with Progress Curiosity. In Proceedings of the International Conference on Machine Learning (Vol. 8).

[16] Kim, K., Sano, M., De Freitas, J., Haber, N., & Yamins, D. (2020). Active World Model Learning with Progress Curiosity. arXiv preprint arXiv:2007.07853.

[17] Riedelbauch, D., & Henrich, D. (2019, May). Exploiting a Human-Aware World Model for Dynamic Task Allocation in Flexible Human-Robot Teams. In 2019 International Conference on Robotics and Automation (ICRA) (pp. 6511-6517). IEEE.

[18] Rosinol, A., Gupta, A., Abate, M., Shi, J., & Carlone, L. (2020). 3D Dynamic Scene Graphs: Actionable Spatial Perception with Places, Objects, and Humans. arXiv preprint arXiv:2002.06289.

[19] Venkataraman, A., Griffin, B., & Corso, J. J. (2019). Kinematically-Informed Interactive Perception: Robot-Generated 3D Models for Classification. arXiv preprint arXiv:1901.05580.

[20] Persson, A., Dos Martires, P. Z., De Raedt, L., & Loutfi, A. (2019). Semantic relational object tracking. IEEE Transactions on Cognitive and Developmental Systems, 12(1), 84-97.

[21] Zuidberg Dos Martires, P., Kumar, N., Persson, A., Loutfi, A., & De Raedt, L. (2020). Symbolic Learning and Reasoning with Noisy Data for Probabilistic Anchoring. arXiv, arXiv-2002.

[22] Chiu, H. P., Samarasekera, S., Kumar, R., Matei, B. C., & Ramamurthy, B. (2020). U.S. Patent Application No. 16/523,313.

[23] Wang, S., Wu, T., & Vorobeychik, Y. (2020). Towards Robust Sensor Fusion in Visual Perception. arXiv preprint arXiv:2006.13192.

[24] Xue, T., Wang, W., Ma, J., Liu, W., Pan, Z., & Han, M. (2020). Progress and prospects of multi-modal fusion methods in physical human-robot interaction: A Review. IEEE Sensors Journal.

BIOGRAPHIES

Evren Daglarli received the B.Sc. from Marmara University in 2004, the M.Sc. degree in mechatronics engineering with a focus on intelligent systems and robotics from Istanbul Technical University (ITU) in 2007, and the Ph.D. degree in control and automation engineering with a focus on computational cognitive-neuroscience, human-robot interaction from ITU in 2019. He has worked as a Research Assistant with the Department of Electrical and Electronics Engineering, Atilim University. Dr. Daglarli has many publications that are published in international journals, conferences/symposiums related to mechatronics, intelligent control systems, and robotic areas. As a Researcher, he took some duties and responsibilities in several national/international projects. Also, he has worked as a Project Engineer and the Department Manager at a private sector technology and engineering company. He is currently working as a Faculty Member and a Lecturer/Instructor with the Computer Engineering Department, Faculty of Computer and Informatics Engineering, Istanbul Technical University. Dr. Daglarli is a member of IEEE.