Design of the Integrated Cognitive Perception Model for Developing Situation-Awareness of an Autonomous Smart Agent

Evren Daglarli

Abstract— **This study explores the potential for autonomous agents to develop environmental awareness through perceptual attention. The main objective is to design a perception system architecture that mimics human-like perception, enabling smart agents to establish effective communication with humans and their surroundings. Overcoming the challenges of modeling the agent's environment and addressing the coordination issues of multi-modal perceptual stimuli is crucial for achieving this goal. Existing research falls short in meeting these requirements, prompting the introduction of a novel solution: a cognitive multi-modal integrated perception system. This computational framework incorporates fundamental feature extraction, recognition tasks, and spatial-temporal inference while facilitating the modeling of perceptual attention and awareness. To evaluate its performance, experimental tests and verification are conducted using a software framework integrated into a sandbox game platform. The model's effectiveness is assessed through a simple interaction scenario. The study's results demonstrate the successful validation of the proposed research questions.**

Index Terms— **Autonomous smart agents, Cognitive perception, Attention modelling, World model.**

I. INTRODUCTION

THE ABILITY to engage with their environment through perception is vital not just for humans but also for autonomous systems equipped with intelligent agents. These abilities rely on representations of the world model in terms of spatial and temporal, as well as perceptual cognition, situation awareness and attentional capabilities [1]. In biological systems, the cortical and cerebral lobes of the human brain play a significant role in providing these functions and characteristics. The cerebral cortex's anatomical structure consists of two primary cortical structures known as the frontal (anterior) and posterior (posterior) lobes [2]. Cognitive abilities related to perception functions are primarily located in the posterior section of the cerebral cortex [2, 3]. This region is further divided into three subregions, namely the occipital, parietal, and temporal lobes [4]. The occipital lobe, housing the primary visual cortex regions, performs various functions on visual stimuli after extracting their features [5, 6]. The temporal lobe is responsible for pattern recognition in visual and auditory stimuli [3, 5]. Spatial perception, on the other hand, is handled by the parietal lobe, which receives visual and somatosensory stimuli [4, 7]. However, autonomous systems comprising intelligent agents, digital assistants, or social robots have encountered significant challenges in implementing these capabilities during human-machine interaction experiments [8, 9]. Therefore, cognitive perception systems are currently critical issues in the fields of human-computer interaction (HCI), aiming to enhance interaction between autonomous agents and humans [10, 11].

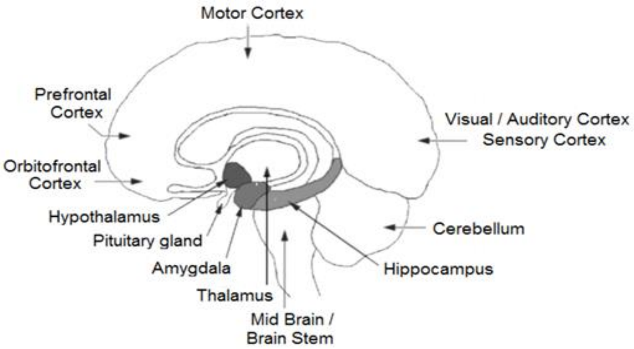

Fig.1. Regions of cerebral cortex in the human brain

In our daily lives, autonomous systems that possess cognitive perceptual abilities require various forms of interaction with their surroundings [12, 13]. Hence, it is necessary for these systems to be endowed with a perceptual system resembling that of humans, enabling them to interpret and sustain representations of the world model. Furthermore, they should be able to assess human-machine interaction through shared attention within a collaborative workspace [1, 14, 15].

Spatial cognition involves being aware of one's environment, encompassing the perception of objects in terms of their spatial aspects such as positions, orientations, distances, and movements [1, 16-18]. On the other hand, temporal cognition focuses on the intermediate processes that involve encoding non-spatial or temporal characteristics of objects, such as color and shape, as well as recognizing

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

https://orcid.org/0000-0002-8754-9527

Manuscript received June 6, 2023; accepted June 20, 2023.
DOI: 10.17694/bajece.1310607

patterns like objects, faces, and spoken words. Accomplishing higher-level cognitive abilities to effectively respond to multimodal perceptual stimuli, engage in pattern recognition, model attention, and exhibit environmental awareness presents a challenging task [1, 19, 20]. Developing perceptual models that represent the environment of an autonomous system, including spatial world models and the interaction of physical behavior models, is among the significant challenges faced [17-20]. Additionally, temporal perception plays a crucial role by incorporating event-based or situation-based representations of the world [21-23]. Errors in representing the world model or coordinating multimodal perceptual stimuli can lead to mistakes in interaction.

Fig.2. Sand-box game platform

The objective of this research is to design a comprehensive multimodal perception system for an intelligent agent that can effectively navigate and explore its surroundings within a virtual gaming platform. By leveraging the computational principles of the posterior neocortex, a software framework can be developed to guide the construction of this cognitive perception system. The proposed solution aims to demonstrate its efficacy in representing dynamic environments with uncertainties. This approach offers several notable contributions. For instance, it incorporates cognitive perceptual functions capable of processing various multimodal stimuli such as visual, auditory, and somatosensory inputs, enabling tasks such as feature extraction, pattern recognition, and spatial perception. Moreover, the system incorporates the ability to coordinate perceptual information, including perceptual association (or sensory fusion) and the management of competition between stimuli, which are vital for modeling perceptual attention. Achieving these objectives involves implementing supervised and unsupervised learning techniques across different modules within the cognitive perception system.

The subsequent section of the paper delves into the second part, where it provides an overview of the relevant research conducted in the field. Section 3 focuses on outlining the design principles governing the computational framework of the comprehensive multimodal perception system, which aims to achieve world model representation and spatial-temporal situational awareness within dynamic and uncertain environments. The article concludes with section 4, which encompasses a discussion of the findings, conclusions drawn from the study, and potential avenues for future research.

II. Related Works

Computational cognitive architectures have emerged as solutions for addressing perceptual and environmental modeling challenges in the context of autonomous aget systems. The number of projects in this domain has been rapidly increasing and is projected to continue growing in the future, signifying its significance. Notably, several commendable examples of computational architectures incorporating cognitive perception principles and facilitating world model representation and attention modeling have been introduced.

In one study, Inceoglu et al. [24] proposed a visual scene representation framework specifically designed for service robots. Their objective was to generate and maintain comprehensive workplace models to facilitate object manipulation. The framework employed various algorithms and vision data flow sources to cater to both humanoid and manipulator systems. It incorporated different detection algorithms that processed visual data, continually improving and updating the world model representation.

Another notable work by Kim et al. [25] explored a curiosity-driven framework called Dynamic World Model Learning (AWML). The study involved the development of a curious agent that constructed models of the world through visual exploration of a rich 3D physical environment [26]. The researchers focused on refining representative real-world agents to drive the AWML framework. They specifically emphasized efficient and adaptive learning progress-based curiosity indicators to guide the exploration process. The study demonstrated that the AWML framework, propelled by such progress-driven controllers, outperformed alternative approaches, including random network distillation and model mismatch, in terms of achieving higher AWML performance. These examples highlight the advancements made in computational architectures for cognitive perception, particularly in relation to world model representation and adaptive learning mechanisms.

Riedelbauch and Henrich proposed an adaptable method tailored for human-robot collaboration, where a robot dynamically selects actions that contribute to a shared objective based on a given behavioral pattern [27]. To gather information about task progress, they constructed a world model using camera images captured from an eye-to-eye perspective. Recognizing that data generated by fractional workspace perceptions can become obsolete over time due to human interaction with resources, they introduced a human-aware world model. This model maintains observations of ongoing human presence and stored item confidence in relation to past assignment progress. Their notable contribution was an action selection mechanism that utilized this confidence measure, combining mission operations with active vision to update the world model. The extensive testing of their system involved simulating various human interests by recreating modernized human models and evaluating the system's performance across different benchmark assignments, resulting in scores associated with various functions.

In a separate study, Rosinol et al. introduced an integrated model termed 3D Dynamic Environmental Networks for dynamic spatial perception [28]. This model represented the scene using environmental networks composed of nodes representing entities such as objects, walls, and rooms, along with the relationships between these nodes. To accommodate moving agents and incorporate dynamic data aiding planning and decision making, they extended this concept with Dynamic Scene Graphs (DSGs). Additionally, they developed an automated Spatial Perception Engine (SPIN) that leveraged visual inertial data to construct a DSG. The researchers focused on state-of-the-art strategies for human and object recognition, posture computation, and perception of objects, robot nodes, and human nodes in crowded environments. Their work incorporated visual-inertial SLAM and dense human network tracking. They also devised algorithms for generating hierarchical models of indoor environments, including places, structures, and rooms, along with their interrelationships. A noteworthy achievement was the demonstration of the spatial perception engine within a photo-realistic Unity-based simulator. The application of the 3D Dynamic Scene Graphics technique had significant implications for planning and decision making, human-robot interaction, long-term autonomy, and scene prediction.

Venkataraman et al. tackled the challenge of generating generic 3D models for original items using a robot capable of decluttering items to enhance organization [29]. Their approach involved creating models of grasped objects through simultaneous manipulation and tracking. These models were processed using a kinematic representation of the robot, which allowed for combining observations from multiple scenes and eliminating background noise. To evaluate their model, they employed a robot equipped with a mobile platform, a manipulator, and an RGBD camera. This setup facilitated the assembly of voxelized representations of unidentified items, which were then classified into new categories.

Persson et al. focused on semantic world representation by combining probabilistic thinking and item binding [30]. Their paradigm adopted a top-down item binding approach based on continuous attribute values obtained from perceptual sensor data. They trained a binding matching model to maintain item entities and validated its performance using a large ground truth dataset of manually labeled real-world items. To handle more complex scenarios, they integrated a high-probability item tracker into the binding architecture, enabling reasoning about the state of unobserved items. The effectiveness of their system was demonstrated through various scenarios, including a shell game scenario that showcased how binding items were preserved through probabilistic reasoning.

Martires et al. aimed to establish a semantic scene representation paradigm based on top-down item connectivity, utilizing an item-induced model of the world [31]. Their approach involved processing continuous perceptual sensor data to maintain perceptual connectivity, which correlated with a symbolic model. They extended the symbol binding model to incorporate binding annotations, enabling the execution of multimodal probability distributions and probabilistic logic reasoning for making inferences. Additionally, they employed statistical associative learning to enable the binding system to acquire symbolic knowledge in the form of probabilistic logic rules from noisy and sub-symbolic sensor input. By leveraging logical rules to reason about the state of indirectly detected items, their system, incorporating perceptual connectivity and statistical associative learning, could maintain a semantic world model of all perceived items over time. They validated the performance of their system by evaluating the framework's probabilistic reasoning on multimodal likelihood and learning probabilistic logical rules from connected items obtained through perceptual observations.

III. Background and Preliminary Materials

Currently, in conventional deep learning techniques, training data comprising input data and corresponding target (class) information can be effectively trained and subsequently evaluated with new data inputs. These deep learning algorithms demonstrate remarkable efficiency in terms of data set size, data set quality, feature extraction methods, hyperparameter selection for deep learning models, activation functions, and optimization algorithms.

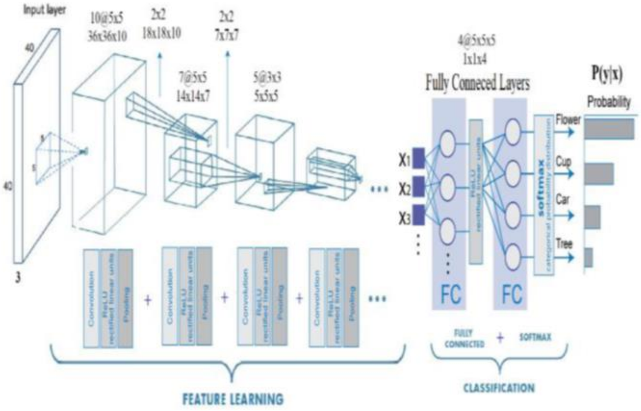

Fig.3. Convolutional neural network

A deep neural network comprising multiple layers enables the system to recognize objects at various levels of abstraction. Given the significant information processing demands associated with performing artificial cognitive functions and executing cognitive tasks, diverse deep neural network architectures such as multilayer perceptron, auto-encoder, convolutional neural network (CNN), and long-short-term memory (LSTM) recurrent neural networks are essential for learning models. These models can be further integrated within a hybrid AI framework, depending on the specific circumstances.

Regarding the traditional convolutional neural network model, which exhibits robust feature learning capabilities similar to high-level abstraction processes in the cortical regions of the human brain, the layer arrangement typically follows the sequence of [input (x) – convolution layer - ReLU

– max pool layer] illustrated in figure 3. The convolution process involves applying convolution filters (weights) arranged as a cubic tensor. The rectified linear unit (ReLU) serves as the activation function. Typically, the neural network is trained using momentum stochastic gradient descent (SGD), facilitating the hierarchical extraction of features to be encoded by the large-scale neural activations within the model [32-34].

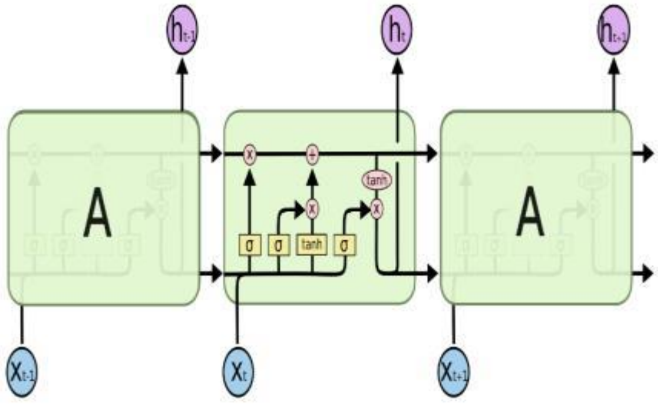

Fig.4. Long-Short Term Memory (LSTM) type neural network [35]

The LSTM (Long Short-Term Memory), a recurrent neural network (RNN) structure depicted in fig. 4, can be engineered using convolutional layers to mimic the memory mechanisms observed in the human brain. This architecture enables the LSTM to maintain short-term memory states over an extended duration, resembling episodic memory formations [32, 33, 35]. The LSTM comprises a fundamental component called the memory cell, which incorporates input, output, and forget gates. Training the neural model may involve employing backpropagation through time calculations [35, 36].

IV. Computational Model of Perceptual Cognition

In this section, we provide an explanation of the cognitive architecture employed in an open-world game/simulation environment, which encompasses the perceptual mechanism of an autonomous intelligent agent engaging with various elements in its surroundings. The key aspect of the developed cognitive integrated perception system is its ability to construct a world model representation for dynamic and uncertain environments, while also supporting the agent's attention model. Throughout this study, the cognitive integrated perception system refers to three core components: the proposed framework, the environmental elements (objects, non-player characters, etc.), and the agent's internal states. The newly proposed structure aims to capture the spatio-temporal relationships and features arising from the dynamic interaction between the autonomous intelligent agent and the world model. To achieve a comprehensive model of the world for the autonomous agent, the attention model is integrated into higher-level perceptual processing, serving as a crucial element for assessing and detecting the level of spatio-temporal state awareness during the agent's interaction with its environment.

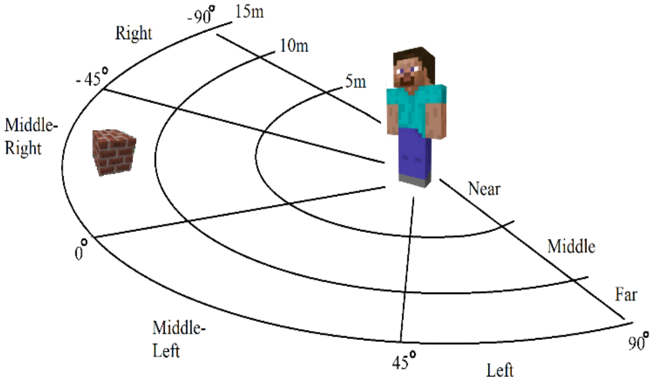

Fig.5. Spatial perception

In order to develop a comprehensive understanding of the environment and navigate through it, the cognitive architecture must create a world model that encompasses all perceptual relationships and incorporates semantic concepts [37-39]. The integration of perceptual data into this model requires the cognitive system to perform complex fusion tasks. Additionally, the attention model, which is an integral part of the cognitive process, enhances the agent's interaction with the environment by promoting situational awareness. However, the presence of non-structural dynamic uncertainties during the creation of the world model can introduce perceptual distortions, leading to limitations in the agent's ability to recognize and attend to different elements in its surroundings.

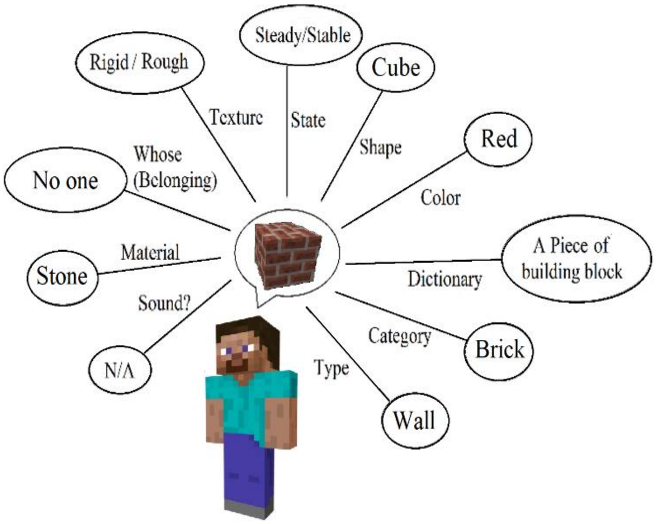

Fig.6. Temporal or non-spatial cognition

The cognitive model presented utilizes real-time image flow to enable visual perception during the autonomous agent's interaction with the environment. The general algorithm follows a sequential data flow, consisting of preprocessing, feature extraction, and basic perceptual operations such as spatial-temporal pattern recognition. The final stage of the perceptual cognition mechanism focuses on achieving situational awareness for the autonomous agent, which involves constructing the world model and implementing the attention model.

Our proposed cognitive perception system offers a comprehensive framework for achieving situational awareness in autonomous robots and intelligent agents. It encompasses various components such as spatio-temporal pattern recognition, world model, and attention model, with a strong emphasis on the relationship between semantic concepts and perceptual connections. The diagram of the cognitive architecture can be seen in Figure 7.

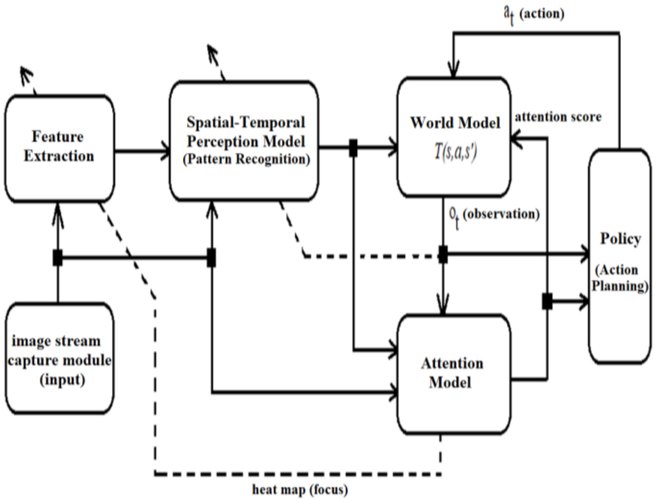

Fig.7. Cognitive architecture of the autonomous agent

Before engaging in cognitive perception processes involving the world model representation and attention model, it is necessary to establish the system parameters for the learning models of the cognitive architecture. This includes performing data attribute processing activities such as segmentation, edge/corner detection, and filtering for feature extraction. The generalization and clustering tasks in this context are guided by unsupervised learning techniques. As a result, two distinct attribute data streams, namely spatial attribute information and temporal attribute information, are obtained for utilization in the spatio-temporal pattern recognition model in figure 8.

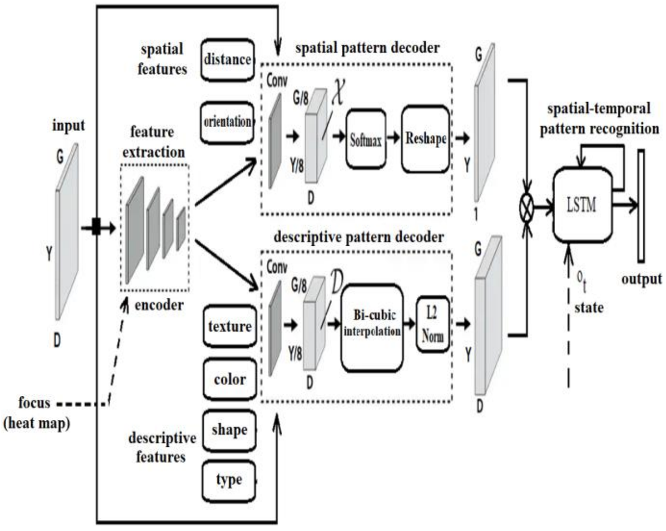

Fig.8. Feature extraction and spatial-temporal perception model

To facilitate the pattern identification process, a convolution filter (weight tensor) size is selected, consisting of a stack of four-time frames with dimensions of 83x158x3, representing the height, width, and depth of the image pattern, respectively. The initial alpha coefficient, which serves as a learning rate parameter, is set to 0.00025. The pattern recognition tasks incorporated in this model primarily employ supervised learning methodologies.

The neural network's weights are continuously updated using the backpropagation algorithm and the stochastic gradient descent optimization method. To extract features from the images and reduce their dimensionality, a VGG-type mesh is employed as the encoder. The encoder comprises convolutional layers, spatial subsampling achieved through maximum pooling, and nonlinear activation functions. The spatial and temporal features obtained from encoding are then separately forwarded to decoders that consist of convolutional layers. In the spatial decoder module, the information from the convolutional layer undergoes the softmax function and subsequently a resize operation. On the other hand, the descriptive decoder module applies bidirectional cubic interpolation, contrasting the spatial decoder, and concludes with the L2 norm on the output information of the convolutional layer. The outputs from the spatial and descriptive decoders are combined with the state information (observation) from the world model and transmitted to the long-short-term memory (LSTM) network. Finally, by adding a classifier layer to the output of this model, the spatio-temporal pattern recognition mechanism produces object recognition results.

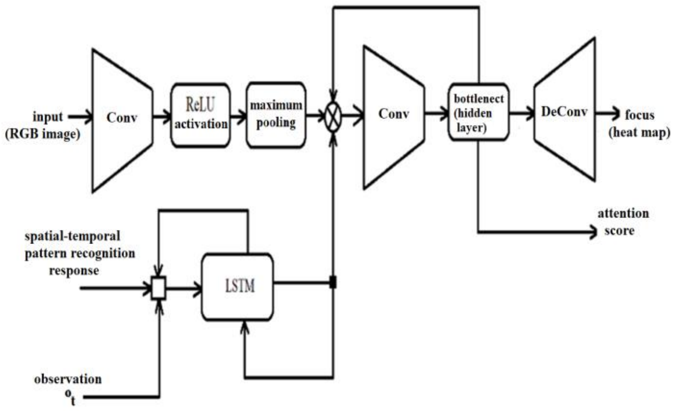

Fig.9. Attention model of the cognitive architecture

The attention model and the world model, crucial components of the situation awareness mechanism, engage in reciprocal interaction. The world model utilizes the attention value computed by the attention model and generates observation information to be relayed back to the attention model. The classification information obtained from the spatio-temporal pattern recognition module serves as a fundamental input for the attention model. Additionally, the captured RGB image stream acts as another input, which undergoes processing through convolutional neural networks, including convolutional, relu activation, and max pooling layers. Subsequently, in conjunction with the spatio-temporal pattern output, the observation response from the world model is conveyed to the long-short-phase memory (LSTM) network. The outputs from this network and the convolutional neural network are merged and transmitted to an auto-encoder network structure. The output of the convolutional layer in this auto-encoder structure is transferred to a hidden layer, also known as the bottleneck, whose output is utilized as attention points in other modules. Furthermore, the output information of the said layer is linked back to the convolutional layer input of the auto-encoder network through internal feedback. In the final stage of the autoencoder within the attention model, the output of the hidden layer is directed to a deconvolutional layer, resulting in the generation of a heat map, denoting focus information to be shared with other modules. The calculation of attention values, crucial for providing situational awareness to the autonomous agent, employs supervised learning methodologies, akin to the spatio-temporal pattern identification model within the attention model. Unsupervised learning methods are employed to obtain the heat map (focus information) as feedback data for feature extraction.

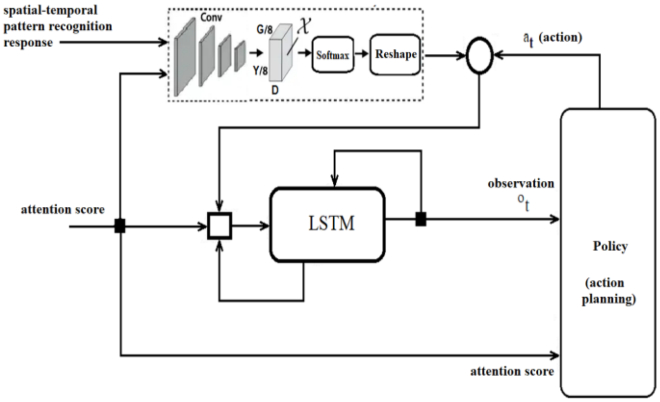

Fig.10. World model of the smart agent

The world model of the autonomous agent, which assumes the role of action planning, relies on both the action information provided by the policy module and the attention score generated by the attention module. Furthermore, the classification information derived from the spatio-temporal pattern recognition module serves as an additional input for the world model. The formation of the world model involves a convolutional layer that combines the attention score and the output of the spatio-temporal pattern recognition module. Subsequently, it undergoes softmax and size reduction operations. The disparity between the resulting output and the action information produced by the policy module is fed into an LSTM-based artificial neural network as input, which includes the attention score. The output of this neural network is utilized as observation information in the policy module, responsible for action planning. Another input to the policy module is the attention score itself. Reinforcement learning methodologies are employed in the policy module to facilitate action planning. As for the learning algorithm of the model in question, the initial discovery probability is set at 1.0, with a decay rate of 0.00001 and a minimum threshold of 0.01. The reward reduction ratio (gamma) is determined as 0.9. Through the fusion of perceptual data using hybrid machine learning tools, this architecture creates a network of semantic relationships, granting it sensing capabilities similar to the human perception system. This enables intelligent agents designed for autonomous systems to engage in continuous learning by establishing a world model and achieving situational awareness.

V. Implementation and Results

This research explores various experimental and simulation environments to enable autonomous navigation for virtual characters. While conducting experiments with physical robots in real-world settings offers realistic evaluation benefits, it also presents challenges in terms of practical implementation. Conversely, utilizing simulators provides significant advantages, including customizable degrees of freedom, noise-free environments, lower costs, and reduced risk compared to deploying mobile robots.

A. Experimental Setup and Application Scenario

In order to implement the application, it is essential to establish the experimental setup and define the initial conditions. The application area serves as the setting for the game installation, where the fundamental mechanics of the game are introduced. Prior to proceeding with the installation, it is necessary to provide an overview of the general setup.

1) General Settings

The computational workload of the system architecture was supported by a workstation PC. The PC specifications include a quad-core Intel i7 CPU running at 3.9 GHz with 8MB cache, 32GB DDR4 RAM operating at 1600MHz, an NVIDIA GeForce GTX1080Ti graphics card with 11GB video memory, and 1TB SSD and 1TB 7200rpm HDD storage. The main framework processes were executed on the Ubuntu 18.04 LTS operating system. TensorFlow, a machine learning framework, was utilized for neural network processing and deep learning applications. OpenCV libraries were chosen for image processing and computer vision tasks.

2) Sand-box game platform

For this research, Minecraft was selected as the simulation environment. It is an open-world, first-person game that revolves around resource collection (such as wood from trees or stone walls) and the construction of structures and items. Players can engage in various actions, including movement, exploration, and building within the three-dimensional voxel space of the Minecraft map [40]. This game offers an infinitely dynamic environment that can be easily modified using a simplified physics engine. It can be played as a single-player or multiplayer open-world game, without any specific objectives. Instead, each player can create their own narrative with diverse sub-objectives, resulting in complex hierarchical structures.

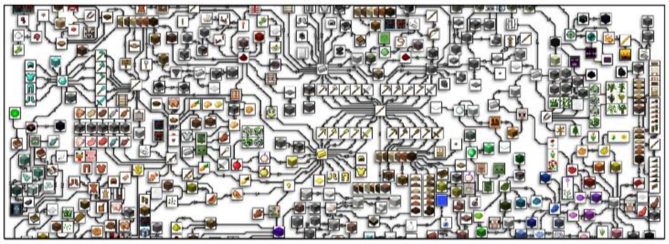

Fig.11. Tool crafting hierarchy of Minecraft game.

The virtual character model of the agent possesses extensive degrees of freedom, granting it a wide range of capabilities. Within the Minecraft environment, the agent can engage in various actions, including movement in any direction, turning, item manipulation (picking up/dropping), chopping, tool selection, and utilization. These actions, arranged in different sequences, form behaviors that represent complex tasks organized hierarchically [41]. Each behavior consists of a combination of interdependent actions that fulfill different needs and prioritize multiple objectives within the Minecraft map or world. Consequently, the difficulty level in Minecraft is influenced by the size and complexity of these hierarchical structures [42]. For instance, a navigation task involves actions such as moving forward/backward, turning right/left, which enable the virtual character model to reach specific locations or avoid obstacles and threats. Another example task, building a structure like a shelter, necessitates several actions such as item manipulation, chopping/destroying, and other equipment-related tasks involving item selection, modification, and usage [41, 42].

Fig.12. Inventory of the smart agent

In addition to engaging in missions that involve item collection and tool crafting, the experimental scenario presents more intricate and abstract hierarchies through various game features, which shape the agent's trajectory. For instance, interactive scenarios like combatting enemies, constructing shelters, and crafting tools from diverse resources for survival require extended durations or open-ended lifetimes to exhibit flexible hierarchies that allow for resource exploration. This facilitates the assembly of numerous resources and situational experiences [43, 44].

Data collection and feature extraction are crucial tasks in this context, necessitating extensive gameplay sessions with a large number of agents/humans [45, 46]. The dataset content encompasses substantial repetitions of memory utilizing observations, rewards, and actions. However, in nature, reward information is implicit and cannot be directly observed. This research paper utilizes the MineRL dataset, which stands as an extensive collection of imitation learning data, containing a staggering number of 60 million frames of human player recordings. This dataset consists of several sub-datasets and is employed to conduct experiments aimed at achieving a model that can adapt to diverse environments. Serving as a meta-dataset, it encompasses a wide range of tasks that showcase challenging problems, including exploration (such as navigation and item collection) and survival (such as tool crafting and combat).

The experimental setting involves creating a 600x600 Minecraft map that incorporates both high and low perceptual overlap. A typical naturalistic Minecraft map comprises elements such as mountains/hills, trenches, caves, valleys, rivers, lakes, trees, vegetation, rocks, and soil. Additionally, buildings like houses, shelters, and warehouses with walls, windows, doors, and furniture can be constructed using materials collected from the generated environment.

B. *Implementation Scenario and Simulation Outputs*

Moving on to the implementation scenario and simulation outputs, the proposed scenarios are realized using application platforms to assess the system's performance and validate its effectiveness in addressing the research questions at hand. To evaluate the system's results, it is necessary to capture a data stream comprising stacked image frames during the application scenarios in the experiments. In the initial stage, a snapshot is taken from the game's video stream, followed by preprocessing operations such as size reduction, grayscale image conversion, and optimization of the sampling rate. Once the system receives sensor data streams containing visual and auditory information, feature extraction tasks are performed to enhance cognitive perception skills. Subsequently, operations related to perceptual cognition, including spatial perceptions and object/event recognition, are conducted on the extracted feature data pertaining to salient attributes like color, texture, size, shape, 3D position, and audio features.

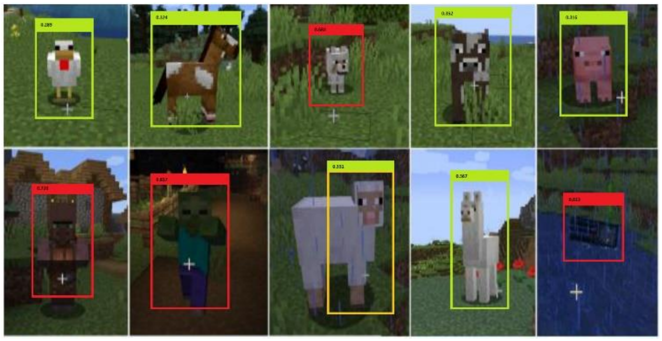

Fig.13. Snapshots from experiments using the Minecraft game platform. Entities and threat levels encountered by the autonomous agent in the experiment.

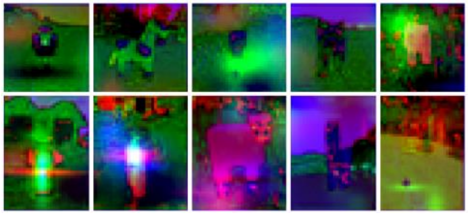

Fig.14. Focus (heat map matrix) information generated in the attention model of the cognitive perception architecture

Screenshots from the experiments implementing the interaction scenarios are presented in Figure 13. Two interaction scenarios such as "exploration" and "survival" were considered for the experiments. The "exploration" scenario involves spatio-temporal detection of hidden objects placed at various points on the map with the help of an attention model that takes into account different environmental stimuli. In doing so, it makes diagnoses based on feature information such as distance, direction, color, shape, size, and current state. The information produced by this process and attention score information contributes to the formation of the autonomous agent's world model. In performance evaluation, the number of objects discovered by the agent in the map is used as the main performance parameter. The "survival" scenario involves spatio-temporal risk detection tasks that evaluate the dangers (enemies, etc.) that the autonomous agent may encounter at various points of the island. Meanwhile, the autonomous agent builds a threat analysis-based world model in its memory using the risk detection data and feedback data from the attention model. The number of enemies defeated and the number of dangerous situations avoided by the agent in the map is used as the main performance criterion for efficiency measurement. During the experiments, the entities (non-player characters) that the autonomous agent encounters while navigating the map and their threat states are shown in Figure-13. Accordingly, the attention levels of the autonomous agent when it encounters zombie, wolf, killer and octopus characters are higher than the other entities and are 0.817, 0.683, 0.726 and 0.613, respectively. The threat levels of other entities were lower and defined as "harmless" for the autonomous agent. The attention levels of the autonomous agent when faced with chicken, cow, horse, sheep, pig, sheep, pig, and llama entities are 0.289, 0.352, 0.324, 0.331, 0.316, and 0.367, respectively. Figure-14 presents the focus (heat map matrix) information produced in the attention model of the cognitive perception architecture. Accordingly, when the autonomous agent encounters species with high threat level and species that can be considered as "harmful", more concentration (brightness in the heat map) is observed in the focus information compared to other species.

Fig.15. Minecraft in-game environmental terrain scene

Depending on the scenarios, at the end of these experiments, interaction data such as various experimental statistics (number of objects discovered on the map, scores and times for how many enemies defeated and how many dangerous situations escaped) as well as information about learning performances (accuracies, costs, scores, etc.) are obtained and presented in tables to illustrate the performance of the autonomous agent's cognitive perception system, which includes spatio-temporal pattern recognition, attention model and situational awareness models.

TABLE I
SPATIAL-TEMPORAL PATTERN RECOGNITION MODEL

Model	Accuracy	Precision	Recall	F score
Proposed model	0,73	0,69	0,71	0,72
Regional Convolutional Neural Network (RCNN)	0,67	0,64	0,63	0,66
VGG16	0,58	0,65	0,63	0,61
AlexNet	0,62	0,66	0,62	0,64

Table-1 shows the performance values for spatio-temporal image recognition. In addition to the model proposed in this study, the performance values of the regional convolutional neural network (RCNN), VGG16, and AlexNet structures were also used for comparison. Accordingly, the model proposed in the paper was found to be advantageous with an accuracy of 73%. In terms of the performance evaluation, the runner-up model was observed to be the RCNN model.

TABLE II
THE ATTENTION MODEL'S EFFICIENCY

Model	Accuracy	Precision	Recall	F score
Proposed model	0,79	0,73	0,67	0,74
Regional Convolutional Neural Network (RCNN)	0,66	0,71	0,64	0,69
LSTM	0,75	0,71	0,62	0,72
ResNet	0,70	0,69	0,63	0,67

Table-2 shows the performance of the attention model, which is a part of the state awareness function. In addition to the model proposed in this study, the performance values of regional convolutional neural network (RCNN), LSTM, and ResNet structures were also used for comparison. Accordingly, the model proposed in the paper was found to be advantageous with an accuracy of 79%. The second-best model was found to be the LSTM model.

TABLE III
PERFORMANCE SCORES OF THE WORLD MODEL

Model	Accuracy	Precision	Recall	F score
Proposed model	0,71	0,66	0,63	0,67
Regional Convolutional Neural Network (RCNN)	0,68	0,65	0,62	0,66
LSTM	0,72	0,61	0,62	0,68
Resnet	0,69	0,57	0,59	0,62

Similarly, as in the evaluation and analysis for the attention model, convolutional neural network (RCNN), LSTM, and ResNet structures were used in addition to the proposed model in order to compare the performance values of the world model, which is another part of the situation awareness function. Although LSTM is slightly ahead of the proposed model in terms of accuracy in the world model, it is worse than the proposed model in terms of sharpness. Apart from these, the sharpness and sensitivity results of ResNet lagged behind the results of the other models.

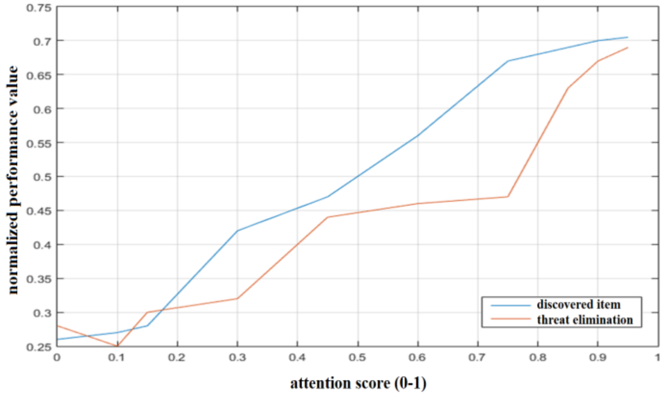

Fig.16. Normalized task performances

The normalized action performances for the scenarios in the experiments are presented in figure-16. One of the parameters related to these actions is the number of hidden objects found in the map for the "exploration" scenario. Another parameter is the number of times the autonomous agent survives the threats encountered in the map for the "survival" scenario. For the convenience of the evaluations, the total number of hidden objects and the number of threatening characters were limited to 50. On the horizontal axis of the figure, the attention score is expressed in percentages between 0-1. The scaled values expressed on the vertical axis are the rate of change over time of the normalized values found by dividing the parameters (actions) in both scenarios by the total number of those parameters. Considering this figure, it is observed that at low attention levels, the parameters in both scenarios are close to each other and as the attention level increases, the number of object discoveries, i.e., the "exploration" task performance, dominates over the "survival" task performance. However, when the attention level reaches the highest levels, it is observed that the "survival" task performance catches up with the "exploration" task performance.

VI. Conclusion

In this study, we investigate the design principles of a novel cognitive perception system for autonomous agents. A cognitive perception system is a framework that includes spatio-temporal pattern recognition, an attention model and a world model.

The spatio-temporal pattern recognition model, which evaluates the 3D spatial environment representation and the dynamics based on the actual event, effectively served as one of the main components of the cognitive perception architecture. The information produced by this architecture was used in the attention model and the world model. The attention model successfully calculated both the attention score and the focus information (heat map matrix) using the information from the world model. The world model, which has a feedback data exchange structure with the attention model, produced both the observation information required for the action planning model.

The experiments focused on two different scenarios such as "survival" and "exploration", and the main performance parameters were the number of dangers avoided in the "survival" task and the number of hidden objects found in the "exploration" task. In addition to the models in the proposed architecture, VGG16, AlexNet, ResNet, LSTM and RCNN models were also used for performance comparison. As a result of the experiments, the superiority and efficiency of the model proposed in the paper compared to other models are presented with the results obtained.

The designed framework represents the perceptual cognition system of autonomous agents in a sand-box game environment. Therefore, it can be used in other intelligent/autonomous systems or by social robots. The presented framework can be further improved in the future by integrating approximate models of other cortical regions of the human brain related to cognitive perception.

Acknowledgment

I would like to express my special thanks to Cognitive Systems Laboratory (CSL) as well as Artificial Intelligence and Data Science Research Center (ITUAI), Istanbul Technical University for their encouragement and opportunity throughout this research.

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] Chin, R., Chang, S. W., & Holmes, A. J. (2022). Beyond cortex: The evolution of the human brain. Psychological Review.

[5] Thiebaut de Schotten, M., & Forkel, S. J. (2022). The emergent properties of the connected brain. Science, 378(6619), 505-510.

[6] Li, B., Solanas, M. P., Marrazzo, G., Raman, R., Taubert, N., Giese, M., ... & de Gelder, B. (2023). A large-scale brain network of species-specific dynamic human body perception. Progress in Neurobiology, 221, 102398.

[7] Devia, C., Concha-Miranda, M., & Rodríguez, E. (2022). Bi-Stable Perception: Self-Coordinating Brain Regions to Make-Up the Mind. Frontiers in Neuroscience, 15, 805690.

[8] 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.

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

[10] Suzuki, R., Karim, A., Xia, T., Hedayati, H., & Marquardt, N. (2022, April). Augmented reality and robotics: A survey and taxonomy for ar-enhanced human-robot interaction and robotic interfaces. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (pp. 1-33).

[11] Farouk, M. (2022). Studying Human Robot Interaction and Its Characteristics. International Journal of Computations, Information and Manufacturing (IJCIM), 2(1).

[12] 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.

[13] 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.

[14] Cangelosi, A., & Asada, M. (Eds.). (2022). Cognitive robotics. MIT Press.

[15] Iosifidis, A., & Tefas, A. (Eds.). (2022). Deep Learning for Robot Perception and Cognition. Academic Press.

[16] 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.

[17] 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).

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

[19] 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.

[20] 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.

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

[22] Gu, R., Jensen, P. G., Poulsen, D. B., Seceleanu, C., Enoiu, E., & Lundqvist, K. (2022). Verifiable strategy synthesis for multiple autonomous agents: a scalable approach. International Journal on Software Tools for Technology Transfer, 24(3), 395-414.

[23] Sakai, T., & Nagai, T. (2022). Explainable autonomous robots: A survey and perspective. Advanced Robotics, 36(5-6), 219-238.

[24] 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.

[25] 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).

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

[27] 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.

[28] 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.

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

[30] 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.

[31] 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.

[32] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436.

[33] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

[34] LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, 3361(10), 1995.

[35] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[36] Sainath, T. N., Vinyals, O., Senior, A., & Sak, H. (2015, April). Convolutional, long short-term memory, fully connected deep neural networks. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4580-4584). IEEE.

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

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

[39] 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.

[40] Guss, W. H., Codel, C., Hofmann, K., Houghton, B., Kuno, N., Milani, S., ... & Wang, P. (2019). Neurips 2019 competition: The minerl competition on sample efficient reinforcement learning using human priors. arXiv preprint arXiv:1904.10079.

[41] MineRL: A Large-Scale Dataset of Minecraft Demonstrations

[42] Frazier, S., & Riedl, M. (2019, October). Improving deep reinforcement learning in Minecraft with action advice. In Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (Vol. 15, No. 1, pp. 146-152).

[43] Aluru, K. C., Tellex, S., Oberlin, J., & MacGlashan, J. (2015, September). Minecraft as an experimental world for AI in robotics. In the 2015 AAAI fall symposium series.

[44] Angulo, E., Lahuerta, X., & Roca, O. (2020). Reinforcement Learning in Minecraft.

[45] Eraldemir, S. G., Arslan, M. T., & Yildirim, E. (2018). Investigation of feature selection algorithms on A cognitive task classification: a comparison study. Balkan Journal of Electrical and Computer Engineering, 6(2), 99-104.

[46] Akinci, T. Ç., & Martinez-Morales, A. A. (2022). Cognitive Based Electric Power Management System. Balkan Journal of Electrical and Computer Engineering, 10(1), 85-90.

BIOGRAPHIES

EVREN DAGLARLI earned his undergraduate degree in electrical and electronics engineering from Marmara University. He pursued a master's degree in mechatronics engineering at Istanbul Technical University (ITU), specializing in intelligent systems and robotics. His Ph.D. degree focused on control and automation engineering, specifically computational cognitive neuroscience and human-robot interaction, also from ITU. He served as a research assistant at Atilim University in the Department of Electrical and Electronics Engineering. Dr. Daglarli has contributed to numerous publications in international journals, conferences, and symposiums within the fields of mechatronics, intelligent control systems, and robotics. He has actively participated in various national and international projects, shouldering responsibilities as a researcher. He has also held positions as a project engineer and department manager in a private technology and engineering company. Dr. Daglarli is a senior member of IEEE. Presently, he serves as a faculty member and instructor in the Computer Engineering Department at Istanbul Technical University's Faculty of Computer and Informatics Engineering. He continues his research work in the Cognitive Systems Laboratory (CSL) and the Artificial Intelligence and Data Science Research Center (ITUAI).