ISSN:2791-1322(Online)
ISSN:2791-1314(Print)
JNCBAE Volume1, Issue2 (April-June 2022)

Working Memory Model for Multi-party Interaction Using Audiovisual Perception

Asma Kanwal[1, *], Ayesha Abdullah[1]
[1]Department of Computer Science, Government College University, Lahore, Pakistan
*Corresponding Author: asmakanwal@gcu.edu.pk

Abstract: - Human mind integrates visual and natural language information simultaneously and perform context switching between different tasks and they also have ability to learn from environment. Interactive systems having cognitive capabilities can learn and acquire knowledge within dynamic environment. Their responses in environment are more human like as compared to the static conversational agents. These systems have limitations to perform multi-party interaction on the basis of cognitive constructs with audiovisual stimuli. Therefore, there is a need to design a working memory model that will serve as an executive control for the required cognitive processing of multisensory modalities perceived within the context of multiparty interaction. This agent will be capable to perform fusion of data collected from different sensors. This may allow the underlying agent to act more humanly due to its subjective nature of selection of context and attention of the afore-mentioned percepts. Such agent that has multiparty capabilities can be applied as an effective conversational agent in any multi-party environment.

Keyword: Perception, interaction, working memory

1 Introduction:

The world is large, excited and full of meanings for human learning and knowledge acquisition that is inferable through the interaction with dynamic environment (1). In the process of interaction, human perceives environment, learn continuously and generate response. Human interaction has multi-modal features e.g., gestures, speech, writing, drawing, gaze and many others to communicate in the environment. So, human interacts and communicates within environment on the basis of attention mechanism to learn and acquire knowledge (2) and perform context switching (3). For this purpose, human uses vision and natural language (4).

Machines can learn and acquire knowledge when human interacts with them. The designing of interactive machine, human-like learning and understanding mechanism should be considered (5). Some multi-modal features e.g., speech is considered in interactive machines but still many aspects (vision-based interaction) of human's vision and natural language processing is neglected (6). Man, and machine both should be aware about each other and for awareness machine should recognize each other through vision, speech, text and gestures (7). So that they can interact by using these multiple features and can understand, communicate and learn environment (8).Thus, in order to design human like interaction in machine architecture, natural language processing and vision processing is essential to understand and recognize visual and audial information (9) (10).

Many attempts have been done to design non-cognitive interactive systems e.g., Eliza (11) and ALICE (12) as these are not capable to meet the criteria. So, for human like interaction, some cognitive interactive architecture has been designed e.g. SOAR (13), IDA (14), LIDA (15) and QuBIC (16) .

There are various studies/ research proposed illustrating the interactive system. Although, all of these have provided well established grounds for the implementation of interactive system with audio and video stimuli but interactive system with inter and intra context switching within dynamic environment in not discussed in various cognitive architectures (13) (17) (18) (19).

It is evident from the all above discussion and literature review that many attempts have been proposed in order to achieve a robust and human-like interaction. The existing architecture/ model does not depict inter and intra context switching and learning within dynamic environment as due to their cognitive and

computational complexities. Thus, this research emphasizes on the modeling/ designing of interactive system in order to achieve inter and intra context switching between different tasks and learning within the dynamic environment.

The study of previous architectures shows that, there is a need to design an interactive system that has human-like cognition. The system that can learn, acquire knowledge, context switch between different tasks and have specific attention. The purpose to design a cognitive interactive system, there is a need of cognitive architecture/ model/ framework that have visual and natural language processing on the basis of which it has human-like cognitive interaction.

2 Literature Review:

Various research/ study in non-cognitive interactive systems have been done (11) (12). This research poses many issues like lack of learning, understanding, knowledge acquisition and context switching. As in Eliza, a non-cognitive architecture that makes use of pattern matching (20).

2.1. IDA:

IDA (Intelligent Distribution Agent) (21) is a conscious software agent in the sense that it implements the global workspace theory, a psychological theory of consciousness (22) (23) It was specially developed for the US Navy. IDA's task is to play the role of detailer and to facilitate the automatic assignment of sailor's duty. The language processing module in IDA is implemented as a Copycat-like architecture (24) (25) with perceptual code lets and a slip net which is a semantic net and stores domain knowledge. The perceptual code lets are triggered by surface features and the slip net passes activation. There is also a pool of perceptual code lets that are specialized to recognize particular pieces of text, and production templates that are used by code lets to build and verify understanding. Collectively they establish an integrated sensing system for IDA that allows it to understand, recognize and categorize.

2.2. Eliza:

Eliza (26) is a program that makes natural language conversation with a computer possible. It was implemented by Joseph Weinbaum in 1966. Eliza program was originally designed to pretend to be a psychiatrist. Eliza is based on a stimulus-response model which is a very simple pattern recognition model. Eliza knowledge is stored as a script in a text file consisting of patterns and corresponding responses. The input is examined for the presence of a keyword. If keyword is found, then input is analyzed on the basis of decomposition rules which are triggered by that keyword. Response is then generated by re-assembly rule that is related to selected decomposition rule. Eliza uses some tricks to successfully perform in conversations. One of its fundamental tricks is personal pronoun transformations. Eliza gives the user an illusion of some understanding. It simply matches the keyword and generates a standard response.

2.3. ALICE

ALICE is an acronym for Artificial Linguistic Internet Computer Entity (27). ALICE knowledge is stored in AIML files (28). AIML files consist of simple stimulus-response modules called categories. These <category> contains a <pattern>, or "stimulus," and a <template>, or "response." The brain of ALICE consists of roughly 41,000 such categories. AIML software stores the stimulus- ALICE by using AIML pretends to be intelligent and self-aware but experience with ALICE indicates that most casual conversation is "stateless," that is, each reply depends only on the current query, without any knowledge of the history of the conversation required to formulate the reply.

2.4. Jabberwocky

Jabberwocky (29) is another top-rated internet dialogue system. Jabberwocky learns from every interaction it has done with its users. It stores everything people have said to it and tries to use again those statements by matching them to the user's input. Since Jabberwocky has no preset rules for conversation. Thus, the only aim of Jabberwocky is to simulate human chat in an interesting, entertaining and amusing way. Response categories in a tree managed by an object called the Graph master which implements a pattern storage and matching algorithm. When a bot client inputs text as a stimulus, the Graph master searches the categories for a matching <pattern>, along with any associated context, and then outputs the

associated <template> as a response. The Graph master is compact in memory and permits efficient pattern matching time.

3 WMM-AVP Model:

This research proposes a working memory model for cognitive agents that used for interaction using audiovisual perception (30) which is capable to interact and collaborate with humans and other machines in human-like way. Thus, making it possible for an agent to wield flexible and adaptive control over action. The proposed detailed architecture (Figure.1) that puts together all the modules and integrates their synergic interrelations is introduced below:

3.1 Working Memory Model:

The limited capacity processing unit that is used for interaction with information either percept from senses or from long term memory (31). Working memory (32) works as buffer for internal activities and provides temporary storage while manipulating information needed for understanding, learning and reasoning for language. It is used to manipulate and organize sensed information in different processes for further proceeding. Working memory surpass this information coming from sensory memory and perceptual memory module to other modules of model. Internally, working memory consists of several other modules that are described as follows:

3.2 Executive Control:

It is the supervisor inside working memory model and responsible for the integration of information between other modules of working memory. It controls the cognitive process and regulates information between them.

3.3 Episodic Buffer Module:

It the subsystem of the working memory and maintains the cues of the incoming perception. This module also analyzes the context behind the generated cues and interrelate them. It has two sub-modules that manage and then analyze the cues.

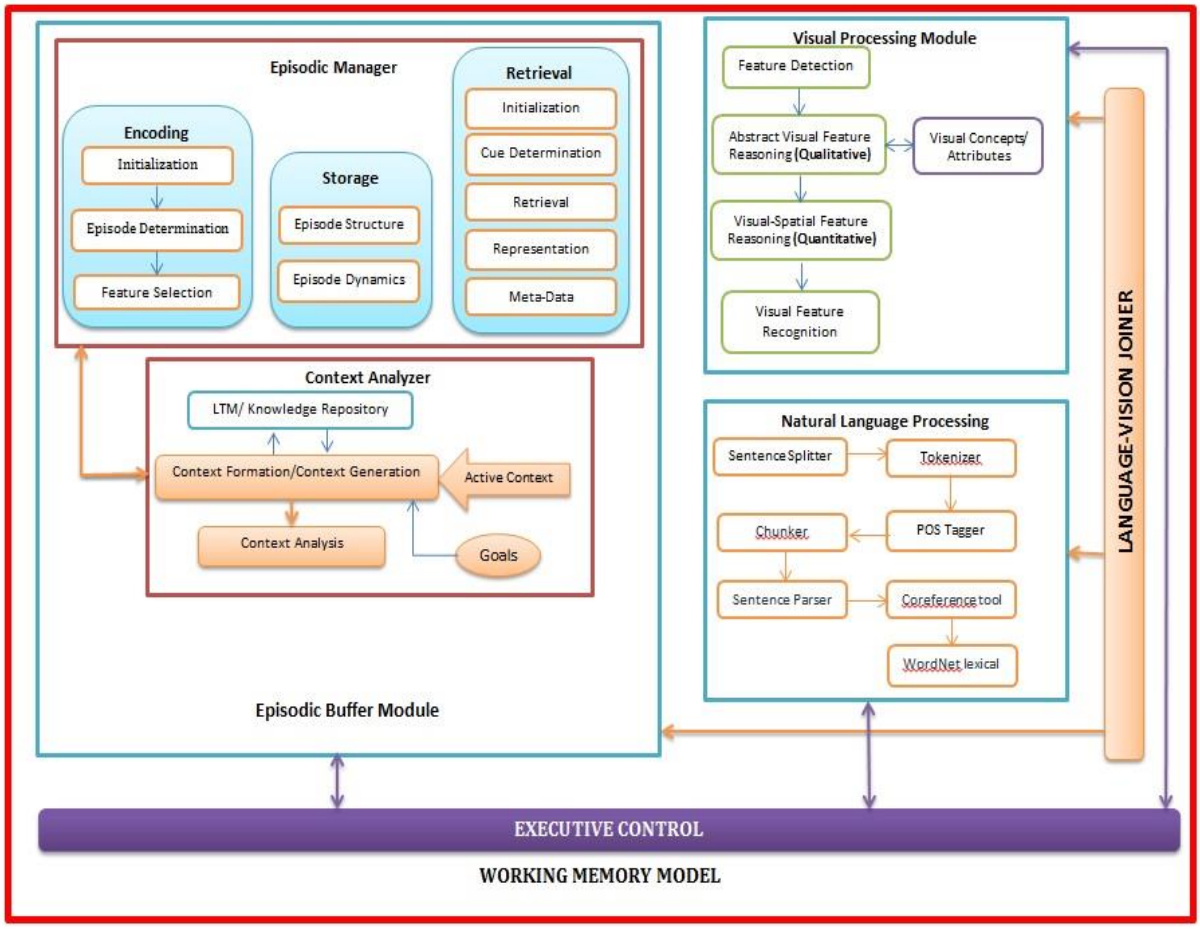

Figure 1:

This figure is defining the working of Working Memory Model, which comprises on different sub modules like Visual Processing Module to detect and recognize the visual input, episodic manager to manage episodic memory, context analyzer for defining the context of ongoing conversation. Natural Language Processing unit for understanding of incoming sentence.

3.3.1 Episodic Manager

This sub-module is like a buffer that manages the incoming stimuli.

3.3.2 Encoding

How the incoming stimuli initiate and what type of features in cues can be extracted, it is the functionality of encoding module. It determines the cues and depending upon the relevant feature this module stores it.

3.3.3 Storage

After the proper encoding of the cues within the buffer episodic manager stores them. The stored cues can be decay and remove from storage module.

3.3.4 Retrieval

It is responsible for providing cues that are being used by other cognitive modules. This module defines which type of key is used to trigger the episode and represents

.

3.3.5 Context Analyzer

This module generates context and analyze them on the behalf of cues from the episodic manager. It formulates context and update the manager so that it could change the feature of selection of cues. For the analysis, this module uses the active context, knowledge repository or Long-Term Memory (LTM) and set of goals.

4 Natural Language Processing Module:

The NLP module process (33) the textual information of the attentive signal in working memory. It processes the audio data and relates it to generated context and other modules of the working memory on the behalf of Language-Vision Joiner. This model processes the audio and textual input in series of following steps:

4.1 Sentence Splitter

It can detect that a punctuation character marks the end of a sentence or not.

4.2Tokenizer

This module segments an input character sequence into tokens. Tokens are usually words, punctuation, numbers, etc.

4.3 POS Tagger

The Part of Speech (POS) Tagger marks tokens with their corresponding word type based on the token itself and the context of the token. A token might have multiple POS tags depending on the token and the context. It uses a probability model to predict the correct POS tag out of the tag set.

4.4 Chunker

Text chunking consists of dividing a text in syntactically correlated parts of words, like noun groups, verb groups, but does neither specify their internal structure, nor their role in the main sentence.

4.5 Sentence Parser

It parses the sentence depending on the semantics of the language.

4.6 Co-reference Tool

The Co-reference Tool links multiple mentions of an entity in a document together. The implementation is currently limited to noun phrase mentions; other mention types cannot be resolved.

4.7 WordNet Lexical

It is the database that contains syntax and semantics of the language and helps the NLP Module to analyze the input perception.

5 Visual Processing Module*:*

The module processes the visual part of the attentive signal using image processing strategies. It switches the vision processed information to executive control for task switching. Mid-Level image processing is done in this module of the visual processing module to extract the attributes of the objects. It applies high level image processing on the extracted objects and creates association.

5.1 Feature Detection

This module gets a part of the image along with the corresponding segmentation mask and returns the features, which are then used for recognition and/or learning.

5.2Visual Concepts/Attributes

This module contains the learned attributes of the images that help the processing system that define the image after the detection.

5.3 Abstract Visual Feature Reasoning

The abstract symbolic visual representation is the neutral, stable medium useful for general reasoning (34). Symbols denote an object, some visual properties of that object, and qualitative spatial relationships

between objects. The meaning of the symbols is dependent on their context and interpretation rather than how the symbols are spatially arranged. The symbols are compostable using universal and existential quantification, conjunction, disjunction, negation, and other predicate symbols.

5.4 Visual-Spatial Feature Reasoning

Visual imagery cannot generate a depictive representation directly from qualitative, abstract symbols without first specifying metric properties, such as location, orientation, and size. Finally, from a computational perspective, there are some spatial reasoning tasks where reverting from qualitative representations to quantitative information is necessary for either efficiency or simply to infer new information.

5.5 Visual Feature Recognition

Abstract Visual Feature Reasoning Module and Spatial-Visual Feature Reasoning Module helps the model to recognize the visual features of the incoming visual stimuli.

5.6 Language-Vision Joiner:

It forms a link between vision and language. The perception generated on the behalf of context analyzer language-vision joiner validates the link.

5.7 Feature Comparison:

Table.1 define the complete feature-based comparison between existing conversational systems and proposed study.

Table 1: Feature Comparison of Interactive Systems

Features	Eliza	ALICE	Jabberwacky	Model
Selective Attention on the behalf of Unified Signal	No	No	No	Yes
Interaction with old and new user in different ways	No	No	No	Yes
Relevant Context Generation	No	Yes	Yes	Yes
Multiparty Interaction	No	No	No	Yes
Interaction on the behalf of A-V Perception	No	No	No	Yes
Commonsense knowledge	No	No	No	Yes
Association Between Audio and Visual	No	No	No	Yes
Learning using A-V Perception	No	No	No	Yes
Memorization of new concepts	No	No	Yes	Yes

6 Results:

A lot of work has been done in the field of cognitive interactive systems but still there is a gap in the work of this field. No model shows the multi-party interaction on the behalf of audio and visual perception. This thesis work proposed a Working Memory model to show how these two input perceptions can be used to perform multi-party interaction.

6.1 Meeting Room Case-Study:

WMM-AVP is deployed in any social and interactive environment and encountered many people with different voices and faces. WMM-AVP has to perform multiparty interaction with in real or virtual environment. Let suppose, WMM-AVP is deployed in Interactive Office Meeting Environment where agent is leading the meeting. Therefore, agent has to perform interaction at a same time with many people that have different voices and faces. Agent can perform multiparty interaction on the behalf of selective attention but have to discriminate the persons. This discrimination between the different persons would have involved perceptual abilities.

In interactive environment, just like Meeting Room, the agent should be effective to keep attendees in constant concentration and appropriate feedback to give them a proper response and as mentioned earlier that agent can perform multiparty interaction so it should have context switching.

In our case study, the proposed solution is dependent on selective attention and a unified signal. The signal contains the content of basic face and voice recognition on the behalf of localization. As this is not in the scope of this work therefore, signals from other cognitive regularities will be assumed accordingly at the development and implementation time, as per our requirement. The steps of a complete cognitive cycle are hierarchically listed down, but only the parts of the proposed architecture for Working Memory Model are elaborated.

When person or agent interact with agent on the basis of audio, the person or agent who is communicating within Meeting Room is located and attention is generated towards them. Now the audio and visual cues received from the environment through speakers and camera deployed in the Meeting Room. Therefore, the agent starts sensing that people are try to interact with him.

6.2 Sensing:

"Sensory input stimuli, external or internal, from real or virtual environment are received through the sensors and not interpreted yet." Audio and visual sensor (Kinect microphone and camera) situated in the agent receives the image frames and acoustic signals emerging from the persons in the environment. This sensory input is raw and saves as nodes in SM. This input contains all the visual object in the visual field and all the sound in its acoustic range. At this stage, there is no differentiation between the human voices and noise. Similarly, all the objects in the visual field of WMM-AVP are sensed and no discrimination between face and non-face objects. Now WMM-AVP is getting all information of Meeting Room including table, chairs, walls and people but not able to distinguish between faces and objects. The sensory stimuli sensed by the sensors are passed through the network.

6.3 Detection and Recognition:

"Distinguishing the stimuli as face and voice on the basis of facial and acoustic features. Decision of whether the input face is known or unknown by consulting the long-term memory."

In detection phase, categorization of face and non-face objects is done, and the detected attendee's faces are passed to the common knowledge base and Long-Term Memory (LTM) where these faces serves as nodes. Parallel to detecting, basic feature extraction is also done from face and voice of the person are extracted (35). These features are responsible for setting attention by unified signal and also helpful for WMM-AVP to recognize and analyze about the context in which the person is communicating.

6.4 Working Memory:

"The percept, including some of the data plus the meaning, is stored in preconscious buffers of IDA's working memory." A buffer is required to manage previous and new generated context between different interactors and analyze them. To analyze the percepts that are in the form of audio and visual model need

an episodic manager and a context analyzer. When WMM-AVP interact in multiparty environment just like Meeting Room Scenario it can perform context switching when different persons are interacting with it.

6.5 Executive Control:

The executive control is responsible for supervision of information integration from other sub-systems of WM and to perform other executive function by regulating and controlling other cognitive process. It monitors the activities of other cognitive modules and may override the response or activity of cognitive modules of unconscious layer." In WMM-AVP, the executive control monitors the context switching between different concepts. It also synchronizes the audio and visual percepts coming during the same time stamp. When agent interact within multiparty real or virtual environment the control over the cognitive modules and model sub-system and led by this module.

6.6 Episodic Manager:

A buffer is required in the model that can maintain the incoming cues and transfer them to the required module or decay them depending on the algorithm (not in the scope of this research word). In WMM-AVP, the input audio and visual stimuli are kept in buffer and one by one transmitted to the context analyzer to perform its task. Episodic manager helps the agent to maintain the incoming cues while memory modules are busy in performing analysis or processing.

6.7 Context Analyzer:

"This module is designed in such a way that it can keep track of different situations and distinguish them by analyzing the context whether it is the matter of talking with an unknown person or a known one. This module helps the agent to behave in different ways by analyzing the context of familiar and strange persons to imitate the real-world behavior." In WMM-AVP, this module works as an analyzer that evaluate the context of the communicated on the behalf of audio and visual perceptions percept from the environment coming from the episodic manager. From the generation of contexts, this module consults the Long-Term Memory (LTM) so that on the behalf of already generated context agent can communicate or may generate new context.

6.8 Language-Vision Joiner:

"The L-V Joiner module is used to generate association between things. In case of users/persons, it can generate association between persons by using audio and visual perception depending on the information available in those user profiles." This module helps the agent WMM-AVP to deduce the new knowledge by joining the audio percepts and visual percepts. When agents perform joining of the language from audio percepts and vision from the visual percepts then it is capable to communicate and interact with the environment. This module in the WMM-AVP increases the interactivity to dialogue on the behalf of audiovisual perception.

7 Conclusion:

Since language is central to so many of human activities, a major research focus has been put by NLP researchers on the ability of intelligent system that can interact in human language. Surveying the literature about the dialogue based interactive agents leads to the conclusion that none of the research is particularly in the work of multiparty interaction using audiovisual perception. Multiparty interaction using audiovisual perception is a paradigm that cannot be ignored since it is an evolutionary advantage emerged in humans. This indicates that it might be beneficial if applied to software agents. The proposed model is an attempt in this regard. Effective utilization of language is entangled with the general cognitive abilities. The proposed model is a cognitive model that works as interactive agent for social interactions with humans. We have found encouraging results, by the experiments based on this approach and we are

in a better position to find some new ways for solving the problem of man-machine communication in natural way.

8 References:

1. Jokinen, K. (2003). Natural Interaction in Spoken Dialogue Systems. New Jersey : Lawrence Eribaum Associates, Human-Computer Interaction: Theory and Practice Part II, 4,730-734.

2. Yousfi-Monod, M and Prince, V. (2007). Knowledge Acquisition Modeling Through Dialogue Between Cognitive Agents. France.

3. Oborski , P. (2005). Man-Machine Interactions In Advanced Manufacturing Systems. Warsaw, Poland.

4. Melichar, M. (2008). Design of Multimodal Dialogue-based Systems. ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE.

5. Te'eni, D., Carey, J. M., and Zhang, P. (2005). Human-Computer Interaction: Developing Effective Organizational Information. John Wiley & Sons.

6. Karray, F., et al. (2008). Human-Computer Interaction: Overview on State of the Art. Waterloo : International Journal On Smart Sensing And Intelligent Systems, 1.

7. Arora, S., Batra, K., and Singh, S. (2009). Dialogue System: A Brief Review. Punjab Technical University, Kapoorthala.

8. Seidel, H., (2009). Multimodal Computing and Interaction – Robust, efficient, and intelligent processing of text, speech and visual data.

9. Harris and Marry, D. (1985). Introduction to Natural Language Processing. USA : Prentice Hall.

10. Hanheide, M. (2006). A Cognitive Ego-Vision System for Interactive Assistance. Bielefeld.

11. Weizenbaum, J. (1966). ELIZA A Computer Program For the Study of Natural Language Communication Between Man and Machine. Cambridge, Mass : Communication of the ACM.

12. Hutchens, J. L. and Alder, M. D. (1998). Introducing MegaHAL. Proceedings of the Human-Computer Communication Workshop. 271-274.

13. Laird, J. E., Newell, A and Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. Intelligent Systems Laboratory Xerox Palo Alto Research Center, University of Michigan. The Artificial Intelligence and Psychology Project, 1-64.

14. Franklin, S. (2000). Learning in “Conscious” Software Agents. Department of Mathematical Sciences, Institute for Intelligent Systems. Memphis.

15. Franklin, S., et al. (2007). LIDA: A computational model of global workspace theory and development learning. AAAI Fall Symposium on AI and Consciousness: Theoretical Foundations and Current Approaches. Memphis.

16. Qazi, W. M. (2013). Modeling Cognitive Cybernetics from Unified Theory Of Mind Using Quantum Neuro-Computing For Machine Consciousness. Lahore, Punjab, Pakistan.

17. Anderson, J. R. and Lebiere, C. (2003). The Newell test for a theory of cognition. Cambridge University Press, Behavioral And Brain Sciences, 587-637.

18. Sun, R. (2004). The CLARION cognitive architecture: Extending cognitive modeling to social simulation. Cambridge University Press: New York., Cognition and Multi-Agent Interaction.

19. Franklin, S. (2007). A foundational architecture for artificial general intelligence. Amsterdam : IOS Press Amsterdam, The Netherlands, The Netherlands, 2007. Proceedings of the 2007 conference on Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms: Proceedings of the AGI Workshop. 36-54.

20. Weizenbaum. (1976). Computer Power and Human Reason: From Judgement to Calculation. San Francisco : W.H. Freeman and Company.

21. Franklin, S. (2003). IDA, A Conscious Artifact?. Journal of Consciousness Studies, 10, 47-66.

22. Baars. (1988). A Cognitive Theory of Consciousness. Cambridge : Cambridge University Press.

23. (1997). In the Theater of Consciousness: The Workspace of the Mind. Oxford : Oxford University Press.

24. Hofstadter, D. and Mitchell, M. (1995). The Copycat Project: A model of mental fluidity and analogy-making. New York : Basic Books, Inc. New York, 2, 205 – 267.

25. Holland, John. H. (1986). A Mathematical Framework for Studying Learning in Classifier Systems. Elsevier Science Publishers B. V. Amsterdam, The Netherlands, The Netherlands, Physica D: Nonlinear Phenomena, 2,307-317.

26. Weizenbaum. (1966). ELIZA-A computer program for the study of natural language communication between man and machine. New York : ACM New York, NY, USA, Communications of the Association for Computing Machinery, 9, 36-45.

27. Wallace, R. S. (2009). The Anatomy of A.L.I.C.E. Bethlehem, PA, USA : Springer Netherlands, 181-210. Part III.

28. Anatomy. (2004). ALICEBOT. A.L.I.C.E. Artificial Intelligence Foundation, Inc, 2004. http://www.alicebot.org/anatomy.html.

29. Carpenter, R. (1997). jabberwacky. http://www.jabberwacky.com/.

30. Khan. M. A, Abbas, S., Raza, S.A., Khan, F., Whangbo, T. K. (2022). Emotion Based Signal Enhancement through Multisensory Integration Using Machine Learning, Computers, Materials and Continua, 71(3), 5911-5931.

31. Barkowsky, T. (2001). Mental Representation and Processing of Geographical Knowledge - A Computational Approach. Cognitive Systems. Bremen : FB Mathematik and Informatik.

32. Baddeley, A. D. (2002). Is Working Memory Still Working? European Psychologist, 7,85-97.

33. Raza, S. A., Kanwal, A., Rehan, M., Khan, K. A., Muhammad, A., and Asif, H. M. S. (2015). ASIA: Attention driven pre-conscious perception for socially interactive agents, 2015 International Conference on Information and Communication Technologies (ICICT), Karachi, Pakistan.

34. Khan, Q., Abbas, S., Khan, M. A., Fatima, A., Alanazi, S., and Sabri, N. (2021). Modelling Intelligent Driving Behaviour Using Machine Learning, Computers, Materials and Continua 68(3):3061-3077.

35. Abbas, S., Alhwaiti, Y, Fatima, A., Khan, M. A., Ghazal, T. G., Kanwal, A., Ahmad, M., and Sabri, N. (2021). Convolutional Neural Network based Intelligent Document Recognition. Computers, Materials and Continua 70(3):4563-4581