Artificial Cognition, from the Ground Up[1]

Dario Pasquali, Carmine Miceli, Giulia Belgiovine, Francesco Rea, Fulvio Mastrogiovanni, Giulio Sandini, Alessandra Sciutti

***Abstract*—We present our work-in-progress efforts toward enabling the humanoid robot iCub to collect and learn from unsupervised first-hand experiences autonomously. Such personal, situated, embodied, and developmental-inspired Artificial Cognition would be crucial to enable social robots to dynamically adapt and interact in everyday life scenarios.**

I. Introduction

For robots introduced in social contexts, it will be mandatory to promptly learn and adapt to the environment, the people therein, and the emerging social dynamics. Most approaches aim at empowering robots with robust, task-specific AI models. However, as toddlers develop skills by interacting with the environment without precise aims, we posit that emergent context-specific abilities should be learned from the continual, first-hand interaction of robots with the environment and the humans in it. Robots should be equipped with the minimal toolset, i.e., a cognitive architecture, to develop over time their personalized, embodied, and situated Artificial Cognition (ACo) [1] incrementally aiming at making the system "survive", i.e., preserving internal consistency [2], rather than fulfilling specific tasks. The recent *iCog Initiative* [4] gathered several international researchers in the joint effort to realize an open-source, developmental-inspired cognitive architecture for the iCub humanoid robot. The focus on development, the centrality of interaction, and the attempt to identify the minimal elements needed to enable the emergence of Artificial Cognition, led us to choose a different solution from the existing cognitive architectures [3]. Hence, we embarked on developing a novel *emergent enactive* cognitive architecture from the ground up. We are applying an incremental approach, designing and integrating one cognitive component at a time, toward building a minimal system able to integrate and learn from multimodal personalized experiences exploitable in everyday activities.

II. Architecture

We designed a preliminary architecture (see Figure 1, center) as a system that actively observes the environment and the agents therein – a fundamental ability in humans' development [2]. Through continual active observation, humans can deduce novelty and regularities, which can support more complex behavior. Similarly, we design our architecture to identify relevant patterns from raw or low-level observations. We embodied the architecture into the ICubHead robot, an actuated 6 DoF head of the humanoid robot iCub, mounted on a stationary 3D-printed upper body (see Figure 1, left). It can look around, moving the neck and eyes. It mounts two cameras, a stereo microphone, a speaker, and LEDs to produce facial expressions. The architecture comprises four components. The ***Multimodal Perception*** module processes raw images (RGB, 640x480, 30 fps) and audio (44.1 kHz) in real-time, extracting five low-level features inspired by humans' early-age development, namely *number of faces, number of people gazing toward iCub (mutual gaze), quantity of motion, illumination* – from the images – and *right and left root mean square (RMS)* – from the audio. We considered such core features the minimal necessary toolset to bootstrap the architecture abilities. The robot behavior is led by a social motive, i.e., seeking other interactive agents, one of the fundamental drives leading infants' behavior [5]. For this purpose, the ***Embodied Behavior*** module controls the ICubHead, alternating between static and gaze-wandering phases looking for human faces to track. The observation of different portions of the environment would also enable the robot to collect more generalized observations. Then, the ***Episode Segmentation*** module integrates and time-synchronizes the raw data, the low-level core features, along with the ICubHead joint values and the embodied behavior state into *Event* objects, timestamped snapshots of the system's external and internal state. It also oversees the aggregation of *Events* into *Episodes*, i.e., sequential events belonging to the same scenario and context. We implemented the episode segmentation by seeking novel perceptual experiences through a curiosity-driven approach, inspired by humans' exploratory motive [2]. The ICubHead tracks the mean and standard deviation of the *number of faces, quantity of motion, and right and left audio RMS*; each new event is compared against the current episode's averages seeking outliers, i.e., observations deviating more than 3 standard deviations from the episode mean. If less than two components are outliers, the new event is kept within the active episode; otherwise, the current episode is encoded in memory, and a new one is started. Lastly, the ***Episodic Memory*** module stores, in an SQLite database, the timestamped event-related core features, the episode-related means and standard deviations, and the paths leading to image frames, and WAV audio files. Also, the memory leverages the SlowFast [6], for the raw images, and the VGGish [7], for the stereo audio, pretrained models to represent episodes in an embedding space efficient to be memorized and compared.

[1] Research supported by the Project *Future Artificial Intelligence Research (FAIR), code PE000013* funded by the European Union - NextGenerationEU PNRR MUR - M4C2 - Investimento 1.3 - Avviso Creazione di *"Partenariati estesi alle università, ai centri di ricerca, alle aziende per il finanziamento di progetti di ricerca di base"*.

D. Pasquali, C. Miceli, G. Belgiovine, F. Rea, and A. Sciutti are with the COgNiTive Architecture for Collaborative Technologies (CONTACT) Unit of the Italian Institute of Technology (IIT), Genoa, Italy.

F. Mastrogiovanni and C. Miceli are with the DIBRIS department of the University of Genova (UniGe), Genoa, Italy.

G. Sandini is with the Robotics Brains and Cognitive Sciences (RBCS) Unit of the Italian Institute of Technology (IIT), Genoa, Italy.

Emails: {name}.{surname}@iit.it

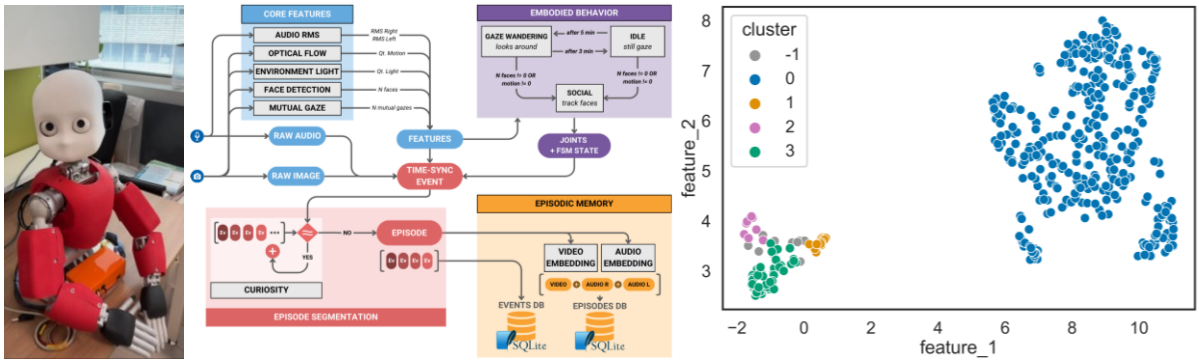

Figure 1. **(Left)** The iCubHead robot; **(Center)** The cognitive architecture; **(Right)** HDBSCAN clustering of UMAP-reduced multimodal embeddings.

Embeddings should highlight the similarities between episodes, letting recurrent and novel experiences emerge. The next development iteration will focus on *(i)* the memory consolidation process, building long-term memory from daily experiences, and *(ii)* the loop closure, leveraging episodes recognition and prediction to support the robot's proactive behavior. However, before going any further, we deemed it necessary to validate the current episode segmentation and memory encoding approach.

III. Validation Experiment

Wwe placed the ICubHead robot in our open office (Figure 1, left). The area hosts four researchers' desks, while up to 10 people could pass by during the day. The room becomes empty at lunchtime and during events involving the group. The ICubHead was positioned close to our break area, where we keep treats to be shared among the group. This zone is highly crowded after lunch, while it is sporadically visited during the day. We kept the ICubHead active for 4 consecutive days, from 10 AM to 4 PM (24 hours in total). Colleagues were asked to keep their habits to ensure the naturalness of the data collection.

IV. Results & Discussion

The architecture collected 449 (*M=112, SD=83)* episodes, distributed resembling the open space occupation of those days: two researchers were present on *day1* (19 eps.); four on *day2* (217 eps.), and *day3* (130 eps.); lastly, on *day4* (83 eps.), a single person was in the room with others coming sporadically. On average, 90% of the episodes were collected between 1 and 2 PM, consistent with our usual after-lunch break. Episodes in such a timeframe were also the shortest – on average, 1-minute long w.r.t. 25 minutes in the remaining time. We speculate that higher population in the room would generate a higher variability in the multimodal perception, causing more outliers and, hence, more scattered episodes. Then, we analyzed the multimodal embeddings produced for the episodes, looking for clusters revealing recurrences in the observations. We normalized and concatenated the video embeddings with the audio ones, averaging between channels, obtaining a 528-long vector. Then, we applied a 2-component dimensionality reduction with UMAP and clustered the resulting features with the HDBSCAN algorithm as it is robust against clusters of different densities and dimensions. The model identified four clusters (Figure 1, right, outliers in gray), achieving a silhouette score of 0.65. To characterize the clusters concerning the low-level multimodal perception features, we fitted a mixed effect model for each one: we considered the outliers-filtered HDBSCAN 'cluster' as a fixed factor and added the random effect of the 'day'. *Cluster 0* showed the highest number of faces (B=1.77, t=17.72, p<0.001) and audio RMS (B=6.86, t=11.59, p<0.001) w.r.t. the other clusters; such scenes could be related to ***populated-loud*** episodes where multiple people socially interact in the room. *Cluster* 1 had the highest illumination (B=0.15, t=7.48, p<0.001), while *cluster 3* presented the lowest quantity of motion (B=-0.44, t=5.61, p<0.001); we speculate they group ***not-populated-quiet*** episodes where either there was nobody in the room or the few ones were at their desks, causing the low amount of motion. Summing up, the low-level multimodal features effectively segmented the episodes and a meaningful representation emerged by unsurprisingly clustering the raw-perception embeddings. Our next step will be improving the episode segmentation robustness, reducing the number of episodes in high-variability periods. An effective architecture should be aware that a novel episode despite being salient, does not necessarily represent the beginning of a new experience. Still, the system was able to learn the primitive distinction between *populated-loud* and *not-populated-quiet* episodes from the ground up. This knowledge will be crucial to equip the architecture with a long-term memory able to categorize the just-experienced episode fostering proactive behavior – e.g., to decide whether it is meaningful to interact – or detecting misalignment compared to the past.

References

[1] G. Sandini, A. Sciutti, and P. Morasso, "Artificial cognition vs. artificial intelligence for next-generation autonomous robotic agents," *Front. Comput. Neurosci.* **18**, 1349408 (2024).

[2] D. Vernon, *Artificial Cognitive Systems: A Primer* (The MIT Press, 2014).

[3] I. Kotseruba and J. K. Tsotsos, "40 years of cognitive architectures: core cognitive abilities and practical applications," *Artif. Intell. Rev.* **53**, 17–94 (2020).

[4] "The iCog Initiative," <https://icog.eu/> (24 June 2024).

[5] C. V. Hofsten, "Action, the foundation for cognitive development," *Scand. J. Psychol.* **50**, 617–623 (2009).

[6] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast Networks for Video Recognition," in *2019 IEEECVF Int. Conf. Comput. Vis. ICCV*, (IEEE, Seoul, Korea (South), 2019).

[7] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, et al., "CNN Architectures for Large-Scale Audio Classification," arXiv:1609.09430 (arXiv, 2017).