1-2023

Demonstrating multi-modal human instruction comprehension with AR smart glass

Mudiyanselage Dulanga Kaveesha WEERAKOON
Singapore Management University, mweerakoon.2019@phdcs.smu.edu.sg

Vigneshwaran SUBBARAJU

Tuan TRAN

Archan MISRA
Singapore Management University, archanm@smu.edu.sg

Citation

WEERAKOON, Mudiyanselage Dulanga Kaveesha; SUBBARAJU, Vigneshwaran; TRAN, Tuan; and MISRA, Archan. Demonstrating multi-modal human instruction comprehension with AR smart glass. (2023). *2023 15th International Conference on COMmunication Systems and NETworkS COMSNETS: Bangalore, January 3-8: Proceedings*. 231-233.
Available at: https://ink.library.smu.edu.sg/sis_research/7797

Demonstrating Multi-modal Human Instruction Comprehension with AR Smart Glass

Dulanga Weerakoon
Singapore Management University
mweerakoon.2019@phdcs.smu.edu.sg

Vigneshwaran Subbaraju
*IHPC, A*STAR, Singapore*
vigneshwaran_subbaraju@ihpc.a-star.edu.sg

Tuan Tran
University of Colorado Boulder
tutr6272@colorado.edu

Archan Misra
Singapore Management University
archanm@smu.edu.sg

***Abstract*—We present a multi-modal human instruction comprehension prototype for object acquisition tasks that involve verbal, visual and pointing gesture cues. Our prototype includes an AR smart-glass for issuing the instructions and a Jetson TX2 pervasive device for executing comprehension algorithms. With this setup, we enable on-device, computationally efficient object acquisition task comprehension with an average latency in the range of 150-330msec.**

***Index Terms*—Human-AI Collaboration, Referring Expression Comprehension, Visual Grounding, Multi-Modal Networks, Pervasive Systems**

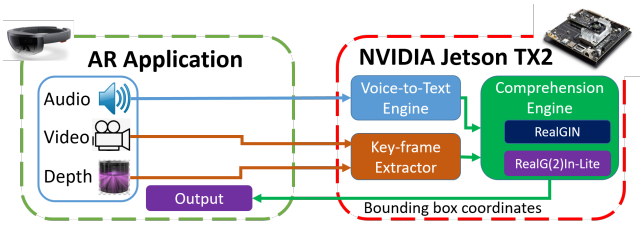

Fig. 1. Overall architecture of the proposed comprehension pipeline

I. Introduction

Target acquisition is a common task in Human-Robot interaction which involves an AI agent or a robot to identify a target object referred through a natural human instruction. When issuing instructions, humans typically use a combination of verbal and gestural cues to identify the target object. In such cases, a common interaction could be captured in an ego-centric viewpoint (As seen from the perspective of the human agent who is issuing the instruction). Comprehending such natural multi-modal instructions using an AI agent is crucial in multiple applications such as virtual shopping assistants, industrial collaborative robots and social robots assisting elderly people. Algorithms designed for comprehending these instructions typically use Referring Expression Comprehension (REC) DNN models such as [1]–[6]. In addition, weerakoon et al. [7] showed that employing an additional pointing gesture improves comprehension accuracy significantly, further emphasizing the benefit of enabling multi-modal human instruction comprehension. Recent trends in AR (Augmented Reality) and VR (Virtual Reality) technologies have paved the way for the introduction of several smart-glass devices such as Microsoft Hololens [8]. Although these smart glasses are equipped with computing resources, executing these complex and large multi-modal DNN algorithms on-device is non-viable with current hardware specifications. To counter that, we introduce a prototype for on-device execution of such comprehension models. Key to our prototype is a hybrid setup with AR-powered smart glass for capturing human instruction and an additional embedded computing device for executing the DNN comprehension models. In particular, our prototype consists of a Hololens device running an AR application which records audio, video and depth sensor frames of issued instruction from the ego-centric viewpoint. These captured multi-modal sensor data are then transferred to an NVIDIA Jetson TX2 [9] for identifying the target object. Jetson TX2 executes a DNN comprehension engine which combines visual, verbal and gestural cues and yields the bounding box coordinates of the target object. The bounding box coordinates are then transferred back to the Hololens for visualization. With this proposed hybrid setup, we enable real-time and computationally efficient execution of object acquisition task comprehension with latency in the range of 150-330msec.

II. Comprehension Pipeline

A. Hardware Components

Figure 1 depicts the overall architecture of the system. Primarily, we use two hardware components in our prototype. Human instructions are captured by using a Microsoft Hololens device from the ego-centric viewpoint. Hololens is an AR smart-glass device with a number of sensors including an RGB camera, short depth sensor and IMU sensors. In addition, Hololens runs on a customized Windows OS, allowing some computational capabilities. We use Hololens's embedded microphone, RGB and short depth sensor to capture user's verbal instruction, visual information and pointing gesture respectively. Current hardware specifications of Hololens are inadequate to run our computationally intensive DNN comprehension pipelines. Hence, we employ NVIDIA TX2 as an additional embedded device specifically to comprehend the human instructions captured by the Hololens device. TX2

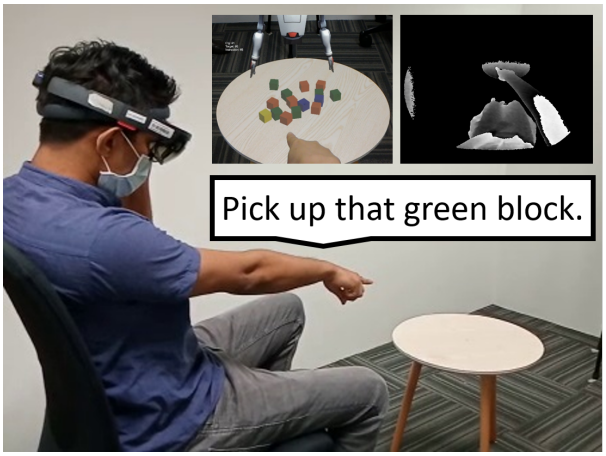

Fig. 2. AR Application: Virtual objects viewed via the Hololens along with the verbal instruction and depth sensor output

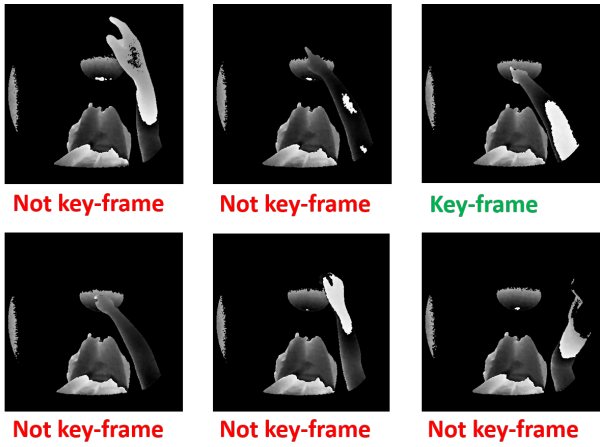

Fig. 3. A series of captured depth frames while issuing an instruction

device is one of the fastest and power efficient embedded AI computing devices with 256 NVIDIA cuda cores, Dual-Core NVIDIA Denver 2 64-Bit CPU, Quad-Core ARM® Cortex®-A57 MPCore and 8GB of LPDDR4 Memory. We will now introduce the rest of the components of the proposed prototype in the following sections.

B. Recording and Transmitting User Instructions at Hololens

We developed an AR application using Unity which executes on Hololens for showcasing a typical table-top object acquisition task. As shown in Figure 2, user wearing the Hololens will see an AR environment with different coloured virtual objects placed on a table. The user then issues a verbal instruction along with an accompanied pointing gesture to pick a particular object from the set of virtual objects. This captured user instruction is then transmitted over TCP network to the TX2 device.

C. Voice-to-Text Engine

Transmitted audio instruction is processed at the Jetson TX2 to convert to text using a real-time speech-to-text model called Picovoice cheetah [10]. Cheetah model allows accurate and on-device speech-to-text conversion while being compact and computationally efficient.

D. Key Frame Extractor

Hololens sends the captured RGB video frames along with the corresponding depth frames to the TX2 device. However, not all of these frames are necessary to capture the target object. Figure 3 shows a series of depth frames captured while a user is issuing an instruction. A typical scenario involves, a) user moving his pointing hand towards the target, b) user steadily pointing towards the target and c) user retracting his pointing hand from the target. A *key-frame* is defined to be the duration in which the user's pointing hand is steadily pointing towards the target object. Identification of these key-frames are done through a lightweight classification network with 3-CNN layers and a consequent 2-neuron fully connected output layer with a softmax activation function. This network takes the depth sensor stream as input and identifies depth *key-frames* and the corresponding image frames. This network is trained for 10 epochs by using a balanced corpus (4000 depth images, 70% used for training and 30% used for testing) of depth frames that are labelled to be either key-frames or not key-frames. This module incurs a latency of 15msec while achieving a classification accuracy of 89.91% on the test split. Rest of the comprehension engine is only executed on these identified key-frames.

E. Comprehension Engine

For the comprehension engine we employ a REC model, which takes the converted text instruction from Voice-to-Text Engine, RGB and depth key-frames to predict the bounding box of the target object. Early work on REC [1], [2] proposes several DNN models suitable for comprehending object acquisition task instructions. However, these works generally utilize multiple stages and incur significantly higher computational requirements on a resource-constrained pervasive device. Contrary to these models, [3] proposed a single-stage architecture termed RealGIN capable of significantly reducing the computational requirements and enabling on-device inferencing on Jetson TX2. However, RealGIN model only accepts verbal and visual image as its input. To further accommodate the additional pointing gesture, [6] proposed RealG(2)In-Lite model; a compact and computationally efficient model which additionally accepts a short depth frame for accommodating pointing gesture input. For our comprehension engine, we will be employing both these models as variants in our prototype. Finally, the predicted bounding box coordinates will be sent back to the Hololens through the TCP network. Let us now further examine the two DNN model variants used in our prototype.

1) RealGIN: This model only supports verbal and visual cues. RealGIN consists of a bi-directional LSTM network [11] for extracting language features, RESNET [12] network-based backbone for extracting visual features, several language-

TABLE I
PERFORMANCE COMPARISON OF REALGIN AND REALG(2)IN-LITE ON JETSON TX2 DEVICE

Model	Accuracy (%)	Latency (msec)	Energy (mJ)
RealGIN	81.7	330	2310
RealG(2)In-Lite	78.8	155	852.5

guided attention layers to fuse language and visual features and a regression network for generating bounding box coordinates. A major percentage of RealGIN's computations are used for the RESNET-based visual backbone.

2) RealG(2)In-Lite: This model was proposed to further reduce the computational complexity of RealGIN and also to accommodate the additional pointing gesture via depth frame input. RealG(2)In-Lite additionally accepts a depth frame as an input which is concatenated along with the RGB frame. To reduce the computational complexity at the visual backbone, this variant uses a Shufflenet network [13] for extracting the visual features. For identifying the pointed location, RealG(2)In-Lite consists of a 2-layer regression network branch. Regressed pointed location from this branch is then used as an input for a gesture-guided attention layer to assign higher attention weights to the region where the user has pointed. The remainder of the language pipeline, language-guided attention layers and regression network for generating bounding box coordinates are identical to that of RealGIN model.

F. Empirical Results

Table I shows the comprehension accuracy vs average latency and energy consumption of RealGIN and RealG(2)In-Lite. Here, accuracy was measured on the COSM2IC dataset [6] with the same accuracy metric that was used in their evaluation. Based on this accuracy, it is evident that both these model variants achieve roughly comparable accuracy. However, we observed that RealG(2)In-Lite runs around 2x faster than RealGIN. Moreover, RealG(2)In-Lite consumes close to 4x less energy than the RealGIN model. Thus, it is evident that RealG(2)In-Lite is significantly more computationally efficient than RealGIN.

III. CONCLUSION

With our hybrid setup of Microsoft Hololens and Jetson TX2, we have demonstrated the ability of real-time execution of object acquisition task comprehension. Object acquisition tasks typically involve multiple modalities with verbal instruction, visual scene image and pointing gestures. To demonstrate such a scenario, we developed an AR application running on Hololens device which projects a set of virtual objects on to a table-top. A user wearing the smart-glass issues a verbal instruction along with an accompanied pointing gesture which is then transmitted over to the TX2 device for comprehension. At the TX2 device, we employ two key comprehension models RealGIN and RealG(2)In-Lite to predict the bounding box coordinates of the target object. Consequently, our proposed system achieves task comprehension with an average latency of 150-330msecs. Although our current prototype does not yet represent a complete wearable device (due to the requirement of an additional embedded device for running DNN models), we expect that future advances in smart glasses may allow running these complex DNN models without the necessity of additional computing devices.

IV. ACKNOWLEDGMENTS

This work was supported partially by 1) National Research Foundation, Singapore under its NRF Investigatorship grant (NRF-NRFI05-2019-0007), 2) Ministry of Education, Singapore under its Academic Research Fund Tier-1 grant (19-C220-SMU-008) and 3) Agency for Science, Technology and Research (A*STAR), Singapore under its AME Programmatic Funding Scheme (Project #A18A2b0046). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore or the Ministry of Education, Singapore or the A*STAR, Singapore.

REFERENCES

[1] S. Yang, G. Li, and Y. Yu, "Dynamic graph attention for referring expression comprehension," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4644–4653.

[2] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, and T. L. Berg, "Mattnet: Modular attention network for referring expression comprehension," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1307–1315.

[3] Y. Zhou, R. Ji, G. Luo, X. Sun, J. Su, X. Ding, C.-W. Lin, and Q. Tian, "A real-time global inference network for one-stage referring expression comprehension," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[4] F. I. Dogan and I. Leite, "Using depth for improving referring expression comprehension in real-world environments," *arXiv preprint arXiv:2107.04658*, 2021.

[5] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 11–20.

[6] D. Weerakoon , V. Subbaraju, T. Tran, and A. Misra, "COSM2IC: Optimizing real-time multi-modal instruction comprehension," vol. 7, no. 4, pp. 10 697–10 704, 2022.

[7] D. Weerakoon, V. Subbaraju, N. Karumpulli, T. Tran, Q. Xu, U.-X. Tan, J. H. Lim, and A. Misra, "Gesture enhanced comprehension of ambiguous human-to-robot instructions," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 251–259.

[8] "Hololens (1st gen) hardware," https://learn.microsoft.com/en-us/hololens/hololens1-hardware, accessed: 2022-12-01.

[9] "Jetson tx2 module," https://developer.nvidia.com/embedded/jetson-tx2, accessed: 2021-10-08.

[10] "Picovoice," https://picovoice.ai/, accessed: 2022-09-12.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[13] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6848–6856.