Towards Designing a Context-Aware Multimodal Voice Assistant for Pronoun Disambiguation: A Demonstration of GazePointAR

Jaewook Lee
University of Washington
Seattle, USA

Jun Wang
University of Washington
Seattle, USA

Elizabeth Brown
University of Washington
Seattle, USA

Liam Chu
University of Washington
Seattle, USA

Sebastian S. Rodriguez
University of Illinois at Urbana-Champaign
Urbana, USA

Jon E. Froehlich
University of Washington
Seattle, USA

Answer: 173 divided by 17 is 10.17647.

Answer: Landfill: waste such as food waste. Recycling... Compost...

Answer: Ursula von der Leyen is the President of the European Union.

Answer: Perhaps there's an ongoing recreational or communal event.

Figure 1: Example interactions with GazePointAR. Pronouns such as "*this*," "*these*," "*she*," and "*there*" are automatically resolved by using real-time gaze tracking, pointing gesture recognition, and computer vision. See demo video for interactive examples.

ABSTRACT

Voice assistants (VAs) like Siri and Alexa have transformed how humans interact with technology; however, their inability to consider a user's spatiotemporal context, such as surrounding objects, drammatically limits natural dialogue. In this demo paper, we introduce *GazePointAR*, a wearable augmented reality (AR) system that resolves ambiguity in speech queries using eye gaze, pointing gesture, and conversation history. With GazePointAR, a user can ask "*what's over there?*" or "*how do I solve this math problem?*" simply by looking and/or pointing. We describe GazePointAR's design and highlight supported use cases.

CCS CONCEPTS

• **Human-centered computing** → **Interaction techniques**; **Mixed / augmented reality**.

KEYWORDS

augmented reality, multimodal input, voice assistants, gaze tracking, pointing gesture recognition, LLM

UIST '23 Adjunct, October 29–November 01, 2023, San Francisco, CA, USA

ACM ISBN 979-8-4007-0096-5/23/10.
https://doi.org/10.1145/3586182.3615819

ACM Reference Format:
Jaewook Lee, Jun Wang, Elizabeth Brown, Liam Chu, Sebastian S. Rodriguez, and Jon E. Froehlich. 2023. Towards Designing a Context-Aware Multimodal Voice Assistant for Pronoun Disambiguation: A Demonstration of GazePointAR. In *The 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23 Adjunct), October 29–November 01, 2023, San Francisco, CA, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3586182.3615819

1 INTRODUCTION

Pronouns such as "*this*", "*it*", "*s/he*", and "*there*" are common in human speech [6] but can cause ambiguity [2, 4] (*e.g.,* "*What is this?*; or "*Who is she?*"). To resolve speech ambiguities, humans employ non-verbal contextual cues such as gaze and gestures [4, 7]. For example, a person may point at an item while asking the question *"How much is this?"*. However, state-of-the-art voice assistants (VAs) such as Amazon Alexa, Google Assistant, and Apple Siri do not yet utilize this spatiotemporal context, which can result in unnatural dialogue or unanswerable queries.

Multimodal speech interaction has long been a focus in HCI, perhaps best marked by Bolt's seminal "*Put That There*" system in 1980 [2]. Recently, researchers have explored multimodal interaction in AR for pronoun disambiguation using various input modalities, including head gaze [8], touch [5], and pointing gesture [10]. However, recent work such as *WorldGaze* [8] and *Nimble* [10] share similar limitations: they rely on Wizard-of-Oz (WoZ) setups, employ only one additional modality alongside speech, and are designed for smartphones rather than always-available head-worn displays.

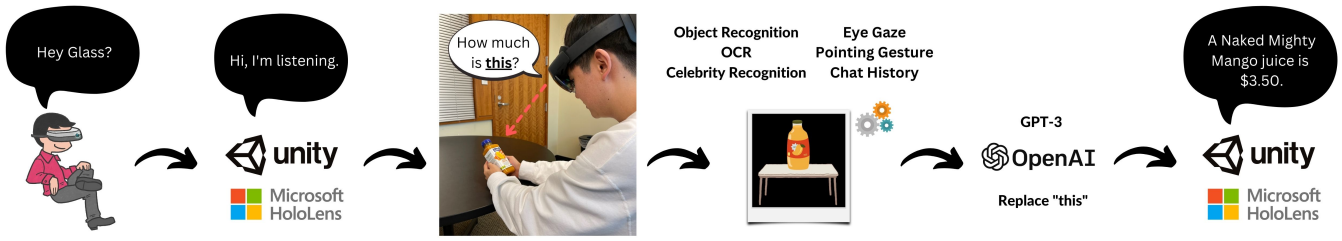

Figure 2: System overview and implementation details of GazePointAR.

In this demo paper, we introduce *GazePointAR*, which leverages advances in real-time computer vision and large language models (LLMs) to create a fully-functional context-aware VA for wearable AR. To disambiguate pronoun usage, GazePointAR utilizes eye gaze, pointing gesture, and conversation history along with an LLM for speech dialogue (Figure 2). GazePointAR supports a variety of pronouns and queries from *"What does this mean?"* to *"What can I cook with these?"* (Figure 1). Our key contributions include: (1) a non-WoZ context-aware VA for pronoun disambiguation, and (2) highlighting use cases of a real-time, always-available wearable VA. At the UIST demo session, we will invite attendees to try their own open-ended queries with our fully-functional prototype.

2 GAZEPOINTAR

We implemented GazePointAR on the *HoloLens 2* with the *Mixed Reality Toolkit* (MRTK[1]). Below, we describe key components.

Activating GazePointAR. To activate GazePointAR, the user first states the trigger phrase, "*Hey Glass*," to which GazePointAR replies, "*Hi, I'm listening.*" Once the user speaks a query, GazePointAR analyzes the input string for pronouns (see Table 1).

Capturing and Analyzing the User's Field-of-View. If a pronoun is detected, GazePointAR takes a 1080p image of the user's field-of-view, which is stored temporarily until the user receives a query response (*i.e.,* one VA cycle). This image is sent concurrently to three machine learning models: *Google Cloud Vision's Object Localization* and *Optical Character Recognition (OCR)* models [3], as well as *Amazon Rekognition's Celebrity Recognition* model [1].

Creating the ML Hierarchy. After receiving JSON responses from the ML services, GazePointAR merges them into a hierarchical structure, with object detection and celebrity recognition results as the parent layer. The child layer, comprised of OCR results, is connected to parent bounding boxes that have 70% pixel overlap (a threshold tuned empirically). Each parent can have up to five children objects ranked by bounding box size. This approach allows GazePointAR to prioritize important information, such as product names, which tend to be larger in the user's field-of-view.

Gaze Tracking and Pointing Gesture Recognition. To capture the user's eye gaze and pointing gesture, we customized MRTK's built-in modules. For gaze, we designed a white sphere that follows the user's gaze from a fixed distance, allowing us to retrieve 3D gaze coordinates and provide visual feedback. For pointing, we implemented a finger-pointing gesture to complement the base palm-pointing gesture [4]. Performing a pointing gesture creates a ray that extends away from the user's hand until a collision is detected.

[1]https://learn.microsoft.com/en-us/windows/mixed-reality/mrtk-unity/mrtk2

Pronouns		
Nominal Demonstrative	Third Person	Adverbial Demonstrative
this	it	here
that	he, him	there
these	she, her	
those	they, them	

Table 1: Pronouns GazePointAR supports.

As the HoloLens captures an image, GazePointAR simultaneously logs the locations of both the user's gaze and pointing gesture. To convert 3D gaze and pointing gesture coordinates to their corresponding pixel locations on the captured image, we project their 3D coordinates onto a normalized 2D space, invert the y coordinate, then scale by the captured image size.

Pronoun Replacement. Using the generated ML hierarchy and pixel coordinates of gaze and pointing gesture, GazePointAR begins assembling a coherent phrase to replace the user-spoken pronoun. If the pronoun is singular, the system computes whether any user input coordinate falls within parent bounding boxes. If so, GazePointAR takes the parent objects' children (*i.e.,* OCR results) and generates the following phrase: "*[parent] with text that says [children]*". Otherwise, GazePointAR finds the five closest OCR results to any user input coordinate to create the following phrase: "*[OCR Result 1] [OCR Result 2] ... [OCR Result 5].*" If the pronoun is plural, each user input coordinate is converted to a bounding box with width and height equivalent to half of the captured image's width and height respectively.

GazePointAR assembles the final query by combining the user-spoken query, the ML-generated phrase, and recent query history (the last five queries). This modified query is processed by *OpenAI's GPT-3* [9] and the result is displayed as text and read aloud.

2.1 Example Usage

Enabled by our techniques, GazePointAR supports a wide range of ambiguous queries, such as: (1) using gaze to solve a math equation ("*Can you solve **this** equation?*"); (2) using gaze and pointing gesture to get information about users' surroundings ("*What's happening over **there**?*"); (3) cooking with GazePointAR (*What can I cook with **this**?*; Is **it** healthy?); and (4) retrieving information about a celebrity (*Who is **s/he**?*"). See the demo video for more.

3 FUTURE WORK AND CONCLUSION

While GazePointAR is capable of answering a wide range of ambiguous queries, we do not yet know how users will interact with a context-aware VA. In the future, we plan to conduct a user study to compare GazePointAR to commercial VA systems and examine when context-aware ambiguous query support is most appropriate and effective. We envision GazePointAR to facilitate more ambiguous, natural human-VA dialogue.

ACKNOWLEDGMENTS

This work has been supported by an NSF GRFP Fellowship.

REFERENCES

[1] Amazon AWS. 2023. Amazon Rekognition. https://aws.amazon.com/rekognition/
[2] Richard A. Bolt. 1980. "Put-That-There": Voice and Gesture at the Graphics Interface. In *Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques* (Seattle, Washington, USA) *(SIGGRAPH '80)*. Association for Computing Machinery, New York, NY, USA, 262–270. https://doi.org/10.1145/800250.807503
[3] Google Cloud. 2023. Vision AI. https://cloud.google.com/vision
[4] Holger Diessel and Kenny R. Coventry. 2020. Demonstratives in Spatial Language and Social Interaction: An Interdisciplinary Review. *Frontiers in Psychology* 11 (2020). https://doi.org/10.3389/fpsyg.2020.555265
[5] Jaewook Lee, Sebastian S. Rodriguez, Raahul Natarrajan, Jacqueline Chen, Harsh Deep, and Alex Kirlik. 2021. What's This? A Voice and Touch Multimodal Approach for Ambiguity Resolution in Voice Assistants. In *Proceedings of the 2021 International Conference on Multimodal Interaction* (Montréal, QC, Canada) *(ICMI '21)*. Association for Computing Machinery, New York, NY, USA, 512–520. https://doi.org/10.1145/3462244.3479902
[6] Geoffrey Leech, Paul Rayson, and Andrew Wilson. 2001. *Word frequencies in written and spoken English: Based on the British National Corpus.* Routledge.
[7] Sharid Loáiciga, Liane Guillou, and Christian Hardmeier. 2017. What is it? Disambiguating the different readings of the pronoun 'it'. In *Conference on Empirical Methods in Natural Language Processing*.
[8] Sven Mayer, Gierad Laput, and Chris Harrison. 2020. Enhancing Mobile Voice Assistants with WorldGaze. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3313831.3376479
[9] OpenAI. 2023. Models. https://platform.openai.com/docs/models/overview
[10] Yevhen Romaniak, Anastasiia Smielova, Yevhenii Yakishyn, Valerii Dziubliuk, Mykhailo Zlotnyk, and Oleksandr Viatchaninov. 2020. Nimble: Mobile Interface for a Visual Question Answering Augmented by Gestures. In *Adjunct Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 129–131. https://doi.org/10.1145/3379350.3416153