"What's this?": Understanding User Interaction Behaviour with Multimodal Input Information Retrieval System

Silang Wang
e0795130@u.nus.edu
Synteraction Lab
School of Computing, National University of Singapore
Singapore

Hyeongcheol Kim
hyeongcheol@u.nus.edu
Synteraction Lab
School of Computing, National University of Singapore
Singapore

Nuwan Janaka
nuwanj@u.nus.edu
Synteraction Lab
Smart Systems Institute, National University of Singapore
Singapore

Kun Yue
yuek20@mails.tsinghua.edu.cn
Tsinghua University
China

Hoang-Long Nguyen
19127463@student.hcmus.edu.vn
VNUHCM, University of Science
Vietnam

Shengdong Zhao*
shengdong.zhao@cityu.edu.hk
Synteraction Lab
School of Creative Media & Department of Computer Science, City University of Hong Kong
Hong Kong, China

Haiming Liu*
h.liu@soton.ac.uk
School of Electronics and Computer Science, University of Southampton
Southampton, United Kingdom

Khanh-Duy Le*
lkduy@fit.hcmus.edu.vn
VNUHCM, University of Science
Vietnam

ABSTRACT

Human communication relies on integrated multimodal channels to facilitate rich information exchange. Building on this foundation, researchers have long speculated about the potential benefits of incorporating multimodal input channels into conventional information retrieval (IR) systems to support users' complex daily IR tasks more effectively. However, the true benefits of such integration remain uncertain. This paper presents a series of exploratory pilot tests comparing **M**ultimodal **I**nput **IR** (*MIIR*) with **U**nimodal **I**nput **IR** (*UIIR*) across various IR scenarios, concluding that *MIIR* offers distinct advantages over *UIIR* in terms of user experiences. Our preliminary results suggest that *MIIR* could reduce the cognitive load associated with IR query formulation by allowing users to formulate different query-component in a unified manner across different input modalities, particularly when conducting complex exploratory search tasks in unfamiliar, in-situ contexts. The discussions stemming from this finding draw scholarly attention and suggest new angles for designing and developing *MIIR* systems.

CCS CONCEPTS

• **Information systems → Search interfaces**; • **Human-centered computing → Empirical studies in ubiquitous and mobile computing**; **User studies**.

KEYWORDS

Information Retrieval; Multimodal Interaction; User Search Behavior; Heads-up Computing

ACM Reference Format:
Silang Wang, Hyeongcheol Kim, Nuwan Janaka, Kun Yue, Hoang-Long Nguyen, Shengdong Zhao, Haiming Liu, and Khanh-Duy Le. 2024. "What's this?": Understanding User Interaction Behaviour with Multimodal Input Information Retrieval System. In *26th International Conference on Mobile Human-Computer Interaction (MOBILEHCI Adjunct '24), September 30–October 03, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3640471.3680230

1 INTRODUCTION

Information retrieval (IR) systems play an indispensable role in modern life, as people rely on them for a wide range of daily information-seeking activities. For instance, for direct and straightforward information retrieval [24], individuals can quickly search for specific answers on IR systems, such as discovering local delicacies during trips. And also, for indirect and investigative information searches [40], they can interact with IR systems more extensively, integrating various pieces of information and developing detailed plans with the integrated information, such as preparing for a camping event. While contemporary commercial search engines (e.g., Google [13] or Bing [25]) and their IR interfaces generally meet many everyday information needs and support various search-related tasks as

*Corresponding Authors.

MOBILEHCI Adjunct '24, September 30–October 03, 2024, Melbourne, VIC, Australia
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0506-9/24/09
https://doi.org/10.1145/3640471.3680230

a system, performing complex exploratory search tasks through these up-to-date IR systems is still challenging [11] compared to simply conducting straightforward look-up search tasks.

As human beings, we naturally communicate in a multimodal fashion to effectively convey rich information [8]. This inherent mode of communication utilizes expressive capabilities through both verbal (e.g., speech) and non-verbal (e.g., gestures) channels. Building on this foundation, existing studies suggest that IR systems supporting users' natural multimodal inputs during their IR query formulation can enhance their ability to perform complex search-related tasks, such as exploratory searches [10, 40].

This paper presents a series of exploratory pilot studies comparing users' query formulation interaction behaviours, between using the wearable-based **M**ultimodal **I**nput **IR** (*MIIR*) system and **U**nimodal **I**nput **IR** (*UIIR*) system, across various IR contexts. *MIIR* systems refer to IR systems that support users' natural multimodal interactions during IR query formulation, contrasting with conventional **U**nimodal **I**nput **IR** (*UIIR*) systems (e.g., Google's text- or image-centered search system). In this paper, the two terms *MIIR* and *UIIR* will be used to denote the respective system's interface, search engine, and the information seeker's IR query formulation behavior. We employed a Wizard-of-Oz protocol to simulate the aimed *MIIR* system, capable of comprehending all explicit and implicit multimodal query formulation interactions from its users. We conducted a series of pilot studies in various IR contexts involving iterative design updates and collected both preliminary qualitative and quantitative data.

Our results indicate that the potential advantage of the *MIIR* system lies not in achieving faster overall performance but rather in enhancing user experiences. Specifically, the *MIIR* system prove promising when information seekers engage in exploratory search tasks [36] within their immediate physical environment or when they have limited prior knowledge in the search domain. In such scenarios, participants found that formulating IR queries based solely on their existing knowledge or personal terms is challenging, and the use of *MIIR* could alleviate cognitive loads by enabling them to compose different parts of the query through their preferred communication channels. This not only streamlines their query formulation process but also contributes to a more satisfying search experience. Our discussions stemming from these findings draw scholarly attention and suggest new angles for designing and developing *MIIR* systems.

This paper's contributions are twofold:

- Conducting pilot tests with iterative design updates to explore preliminary insights into human behaviors in IR query formulation via *MIIR* compared to *UIIR*;
- Warranting discussions for the necessity of further exploring *MIIR* systems based on study results.

2 RELATED WORK

This section starts with background on IR tasks, highlights advanced IR interface importance, explores current design trends driven by tech advances, and introduces a vision for *MIIR* interface design.

2.1 Information Retrieval, Process, Tasks, and Advanced Interfaces

In the realm of Human-Computer Interaction (HCI) and information science, "information retrieval" (IR) tasks encompass a broad range of activities where individuals seek specific information from vast structured or unstructured data sources [21]. The widespread adoption of personal mobile devices, such as smartphones, facilitates seamless engagement in IR tasks during daily activities, from looking up local delicacies to searching for images based on artwork or exploring transportation options. This underscores the critical role of efficient IR task performance in everyday life.

Recent studies have concentrated on enhancing IR task efficiency through categorizing task types [24, 30, 31], understanding the IR process [37, 38], and improving user interfaces (UI) for IR tasks [15, 17, 33]. These efforts have delineated IR tasks into two primary forms: 'lookup' and 'exploratory' search tasks [24, 30, 31]. The 'lookup' task involves precise searches with clear initial queries that lead directly to the search goal (e.g., fact retrieval, known-item search). In contrast, the 'exploratory' search task is open-ended, offering multiple pathways to desired information, often used for learning or investigative purposes.

In terms of the IR process, prior research [18, 38, 39] has identified four typical stages: 1) formulating a search query, 2) interacting with interfaces to input the query, 3) evaluating initial results, and 4) iteratively refining queries to meet specific information needs. This process applies to users' various IR tasks regardless of whether they are 'search' or 'exploratory' IR tasks.

Efforts to design and develop effective search interfaces for diverse IR tasks have historically centered on desktop-based web interfaces [15, 17, 22, 23, 33], driven by the emergence of new search engines like Google and the rapid expansion of web IR technologies. However, the proliferation of new mobile platforms (e.g., smartphones, smart glasses) and advancements in sensor technologies have prompted a shift in research focus. Researchers are now exploring the enhanced capabilities of mobile devices within the context of daily life (e.g., [37, 42]), aiming to enable seamless and convenient access to comprehensive digital resources across diverse daily contexts.

2.2 Emerging Trends in IR Interface

Recent advancements in information processing capacities and sensor technologies have spurred extensive research into evolving IR interface designs. These efforts increasingly prioritize integrating multimodal inputs, expanding the scope of IR tasks beyond traditional desktop-based textual queries to encompass a diverse range of information sources and input modalities. Hearst [16] identifies emerging trends in search interface design, emphasizing device independence, mobile compatibility, and the facilitation of multimodal search queries combining text, multimedia, and real-world data to enhance result precision. Additionally, Nigay et al. [27] propose concepts for multimodal interaction to promote user acceptance in developing multimodal search interfaces. Rigas [34] demonstrates that integrating multimodal input channels such as speech significantly enhances search engine usability. To support these innovations, various frameworks have been proposed. For instance, the World Wide Web Consortium (W3C) introduced the

‘Multimodal Architecture and Interfaces’ standard [1] for web-based multimodal IR tasks. Furthermore, Serrano et al. [35] developed an open interface framework facilitating the creation of flexible interaction pipelines integrating diverse input channels like speech and touch.

Following these frameworks, new IR interface designs and developments have been both explored and implemented, focusing on enhancing multimodal IR tasks. In their early work, J. Etzold et al. [7] introduced the ‘Multimodal Search Bag’ concept, which allows for diverse user multimodal inputs for query formulation within I-SEARCH, a search box in a segmented manner. Recent advancements in voice and gesture recognition accuracy have further propelled these design efforts, leading to a deeper integration of artificial intelligence in IR interface design and development. For instance, Zhang et al.[41] proposed ‘Vroom!’, a system that enables users to retrieve specific audio content from its search engine by imitating it vocally. Similarly, S. Y. Peng et al. [32] designed and developed a real-time hand gesture recognition IR interface and system for daily online information retrieval.

2.3 New Vision for *MIIR* Interface

Recent advancements in interactive retrieval (IR) interfaces explore novel approaches to enhance search capabilities through consolidated multimodal inputs in complex user contexts. For instance, ShapeFindAR [37] enables users to search for similar 3D objects during design tasks by tracing nearby objects with gestures or drawing shapes in the air. It supports voice-guided navigation of 3D models and gesture-based textual searches via photo capture. This proof-of-concept expands IR scenarios where users reference nearby objects or colors, aiding design exploration without visual or manual distraction, contrasting with traditional **U**nimodal **I**nput **IR** (*UIIR*) interfaces.

Building on these innovations and the rise of smart glasses, we envision new opportunities for **M**ultimodal **I**nput **IR** (*MIIR*) interfaces, aligned with users’ daily activities, such as described in [42]. To pursue this vision, understanding user responses to MIIR interfaces and their IR query behaviors across contexts is crucial. Hence, we explore these questions through pilot studies comparing MIIR with smartphone-based UIIR systems.

3 USER STUDIES AND RESULTS

This section explores specific IR scenarios where the *MIIR* system may offer advantages, identifies its limitations, and compares it with the *UIIR* system in areas where it shows potential. We present this through four iteratively designed pilot studies, with each pilot’s findings informing the design of the next. The latter part of this section outlines the common elements across all studies, followed by detailed descriptions, results, and insights from each pilot.

Participants: For each pilot study, we recruited three participants, totaling twelve participants (P1-P12). These participants were tech-savvy university students (mean age: 24 years; 3 females, 9 males), fluent in English, and accustomed to using search engines like Google daily. IRB approval from the university was obtained before the commencement of the studies.

Apparatus: In each pilot study, the *UIIR* system utilized was Google Search, accessed via an iPhone 12 smartphone. This allowed unimodal input query formulation through **either** text typing, voice dictation, **or** image capturing, with the output displayed as (Google) Search Engine Results Pages (SERPs) on the same device (browser). Conventional IR system on smartphones was chosen as the *UIIR* system based on the premise that participants were familiar with such systems, thereby providing more reliable results if the *MIIR* system was preferred.

The *MIIR* system was a simulated (Wizard-of-Oz [4]) IR system operated on a Microsoft HoloLens 2 (HL2) [26], managed by a hidden ‘wizard’. The wizard used the HL2 front camera [19] to monitor participants’ multimodal input query formulation in real time and translated it into textual queries following the *Wizarding Protocol*. These queries were then issued to Google Search by the wizard on a separate laptop for relevant SERPs. The links to these SERPs were sent back to the HL2 via Google Meet [12], accessed by the participants as the output interface on their browser. Although the *MIIR* system provided no visual feedback for user inputs during query formulation, participants were informed that the input interface was always active, capable of sensing their physical surroundings and recognizing both verbal and nonverbal expressions. Participants were encouraged to express their information needs naturally to the *MIIR* system without any restrictions, such as the order or simultaneity of their expressions.

Wizarding Protocol of the MIIR System: The wizard, pre-trained to follow a specific protocol, translates participants’ multimodal expressions into textual queries. Recognizing that speech often conveys explicit expressions in human communication [9], the wizard begins the translation process only when the participant explicitly vocalizes their information needs (e.g., “*search/look for...*”, “*find something like...*”). All verbal expressions are transcribed verbatim and entered into the Google Search bar. For non-verbal expressions, such as the user’s gaze and gestures, the wizard captures an image of the user’s first-person view, annotates it with points indicating gaze and gesture locations and then inputs into ChatGPT (GPT-4V) [28] to interpret what the user is attempting to express. The textual interpretation of these non-verbal expressions is combined with the transcribed verbal expressions to form a single combined textual query, before entering into the Google Search bar. For more information on the MIIR system, please refer to our video.

Procedure: After obtaining informed consent, participants were trained on each system until they became familiar with its use. Specifically for the *MIIR* system, training included accessing links provided by the wizard and navigating the search engine results pages (SERPs) by scrolling and clicking within the HL2 browser. Semi-structured interview was conducted after participants completed search tasks with both systems, and qualitative feedbacks were gathered.

3.1 Pilot Study 1

Motivation: Existing literature indicates that information seekers are likely to benefit from multimodal input query formulation, especially when queries involve descriptions of spatial information, such as location, number, or size [29]. Therefore, a simulated search task for hotel booking was chosen, as it often requires specifying varied types of information, including spatial aspects like hotel location or bed sizes.

Objective: Qualitatively assess the usefulness of *MIIR* system in exploratory search tasks requiring spatial input.

Design: Three participants (P1-P3) were instructed to search for hotels for their holiday travel to two different destinations of their choice. Each participant conducted two sessions: one using the *MIIR* system and the other using the *UIIR* system, with no time limit imposed. They were encouraged to thoroughly explore all available options and base their hotel searches on personal preferences, such as preferred location, price, or brand.

Results: None of the participants found the *MIIR* system more useful than the *UIIR* system in this study. They seldom used non-verbal expressions for query formulation, considering it "*unnecessary*". P2 remarked, "*If I know what I want to search for, words will suffice.*" When inquired, some participants expressed difficulty in utilizing non-verbal expressions, as P1 noted, "*If I need to input the preferred hotel location, I would just use words like 'near' or 'around' with a known landmark. I find non-verbal expressions more challenging to carry out during the search.*"

Insight: The *MIIR* system may have limited usefulness in exploratory search tasks that involve *familiar ("known")* spatial information input. In this scenario, unimodal input through speech alone appeared sufficient to convey all necessary information. This suggests that participants found verbal input more intuitive for this particular IR scenario.

3.2 Pilot Study 2

Motivation: Besides spatial input, existing research indicates that multimodal input query formulation can be advantageous for selecting objects from the surrounding environment [29]. This insight prompted us to integrate a physical search context into our second pilot study, anticipating that participants might use non-verbal references to specify objects around them. We define this situated physical environment associated with search query formulation as an in-situ search context, in contrast to an ex-situ search context where such direct references are not applicable.

Objective: Qualitatively assess the usefulness of *MIIR* system in familiar in-situ exploratory search tasks.

Design: Three participants (P4-P6) were tasked with finding two pieces of matching apparel based on their own clothing. Conducted in two sessions, one task involved searching for a top apparel that matched the bottom clothing they were wearing, and vice versa in the other session, using either the *MIIR* or the *UIIR* system. With no time constraints, participants were encouraged to explore various options and choose apparel based on their personal preferences.

Results: Participants found the *MIIR* system more intuitive for interaction than the *UIIR* system, particularly when referencing items in their immediate surroundings as part of their query formulation. For instance, P5 noted, "*The second system [UIIR] couldn't interpret 'this', but with the first system [MIIR], 'this' could refer to any feature on your clothing. It [MIIR] was definitely more natural.*" However, regarding overall usefulness, they reported no significant difference between the systems. This might be due to their familiarity with the search task, as P5 explained, "*I am very familiar with buying clothes online and have a fixed style, so words are enough for me to search for everything I want.*"

Insight: The *MIIR* system may offer a more natural interaction in in-situ exploratory search tasks, particularly through the ease of referring to elements in the immediate environment using deictic gestures or voice (e.g., "*this*"). Yet, users' familiarity with the search task could diminish the perceived advantages of multimodal input, leading them to rely predominantly on familiar modalities like text or speech.

3.3 Pilot Study 3

Motivation: Insights from prior studies suggested that the usefulness of multimodal query formulation for in-situ exploratory search tasks might vary given the participant's familiarity with the tasks. Noting that the *MIIR* system had limited usefulness for familiar search tasks, we aimed to investigate its usefulness in unfamiliar search scenarios.

Objective: Qualitatively evaluate search experience with *MIIR* system in unfamiliar in-situ exploratory search tasks.

Design: Three participants (P7-P9) were assigned to search for six pieces of furniture suitable for their existing room layout (e.g., a chair for a desk and a new bookshelf to replace an old one). Three items were to be searched using the *MIIR* system, and the remaining three with the *UIIR* system. The study was conducted over two sessions, with no time limit imposed, and participants were encouraged to explore numerous options based on personal preferences.

Results: All participants perceived the *MIIR* system as more capable and human-like in understanding their information needs compared to the *UIIR* system. Participant P7 noted, "*Initially, I treated it [MIIR] like Google. But when I couldn't find what I wanted using keywords, I just asked it to 'Find me a chair that fits the color of this table' while pointing at my table, and it worked! It felt more like talking to an interior designer.*" Participants also found the *MIIR* system more natural and efficient to interact with, echoing similar sentiments from Pilot Study 2. P8 mentioned, "*It was much faster to query compared to the phone.*" P7 added that the *MIIR* system felt less taxing to use, "*When searching for new furniture, I could easily and accurately express my needs without much effort.*"

Insight: In unfamiliar in-situ search tasks, the *MIIR* system appears to facilitate easier and quicker query formulation for users. This efficiency seems to stem from users' awareness of the *MIIR* system's capability to understand their multimodal expressions, allowing them to convey their information needs as naturally as they would with a human.

3.4 Pilot Study 4

Motivation: Building on the potential benefits identified in previous pilot studies, this study aims to deepen our understanding of the *MIIR* system's impact on the IR process.

Objective: Quantitatively evaluate search experience with *MIIR* system in unfamiliar in-situ exploratory search tasks.

Design: Three participants (P10-P12) were assigned to search for three new animal specimens in a local museum, conducted over two sessions: one with the *MIIR* system and the other with the *UIIR* system. Each session occurred in different exhibitions, with the task of finding animal specimens that complemented the theme of each exhibition. Participants were encouraged to immerse themselves in the exhibitions to understand their arrangements before identifying suitable new specimens. Each session was limited to a maximum duration of 20 minutes.

Measures	Definition
Objective Measures [2]	*Task Completion Time* = The duration from start to end of task
	Total Query Formulation Time = Sum of all queries formulation time
	Issued Query Count = The number of issued queries in a task
	Average Single Query Formulation Time = $\frac{\textit{Total Query Formulation Time}}{\textit{Issued Query Count}}$
Subjective Measures [14]	Raw NASA-Task Load (RTLX, 0-100)

Table 1: The measures used in the Pilot Study 4

Preliminary Hypotheses: Table 1 outlines the measures [2, 14] used in this pilot study. We hypothesize that efficient input modalities for query formulation will reduce task completion time, enabling users to decrease both the time spent formulating queries and the number of queries needed to retrieve the desired search results. Similarly, we anticipate that the task load will be lower with more efficient input modalities, as users can formulate queries more easily. Consequently, we expect the *MIIR* system to provide a better user experience than the *UIIR* system.

Results: *Search Efficiency:* The average total query formulation time (N = 3 participants) was similar in both *MIIR* (72.5 seconds) and *UIIR* (78 seconds) sessions. However, the *MIIR* system showed greater efficiency per query, with an average input time of 7.25 seconds compared to the 12 seconds required for *UIIR* queries. The average task completion times were comparable, with multimodal sessions averaging 810 seconds and unimodal sessions 825 seconds.

Cognitive Load: According to the NASA TLX questionnaire results, the *MIIR* system incurred a lower cognitive workload compared to the *UIIR* system across all six evaluated aspects. The mean score for *UIIR* was 39.67, considerably higher than *MIIR*'s 26.89. Notable disparities were observed in "Mental Demand" (33.33 for *UIIR* vs. 16.67 for *MIIR*), "Temporal Demand" (46.67 vs. 28.33), "Effort" (35 vs. 20), and "Frustration" (41.67 vs. 20), indicating a more user-friendly experience with the *MIIR* system in terms of reduced workload and enhanced performance.

Insight: Insights from this pilot study, along with the qualitative findings gathered so far, will be discussed in the following section to provide a comprehensive understanding of the *MIIR* system's usefulness.

4 DISCUSSION

4.1 Differences in Mental Models during Query Formulation

Different mental models toward various systems can affect user experience to varying degrees [6], and they are crucial in shaping users' behaviors, especially in complex environmental situations [3]. Our preliminary findings suggest potential disparities in users' mental models when interacting with the *MIIR* system for unfamiliar in-situ search tasks compared to the *UIIR* system, leading to the qualitative and quantitative differences observed in user behavior.

In unfamiliar in-situ search tasks, information seekers using the *UIIR* system appeared first to undergo a mental process of carefully formulating their input queries. This process often required significant mental effort to translate the users' thoughts into system-understandable words, such as summarizing in-situ elements into a few keywords. We analogize this interaction with the *UIIR* system to seeking advice from a visually-impaired interior designer for refurbishment, where users need to articulate their thoughts in precise, descriptive language to accommodate for the listener's perceptual limitation, hence leading to a diminished user experience.

Conversely, with the *MIIR* system, our results indicate that users tended to communicate their thoughts to the system directly, bypassing the deliberate step of transforming their thoughts into system-friendly utterances. This ease of interaction might stem from the *MIIR* system's ability to perceive various forms of user expressions, allowing users to focus more on other aspects of the search task, like refining their questions through point-and-speech multimodal queries. We draw an analogy between interacting with the *MIIR* system and consulting a 'magic mirror' for renovation tips, where the mirror can sense the entire room and respond to questions expressed in any manner, leading to an effortless user experience.

4.2 Difference in Cognitive Load During Query Formulation

Our results suggest that users experienced lower cognitive load when using the *MIIR* system than the *UIIR* system in unfamiliar in-situ search tasks, when users want to search based on their in-situ context without specifically naming the in-situ elements of interest, during their input query formulation. This could be due to the users **(1)** not knowing the exact label of in-situ information cues, **(2)** finding certain in-situ information cues incompatible with speech description alone, or **(3)** finding certain in-situ information cues too tedious to fully describe verbally. Hence, they adopt a combination of verbal expression, which can easily express general information needs (e.g., "*What is the price of...*", and non-verbal expression, which serves to intuitively identify the unknown in-situ element (e.g., gazing or pointing at a nearby unknown object), describe speech-incompatible concepts (e.g., gesture-based on the existing dimension of nearby objects to represent new dimensions) or quickly express information which is significantly tedious for single expressive modality (e.g., gesture at a large number of nearby objects to include all of them in the input query) [20].

4.3 Effects of Multimodal vs. Unimodal Input During Query Formulation

Our result showed that the average individual multimodal input query formulation is faster than unimodal input. We believe this is due to the efficient simultaneous combination of verbal and non-verbal expression during input query formulation, as discussed in Sec 4.2, which allowed them to issue a query more efficiently.

On average, users issued more queries in the *MIIR* system compared to the *UIIR* system in unfamiliar in-situ search tasks. We believe the difference in mental models (Sec 4.1) and cognitive load (Sec 4.2) when using the *MIIR* system likely encourages users to issue more queries compared to the *UIIR* system. Due to the users' awareness of the increased capability of the *MIIR* system at understanding user's natural expression, users could be encouraged to issue more multimodal queries that incur less cognitive workload, which in turn promotes them to issue more such queries.

The overall task completion time is comparable across the two systems, which is not surprising given the higher number of queries issued and the shorter individual query formulation time with the *MIIR* system.

5 CONCLUSION AND FUTURE WORK

In our preliminary investigation, we found that IR behaviors vary between *MIIR* and *UIIR* systems. We also identified potential advantages of multimodal input for query formulation during exploratory search tasks, summarized in Table 2.

Exploratory IR Scenarios	**Sample Search Task**	**Usefulness of *MIIR***	**Reason**
Ex-situ IR task with spatial input	Searching for hotel (Pilot Study 1)	**Low**	The absence of in-situ search context renders non-verbal expression rarely utilized for input
In-situ IR task with spatial input	Searching for clothing (Pilot Study 2)	**Moderate**	The presence of in-situ search context allows synergistic usage of verbal and non-verbal expression for input
Unfamiliar in-situ IR task with spatial input	Searching for furniture and exhibition items (Pilot Study 3 and 4)	**High**	The unfamiliar nature of the search task encourages and benefits more from cognitively-effortless multimodal input

Table 2: Summary of the findings from Four Pilot Studies.

However, our pilot studies were conducted with small sample sizes, involving tech-savvy users fluent in English, over short periods with potential novelty bias. We targeted tech-savvy users, anticipating them as potential early adopters of the *MIIR* systems. Therefore, this initial exploration requires further formal, longitudinal studies with larger, more diverse user populations and extended usage periods to more comprehensively validate the results. Additionally, incorporating more measures, such as detailed query formulation time, result browsing time [2], number of search results clicked, search query effectiveness and search satisfaction would help deepen our understanding of the differences between MIIR and UIIR systems. However, we recognize that systematic evaluation methods for multimodal exploratory search scenarios are still lacking. One approach to address this challenge is to evaluate various usage scenarios (similar to our approach) and develop guidelines (e.g., Table 2) for effective usage and design of multimodal input interfaces.

Acknowledging the potential biases associated with the Wizard-of-Oz approach [4, 5], we attempted to mitigate them by training the wizard to follow a specific protocol (Sec 3). Nonetheless, this method may have introduced response delays and might not accurately reflect the technical limitations of a real system. In some instances, the wizard struggled to identify suitable keywords for searching depicted gestures and abstract concepts (e.g., size), a challenge that realistic multimodal IR systems are likely to encounter. Future research should explore how to convert multimodal inputs into formats that these systems can understand, including the optimal representation of gestures and gaze, and the effective integration of these with verbal or textual formats (e.g., actions coupled with text). Additionally, developing appropriate evaluation metrics to assess such conversions and optimizations is essential.

ACKNOWLEDGMENTS

This paper acknowledges the use of ChatGPT for sentence refinement. We would also like to express our gratitude to the volunteers who participated in our studies and to the anonymous reviewers for their valuable time and insightful comments.

REFERENCES

[1] 2023. W3C Multimodal user interface framework. https://www.w3.org/TR/mmi-arch/. Accessed: 2023-01-23.

[2] Kumaripaba Athukorala, Dorota Głowacka, Giulio Jacucci, Antti Oulasvirta, and Jilles Vreeken. 2016. Is exploratory search different? A comparison of information search behavior for exploratory and lookup tasks. *Journal of the Association for Information Science and Technology* 67, 11 (2016), 2635–2651. https://doi.org/10.1002/asi.23617

[3] Susan Carey. 1986. Cognitive science and science education. *American psychologist* 41, 10 (1986), 1123.

[4] N. Dahlbäck, A. Jönsson, and L. Ahrenberg. 1993. Wizard of Oz studies — why and how. *Knowledge-Based Systems* 6, 4 (Dec. 1993), 258–266. https://doi.org/10.1016/0950-7051(93)90017-N

[5] S. Dow, B. MacIntyre, J. Lee, C. Oezbek, J.D. Bolter, and M. Gandy. 2005. Wizard of Oz support throughout an iterative design process. *IEEE Pervasive Computing* 4, 4 (Oct. 2005), 18–26. https://doi.org/10.1109/MPRV.2005.93 Conference Name: IEEE Pervasive Computing.

[6] Yuemeng Du, Jingyan Qin, Shujing Zhang, Sha Cao, and Jinhua Dou. 2018. Voice user interface interaction design research based on user mental model in autonomous vehicle. In *Human-Computer Interaction. Interaction Technologies: 20th International Conference, HCI International 2018, Las Vegas, NV, USA, July 15–20, 2018, Proceedings, Part III 20.* Springer, 117–132.

[7] Jonas Etzold, Arnaud Brousseau, Paul Grimm, and Thomas Steiner. 2012. Context-aware querying for multimodal search engines. In *International Conference on Multimedia Modeling.* Springer, 728–739.

[8] C. Ailie Fraser, Julia M. Markel, N. James Basa, Mira Dontcheva, and Scott Klemmer. 2020. ReMap: Lowering the Barrier to Help-Seeking with Multimodal Search. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology (UIST '20).* Association for Computing Machinery, New York, NY, USA, 979–986. https://doi.org/10.1145/3379337.3415592

[9] Marlen Fröhlich, Christine Sievers, Simon W. Townsend, Thibaud Gruber, and Carel P. van Schaik. 2019. Multimodal communication and language origins: integrating gestures and vocalizations. *Biological Reviews* 94, 5 (2019), 1809–1829. https://doi.org/10.1111/brv.12535

[10] Andrew Gambino, Jesse Fox, and Rabindra A. Ratan. 2020. Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication* 1 (Jan. 2020), 71–85. https://doi.org/10.3316/INFORMIT.097034846749023 Publisher: Communication and Social Robotics Labs.

[11] Gene Golovchinsky, Abdigani Diriye, and Tony Dunnigan. 2012. The future is in the past: designing for exploratory search. In *Proceedings of the 4th Information Interaction in Context Symposium (IIIX '12)*. Association for Computing Machinery, New York, NY, USA, 52–61. https://doi.org/10.1145/2362724.2362738

[12] Google. 2024. Google Meet. https://meet.google.com/

[13] Google. 2024. Google Search. https://www.google.com

[14] Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9 (2006), 904–908. https://doi.org/10.1177/154193120605000909

[15] Marti A. Hearst. 2009. *Search User Interfaces* (1st ed.). Cambridge University Press, USA.

[16] Marti A Hearst. 2011. Emerging trends in search user interfaces. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*. 5–6.

[17] Shih-Ting Huang, Tsai-hsuan Tsai, and Hsien-tsung Chang. 2009. The UI issues for the search engine. In *2009 11th IEEE International Conference on Computer-Aided Design and Computer Graphics*. IEEE, 330–335.

[18] Harvey Hyman, Terry Sincich, Rick Will, Manish Agrawal, Balaji Padmanabhan, and Warren Fridy. 2015. A process model for information retrieval context learning and knowledge discovery. *Artificial Intelligence and Law* 23 (2015), 103–132.

[19] Nuwan Janaka, Runze Cai, Ashwin Ram, Lin Zhu, Shengdong Zhao, and Yong Kai Qi. 2024. PilotAR: Streamlining Pilot Studies with OHMDs from Concept to Insight. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* (Sept. 2024). https://doi.org/10.1145/3678576

[20] 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 *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–3.

[21] Haiming Liu, Suzanne Little, and Stefan Rüger. 2011. Multimedia: behaviour, interfaces and interaction. (2011).

[22] Haiming Liu, Paul Mulholland, Dawei Song, Victoria Uren, and Stefan Rüger. 2010. Applying information foraging theory to understand user interaction with content-based image retrieval. In *Proceedings of the third symposium on Information interaction in context*. 135–144.

[23] Haiming Liu, Paul Mulholland, Dawei Song, Victoria Uren, and Stefan Rüger. 2011. An information foraging theory based user study of an adaptive user interaction framework for content-based image retrieval. In *Advances in Multimedia Modeling: 17th International Multimedia Modeling Conference, MMM 2011, Taipei, Taiwan, January 5-7, 2011, Proceedings, Part II 17*. Springer, 241–251.

[24] Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (apr 2006), 41–46. https://doi.org/10.1145/1121949.1121979

[25] Microsoft. 2024. Microsoft Bing. https://www.bing.com

[26] Microsoft. 2024. Microsoft Hololens 2. https://www.microsoft.com/en-us/hololens

[27] Laurence Nigay. 2004. Design space for multimodal interaction. In *Building the Information Society: IFIP 18th World Computer Congress Topical Sessions 22–27 August 2004 Toulouse, France*. Springer, 403–408.

[28] OpenAI. 2024. ChatGPT. https://chat.openai.com/

[29] Sharon Oviatt. 1999. Ten myths of multimodal interaction. *Commun. ACM* 42, 11 (nov 1999), 74–81. https://doi.org/10.1145/319382.319398

[30] Emilie Palagi, Fabien Gandon, Alain Giboin, and Raphaël Troncy. 2017. A survey of definitions and models of exploratory search. In *Proceedings of the 2017 ACM workshop on exploratory search and interactive data analytics*. 3–8.

[31] Patrick Cheong-Iao Pang, Karin Verspoor, Shanton Chang, and Jon Pearce. 2015. Conceptualising health information seeking behaviours and exploratory search: result of a qualitative study. *Health and Technology* 5 (2015), 45–55.

[32] Sheng-Yu Peng, Kanoksak Wattanachote, Hwei-Jen Lin, and Kuan-Ching Li. 2011. A real-time hand gesture recognition system for daily information retrieval from internet. In *2011 Fourth International Conference on Ubi-Media Computing*. IEEE, 146–151.

[33] W Quesenbery, C Jarrett, I Roddis, S Allen, and V Stirling. 2008. Designing for Search: Making Information Easy to Find. In *55th Conference" STC", the Society for Technical Communication, Philadelphia [online], http://stc-access. org/conference-session-materials*.

[34] Dimitrios Rigas and Antonio Ciuffreda. 2007. An empirical investigation of multimodal interfaces for browsing internet search results. In *Proceedings of the 7th WSEAS International Conference on Applied Informatics & Communication*. 194–199.

[35] Marcos Serrano, Laurence Nigay, Jean-Yves L Lawson, Andrew Ramsay, Roderick Murray-Smith, and Sebastian Denef. 2008. The openinterface framework: A tool for multimodal interaction. In *CHI'08 Extended abstracts on human factors in computing systems*. 3501–3506.

[36] Ayah Soufan, Ian Ruthven, and Leif Azzopardi. 2022. Searching the Literature: An Analysis of an Exploratory Search Task. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval* (Regensburg, Germany) *(CHIIR '22)*. Association for Computing Machinery, New York, NY, USA, 146–157. https://doi.org/10.1145/3498366.3505818

[37] Evgeny Stemasov, Tobias Wagner, Jan Gugenheimer, and Enrico Rukzio. 2022. ShapeFindAR: Exploring in-situ spatial search for physical artifact retrieval using mixed reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–12.

[38] Don R Swanson. 1977. Information retrieval as a trial-and-error process. *The Library Quarterly* 47, 2 (1977), 128–148.

[39] Pertti Vakkari. 2001. A theory of the task-based information retrieval process: A summary and generalisation of a longitudinal study. *Journal of documentation* 57, 1 (2001), 44–60.

[40] Ryen W. White and Resa A. Roth. 2009. *Exploratory Search: Beyond the Query-Response Paradigm*. San Rafael, Calif.

[41] Yichi Zhang, Junbo Hu, Yiting Zhang, Bryan Pardo, and Zhiyao Duan. 2020. Vroom! a search engine for sounds by vocal imitation queries. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 23–32.

[42] Shengdong Zhao, Felicia Tan, and Katherine Fennedy. 2023. Heads-Up Computing Moving Beyond the Device-Centered Paradigm. *Commun. ACM* 66, 9 (aug 2023), 56–63. https://doi.org/10.1145/3571722