MRTranslate: Bridging Language Barriers in the Physical World Using a Mixed Reality Point-and-Translate System

Florian Mathis
florian.mathis@unisg.ch
University of St. Gallen
St. Gallen, Switzerland

Yu Sun
yu.sun@unisg.ch
University of St. Gallen
St. Gallen, Switzerland

Adrian Preussner
adrian.preussner@unisg.ch
University of St. Gallen
St. Gallen, Switzerland

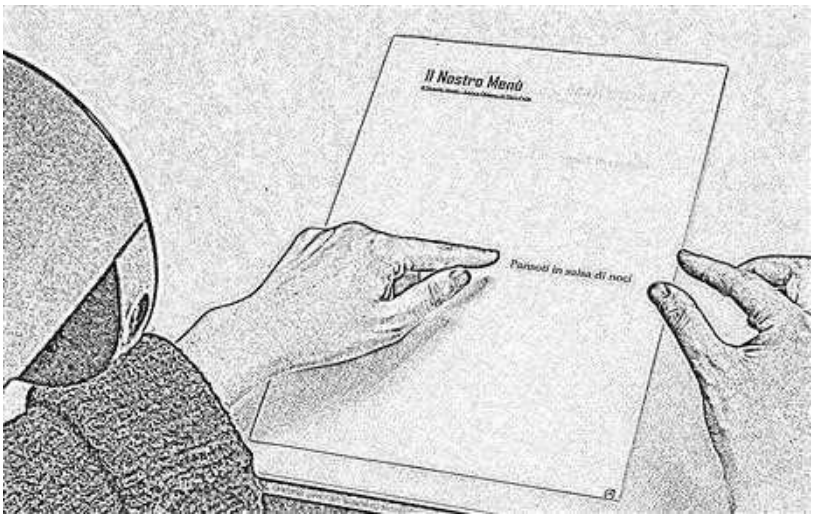

Figure 1: We present *MRTranslate*, an assistive mixed reality prototype that allows users to translate text by pointing at it in the real world. *MRTranslate* makes use of human asymmetric bimanual cooperation [20, 21] to point at the information in the real world and then translate it, envisioning a future pervasive augmented reality [19, 36].

ABSTRACT

Language barriers pose significant challenges in our increasingly globalized world, hindering effective communication and access to information. Existing translation tools often disrupt the current activity flow and fail to provide seamless user experiences. In this paper, we contribute the design, implementation, and evaluation of *MRTranslate*, an assistive Mixed Reality (MR) prototype that enables seamless translations of real-world text. We instructed 12 participants to translate items on a food menu using *MRTranslate*, which we compared to state-of-the-art translation apps, including *Google Translate* and *Google Lens*. Findings from our user study reveal that when utilising a fully functional implementation of *MRTranslate*, participants achieve success in up to 91.67% of their translations whilst also enjoying the visual translation of the unfamiliar text. Although the current translation apps were well perceived, participants particularly appreciated the convenience of not having to grab a smartphone and manually input the text for translation when using *MRTranslate*. We believe that *MRTranslate*, along with the empirical insights we have gained, presents a valuable step towards a future where MR transforms language translation and holds the potential to assist individuals in various day-to-day experiences.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in interaction design**; **Mixed / augmented reality**.

KEYWORDS

Mixed Reality, Assistive MR Artefacts, Empirical Research

ACM Reference Format:
Florian Mathis, Yu Sun, and Adrian Preussner. 2024. MRTranslate: Bridging Language Barriers in the Physical World Using a Mixed Reality Point-and-Translate System. In *International Conference on Advanced Visual Interfaces 2024 (AVI 2024), June 03–07, 2024, Arenzano, Genoa, Italy.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3656650.3656652

AVI 2024, June 03–07, 2024, Arenzano, Genoa, Italy

ACM ISBN 979-8-4007-1764-2/24/06
https://doi.org/10.1145/3656650.3656652

1 INTRODUCTION

In an increasingly interconnected and diverse global society, the ability to consume and understand information presented in different languages is paramount. As individuals navigate the real world, they are often faced with language barriers [46], whether it is understanding street signs, restaurant menus, or product labels. Overcoming these language barriers is crucial for safety, effective

communication, unrestricted access to information, and counteracting negative emotional and cognitive reactions. The rapid advancement of technology, exemplified by tools like Google Lens [17], an app that uses the magic lens paradigm [3, 48] to recognise the visual information at which a user points their mobile device's camera and then translates the text, has led to a plethora of transformative solutions that are capable of addressing language-related challenges. However, such experiences often require users to move their attention from reality to a small screen. Mixed Reality (MR), capable of merging the physical and digital worlds [41, 51], can enable experiences beyond device-centred paradigms [36, 55]. For example, in line with Weiser's notion of computers being transparently incorporated into people's daily lives [53], Feiner asserted the overlaid information of MR systems will become "part of what we expect to see at work and at play" [13].

In this paper, we harness the potential of everyday MR to bridge language barriers and empower individuals to understand unfamiliar text in the real world. We contribute empirical insights into the use of MR to translate text and shed light on users' experiences and perceptions using *MRTranslate*, an MR prototype that allows users to point at text in the real world and translate it in real-time (see Figure 1 and Figure 3). Our study reveals that users find interacting with *MRTranslate* highly enjoyable. They appreciated that they were not obliged to manually input the text they wished to translate on their mobile device, which freed their hands from holding and tapping the phone. Furthermore, participants voiced that the seamless integration of translations into traditional MR-capable glasses will abolish the need for an extra device like a mobile phone, which they would typically need to retrieve from their pocket. Aside from showing promising results, our study underscores the significant influence of technology on user perception of usability and satisfaction when it comes to processing real-world content. A comparison of a fully functional *MRTranslate* prototype with a Wizard-of-Oz (WoZ) experience [30, 39] resulted in differences in system usability scores and perceived workloads.

Contributions. Per definition by Wobbrock and Kientz [54], our contributions encompass both artefacts and empirical findings.

(1) We contribute *MRTranslate*, an MR implementation that enables researchers and practitioners to sense the real world, process visual information through a client-server architecture, and then present information to users according to their preference (see Figure 2). *MRTranslate*, including all required hardware and software implementations, is publicly available under https://github.com/FlorianMathis/MRTranslate.
(2) We present insights from a lab study conducted with a WoZ implementation and a fully functional implementation of *MRTranslate*, outlining the advantages and limitations of using MR for real-world translations compared to Google Translate [18] and Google Lens [17].

2 RELATED WORK

We review existing research on translation systems and explore works within the context of a pervasive augmented reality.

2.1 Translating Real-World Information

A substantial body of Human-Computer Interaction (HCI) research has delved deeply into the domain of translation systems. An illustrative example is the work by Gao et al., which explored the potential of machine translations to facilitate communication among individuals with varying language skills [15, 16], and the prototype by Shilkrot et al. [50], a finger-worn device that allows people with visual impairments to read printed text on-the-go. Existing translation applications offer a diverse range of modalities for both input and output, encompassing text entry, speech recognition, and more. Aligning with the work by Gao et al. [15], research by Hara and Iqbal [22] revealed user preferences for visual output (i.e., translated text) over auditory output (i.e., text-to-speech) as visual output facilitates the identification of recognition errors with greater ease. In work by Ibrahim et al. [27], a HoloLens setup was employed to augment objects in the user's surroundings by displaying corresponding words and providing audio samples. Leading technology companies like Google have propelled the application of MR in real-time translations. Google Lens [17], utilised by billions of people worldwide, exemplifies this advancement by leveraging a smartphone's camera to translate text seamlessly.

The reviewed works represent significant advancements beyond conventional translation applications, such as Google Translate [18]. Noteworthy systems like Google Lens [17] and the prototype developed by Draxler et al. [10] have leveraged smartphone technology for translations. While there are prototypes similar to MRTranslate available (e.g., [52]), *MRTranslate* extends upon these existing efforts by exploring the integration of MR alongside asymmetric bimanual point-and-translate interaction. By focusing on the seamless integration of MR into everyday life, *MRTranslate* adds to the broader objective of a *Pervasive Augmented Reality*, as introduced by Grubert et al. [19], which we review in more detail below.

2.2 A Pervasive Augmented Reality

Future everyday MR interfaces will be pervasive and omnipresent, enabling individuals to continuously augment, alter or diminish their everyday experiences [19, 42, 49]. One of the pioneering works in the domain of everyday MR is the concept of *Pervasive Augmented Reality* by Grubert et al. [19], a continuous and universal augmented interface to information in the physical world.

Following the footsteps of mobile devices, MR technology will be capable of supplying users with information directly through their field of view without being required to pull out a separate device anytime and anywhere. For example, Davari et al. [9] investigated context-aware AR interfaces that minimise intrusiveness whilst providing fast and easy information access during social contexts. Lu and Bowman [32] investigated in-situ AR interfaces to extend physical monitors through glanceable widgets residing at the edge of the display or to support cooking through augmented recipe and timer widgets. Google Research introduced opportunistic interfaces (i.e., an extension of the concept of opportunistic controls where a semantic matching of virtual content is associated with physical objects [24]) to grant individuals complete freedom to summon augmented interfaces on everyday objects via voice commands or tapping gestures [11]. They showcase how such interfaces can

lead to ad hoc map interfaces on a transportation card or to live transcriptions on a user's open palm.

Glanceable AR by Lu et al. [33] enables users to access weather information, news, activities, and many more information on top of their real-world environment. In an in-the-wild study, they found that using *Glanceable AR* compared to a phone/watch led to more unobtrusive experiences (e.g., *"While having conversations with people, it's way nicer to use the* Glanceable *apps for quick checks [...] compared to taking out my phone or even looking at my watch. If I had used my phone/watch the same amount as the Glanceable apps, people would have been annoyed with me, or thought I wasn't paying attention to them (P2)."* [33].

Essentially, the literature highlights the remarkable research efforts devoted to exploring users' perceptions and usages of MR technology in everyday life scenarios. All of these works, including research efforts on learning case grammar from real-world interactions [10], using AR to learn about real-world objects [31], or extending blind or low-vision people's capabilities in social situations [56], motivated us to design, implement and empirically evaluate *MRTranslate*, contributing to, and pushing forward, a supportive pervasive everyday augmented reality [19, 38].

3 CONCEPT OF *MRTRANSLATE*

MRTranslate is designed and implemented as a fully functional prototype to allow the implementation of the experience pipeline visualised in Figure 2. The concept of *MRTranslate* is motivated as follows: the interaction is derived from the use of Guiard's kinematic chain model for human asymmetric bimanual cooperation [20, 21], where the user sets the frame of reference using their non-dominant hand (pointing on information in the real world) and then performs the subsequent action using their dominant hand (confirming the pointing and initiating the translation pipeline, Figure 1 and Figure 3). In previous works, Buxton and Myers [8] and Kabbash et al. [29] highlighted the advantages of two-handed interactions, which can be designed to leverage existing skills. In *MRTranslate*, we make use of natural pointing on the non-dominant hand to indicate that this is the area of interest (i.e., the text to translate) and the dominant hand to confirm the selection (i.e., pinch to confirm the translation of the specific text). We decided to implement the confirmation of the selection using a pinch gesture

Figure 3: A user's perspective when they translate the term *"pesce"* from Italian to English using *MRTranslate*.

as this is the most common gesture in MR experiences [45, 47]. Furthermore, extending the index finger is one of the most natural gesture when people want to draw attention to something and is considered to play a foundational role in human language [44]. Such a pointing technique is also applied in Google Lens, where users use their mobile device to "point" on an object in their real world to then translate it to their preferred language using AR [17].

In essence, *MRTranslate*'s underlying concept integrates established MR interaction techniques, such as pinching, with natural pointing gestures towards information in the real world. This envisions a future of a pervasive augmented reality [19] wherein individuals wearing traditional glasses with MR functionality can seamlessly point at unfamiliar text in the real world and translate it with little to no effort.

4 PROTOTYPE IMPLEMENTATION

MRTranslate consists of hardware components and software elements. All implementations, including the schematics for the hardware, the client-server Python architecture, and the Unity 3D (C#) code, are available in a public repository for future research: https://github.com/FlorianMathis/MRTranslate. Below, we provide an overview of *MRTranslate*'s technical contribution.

4.1 Sensing of the Real-World Information

To enable users to sense their real-world surroundings, i.e. translate unfamiliar information, we used a Raspberry Pi 3 camera (Sony IMX708 image sensor) which we attached to a Raspberry Pi 2B using a 60cm flexible ribbon cable. Additionally, a small push button was integrated into the circuit to give users full control of the real-world sensing (Figure 3). Both the camera and the button were attached to

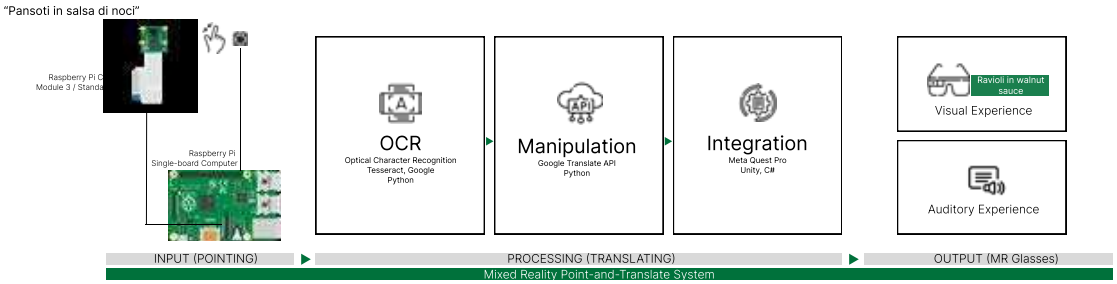

Figure 2: Our pipeline consists of a) *user input*, i.e., pointing to the text in the real world the user wants to translate; b) a *server-client processing*, i.e., taking the user-generated image as input, applying optical character recognition to extract the text (*"Pansoti in salsa di noci"*), and then translating the text using the Google Translate API (*"Ravioli in walnut sauce"*); and c) *output*, i.e., translating the real-world text and outputting it either visually or auditory on the user's MR glasses.

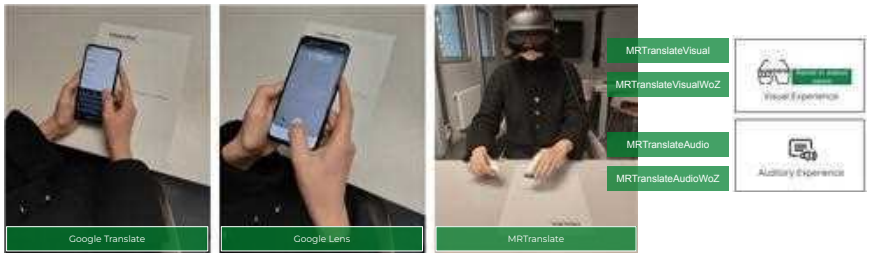

Figure 4: We compare four variations of *MRTranslate* (right) with Google Translate [18] (left) and Google Lens [17] (middle).

finger gloves to allow for human asymmetric bimanual interaction [20, 21], enable point and pinch gesture input, and realise a fully functional implementation of the concept of using MR to point and translate real-world information. To sense the real world, i.e. take a picture of the reality, the user points at the term they want to translate, which is supported through visual guidance of a red line. They then perform a pinch gesture, as shown in Figure 3.

4.2 Processing of the Real-World Information

Processing real-world information and waiting for its modified version to appear on the MR glasses involves three core steps:

4.2.1 Step 1: Image Transmission. After the user took a picture of the real-world information they want to translate, the Raspberry Pi sends the image through a Python-based socket connection to a server, which in our case is a local PC (*i9-11900K@3.50GHz, 8 Cores; 64GB RAM; NVIDIA GeForce RTX 3090*) acting as both server and processing pipeline for the user study.

4.2.2 Step 2: Image-to-Text Using Optical Character Recognition (OCR). On the server, we run a Python script that processes the image according to the processing pipeline in Figure 2. The image is first preprocessed using OpenCV [28] (i.e., transformed into gray scale, applied dilation) and then processed using pytesseract [25], an OCR implementation for Google's Tesseract-OCR Engine, to extract text from the image.

4.2.3 Step 3: Translation and Integration. The output, i.e., the extracted text in the image, is then sent to Google Translate using the deep_translator library [1]. Depending on the output modality, visual or auditory, the output of Google Translate is further processed using Google text-to-speech (gTTS) [12], a Python library to interface with the Google Translate text-to-speech API. Finally, a C# script using Unity 3D presents the translated real-world information on the user's MR glasses, either of visual nature or of auditory nature, see Figure 2-OUTPUT.

5 USER STUDY

We evaluated *MRTranslate* in a user study to provide initial answers to the following research question, **RQ:** *How does* MRTranslate*, a mixed reality point-and-translate system, compare to Google Translate and Google Lens in terms of usability and efficiency when translating real-world text?* The study followed a within-subjects experiment with 12 participants (6 conditions × 2 latin squares = 12 participants), ensuring a perfect counter-balancing of the conditions [5]. Each participant translated overall six terms, see supplementary material. Participants were recruited through word of mouth and mailing lists. Based on our translation tasks (i.e., Italian → English), we aimed to recruit participants with a considerable high English language proficiency and no or limited knowledge of the Italian language. All participants used a Meta Quest Pro and the Google Pixel 4a to participate. Both devices were provided by us to contribute towards internal validity. Each session lasted for a maximum of 1 hour. The study went through an Ethics checklist at the University of St. Gallen and was exempted from an additional full ethics review.

5.1 Conditions

Below, we describe our six conditions, including Google Translate [18] and Google Lens [17] (baselines), and four MR conditions, covering visual (*MRTranslateVisual* and *MRTranslateVisualWoZ*) and auditory output (*MRTranslateAudio* and *MRTranslateAudioWoZ*).

5.1.1 ***Google Translate (first baseline):*** A digital dictionary, a dictionary whose data exists in digital form and can be accessed using a mobile device, is frequently used when translating text across languages. When using a digital dictionary, the user has to typewrite the unfamiliar term on, e.g., their mobile device, to then receive the translation. Due to the popularity of Google Translate, we decided to compare *MRTranslate* against the pre-installed Google Translate app on Android devices (see Google Translate in Figure 4).

5.1.2 ***Google Lens (second baseline):*** Our second baseline depicts Google Lens [17], a visual recognition technology that allows users to translate information in the real world by using a smartphone's camera (see Google Lens in Figure 4). Compared to *Google Translate*, in *Google Lens* users do not have to typewrite the unfamiliar term; instead, they hover their smartphone's camera over the information in the real world and then receive an in situ translation. In line with *Google Translate*, we compare *MRTranslate* against *Google Lens* as it has already found widespread application with rising usage numbers (> 10 billion requests per month [4]).

5.1.3 ***MRTranslate (visual and auditory):*** For *MRTranslate*, we are interested in the impact of the output modality on perceived usability and workload. Particularly, we explore *MRTranslate* with a visual and an auditory output. In both variants, the input and processing pipeline remains the same, as depicted in Figure 2-PROCESSING. However, whilst in *MRTranslateVisual* the translated real-world information is visually rendered on the users' MR glasses for five seconds (visual ⇒ visual), in *MRTranslateAudio* the translated real-world information is transformed into audio and then played on

the integrated audio speakers on the MR glasses (visual ⇒ auditory). Our additional investigation of a modality switch, i.e., from visual information to information that is presented through audio, enables us to explore how such a modality switch impacts users' experiences and preferences.

For both *MRTranslateVisual* and *MRTranslateAudio*, we additionally investigated the impact of a Wizard-of-Oz implementation (i.e., an experience that always returned the correct translation after a pinch gesture) on users' perceived workload and usability. In *MRTranslateVisualWoZ* and *MRTranslateAudioWoZ*, the point-and-translate system worked without any inaccuracies that may be present in a real environment where low picture quality or low OCR-detection may exist. It can be expected that with the advancement of technologies, noise and inaccurate real-world sensing can be overcome. As such, the additional WoZ experiences represent such a future scenario. Additionally, comparing a fully functional implementation with a WoZ implementation of *MRTranslate* enables us to assess the influence of system-induced inaccuracies on users' perceptions and preferences. This comparison underscores the significance of not exclusively depending on WoZ in pioneering system research, especially when technology is not yet mature enough, underscoring the need to account for the potential influence of a fully functional prototype on user experience.

5.2 Measures

We asked participants after each translation experience about their a) perceived workload using the raw NASA-TLX [23] and b) perceived system usability using the SUS [6]. Furthermore, we measured their performance (i.e., correct translations and number of translation trials) and asked about their level of confidence in completing the translation (5-point Likert scale, from strongly disagree to strongly agree). To conclude the study, we asked the participants about their preferred translation mechanism and engaged with them in a semi-structured interview to learn more about their experience when using *MRTranslate* to translate real-world information.

5.3 Study Task

We first welcomed the participants by introducing them to the study. They then filled in the demographics, which are reported in Section 5.4. Participants were then introduced to the six conditions when translating real-world content along with a training session to get familiar with *Google Translate*, *Google Lens*, and *MRTranslate*. This ensured that all participants were familiar with the study equipment, including the Google Pixel 4a as well as the Meta Quest Pro. We then applied storytelling, a process where participants are told to be part of a specific environment for the duration of the study. In our case, the study setting depicted a scenario where the participants experienced difficulties reading the menu of an Italian restaurant due to their unfamiliarity with the Italian language. Participants then went through one of the conditions, depending on the Latin square. Their task was to translate a specific term on the Italian menu (e.g., *"coniglio"*) and then write down the English translation (e.g., *"rabbit"*), along with their level of confidence. We used translations into English across all participants to contribute towards internal validity. After each experience, the participants reported their perceived workload [23] and rated the system's usability [6]. After each condition, participants went through a usage preference ranking, i.e., which experience they liked most and which they liked least, and took part in a semi-structured interview.

The six translation tasks and the core interview questions are available in our supplementary material for reproducibility.

5.4 Participants

Our sample consisted of 12 participants (5 female, 7 male) with an average age of 30.0 (SD = 7.76). We used the Common European Framework of Reference for Languages (CEFR) when asking the participants about their language skills for both Italian and English. All participants, except two (one beginner knowledge, one elementary knowledge), reported to have no prior experience of the Italian language (i.e., below A1). For the English language, five participants reported to have proficient language knowledge (C2), five advanced language knowledge (C1), and two upper-intermediate knowledge (B2). Participants' experience with mixed reality on a 5-Point Likert scale was 3.25 (SD = 1.29). Their familiarity with Google Translate and Google Lens was M = 4.50 (SD = 0.52) and M = 3.42 (SD = 1.51), retrospectively. Their affinity technology score measured using the ATI questionnaire [14] was M = 4.86 (SD = 0.63).

6 RESULTS

Unless otherwise stated, we ran Friedman tests on non-parametric data and on non-normal distributed parametric data. Post-hoc tests were Bonferroni corrected to correct for multiple comparisons. Effect sizes are reported using *Kendall's W*, with 0.1 - < 0.3 indicating a small effect, 0.3 - < 0.5 a moderate effect, and >= 0.5 a large effect.

6.1 Translation Performance and Confidence

Results of a Cochran's Q test indicated no significant differences among the conditions in terms of translation performance, $\chi^2(5) = 6.612$, $p = 0.2511$. However, compared to *Google Translate* and *MRTranslateVisualWoZ* (100% correct translations in both), *MRTranslateAudio* and *MRTranslateAudioWoZ* yielded a success rate of correct translations of 83.33% and 75%, respectively. Note that the numbers do not represent the system's translation accuracy; instead, they reflect the extent to which participants were able to provide the correct translation based on the system output they received. In both auditory conditions, i.e., *MRTranslateAudio* and *MRTranslateAudioWoZ*, the participants' levels of confidences were slightly lower (M = 3.92, SD = 1.24; M = 3.92,SD = 1.31) compared to the smartphone conditions (*Google Translate*: M = 4.83, SD = 0.39; *Google Lens*: M = 4.42, SD = 0.79) and *MRTranslate* with visual output (*MRTranslateVisual*: M = 4.83, SD = 0.39; *MRTranslateVisualWoZ*: M = 4.83, SD = 0.39). Table 1 shows all scores and the number of translation trials for each of the *MRTranslate* conditions. For *Google Translate* and *Google Lens*, participants used the corresponding app always once, i.e., they entered the Italian term one time in *Google Translate* and opened the *Google Lens* app once to hover over the Italian term in reality.

6.2 NASA-TLX and SUS

Results of a Friedman test on the raw NASA-TLX values indicated a significant difference among the experiences, $\chi^2(5) = 18.29$, $p < 0.05$, $W = 0.305$. Participants' perceived workloads were statistically

Table 1: The table shows the participants' translation performance, their confidence in their translations, the raw NASA-TLX values, and system usability scores. Reporting is in the form of *mean (standard deviation)*. $p < 0.05$ highlighted.

	(1) *Google Translate*	(2) *Google Lens*	(3) *MRTranslateVisual*	(4) *MRTranslateVisualWoZ*	(5) *MRTranslateAudio*	(6) *MRTranslateAudioWoZ*	Statistical Analysis	p<0.05
Correct Translations (in %)	100%	91.67%	91.67%	100%	83.33%	75%	$\chi^2(5) = 6.612$, p = 0.2511	n/a
Number of Trials	-	-	5.0 (6.22)	1.33 (0.88)	5.92 (4.19)	2.0 (1.35)	$\chi^2(3) = 19.567$, $p < 0.05$, W = 0.544	4-5;4-6
Level of Confidence	4.83 (0.39)	4.42 (0.79)	4.33 (1.15)	4.83 (0.39)	3.92 (1.24)	3.92 (1.31)	$\chi^2(5) = 13.84$, $p < 0.05$, W = 0.618	n/a
Mental Demand	16.67 (23.68)	17.50 (22.41)	19.17 (25.30)	15.83 (24.76)	28.75 (26.04)	30.00 (36.93)	$\chi^2(5) = 10.199$, p = 0.06979, W = 0.170	n/a
Physical Demand	13.33 (16.14)	12.92 (15.44)	33.75 (32.48)	21.67 (24.53)	34.58 (29.88)	24.17 (27.87)	$\chi^2(5) = 18.528$, $p < 0.05$, W = 0.309	n/a
Temporal Demand	16.67 (16.56)	16.67 (20.04)	15.83 (18.44)	15.83 (20.65)	13.75 (15.97)	18.75 (25.24)	$\chi^2(5) = 2.2727$, p = 0.8103, W = 0.038	n/a
Performance	3.33 (6.15)	7.50 (14.54)	16.25 (22.27)	5.83 (9.00)	29.17 (24.29)	23.33 (29.10)	$\chi^2(5) = 22.5$, $p < 0.05$, W = 0.375	n/a
Effort	17.08 (19.12)	12.92 (16.02)	23.75 (20.68)	14.58 (20.94)	30.00 (24.12)	24.58 (28.72)	$\chi^2(5) = 15.222$, $p < 0.05$, W = 0.254	n/a
Frustration	14.17 (19.75)	12.50 (17.77)	20.00 (22.16)	10.00 (16.38)	26.25 (26.38)	26.25 (28.45)	$\chi^2(5) = 15.685$, $p < 0.05$, W = 0.261	n/a
Overall RAW NASA-TLX [23]	13.54 (17.77)	13.33 (17.58)	21.46 (23.92)	13.96 (20.03)	27.08 (24.79)	24.51 (28.74)	$\chi^2(5) = 18.29$, $p < 0.05$, W = 0.305	2-5
SUS [6]	85.63 (12.21)	80.21 (14.36)	73.13 (13.99)	81.46 (17.60)	65.00 (20.39)	67.92 (24.14)	$\chi^2(5) = 30.4$, $p < 0.05$, W = 0.506	1-5;4-6

significantly higher in *MRTranslateAudio* (M = 27.08, SD = 24.79) than in *Google Lens* (M = 13.33, SD = 17.58) ($p < 0.05$). Table 1 and Figure 5 summarise the data. For the system usability scores, we noticed a significant main effect, $\chi^2(5) = 30.4$, $p < 0.05$, $W = 0.506$. Post-hoc tests revealed significant different usability scores between *Google Translate* (M = 85.63, SD = 12.21) and *MRTranslateAudio* (M = 65.00, SD = 20.39), and between *MRTranslateVisualWoZ* (M = 81.46, SD = 17.60) and *MRTranslateAudioWoZ* (M = 67.92, SD = 24.14). No other pairs were significant. *Google Translate* yielded an "excellent" usability [2], whereas *Google Lens*, *MRTranslateVisual*, and *MRTranslateVisualWoZ* yielded "good" usability. Both *MRTranslateAudio* and *MRTranslateAudioWoZ* achieved an "OK" usability according to Bangor et al. [2]. Figure 5 provides an overview of the usability scores. The comparable low usability scores of *MRTranslateAudio* and *MRTranslateAudioWoZ* also reflect the participants' usage preferences, as described in Section 6.3 and Section 6.4.

6.3 Usage Preference

We asked participants to rank the experiences based on usage preference (1 = best; 6 = worst). *Google Lens* yielded the highest weighted score (57), followed by *MRTranslateVisualWoZ* (53), *Google Translate* (51), *MRTranslateVisual* (41), *MRTranslateAudio* (26), and *MRTranslateAudioWoZ* (22). This means that contradictory to the SUS, where *Google Translate* achieved the highest score, *Google Translate* was less preferred compared to *Google Lens* and *MRTranslateVisualWoZ*. In fact, *Google Lens* was ranked five times as the most preferred experience (and 2× as the second most preferred), with *MRTranslateVisualWoZ* being ranked three times as the most preferred experience (and 4× as the second most preferred).

6.4 Semi-Structured Interviews

The interviews were audio recorded and literally transcribed. Then, the interviews were split into 158 meaningful excerpts, which were analysed using an affinity diagram [35] by three researchers.

Theme 1: System Familiarity and Convenience in Everyday Translations are Key. While not surprising, it became evident that participants' familiarity with *Google Translate* and *Google Lens* influenced their preference for usage. For example, P4 expressed a clear preference for *Google Translate*, stating, *"for the simple reason it is what I am used to."* (P4). Similarly, P1 voiced that they *"generally liked the Google Lens because [they] use it themselves in everyday tasks."* (P1). Although P4 argued that *"the basic way to translate something with the state of technology is to write it down."* (P4), there was a consensus among the participants that *"scanning"* unfamiliar text in the real world is more convenient than typing it, a sentiment applicable to both *Google Lens* and *MRTranslate*. P8 articulated this viewpoint quite well: *"Google Translate, yes, it is great, but you have to type it which is more effort."* (P8). Others expanded on the drawbacks of Google Translate, noting concerns such as potential mistyping in foreign languages (P2) and asserting that *"typing is more clumsy than pointing with the camera."* (P12). Participants also raised issues related to typing lengthy texts for translation using *Google Translate*, emphasising the time-consuming nature of the process. Additionally, they pointed out challenges in typing unfamiliar characters (e.g., French accent marks or Chinese characters).

Theme 2: Enhanced Translation Experience: Why* MRTranslate *with Visual Output Outperforms. Our qualitative data revealed two core elements that contributed to the superiority of

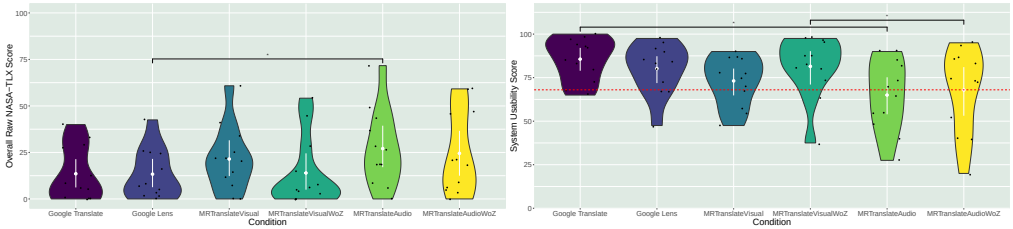

Figure 5: Violin plots of participants' workloads using the NASA-TLX by Hart [23] and the system usability scores of our six conditions using the SUS by Brooke et al. [6]. Red line shows a usability score of 68, which is considered above average [6].

MRTranslate with visual output over *Google Translate*, *Google Lens*, and *MRTranslate* with auditory output.

First, participants highlighted the enhanced ease and seamlessness of translating text in the real world using *MRTranslate*. P2 expressed the convenience of having it integrated into their glasses, stating, *"it could be really convenient when it works well and I have it on my glasses. Instead of getting out the phone, opening the app, and everything."* (P2). P6 echoed this sentiment and argued that they *"prefer the visual feedback simple because it is really quickly, you can just point at something [...] you don't have, for example, to pull out your phone. [...] if I can wear a convenient headset, like glasses, and can point at things and get translations I would use it, more than Google Lens and Google Translate."* (P6). Concerns were raised about the anxiety of dropping phones when using *Google Lens* for real-world text translation, as P4 pointed out, *"I have the fear of dropping [the phone]. In everyday life, I just don't like to take pictures with my phone. I am always afraid to drop it."* (P4).

Second, there was a consensus that visually rendering translated text is the preferred method over auditory rendering. Participants expressed reservations about the modality change during translation, where text is transformed into speech (i.e., text → speech). For example, P1 voiced *"for the auditory, it didn't feel natural because it is text to speech. It feels too unnatural, it feels hard to recognize what they say."* (P1). Another participant, P9, highlighted one of the strengths of visual rendering: to read the translated text multiple times within a certain time. For *MRTranslateAudio* and *MRTranslateAudioWoZ*, the translated term is translated and aurally rendered on the glasses only once, which would require another user input to re-trigger the translation. P2 emphasised the importance of seeing the translated word in text, stating, *"[they] would like to see the word in text, so it is much easier to verify the word. if the word is maybe a correct translation..."* (P2). Despite acknowledged shortcomings of *MRTranslate* with auditory rendering, some participants noted positive aspects. For example, P11 voiced *"the positive thing about only hearing it is that maybe [they] can link both, the thing that [they] hear and [they] see in [their] brain and maybe to learn the language better.* (P11). P3 expressed the necessity of auditory output for communication, saying, *"if I want to communicate with others, then I need the sound in Italian. I want to hear the original term in Italian for communication."* (P3).

Theme 3: Enhancements for* MRTranslate*: Synchronised Multimodal Output and Implicit Real-World Scanning. When asked about potential improvements for *MRTranslate*, we received a large number of comments about the desire of synchronised multimodal output and implicit real-world scanning. A prevailing preference emerged for simultaneous visual and auditory outputs. P1 voiced that *"if you are in a rush, then it is good to have output via two sensory inputs (visual and aural)"* (P1). P9 commented to would *"find it cool if it is not only written down or spoken by an assistant, so if both would be there. So you hear it and read it that would be nice."* (P9). P7 envisioned an additional layer that visually renders the unfamiliar text into a visual representation (e.g., an "augmented fish hologram" next to the term "Pesce").

Concerning implicit real-world scanning, participants advocated for a departure from physical pointing on real-world text. P3 proposed an alternative, suggesting, *"use eye-gaze pointing instead? [...] just gaze at the word"* (P3). There was a general preference of directly translating text using the glasses, without taking an explicit interaction. P10 envisioned a seamless system where they *"just look at the word and then it translates on the lens"* (P10). P5 pondered if they *"can use [their] eyes, like [they] stare at it for 5 seconds, after 5 seconds it will show [them] something."* (P5).

Additional suggestions centered around incorporating a state label indicating the system's confidence in translation, the ability to change translation languages, and augmented visual guidance for pointing at distant objects (e.g., an augmented crosshair).

7 DISCUSSION

We first discuss how *MRTranslate* compares to Google Translate [18] and Google Lens [17], contributing answers to our research question. We then outline promising future research directions and discuss a few limitations that are worth discussing.

From our empirical evaluation, we found that *Google Translate*, *Google Lens* and *MRTranslateVisualWoZ* were perceived as most usable. This is evidenced by participants' task performance, their usage preference, the usability scores, and the perceived workloads during the study tasks (see Figure 5 and Table 1). For example, although participants voiced that *Google Translate* and *Google Lens* require an additional interaction step, i.e., taking out the phone, both experiences resulted in slightly lower raw NASA-TLX scores compared to all experiences using *MRTranslate*.

Both WoZ experiences (i.e., *MRTranslateVisualWoZ* and *MRTranslateAudioWoZ*) were perceived as less demanding and achieved higher usability scores than their fully functional equivalents (i.e., *MRTranslateVisual* and *MRTranslateAudio*). However, both the auditory experiences were perceived as more demanding and yielded a lower usability score than the visual experiences. This finding highlights a clear tendency towards the use of *MRTranslate* in combination with visual output rather than auditory. We observed that users prefer the output modality of *MRTranslate* to be in line with how they experience the original term in the real world. For example, if a user experiences difficulties in understanding text in the real world, then *MRTranslate* should visually render the translated term on the MR glasses rather than translating the term and transforming it into speech. This raises interesting questions about the next steps of assistive MR interfaces. For example, in an educational context where users make use of AR to learn unfamiliar terms, Draxler et al. [10] applied AR labels onto the physical objects in space, whereas Ibrahim et al. [27] combined virtual labels with optional audio output. We observed that users of *MRTranslate* express a strong preference for synchronous multimodal output, i.e., the desire to see the translation displayed on their MR glasses while simultaneously hearing the translated term. As reported in Section 6.4, P1 voiced *"if you are in a rush, then it is good to have output via two sensory inputs (visual and aural)"* (P1). P7 even envisioned an additional visual layer that renders the unfamiliar text into a visual representation of it (e.g., a hologram next to the text).

The key message is that *MRTranslateVisual* and *MRTranslateVisualWoZ* resulted in a positive user experience and usability, close, and sometimes even superior, to Google Translate [18] and Google Lens [17]. Yet, *MRTranslate* with auditory output (e.g., *MRTranslateAudio*) was less preferred.

7.1 Future Research

We discuss three promising research directions for assistive MR.

7.1.1 Design Space of Assistive MR Prototypes. We proposed and evaluated an exemplary assistive MR prototype that, as soon as everyday MR glasses will find widespread adoption, could be beneficial for many people. However, a logical next step is to synthesise a design space of assistive MR experiences, including application scenarios, sensing capabilities, and output modalities. We encourage researchers to work on investigations around synchronised multimodal output, where output is rendered visually and auditory at the same time, as voiced by some participants when using *MRTranslate*. Furthermore, substantial more work is required to capture the rich set of scenarios where assistive MR prototypes can be beneficial for people in their day-to-day tasks [37], which we will discuss in the realm of OpenAI's GPT-4V and Humane's AI Pin in the next section.

7.1.2 There is more to Augment. Building upon *MRTranslate* and a future design space of assistive MR prototypes [37], OpenAI's GPT-4 with vision, often referred to as GPT-4V[1], and Humane's AI Pin [26], a wearable device that projects a visual interface onto a person's palm and comes with a virtual ChatGPT-like assistant [26], pave the way for even more advanced visual understanding and processing of real-world information. Combining these powerful concepts with *MRTranslate* and existing works on olfactory output [34] could result in novel, impactful experiences that advance human perception. For example, smell, taste, and temperature interfaces [7, 34] could be used to augment someone's reality in various scenarios. MR glasses or Humane's AI Pin that can sense the real-world environment may want to augment unfamiliar objects (i.e., a dragon fruit) with the object's actual taste, potentially moving to a new era of human-computer interfaces. While building such systems introduces additional engineering challenges, the HCI community is already designing and implementing low-cost prototypes that have the potential to integrate such experiences into real-world scenarios in the near future [7]. Whilst beyond the scope of our work, such changes in object representations (e.g., text → olfactory experiences) could support people with disabilities in their everyday life [38]. Prototypes like FingerReader [50], designed to help people with visual impairments read printed text on-the-go, could be enhanced by incorporating additional olfactory output to assist individuals in better comprehending real-world information.

7.1.3 Societal Impact of Assistive MR. A future filled with assistive MR experiences prompts important questions about its societal implications and the ethical boundaries of capturing and processing real-world information, especially when it involves bystanders who have not explicitly consented. O'Hagan et al. [43] investigated the concerns of bystanders regarding privacy-invading activities in augmented reality. Their findings emphasised the necessity for AR technologies to consider the nature of the activity and the relationship with bystanders to safeguard the privacy of both users and those inadvertently affected.

While the primary objective of *MRTranslate*, and more broadly, assistive MR technologies, is to enhance everyday life and assist individuals in their day-to-day tasks, it is inevitable to conduct research on the broader societal realm and to design these assistive MR systems with privacy in mind. It is therefore essential to engage in future research and industry deployments around wearable everyday MR experiences that consider the societal impact of systems similar to *MRTranslate*. Equally, it is inevitable to implement potential safeguards against always-on (i.e.,anywhere and at any time) capturing and processing of real-world information. Future research is encouraged to delve into the societal implications of employing assistive MR technology in real-world settings, such as ordering food at a restaurant, as opposed to simulating such scenarios in controlled laboratory settings. Furthermore, it would be interesting to investigate if social contexts, such as emotions and various environmental stimuli, impact how people utilize assistive MR in their everyday life (e.g., are they worried about their gestural interactions when using systems such as *MRTranslate*?).

[1]https://platform.openai.com/docs/guides/vision, last accessed 13/03/2024

7.2 Limitations

First, with *MRTranslate*, we provide an open-source prototype that is capable of sensing and processing real-world information. However, accessing and processing camera streams on MR glasses is often disabled due to privacy [40]. Future work is encouraged to investigate how additional privacy layers can be integrated into experience pipelines to maintain users' and bystanders' privacy when using MR to assist in everyday life. Second, we acknowledge that future experiences should not require users to wear finger gloves for the sensing of the real world, as implemented in *MRTranslate* and in other prototypes in neighbouring fields, such as FingerReader [50]. However, building *MRTranslate* was a necessary and valuable step towards providing a fully functional prototype that already tests the idea of using MR to assist in translation tasks and compares *MRTranslate* to existing translation experiences. Finally, despite an initial training session where we introduced participants to the different experiences, our results might be influenced by participants' high familiarity with Google Translate [18], the current state-of-the-art when translating text.

8 CONCLUSION

In this paper, we introduced and empirically evaluated *MRTranslate*, a MR experience that enables users to translate non-familiar text in their real world. We investigated *MRTranslate* in a user study and compared user experience and usability against state-of-the-art translation apps, including Google Translate [18] and Google Lens [17]. Our study highlighted the strengths of everyday MR glasses for real-world translations, not requiring users to carry around an extra device they would need to retrieve from their pocket. We encourage further exploration and industry advancements in the realm of wearable everyday MR experiences to thoroughly assess the societal consequences of systems like *MRTranslate* and how they can make a positive contribution to society.

ACKNOWLEDGMENTS

This research was supported by the International Postdoctoral Fellowships (IPF, University of St. Gallen). We thank the reviewers for their feedback and the participants for their valuable input.

REFERENCES

[1] Nidhal Baccouri. 2023. *DeepTranslator.* Retrieved October 24, 2023 from https://pypi.org/project/deep-translator/

[2] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual SUS scores mean: Adding an adjective rating scale. *J. of usability studies.*

[3] Eric A. Bier, Maureen C. Stone, Ken Pier, William Buxton, and Tony D. DeRose. 1993. Toolglass and Magic Lenses: The See-through Interface. In *Proc. of the 20th Annual Conference on Computer Graphics and Interactive Techniques* (Anaheim, CA) *(SIGGRAPH '93)*. ACM, New York, NY, USA.

[4] Mike Boland. 2023. *Google Lens Reaches 10 Billion Monthly Searches.* Retrieved October 24, 2023 from https://arinsider.co/2023/02/13/google-lens-reaches-10-billion-monthly-searches

[5] James V Bradley. 1958. Complete counterbalancing of immediate sequential effects in a Latin square design. *J. Amer. Statist. Assoc.* 53, 282 (1958).

[6] John Brooke et al. 1996. SUS-A quick and dirty usability scale. (1996).

[7] Jas Brooks and Pedro Lopes. 2023. Smell & Paste: Low-Fidelity Prototyping for Olfactory Experiences. In *Proc. of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. ACM, New York, NY, USA.

[8] W. Buxton and B. Myers. 1986. A Study in Two-Handed Input. In *CHI* (Boston, Massachusetts, USA) *(CHI '86)*. ACM, New York, NY, USA.

[9] Shakiba Davari, Feiyu Lu, and Doug A. Bowman. 2022. Validating the Benefits of Glanceable and Context-Aware Augmented Reality for Everyday Information Access Tasks. In *IEEE VR*.

[10] Fiona Draxler, Audrey Labrie, Albrecht Schmidt, and Lewis L. Chuang. 2020. Augmented Reality to Enable Users in Learning Case Grammar from Their Real-World Interactions. In *Proc. of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. ACM, New York, NY, USA.

[11] Ruofei Du, Alex Olwal, Mathieu Le Goc, Shengzhi Wu, Danhang Tang, Yinda Zhang, Jun Zhang, David Joseph Tan, Federico Tombari, and David Kim. 2022. Opportunistic Interfaces for Augmented Reality: Transforming Everyday Objects into Tangible 6DoF Interfaces Using Ad Hoc UI. In *Ext. Abstracts of the CHI Conference* (New Orleans, LA, USA) *(CHI EA '22)*. ACM, New York, NY, USA.

[12] Pierre Nicolas Durette. 2023. *gTTS (Google Text-to-Speech).* Retrieved October 24, 2023 from https://pypi.org/project/gTTS/

[13] Steven K Feiner. 2002. Augmented reality: A new way of seeing. (2002).

[14] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *Int. J. of Human–Computer Interaction* (2019).

[15] Ge Gao, Hao-Chuan Wang, Dan Cosley, and Susan R. Fussell. 2013. Same Translation but Different Experience: The Effects of Highlighting on Machine-Translated Conversations. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. ACM, New York, NY, USA, 10 pages.

[16] Ge Gao, Naomi Yamashita, Ari MJ Hautasaari, Andy Echenique, and Susan R. Fussell. 2014. Effects of Public vs. Private Automated Transcripts on Multiparty Communication between Native and Non-Native English Speakers. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. ACM, New York, NY, USA.

[17] Google. 2023. *Google Lens - Search What You See.* Retrieved October 24, 2023 from https://lens.google/

[18] Google. 2023. *Google Translate.* Retrieved October 24, 2023 from https://translate.google.com/about/

[19] Jens Grubert, Tobias Langlotz, Stefanie Zollmann, and Holger Regenbrecht. 2017. Towards Pervasive Augmented Reality: Context-Awareness in Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics* 23, 6 (2017).

[20] Yves Guiard. 1987. Asymmetric division of labor in human skilled bimanual action: The kinematic chain as a model. *Journal of motor behavior* 19, 4 (1987).

[21] Yves Guiard. 1988. The kinematic chain as a model for human asymmetrical bimanual cooperation. In *Advances in Psychology*. Vol. 55. Elsevier.

[22] Kotaro Hara and Shamsi T. Iqbal. 2015. Effect of Machine Translation in Interlingual Conversation: Lessons from a Formative Study. In *Proc. of the CHI Conference* (Seoul, Republic of Korea) *(CHI '15)*. ACM, New York, NY, USA.

[23] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proc. of the human factors and ergonomics society annual meeting*. Sage publications.

[24] Steven Henderson and Steven Feiner. 2010. Opportunistic Tangible User Interfaces for Augmented Reality. *IEEE TVCG* 16, 1 (2010).

[25] Samuel Hoffstaetter. 2023. *Python-tesseract.* Retrieved October 24, 2023 from https://pypi.org/project/pytesseract/

[26] Humane. 2023. *Humane AI Pin: Beyond touch, beyond screens.* Retrieved November 10, 2023 from https://hu.ma.ne/aipin

[27] Adam Ibrahim, Brandon Huynh, Jonathan Downey, Tobias Höllerer, Dorothy Chun, and John O'donovan. 2018. Arbis pictus: A study of vocabulary learning with augmented reality. *IEEE TVCG* (2018).

[28] Itseez 2014. *The OpenCV Reference Manual* (2.4.9.0 ed.). Itseez.

[29] Paul Kabbash, William Buxton, and Abigail Sellen. 1994. Two-Handed Input in a Compound Task. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, Massachusetts, USA). ACM, New York, NY, USA.

[30] Thomas K Landauer. 1986. Psychology as a mother of invention. *ACM* (1986).

[31] Thibault Louis, Jocelyne Troccaz, Amélie Rochet-Capellan, Nady Hoyek, and François Bérard. 2020. When High Fidelity Matters: AR and VR Improve the Learning of a 3D Object. In *Proceedings of the International Conference on Advanced Visual Interfaces* (Salerno, Italy) *(AVI '20)*. ACM, New York, NY, USA.

[32] Feiyu Lu and Doug A. Bowman. 2021. Evaluating the Potential of Glanceable AR Interfaces for Authentic Everyday Uses. In *2021 IEEE VR*.

[33] Feiyu Lu, Leonardo Pavanatto, and Doug A. Bowman. 2023. In-the-Wild Experiences with an Interactive Glanceable AR System for Everyday Use. In *Proc. of the 2023 ACM SUI* (Sydney, NSW, Australia) *(SUI '23)*. ACM, New York, NY, USA.

[34] Jasmine Lu, Ziwei Liu, Jas Brooks, and Pedro Lopes. 2021. Chemical Haptics: Rendering Haptic Sensations via Topical Stimulants. In *ACM UIST* (Virtual Event, USA) *(UIST '21)*. ACM, New York, NY, USA.

[35] Andrés Lucero. 2015. Using Affinity Diagrams to Evaluate Interactive Prototypes. In *Human-Computer Interaction – INTERACT 2015*, Julio Abascal, Simone Barbosa, Mirko Fetter, Tom Gross, Philippe Palanque, and Marco Winckler (Eds.). Springer International Publishing, Cham, 231–248.

[36] Wendy E. Mackay. 1998. Augmented Reality: Linking Real and Virtual Worlds: A New Paradigm for Interacting with Computers. In *Proceedings of the Working Conference on Advanced Visual Interfaces* (L'Aquila, Italy) *(AVI '98)*. Association for Computing Machinery, New York, NY, USA.

[37] Florian Mathis. 2024. Everyday Life Challenges and Augmented Realities: Exploring Use Cases For, and User Perspectives on, an Augmented Everyday Life. In *Proceedings of the Augmented Humans International Conference 2024 (AHs '24)*. Association for Computing Machinery, New York, NY, USA.

[38] Florian Mathis, Jolie Bonner, Joseph O'Hagan, and Mark McGill. 2023. Breaking Boundaries: Harnessing Mixed Reality to Enhance Social Engagement.

[39] David Maulsby, Saul Greenberg, and Richard Mander. 1993. Prototyping an intelligent agent through Wizard of Oz. In *Proc. of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*.

[40] Meta. 2023. *Meta Quest Pro: Built with Privacy in Mind.* Retrieved October 24, 2023 from https://www.meta.com/de-de/blog/quest/meta-quest-pro-privacy/

[41] Paul Milgram and Fumio Kishino. 1994. A taxonomy of mixed reality visual displays. *IEICE TRANSACTIONS on Information and Systems* 77, 12 (1994).

[42] Shohei Mori, Sei Ikeda, and Hideo Saito. 2017. A survey of diminished reality: Techniques for visually concealing, eliminating, and seeing through real objects. *IPSJ Transactions on Computer Vision and Applications* 9, 1 (2017).

[43] Joseph O'Hagan, Pejman Saeghe, Jan Gugenheimer, Daniel Medeiros, Karola Marky, Mohamed Khamis, and Mark McGill. 2023. Privacy-Enhancing Technology and Everyday Augmented Reality: Understanding Bystanders' Varying Needs for Awareness and Consent. *IMWUT* (2023).

[44] Cathal O'Madagain, Gregor Kachel, and Brent Strickland. 2019. The origin of pointing: Evidence for the touch hypothesis. *Science Advances* 5, 7 (2019).

[45] Ken Pfeuffer, Jason Alexander, and Hans Gellersen. 2016. Partially-Indirect Bimanual Input with Gaze, Pen, and Touch for Pan, Zoom, and Ink Interaction. In *Proc. of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. ACM, New York, NY, USA.

[46] Ingrid Piller. 2022. *Language Barriers to Social Participation.* Retrieved October 24, 2023 from https://www.languageonthemove.com/language-barriers-to-social-participation/

[47] Thammathip Piumsomboon, Adrian Clark, Mark Billinghurst, and Andy Cockburn. 2013. User-Defined Gestures for Augmented Reality. In *CHI '13 Ext. Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, USA.

[48] Michael Rohs, Johannes Schöning, Martin Raubal, Georg Essl, and Antonio Krüger. 2007. Map Navigation with Mobile Devices: Virtual versus Physical Movement with and without Visual Context. In *Proc. of the Int. Conference on Multimodal Interfaces* (Nagoya, Aichi, Japan). ACM, New York, NY, USA.

[49] Hanna Kathrin Schraffenberger. 2018. *Arguably augmented reality: relationships between the virtual and the real.* Ph. D. Dissertation. Leiden University.

[50] Roy Shilkrot, Jochen Huber, Wong Meng Ee, Pattie Maes, and Suranga Chandima Nanayakkara. 2015. FingerReader: A Wearable Device to Explore Printed Text on the Go. In *Proc. of the ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. ACM, New York, NY, USA.

[51] Maximilian Speicher, Brian D. Hall, and Michael Nebeling. 2019. What is Mixed Reality?. In *Proc. of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA.

[52] Takumi Toyama, Daniel Sonntag, Andreas Dengel, Takahiro Matsuda, Masakazu Iwamura, and Koichi Kise. 2014. A mixed reality head-mounted text translation system using eye gaze input. In *Proc. of the International Conference on Intelligent User Interfaces (IUI '14)*. ACM, New York, NY, USA.

[53] Mark Weiser. 1991. The Computer for the 21 st Century. (1991).

[54] Jacob O. Wobbrock and Julie A. Kientz. 2016. Research Contributions in Human-Computer Interaction. *Interactions* 23, 3 (apr 2016).

[55] Shengdong Zhao, Felicia Tan, and Katherine Fennedy. 2023. Heads-Up Computing Moving Beyond the Device-Centered Paradigm. *Commun. ACM* (2023).

[56] Annuska Zolyomi, Anushree Shukla, and Jaime Snyder. 2017. Technology-Mediated Sight: A Case Study of Early Adopters of a Low Vision Assistive Technology. In *Proc. of the 19th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '17)*. ACM, New York, NY, USA.