Designing Multimodal Interactions in Medical Augmented Reality

Laura Schütz
Chair for Computer Aided Medical Procedures and Augmented Reality
Technical University of Munich
Munich, Germany
laura.schuetz@tum.de

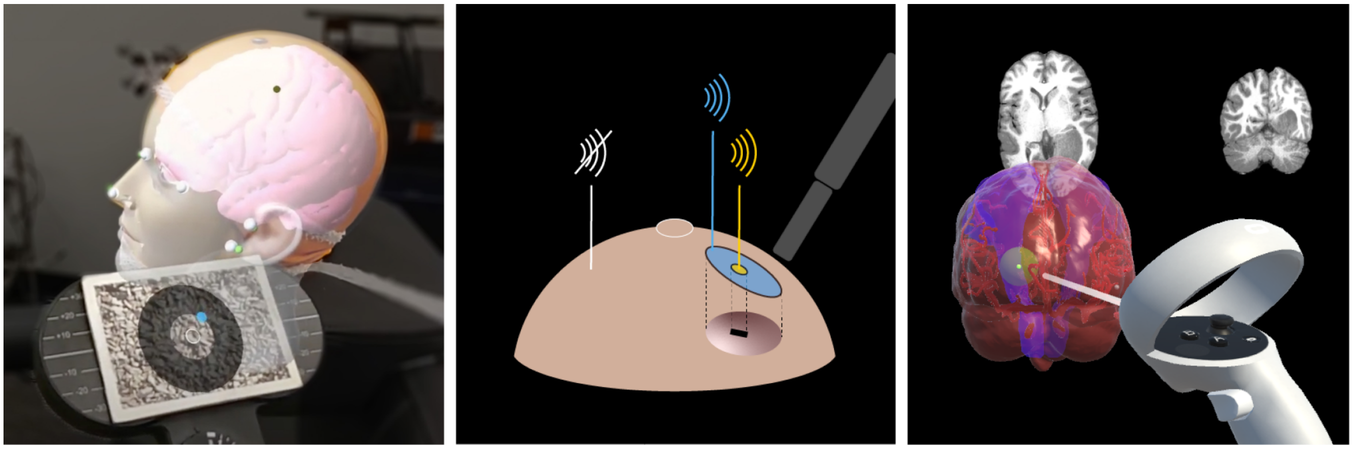

Figure 1: Medical augmented reality projects using audiovisual and audio feedback; Left: Audiovisual augmented reality system for coil placement during transcranial magnetic stimulation; Middle: Shape sonification for breast cancer localization; Right: Multimodal medical image interaction for brain tumor localization

Abstract

Surgical procedures demand exceptional spatial and temporal coordination, precision, and advanced motor skills, challenging surgeons' cognitive abilities. Although numerous augmented reality (AR) systems have been developed to assist in the precise localization of target regions through anatomical visualizations, they often add visual complexity and increase the information burden during high-intensity tasks. In these situations of information overload, multisensory feedback can help to disambiguate unisensory representations, resulting in more robust interpretations and enhancing task performance. This work aims to design and evaluate multimodal interactions in AR that enhance user performance and reduce cognitive load during surgical procedures, focusing primarily on audiovisual interactions. Preliminary findings indicate that audiovisual interactions improve accuracy and reduce cognitive load during surgical alignment and localization tasks. Future work might explore cognitive load-adaptive audiovisual augmentations, which could further enhance surgeon performance and user experience.

CCS Concepts

• **Human-centered computing** → **Mixed / augmented reality**; **Virtual reality**; **Auditory feedback**; **User studies**.

Keywords

multimodal interaction, multisensory feedback, audiovisual feedback, sonification, user interface design, augmented reality, virtual reality, surgical guidance, computer-aided surgery

ACM Reference Format:
Laura Schütz. 2025. Designing Multimodal Interactions in Medical Augmented Reality. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3706599.3707613

1 Introduction

The medical field is a prominent area of application of augmented reality (AR), as diagnostic and treatment-relevant information can be superimposed atop the patient [18]. Many medical AR user interfaces (UI) have focused on unimodal visual interaction between the user and the system. However, the human way of interacting with the world is multimodal in its nature, as we use multiple senses to explore our environment [31]. Research has shown that causally linked multisensory stimuli can improve task performance and reduce cognitive load [28]. This makes multimodal interactions especially promising for designing medical AR interfaces, as they can enhance efficiency, reduce errors, and offer a variety of sensory feedback options to accommodate users' diverse needs and use patterns

CHI EA '25, Yokohama, Japan
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1395-8/25/04
https://doi.org/10.1145/3706599.3707613

[20]. This freedom of choice is particularly suitable for surgical applications, as the operating room is a rich sensory environment. The surgeon must integrate multiple simultaneous streams of visual, auditory, and haptic stimuli. Visual information from the surgical site and the medical images, auditory stimuli from conversations, and sonified physiological signals such as heart rate monitors, as well as haptic feedback from the interaction of the surgical tool with the patient's anatomy, need to be processed and integrated. Although visual, audio, and haptic feedback could potentially be used in the operating room, haptic feedback, for example, provided by wrist-worn devices [33], is not ideal for augmenting precise surgical tasks, as it can interfere with delicate haptic sensations crucial for detecting tool-tissue interactions. Visual feedback is well-suited for conveying spatial information, as the human visual system excels at organizing objects efficiently in space. However, it struggles with tracking multiple dynamic objects simultaneously [7]. In AR environments, visual augmentations may also distract from the primary task if not designed properly. In contrast, auditory perception excels at detecting subtle changes in complex auditory stimuli. Audio signals are processed quickly and do not require direct visual attention, as they operate omnidirectionally.

Therefore, my work aims to leverage the benefits of the visual and the auditory domain through multimodal audiovisual interactions. Best practices for the human-centered design of audiovisual interactions in AR have yet to be explored for many medical procedures.

2 Related Work

2.1 Multisensory Perception

In sensory overload situations, such as busy operating rooms, multisensory feedback can help clarify ambiguous unisensory inputs, leading to more robust interpretations [30]. Multisensory cues are also more effective at capturing spatial attention compared to unisensory cues, especially in situations involving concurrent perceptual load or multitasking [23]. The combination of independent but causally related information sources has been shown to enhance information processing [28]. Research suggests that integrating multiple senses, particularly when auditory cues are embedded in complex sensory environments, improves performance across a variety of tasks [19, 32]. This improvement is linked to attentional mechanisms [2, 15] and modulatory effects on working memory [5]. Additionally, studies on learning show that distributing information across multiple sensory channels, rather than overloading a single one, results in better performance and recall [21]. The nature of this multisensory integration depends on several factors, such as whether the sensory modalities provide redundant information about an event and whether there is competition for cognitive resources during response or decision-making [8, 12]. By averaging signals, the nervous system can reduce noise and expand the perceptual space, allowing for more refined distinctions between sensory inputs.

2.2 Audiovisual Medical Augmented Reality

Several studies have explored sonification as an alternative or addition to visual surgical guidance. Roodaki et al., for instance, found that auditory feedback improved angle alignment accuracy over visual guidance in needle placement for eye surgery [22]. Matinfar et al. used auditory stimuli for tool navigation, including a four-dimensional sonification system for surgical instrument alignment during screw placement in spine surgery [16, 17]. Other relevant works include Black et al., who implemented auditory feedback using pitch and stereo panning for needle placement [3], and Ziemer et al., who mapped the spatial position of a surgical tool tip relative to a target, encoding direction through sound pitch and distance through beat frequency [34]. These studies collectively show that sonification can be as effective as visualization in conveying tool location and angle information. Among studies evaluating audiovisual interactions is a work by Bork et al., whose audiovisual AR system improved 3D localization perception and needle placement accuracy using visuotemporal guidance [4].

3 Research Objectives

While prior research highlights the advantages of audio and audiovisual feedback over visual feedback in medical AR, the potential of multimodal interactions stays underexplored. Many open questions on the design of audiovisual feedback remain, e.g., regarding redundancy or complementarity of the multisensory cues and what task or context-specific factors make multimodal interactions more effective than unimodal ones. This research aims to answer the following set of open questions:

- Q1: Can multimodal interactions in medical AR increase task performance and reduce cognitive load during surgical localization tasks?
- Q2: Which information is best conveyed visually, auditorily, or audio-visually during surgical localization tasks?
- Q3: Can modality-adaptive user interfaces based on physiological cognitive load measurements enhance the performance and user experience during surgical localization tasks?

4 Research Approach

When first approaching the user interface design for a surgical procedure, design methods such as user observations, semi-structured user interviews, and hierarchical task analysis are used to understand the medical user's context, tasks, and needs. Research ideas and concepts are generated upon learnings from the preceeding phase of user research. Design solutions are then continuously refined and adapted via user testing. Finally, the UI is evaluated in terms of quantitative and qualitative measures in a simulated surgery setting. Quantitative measures, such as task time and localization accuracy, are used to assess user performance. Physiological measures captured using eye tracking and EEG are used to determine cognitive load. Qualitative feedback is gathered using the NASA-TLX [11] and System Usability Scale [13] questionnaires to evaluate the usability of the proposed UIs.

From a UI design point of view, the following high-level questions need to be considered throughout the projects. These simple yet essential questions in reference to the framework for model-based UI adaptation by Abrahão et al. [1] present additional guidelines for the research:

- What: Which information is relevant to the user?
- When: At which point in time does the user require the information?

- How: Which feedback modality/modalities is/are the most suitable to convey this information?
- Where: How are feedback elements placed and adapted to achieve the best usability?

5 Results

5.1 Complimentary audiovisual feedback is modality invariant.

In an initial research project, we developed an audiovisual guidance system for Transcranial Magnetic Stimulation (TMS), a treatment for major depressive disorder. TMS involves positioning an electromagnetic coil on a patient's head to deliver repetitive stimulation to a specific brain region [14]. Accurate placement of the coil is crucial for effective treatment outcomes [6]. While neuronavigation systems are considered the gold standard for ensuring precise stimulation [9], I proposed an alternative: an audiovisual Augmented Reality (AR) system for coil positioning (Figure 1 - left) [26]. This system provides real-time sonification and visualization of the translational and rotational differences between the target location on the head and the coil's current position. To evaluate the system's effectiveness, a user study was conducted to assess the impact of cross-modal integration on usability and targeting precision. The study compared two multimodal audiovisual AR interfaces against purely auditory and visual feedback conditions. The results demonstrated significant reductions in task completion time for both multimodal AR conditions compared to visual neuronavigation. Notably, the auditory-only condition performed similarly to the audiovisual interfaces. There were no significant differences between the two audiovisual conditions, which varied the assignment of modality to spatial parameters. These findings suggest that due to cross-modal integration, complementary audiovisual feedback regarding distance and angle is modality invariant.

5.2 Sonification can precisely convey shape information.

A second project on sound feedback in surgical interventions was motivated by the high breast cancer reoperation rates due to imprecise tumor margin localization. We developed an auditory display using shape sonification to improve tumor margin localization (Figure 1 - middle) [25]. Accuracy and usability of the interactive shape sonification were determined on models of the female breast in three user studies with both breast surgeons and non-clinical participants. The comparative studies showed a significant increase in usability ($p<0.05$) and localization accuracy ($p<0.001$) of the shape sonification over the auditory feedback currently used in surgery.

5.3 Physics-based multimodal feedback enhances localization accuracy.

Having shown the benefits of audiovisual interaction in medical AR, a subsequent work used these findings to build a framework that enables audiovisual interaction with any part of the human body during surgical procedures. Clinicians often face challenges in forming a dynamic mental model of 3D tissue location during surgery, despite the availability of advanced imaging technologies like computed tomography (CT) and magnetic resonance imaging (MRI) [10]. To tackle this issue, we created the Multimodal Medical Image Interaction (MMII) framework [27]. This framework dynamically visualizes human tissue in a 3D virtual reality environment and provides physics-based, real-time audiovisual feedback (Figure 1 - right). MMII employs a model-based sonification approach to generate sounds based on the geometry and physical properties of tissue, eliminating the need for hand-crafted sound design. We conducted two user studies with 34 general experts and nine clinical specialists to evaluate the framework's learnability, usability, and accuracy. The results demonstrated excellent learnability of the audiovisual correspondences, with a significant increase in correct associations ($p < 0.001$) throughout the study. Furthermore, MMII achieved superior accuracy in brain tumor localization ($p < 0.05$) compared to conventional medical image interaction methods.

6 Next Steps

Research Question 1 has been addressed, confirming the performance and cognitive load benefits of multimodal interactions in medical AR through two research projects (5.1, 5.3). Research Question 2 has been partially explored (5.1) but requires further investigation into cognitive psychology and controlled experiments that expose participants to the same information presented purely visually, auditorily, or through redundant audiovisual means. Defining an abstract localization task that allows generalized insights for all surgical procedures remains an open challenge.

To address Research Question 3, expressive and reliable physiological measures of cognitive load for surgical localization tasks must be identified. Eye-tracking, electrocardiography, and electroencephalography data have already been collected during virtual reality-based eye surgery simulations. Preliminary results from a pilot study indicate that these measures can effectively assess cognitive load. The following steps involve analyzing and fusing the multimodal physiological data, implementing and refining a UI adaptation algorithm, and conducting a user study with ophthalmologists. This study will investigate the effects of cognitive load-adaptive multimodal AR on user performance and experience in eye surgery, determining whether adaptive multimodal UIs can outperform traditional multimodal AR in terms of efficiency (task time), effectiveness (task performance), and overall user experience.

7 Long-Term Goals

Looking ahead, the ultimate pursuit of this research is to design multimodal UIs that seamlessly integrate with the user and their environment. UIs should automatically adapt to the user's context, cognitive, and emotional state to provide the right information at the right time. Advances in computer vision that allow for semantic scene understanding can enable sophisticated context-aware UIs [29]. Emotional awareness can, for example, be achieved through analysis of facial expressions or acoustic features in speech [24]. This rich user information captured to enable adaptive user interfaces can help design effective surgeon-anatomy interactions. These multilayered adaptation properties promise the advent of truly natural human-computer interaction by eliminating user frustration due to mental model misalignment when interacting with spatial computing systems.

References

[1] Silvia Abrahão, Emilio Insfran, Arthur Sluÿters, and Jean Vanderdonckt. 2021. Model-based intelligent user interface adaptation: challenges and future directions. *Software and Systems Modeling* 20, 5 (2021), 1335–1349.

[2] Lucas Battich, Merle Fairhurst, and Ophelia Deroy. 2020. Coordinating attention requires coordinated senses. *Psychonomic bulletin & review* 27 (2020), 1126–1138.

[3] David Black, Julian Hettig, Maria Luz, Christian Hansen, Ron Kikinis, and Horst Hahn. 2017. Auditory feedback to support image-guided medical needle placement. *International Journal of Computer Assisted Radiology and Surgery* 12 (2017), 1655–1663.

[4] Felix Bork, Bernhard Fuers, Anja-Katharina Schneider, Francisco Pinto, Christoph Graumann, and Nassir Navab. 2015. Auditory and Visio-Temporal Distance Coding for 3-Dimensional Perception in Medical Augmented Reality. In *2015 IEEE International Symposium on Mixed and Augmented Reality*. IEEE, New York, NY, USA, 7–12. https://doi.org/10.1109/ISMAR.2015.16

[5] Fabiano Botta, Valerio Santangelo, Antonino Raffone, Daniel Sanabria, Juan Lupiáñez, and Marta Olivetti Belardinelli. 2011. Multisensory integration affects visuo-spatial working memory. *Journal of Experimental Psychology: Human Perception and Performance* 37, 4 (2011), 1099.

[6] Stewart Denslow, Daryl E. Bohning, Peter A. Bohning, Mikhail P. Lomarev, and Mark S. George. 2005. An increased precision comparison of TMS-induced motor cortex BOLD fMRI response for image-guided versus function-guided coil placement. *Cognitive and behavioral neurology* 18, 2 (2005), 119–126.

[7] Benjamin J. Dixon, Michael J. Daly, Harley Chan, Allan Vescan, Ian J. Witterick, and Jonathan C. Irish. 2014. Inattentional blindness increased with augmented reality surgical navigation. *American journal of rhinology & allergy* 28, 5 (2014), 433–437.

[8] Marc O. Ernst and Massimiliano Di Luca. 2011. Multisensory perception: from integration to remapping. *Sensory cue integration* 15 (2011), 224–250.

[9] Paul B. Fitzgerald, Kate Hoy, Susan McQueen, Jerome J. Maller, Sally Herring, Rebecca Segrave, Michael Bailey, Greg Been, Jayashri Kulkarni, and Zafiris J. Daskalakis. 2009. A randomized trial of rTMS targeted with MRI based neuronavigation in treatment-resistant depression. *Neuropsychopharmacology* 34, 5 (2009), 1255–1262.

[10] Jung-Leng Foo, Marisol Martinez-Escobar, Bethany Juhnke, Keely Cassidy, Kenneth Hisley, Thom Lobe, and Eliot Winer. 2013. Evaluating mental workload of two-dimensional and three-dimensional visualization for anatomical structure localization. *Journal of Laparoendoscopic & Advanced Surgical Techniques* 23, 1 (2013), 65–70.

[11] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*. Vol. 52. Elsevier, Amsterdam, Netherlands, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[12] Hannah B. Helbig and Marc O. Ernst. 2008. *Haptic perception in interaction with other senses*. Birkhäuser Basel, Basel, Switzerland, 235–249. https://doi.org/10.1007/978-3-7643-7612-3_18

[13] James R. Lewis. 2018. The System Usability Scale: Past, Present, and Future. *International Journal of Human–Computer Interaction* 34, 7 (2018), 577–590.

[14] Colleen K. Loo and Philip B. Mitchell. 2005. A review of the efficacy of transcranial magnetic stimulation (TMS) treatment for depression, and current and future strategies to optimize efficacy. *Journal of affective disorders* 88, 3 (2005), 255–267.

[15] Jessica Lunn, Amanda Sjoblom, Jamie Ward, Salvador Soto-Faraco, and Sophie Forster. 2019. Multisensory enhancement of attention depends on whether you are already paying attention. *Cognition* 187 (2019), 38–49.

[16] Sasan Matinfar, Shervin Dehghani, Michael Sommersperger, Koorosh Faridpooya, Merle Fairhurst, and Nassir Navab. 2024. Ocular Stethoscope: Auditory Support for Retinal Membrane Peeling. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel (Eds.). Springer Nature Switzerland, Cham, 433–443.

[17] Sasan Matinfar, Mehrdad Salehi, Daniel Suter, Matthias Seibold, Shervin Dehghani, Navid Navab, Florian Wanivenhaus, Philipp Fürnstahl, Mazda Farshad, and Nassir Navab. 2023. Sonification as a reliable alternative to conventional visual surgical navigation. *Scientific Reports* 13, 1 (2023), 5930.

[18] Nassir Navab, Alejandro Martin-Gomez, Matthias Seibold, Michael Sommersperger, Tianyu Song, Alexander Winkler, Kevin Yu, and Ulrich Eck. 2023. Medical Augmented Reality: Definition, Principle Components, Domain Modeling, and Design-Development-Validation Process. *Journal of Imaging* 9, 1 (2023), 1–21. https://doi.org/10.3390/jimaging9010004

[19] Mary Kim Ngo and Charles Spence. 2010. Auditory, tactile, and multisensory cues facilitate search for dynamic visual stimuli. *Attention, Perception, & Psychophysics* 72, 6 (2010), 1654–1665.

[20] Sharon Oviatt. 1999. Ten myths of multimodal interaction. *Commun. ACM* 42, 11 (1999), 74–81.

[21] A Ravishankar Rao. 2016. A spatio-temporal model of multi-sensory learning that demonstrates improved object recall. In *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE Computer Society, Vancouver, BC, Canada, 5043–5050.

[22] Hessam Roodaki, Navid Navab, Abouzar Eslami, Christopher Stapleton, and Nassir Navab. 2017. Sonifeye: Sonification of visual information using physical modeling sound synthesis. *IEEE transactions on Visualization and Computer Graphics* 23, 11 (2017), 2366–2371.

[23] Olli Rummukainen and Catarina Mendonça. 2016. Task-relevant spatialized auditory cues enhance attention orientation and peripheral target detection in natural scenes. *Journal of Eye Movement Research* 9, 1 (2016), 1–10.

[24] Björn Schuller, Manfred Lang, and Gerhard Rigoll. 2002. Multimodal emotion recognition in audiovisual communication. In *Proceedings. IEEE international conference on multimedia and expo*, Vol. 1. IEEE Computer Society, Lausanne, Switzerland, 745–748.

[25] Laura Schütz, Trishia El Chemaly, Emmanuelle Weber, Anh Thien Doan, Jacqueline Tsai, Christoph Leuze, Bruce Daniel, and Nassir Navab. 2024. Interactive Shape Sonification for Tumor Localization in Breast Cancer Surgery. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 885, 15 pages. https://doi.org/10.1145/3613904.3642257

[26] Laura Schütz, Emmanuelle Weber, Wally Niu, Bruce Daniel, Jennifer McNab, Nassir Navab, and Christoph Leuze. 2023. Audiovisual augmentation for coil positioning in transcranial magnetic stimulation. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 11, 4 (2023), 1158–1165. https://doi.org/10.1080/21681163.2022.2154277

[27] Laura Schütz, Sasan Matinfar, Gideon Schafroth, Navid Navab, Merle Fairhurst, Arthur Wagner, Benedikt Wiestler, Ulrich Eck, and Nassir Navab. 2024. A Framework for Multimodal Medical Image Interaction. *IEEE Transactions on Visualization and Computer Graphics* 30, 11 (2024), 7419–7429. https://doi.org/10.1109/TVCG.2024.3456163

[28] Ladan Shams and Aaron R. Seitz. 2008. Benefits of multisensory learning. *Trends in cognitive sciences* 12, 11 (2008), 411–417.

[29] Zinovia Stefanidi, George Margetis, Stavroula Ntoa, and George Papagiannakis. 2022. Real-time adaptation of context-aware intelligent user interfaces, for enhanced situational awareness. *IEEE Access* 10 (2022), 23367–23393.

[30] Barry E. Stein. 2012. *The New Handbook of Multisensory Processing*. The MIT Press, Cambridge, MA. https://doi.org/10.7551/mitpress/8466.001.0001

[31] Matthew Turk. 2014. Multimodal interaction: A review. *Pattern recognition letters* 36 (2014), 189–195.

[32] Erik Van der Burg, Christian Olivers, Adelbert Bronkhorst, and Jan Theeuwes. 2008. Pip and pop: nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology: Human Perception and Performance* 34, 5 (2008), 1053.

[33] Kyle T. Yoshida, Zane A. Zook, Hojung Choi, Ming Luo, Marcia K. O'Malley, and Allison M. Okamura. 2024. Design and Evaluation of a 3-DoF Haptic Device for Directional Shear Cues on the Forearm. *IEEE Transactions on Haptics* 17, 3 (2024), 483–495. https://doi.org/10.1109/TOH.2024.3365669

[34] Tim Ziemer, David Black, and Holger Schultheis. 2017. Psychoacoustic sonification design for navigation in surgical interventions. In *Proceedings of Meetings on Acoustics*, Vol. 30. Acoustical Society of America, Melville, NY, USA, 050005.