The Role of AI in Mixed and Augmented Reality Interactions

Tanya R. Jonker
Ruta Desai
Kevin Carlberg
James Hillis
Sean Keller
Hrvoje Benko
Facebook Reality Labs
Redmond, WA 98052, USA
tanya.jonker@fb.com
rutadesai@fb.com
carlberg@fb.com
jmchillis@fb.com
seankeller@fb.com
benko@fb.com

CHI 2020 Extended Abstracts, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6819-3/20/04.
DOI: https://doi.org/10.1145/3334480.XXXXXXX
*update the above block & DOI per your rightsreview confirmation (provided after acceptance)

Abstract
Mixed and augmented reality (XR) devices blur the boundaries between the physical and the digital worlds, resulting in user inputs that are noisy and unreliable and an interaction environment that is not fully known to the system. And yet these devices can display information to the user in a low-friction manner that is more tightly coupled to one's body and the physical environment, exposing the opportunity for persistent "always-on" assistance. We argue that to build effective XR interactions, we must (1) reduce system uncertainty by understanding the user and environment, and (2) effectively adapt the interface to engage the user in uncertainty reduction and to allow for online learning and personalization. Modern methods in AI and machine learning are important to achieve these goals.

Author Keywords
Machine learning; artificial intelligence; augmented reality; mixed reality; context-aware computing

CSS Concepts
• **Human-centered computing~Human computer interaction (HCI)**; *HCI theory, concepts and models*

Introduction
Head-mounted mixed and augmented reality (XR) devices provide the opportunity for an entirely new era

of personal computing. For the first time, XR technologies may allow users to benefit from persistent "always-on" assistance that is integrated seamlessly into their physical world, personalized to their goals, and assistive without being disruptive. In contrast to current personal computing devices, such as a laptop or smartphone, XR technologies are characterized by four distinguishing attributes: (1) XR devices can display outputs in a more accessible, lower-friction manner that is more tightly coupled to the physical environment (e.g., labels placed on real-world objects); (2) the interaction environment cannot be fully known or prespecified by the designer (e.g., virtual content is overlaid on the unknown real world); (3) user inputs are ambiguous, low precision, and noisy (e.g., hand gestures, speech); and (4) XR devices can ingest large volumes of data that is highly informative of user context (e.g., egocentric video, gaze, audio).

These important differences highlight the need for a completely new interaction paradigm beyond what is currently used in personal computing. In particular, the first attribute exposes an exciting opportunity for XR devices by enabling *low-friction, "always-on" assistance*; this contrasts with the traditional "opt-in" paradigm, wherein the user explicitly initiates interaction with the device by picking it up and launching the app of interest. However, the second and third attributes indicate that XR-interaction research must focus on *reducing uncertainty about user context and intent*. To this end, the fourth attribute implies that the device likely has access to enough data to make each of these efforts viable through the application of modern AI and machine-learning techniques.

We posit that AI and machine learning are important for *reducing uncertainty about user context and intent*, which will ultimately allow XR to provide *low-friction, "always-on" assistance.* We view these as the most important problems facing XR interaction researchers. In service of these goals, we share a framework that leverages AI and machine learning to (1) gain an understanding of the user's context, (2) sense intentional input to the system, (3) infer the user's interaction goals within the environment, (4) engage the user in uncertainty reduction as needed, and (5) personalize and refine these interactions by learning from past user actions.

Robust interaction through input and context modeling

We propose a general framework for how XR systems might leverage machine learning, AI, and an adaptive user interface (UI) to produce high quality XR interactions, which is shown in Figure 1.

One important aspect of the proposed framework is to leverage the large volume of data ingested by the XR device, such as eye-gaze on targets of interest or knowledge of the current location (e.g., office), to imbue the system with **context understanding**. Here, without any explicit user input, machine-learning methods can be deployed to infer important contextual elements of the user's environment, such as the objects of interest and the user's activity. Such contextual features, in turn, are highly informative of the set of interactions the user is most likely to desire. Context understanding enables the system to increase understanding of the unknown interaction environment, and it can be used to enhance input sensing, which is also a critical component of our framework.

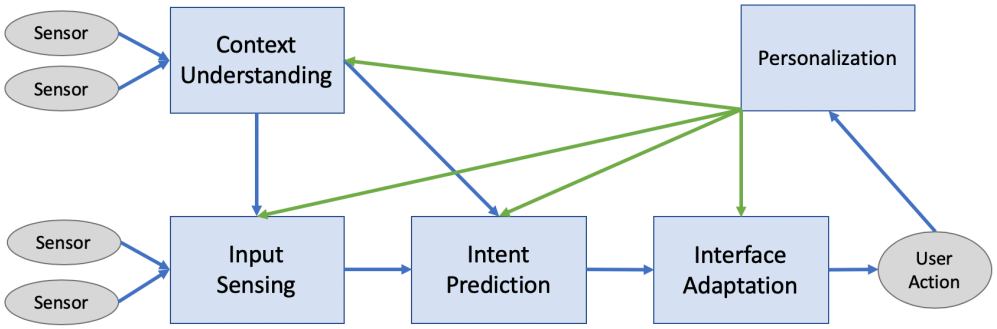

Figure 1: A visual diagram of our AI-assisted interaction framework for XR systems. This framework represents an example of how different models (shown in blue) can interact to produce compelling XR interactions.

Core **input sensing** captures sensing of explicit system input, such as gesture recognition or voice commands. In the real world, these sensing modalities are noisy and may not always be driven by intentional interaction with the system (e.g., pointing for emphasis during conversation rather than selection in the system). Thus, one important role of machine learning is to disambiguate explicit commands from natural movements. Note that this pipeline can leverage data not only from multiple sensors, but also from context-understanding models, which can lend insight into whether natural movements are expected given the present context. For example, if the system detects that the user is engaged in conversation, the system might adjust the threshold for detection of pointing gestures in input sensing to avoid false positives.

Intent prediction involves predicting the interaction goals of the user. For example, Henrikson et al. [3] used hand motion, head motion, and eye gaze to make predictions about where a user might point in VR. Similarly, Desai et al. [1] enabled users to specify their high-level intent explicitly in an interactive design task. The specified high-level intent combined with user interactions over time allowed the system to fine-tune and customize the design suggestions for the user. These two examples involve prediction of intentional system interactions, but a system might also infer implicit intent. For example, Gebhardt et al. [2] used reinforcement learning to show an object's label in a visual search task based on eye movement data (see Figure 2). This model can learn the categories of features that are of interest to a user, effectively predicting where they might gaze next.

Robust interaction through UI adaptation, user feedback, and personalization

Building models of input sensing, context understanding, and intent inference is a massive and challenging endeavor, and these models will likely never provide perfect predictions of the user's precise interaction goal. To ensure that XR interactions produce

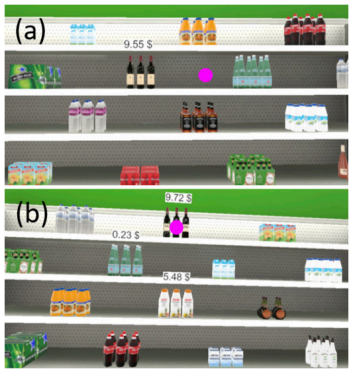

Figure 2: A gaze-driven intent prediction example from Gebhardt et al. [2]: Behaviour of two different policies trained and tested in a supermarket scenario. (A) Policy trained on data where the user was instructed to search for wine. Here, the policy correctly displays only labels of a single item of interest (B) Policy trained on data where the user was instructed to search for wine, water, and juice. Here, the policy displays the labels of multiple items from the target categories while hiding other irrelevant drinks.

a positive user experience despite this, we propose two system features: (1) the UI adapts to underlying uncertainty in the system, engaging the user to reduce uncertainty as needed; and (2) user feedback is used to refine and personalize the underlying models.

Interface adaptation is inspired by principles of the mixed-initiative systems proposed by Horvitz [4], wherein the AI models and users collaborate efficiently to achieve the user's goals. The adaptations explicitly engage the user to provide feedback about the errors that may be produced by the net uncertainty of the AI models. For example, the UI might invoke a dialog for confirmation, or it might prompt the user to dwell longer during pointing to reduce uncertainty.

These UI adaptations will lead to **user actions**, which can provide a feedback signal to models for iterative improvement and, ultimately, **personalization**. There are three types of user actions that a system might leverage. *Confirmation actions* involve a positive or negative response from the user to the UI prompt. For example, when the UI produces a dialogue, the user can finalize the interaction by confirming, which can be used to execute the action and as reinforcement feedback to the underlying models. During *corrective actions*, the user corrects an output from the system. For example, a user might undo the spelling autocorrection. Finally, during *rejection actions*, the user fails to engage with a system's suggestion because it is inappropriate. For example, a user might ignore a smartphone's suggestion to try a new app. These feedback actions are a powerful implicit signal for adaptation and improvement of the underlying AI models.

Conclusion

XR presents the opportunity to enable *low-friction, "always-on" assistance,* but given that it operates in the real world, it also requires that a system *reduce uncertainty about user context and intent*. AI and machine learning are important for a positive user experience. We propose a novel framework for structuring several classes of models for user interaction in XR.

References

[1] Ruta Desai, Fraser Anderson, Justin Matejka, Stelian Coros, James McCann, George Fitzmaurice, and Tovi Grossman. 2019. Geppeto: Enabling semantic design of expressive robot behaviors. In *Conference on Human Factors in Computing Systems - Proceedings*, Association for Computing Machinery. DOI:https://doi.org/10.1145/3290605.3300599

[2] Christoph Gebhardt, Brian Hecox, Bas Van Opheusden, Daniel Wigdor, James Hillis, Otmar Hilliges, and Hrvoje Benko. 2019. Learning cooperative personalized policies from gaze data. In *UIST 2019 - Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, Association for Computing Machinery, Inc, 197–208. DOI:https://doi.org/10.1145/3332165.3347933

[3] Rorik Henrikson, Tovi Grossman, Sean Trowbridge, Daniel Wigdor, and Hrvoje Benko. 2020. Head-Coupled Kinematic Template Matching: A Prediction Model for Ray Pointing in VR. In *Conference on Human Factors in Computing Systems - Proceedings CHI 2020*.

[4] Eric Horvitz. 1999. Principles of Mixed-Initiative User Interfaces. In *Conference on Human Factors in Computing Systems - Proceedings*, 159–166. DOI:https://doi.org/10.1145/302979.303030