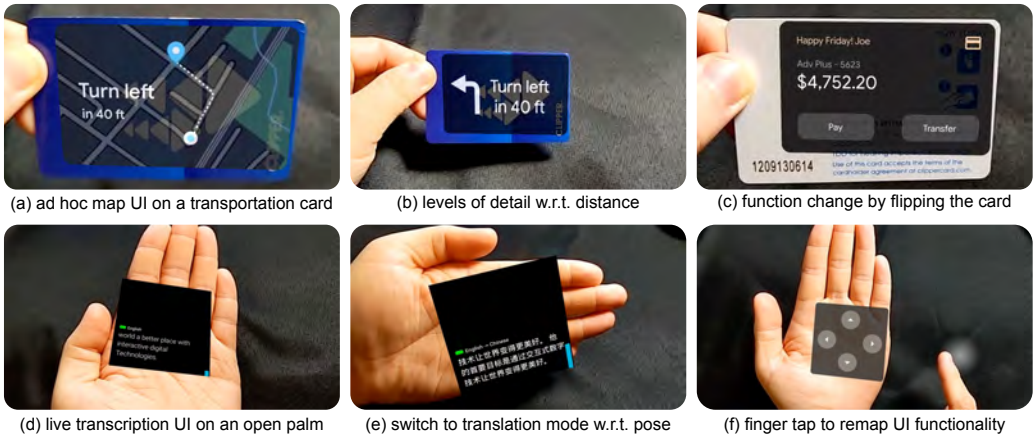
Opportunistic Interfaces for Augmented Reality: Transforming Everyday Objects into Tangible 6DoF Interfaces Using Ad hoc UI

RUOFEI DU, ALEX OLWAL, MATHIEU LE GOC, SHENGZHI WU, DANHANG TANG, YINDA ZHANG, JUN ZHANG, DAVID JOSEPH TAN, FEDERICO TOMBARI, DAVID KIM, Google Research



(a) ad hoc map UI on a transportation card (b) levels of detail w.r.t. distance (c) function change by flipping the card

(d) live transcription UI on an open palm (e) switch to translation mode w.r.t. pose (f) finger tap to remap UI functionality

Fig. 1. Examples using the Ad hoc UI toolkit and its interactive techniques: (a) shows a map widget on a transportation card triggered by voice command. (b) shows the level of detail change with respect to distance. (c) shows a wallet interface after the user flips the card. (d) shows a live transcription UI on an open palm. (e) shows mode change from transcription to translation after the user changes the pose. (f) shows remapping the UI functionally with voice command and tapping gestures.

Real-time environmental tracking has become a fundamental capability in modern mobile phones and AR/VR devices. However, it only allows user interfaces to be anchored at a static location. Although fiducial and natural-feature tracking overlays interfaces with specific visual features, they typically require *developers* to define the pattern before deployment. In this paper, we introduce opportunistic interfaces to grant *users* complete freedom to summon virtual interfaces on everyday objects via voice commands or tapping gestures. We present the workflow and technical details of Ad hoc UI (AhUI), a prototyping toolkit to empower users to turn everyday objects into opportunistic interfaces on the fly. We showcase a set of demos with real-time tracking, voice activation, 6DoF interactions, and mid-air gestures and prospect the future of opportunistic interfaces.

CCS Concepts: • **Human-centered computing** → **Mixed / augmented reality**; • **Computing methodologies** → *Mixed / augmented reality*.

Additional Key Words and Phrases: augmented reality, everyday objects, tangible user interface, 3D user interface, 6 DoF, spatial interaction, markerless tracking, tangible interaction, hand gestures

ACM CHI 2022 Interactivity. DOI: 10.1145/3491101.3519911

ACM Reference Format:
Ruofei Du, Alex Olwal, Mathieu Le Goc, Shengzhi Wu, Danhang Tang, Yinda Zhang, Jun Zhang, David Joseph Tan, Federico Tombari, David Kim. 2022. Opportunistic Interfaces for Augmented Reality: Transforming Everyday Objects into Tangible 6DoF Interfaces Using Ad hoc UI. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '22 Extended Abstracts), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3491101.3519911

1 INTRODUCTION

Recent advances in augmented reality (AR) have promoted the interests of both customers and enterprises to use 3D user interfaces in their everyday lives. Toolkits for augmented reality on mobile devices (*e.g.*, ARCore[1], ARKit[2]) and head-worn displays (*e.g.*, Microsoft HoloLens[3], MagicLeap[4]) empower users to anchor virtual interfaces to a specific location, so that they appear to be an integral part of the real world. Users may utilize a dedicated 6DoF controller or hand tracking on powerful headset devices to manipulate and directly interact with spatial interfaces or AR content [15, 16]. However, mobile phones and lightweight AR glasses lack rich spatial input mechanisms due to their portability and power consumption requirements, while only providing a limited sense of physicality to users [1].

Another way to interact with virtual interfaces is to track the real-world object to which we anchor content. Extensive research on tangible user interfaces (TUIs) has demonstrated the benefits of physicality, resulting in enhanced effectiveness [11], reduced cognitive load, and improved collaboration [8]. Nevertheless, transforming everyday objects into tangible user interfaces is under-explored in the current generation of spatial tracking technologies. While tracking of fiducials and natural features has been successfully demonstrated in prior art (*e.g.*, ARToolKit [13], ARTag [5], ArUco [10], and GOTURN [6]), it is typically required that developers or content creators prepare and predefine object tracking patterns in advance.

In this work, our goal is to allow users to instantly transform arbitrary everyday objects into TUIs without relying on a fixed set of physical objects that developers have predefined. Moreover, we aim to enrich tangible interaction by providing built-in support for multimodal interaction through voice and gesture.

To achieve this goal, we extend the concept of *opportunistic controls* [7] where semantically matching virtual content is associated with physical objects. Our *opportunistic interfaces* expand on this concept by enabling users to directly associate, activate and interact with widgets on everyday objects without contextual limitations. As a proof-of-concept, we have developed Ad hoc UI (AhUI), a prototype toolkit to empower users to turn everyday objects into opportunistic interfaces on the fly. Using AhUI, we demonstrate summoning the desired UI via voice activation, anchoring UIs to textured objects while tracking their 6DoF poses, and interacting with UIs via a set of tangible interactions. We couple our visual output with real-time tracking technologies to build user interfaces with multi-modal interaction.

2 USER JOURNEY OF OPPORTUNISTIC INTERFACES

We present the workflow of the proposed opportunistic interfaces by demonstrating an example user journey in Ad hoc UI. Figure 2 shows: a first-time user picks up a transportation card and says: *"Show me today's weather on the card."*. The system learns the card's visual features, starts tracking its pattern, computes the 6DoF pose, and associates the card with the *weather* widget. When the user moves the card, the system responds to its static and dynamic 6DoF poses. In this case, the rendered "weather" widget changes its level of detail depending on how far away the card is

[1] ARCore: https://arvr.google.com/arcore
[2] ARKit: https://developer.apple.com/augmented-reality
[3] Microsoft HoloLens: https://microsoft.com/hololens
[4] MagicLeap: https://www.magicleap.com

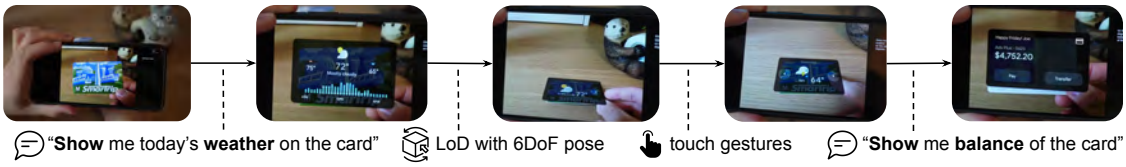


Fig. 2. An example user journey of opportunistic interfaces in Ad hoc UI system. See the supplementary video for complete demos.

from the user. When the user touches the "next" button on the card, AdUI recognizes the fingertip position, the touch event, and then renders the next day's weather information. Finally, the user may flip the card and register the "Wallet" app to the back of the card by saying "Show me the balance of the card.". This way, the next time the user picks up the card both sides will show different widgets and allow their own unique tangible interactions.

We envision that future opportunistic interfaces could also summon new interfaces by recognizing the object that the user is pointing to or gazing at, and extract the essential pattern with orthogonal re-projections. While our presented example is limited by current off-the-shelf mobile phone hardware, our system could be adapted to other form-factor devices, such as wearables with advanced eye tracking and active depth sensors.

3 AD HOC UI SYSTEM DESIGN

The AhUI toolkit consists of three key components: perception, representation, and interaction (Figure 3). We developed the AhUI toolkit in Unity 2020.3 for cross-platform compatibility, with some parts using C++ plugins and Java software.

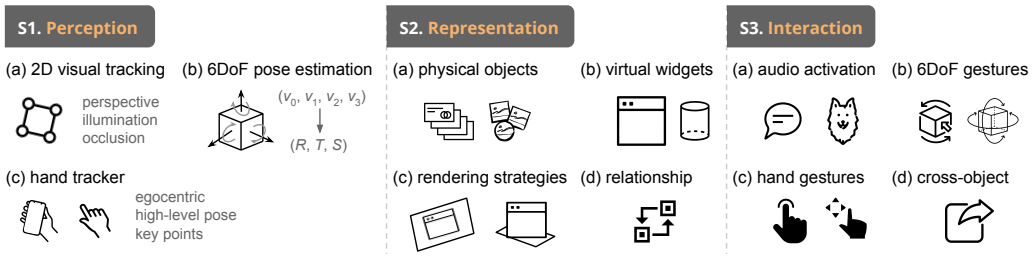


Fig. 3. Main components of AhUI toolkit: perception, representation, and interaction modules.

3.1 Perception Modules

The perception component comprises three submodules: 2D visual tracking, 3D pose estimation, and hand tracking. We implement these algorithms in C++ for optimal real-time performance and package them as Unity plugins.

2D Visual Tracking Our system leverages a real-time 2D template tracker in C++ based on KLT [9, 12]. For real-time performance and robustness, we tune the tracker to handle grayscale images of 640×480 resolution. Users bring an object close to the camera to initiate tracking, which captures a template image, currently configured to the central camera "region" of 250×250 pixels (~20% of the image). Users can thus pick up an object and move it in front of the camera, or move closer to objects of interest. As shown in Figure 4, after initialization within 100ms on a Pixel 4, we are able to track objects with perspective transformations, under strong illumination changes, and with partial occlusion (*e.g.*, when occluded by the fingertip) in real time.

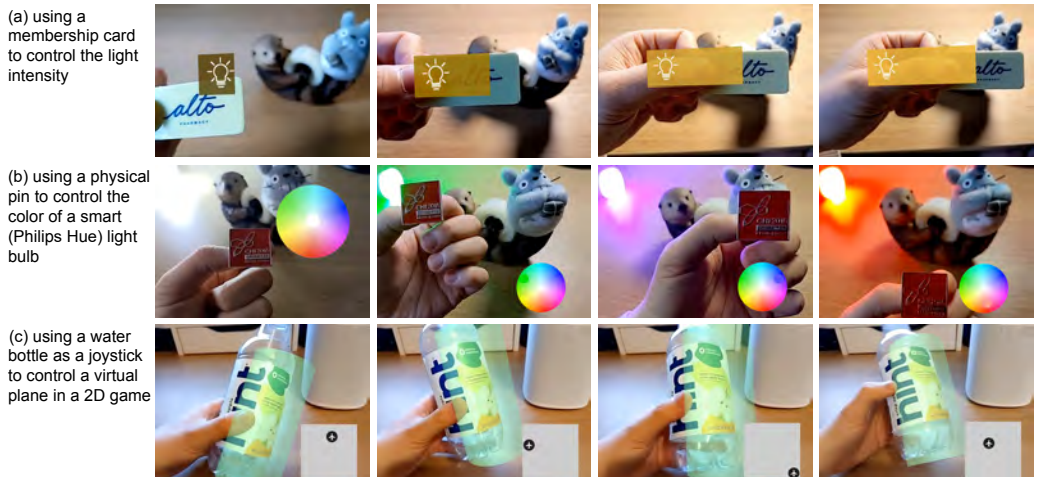


Fig. 4. Example applications enabled by the AhUI toolkit.

6DoF Pose Estimation For geometrical registration of textures (UV mapping) to real-world objects, we leverage the infinitesimal plane-based pose estimation [3] approach to compute the 6DoF pose in real time. In the AhUI toolkit, we minimize latency and programming complexity by directly bundling the 6DoF pose given an input image, a template to be tracked, and the rendering screen's orientation and resolutions, unifying computer vision and graphics coordinate systems. We apply the 1€ filter [2] to the position and quaternion to reduce undesirable jitter.

Hand Tracker Our custom hand tracker consists of a hand detector and an index-fingertip regressor. Both are Deep Neural Network-based models trained with a dataset of 750 ego-centric videos. The hand detector produces a bounding box and classifies the region of interest into one of the following four categories: *no hands*, *holding*, *pointing*, and *other hand poses in sight*. When the hand is pointing, the region of interest is cropped and fed into the fingertip regressor, which estimates the index fingertip location. Based on the trajectory of the fingertip, our pipeline can detect whether a tapping gesture is performed. Our standalone hand tracker can run at over 90 FPS. However, to lower the on-device power consumption, in the AhUI toolkit, we limit the hand detector to 2 FPS, and the fingertip regressor to 15 FPS, such that a smooth rendering of 30 FPS on a commodity smartphone (*e.g.*, Pixel 4) is achieved.

3.2 Representation Gallery

Our representation gallery consists of visual patterns, virtual widgets, and relationships.

Visual Patterns AhUI allows users to register their everyday objects as visual patterns for ad hoc input devices or virtual interfaces. We focus on 2D flat, rigid, and textured objects to enable a variety of AR applications. Some textured 3D objects (*e.g.*, bottles) can also be tracked robustly, as demonstrated in the supplementary video.

Virtual Widgets We prepare a variety of virtual widgets in the AhUI toolkit, including both visual elements and input proxies. Visual elements are representations of common functionality, such as weather, time, music, maps, and electronic wallets. Input proxies refer to gamepads, 6DoF controllers, volume controllers, and color pickers.

Relationship The relationship submodule maintains the association between different tracked objects. For example, linking a metro card to an analog clock, could show an AR overlay with arrival times for the upcoming trains.

3.3 Interaction Techniques

To enable interaction with everyday objects, the AhUI toolkit supports interaction techniques based on combinations of 6DoF poses, hand gestures, and audio activation.

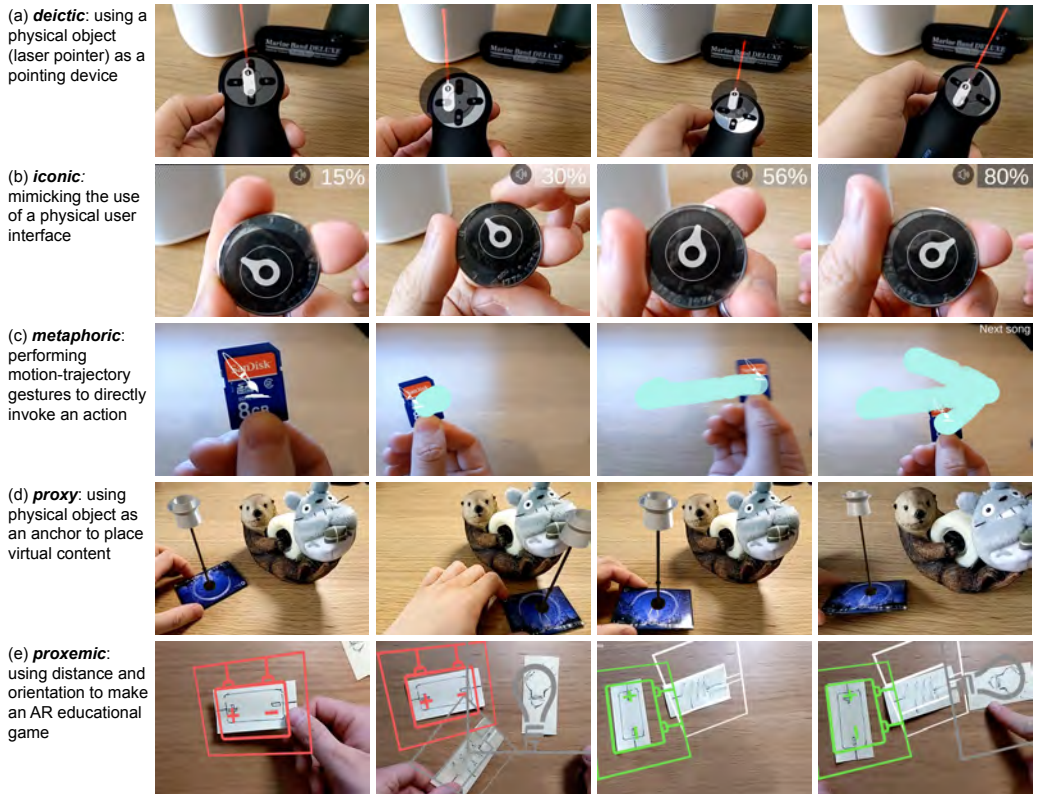


Fig. 5. Tangible interaction with the 6DoF poses of physical objects for opportunistic interfaces.

Audio Activation We leveraged a variant of *The Live Transcribe Speech Engine*[5] with UDP broadcasting to obtain users' speech. We perform regular-expression matching of each finalized sentence in the transcripted text to allow the user to register a new widget to an object or to show the memorized widget. Following are several examples queries:

- "**Show** me today's *weather* on the book" registers a weather widget on the current observed object.
- "**Show** me *more information* about the card" will pull the memorized widget from the relationship database.
- "**Use** this coupon to *control* the *map*" turns an object into a 6DoF controller for other virtual widgets.
- "**Bring** my *calendar* to the book" changes the functionality of a physical object into a tangible calendar.
- "**Close** the *shopping list* on the magnet" stops tracking of the opportunistic interface on a physical object.

6DoF Poses As shown in Figure 5, we implemented a set of interactions using both static and dynamic 6DoF poses of the object itself. Static poses map the raw 6DoF information values to translation and rotation vectors in the world space to obtain position and orientation (proxy), pointing direction (deictic), and position relative to other objects

[5]The Live Transcribe Speech Engine: https://github.com/google/live-transcribe-speech-engine

(proxemic). Dynamic poses use first and/or second derivatives to extract motion information and recognize unistroke gestures recognition [14] (metaphoric), or mimick well-known manipulations (iconic).

Hand Manipulations Building on prior research in tangible interaction and the capabilities of our hand tracker, we enable grasping, pointing, and tapping in the AhUI toolkit. Grasping can activate an opportunistic interface or increase the level of detail, while pointing allows hovering effects for virtual buttons. Tapping activates virtual buttons or interacts with virtual sliders on the tangible objects.

4 CONCLUSION

In this paper, we introduce *opportunistic interfaces* to enable users to summon desired widgets on everyday objects with multi-modal interactions. We presented Ad hoc UI, a prototyping system to empower users to enable on-the-fly interaction with arbitrary physical objects. We hope that our toolkit will accelerate and inspire the work of AR researchers, developers, and enthusiasts by providing a new set of building blocks for interaction with everyday objects, in addition to our prior work on depth-based library [4] for AR.

REFERENCES

[1] Emily Bennett and Brett Stevens. 2005. The Effect That Touching a Projection Augmented Model Has on Object-Presence. In *Ninth International Conference on Information Visualisation (IV'05)*. IEEE, 790–795. https://doi.org/10.1145/1056808.1056834

[2] Géry Casiez, Nicolas Roussel, and Daniel Vogel. 2012. 1€Filter: a Simple Speed-Based Low-Pass Filter for Noisy Input in Interactive Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2527–2530. https://doi.org/10.1145/2207676.2208639

[3] Toby Collins and Adrien Bartoli. 2014. Infinitesimal Plane-Based Pose Estimation. *International Journal of Computer Vision* 109, 3 (2014), 252–286. https://doi.org/10.1007/s11263-014-0725-5

[4] Ruofei Du, Eric Turner, Maksym Dzitsiuk, Luca Prasso, Ivo Duarte, Jason Dourgarian, Joao Afonso, Jose Pascoal, Josh Gladstone, Nuno Cruces, Shahram Izadi, Adarsh Kowdle, Konstantine Tsotsos, and David Kim. 2020. DepthLab: Real-Time 3D Interaction With Depth Maps for Mobile Augmented Reality. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology (UIST)*. ACM, 829–843. https://doi.org/10.1145/3379337.3415881

[5] Mark Fiala. 2005. ARTag, a Fiducial Marker System Using Digital Techniques. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 2. IEEE, 590–596. https://doi.org/10.1109/CVPR.2005.74

[6] David Held, Sebastian Thrun, and Silvio Savarese. 2016. Learning to Track at 100 Fps With Deep Regression Networks. In *European Conference on Computer Vision*. Springer, Springer, 749–765. https://doi.org/10.1007/978-3-319-46448-0_45

[7] Steven Henderson and Steven Feiner. 2009. Opportunistic Tangible User Interfaces for Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics* 16, 1 (2009), 4–16. https://doi.org/10.1145/1450579.1450625

[8] Eva Hornecker and Jacob Buur. 2006. Getting a Grip on Tangible Interaction: a Framework on Physical Space and Social Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 437–446. https://doi.org/10.1145/1124772.1124838

[9] Bruce D Lucas, Takeo Kanade, et al. 1981. An Iterative Image Registration Technique With an Application to Stereo Vision. Vancouver, British Columbia, Vancouver, British Columbia. https://doi.org/10.5555/1623264.1623280

[10] Francisco J Romero-Ramirez, Rafael Muñoz-Salinas, and Rafael Medina-Carnicer. 2018. Speeded Up Detection of Squared Fiducial Markers. *Image and Vision Computing* 76 (2018), 38–47. https://doi.org/10.1016/j.imavis.2018.05.004

[11] Lucia Terrenghi, David Kirk, Abigail Sellen, and Shahram Izadi. 2007. Affordances for Manipulation of Physical Versus Digital Media on Interactive Surfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1157–1166. https://doi.org/10.1145/1240624.1240799

[12] Carlo Tomasi and Takeo Kanade. 1991. Detection and Tracking of Point. *Int J Comput Vis* 9 (1991), 137–154. https://doi.org/10.1109/CVPR.1994.323794

[13] Daniel Wagner and Dieter Schmalstieg. 2007. ARToolKitPlus for Pose Tracking on Mobile Devices. (2007).

[14] Jacob O Wobbrock, Andrew D Wilson, and Yang Li. 2007. Gestures Without Libraries, Toolkits or Training: A $1 Recognizer for User Interface Prototypes. In *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology*. 159–168. https://doi.org/10.1145/1294211.1294238

[15] Robert Xiao, Chris Harrison, and Scott E Hudson. 2013. WorldKit: Rapid and Easy Creation of Ad-Hoc Interactive Applications on Everyday Surfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 879–888. https://doi.org/10.1145/2470654.2466113

[16] Robert Xiao, Julia Schwarz, Nick Throm, Andrew D Wilson, and Hrvoje Benko. 2018. MRTouch: Adding Touch Input to Head-Mounted Mixed Reality. *IEEE Transactions on Visualization and Computer Graphics* 24, 4 (2018), 1653–1660. https://doi.org/10.1109/TVCG.2018.2794222