AirSnip: Gesture-Based AR Screenshot Enhanced by GenAI

Robson Oliveira
Voxar Labs, Centro de Informática
Universidade Federal de Pernambuco
Recife, Pernambuco
ropj@cin.ufpe.br

Francisco Simões
Voxar Labs, Centro de Informática
Universidade Federal de Pernambuco
Recife, Pernambuco
fpms@cin.ufpe.br

***Abstract*—This paper presents the development of an interactive tool for image capture in Augmented Reality (AR) environments, based on a custom mid-air hand gesture that mimics the behavior of a snipping tool. The system allows users to delimit a rectangular area in space using a gesture, capture a real-world view through the Meta Quest 3's passthrough camera, and generate a 3D virtual frame with the resulting image. These frames can be manipulated using natural gestures (e.g., push, pinch) and are enhanced with personalized descriptions generated by a Large Language Model (LLM) via Google Gemini, with audio output through Text-to-Speech (TTS). Built using Unity and the Meta XR SDK, the prototype offers a seamless and intuitive interface for storytelling, education, and spatial memory. Although user studies are planned for future work, the tool demonstrates the technical feasibility of integrating gesture recognition, spatial rendering, and Generative AI in a unified AR experience.**

I. Introduction

Augmented Reality (AR) is a promising technology for transforming the way we perceive and interact with virtual content. It is increasingly described as one of the key technologies of the 21st century and as a pillar of the new industrial revolution envisioned by Industry 4.0 [1]. Despite current technological limitations, recent devices such as the Meta Quest 3 demonstrate considerable advancements, enabling the creation and execution of increasingly immersive applications and opening up a wide range of possibilities for the development of more natural and intuitive interfaces.

However, there is still limited progress in developing user-centered AR experiences. Interaction management often relies on manual controllers and barely explores how natural gestures could enhance the fluidity of the experience and broaden the applicability of actions in dynamic contexts. In this scenario, the study and detection of 'natural' gestures emerge as a promising alternative to improve the adaptation of new users in both virtual and augmented environments.

In this work, we propose **AirSnip**, a mixed reality headset-based application that enables the capture and transformation of real-world scenes into virtual frames integrated with Generative AI, allowing use in various contexts and narratives. The scene is captured through the detection and framing of a custom hand gesture, mimicking a 'snipping tool.' With the resulting image, virtual frames are created that can be stored, downloaded, deleted, moved, and rotated using only hand gestures. Moreover, the AirSnip allows personalized use of Generative AI to describe captured scenes. Users can move freely through space, capture moments, build environments, and create micronarratives for a wide range of purposes.

As such, the AirSnip can be applied in various fields such as education, storytelling, personal records, tourism, etc. The main goals of this work are: **(I)** to propose a method for screen capture in augmented reality using a natural and intuitive gesture, and **(II)** to develop a functional prototype that integrates this capture method with Generative AI to explore its potential applications.

II. Related Work

The use of gestures as a form of interaction in immersive environments has been explored in previous research, highlighting approaches aimed at making the experience more natural and fluid. Studies such as Figueiredo et al. [3] presented catalogs of gestures inspired by science fiction works, evaluating their practical applicability in gesture-based interfaces. These contributions provide a foundation for developing new forms of control in AR that feel natural even to non-expert users.

Regarding scene capture and cropping in physical space, Phursule et al. [2] proposed the Augmented Reality Snipping Tool, a tool that enables the capture of real-world objects using a smartphone camera and the application of automatic segmentation techniques powered by models such as U2-Net. Although functional, this approach relies on click-based interaction and does not explore real-time gesture-based capture in immersive 3D environments.

Another relevant contribution is SceneAR by Chen et al. [4], which enables the creation of AR micronarratives based on scenes composed of 3D elements and dialogue balloons. While the focus is on content authoring and remixing in AR, interaction is limited to mobile devices and does not support custom gestures or AI integration. In contrast, our system allows for the creation of visual narratives in virtual space through manual gestures, combining natural interaction with intelligent scene description generation.

Finally, Lyu [5] highlights the application of Generative AI in AR / VR environments, emphasizing the potential of these technologies to enrich digital experiences through automatic

scene description, object generation, and contextual adaptation. This perspective reinforces the uniqueness of the system presented here, which combines gesture-based capture, virtual frame generation, and automatic AI-powered description.

These studies indicate that, although there have been advances in visual capture, storytelling, and the application of AI in AR, there is still a gap in integrating these elements into a single experience centered on gesture as the primary input method—especially on mixed reality platforms.

III. Methodology

This research adopts an exploratory approach focused on the development and evaluation of an interactive tool for personalized image capture in Augmented Reality (AR) environments, based on a custom hand gesture also proposed in this study. The main objective is to investigate the technical feasibility and usability of a gesture capable of simulating a "snipping tool" in three-dimensional space, while exploring the potential applications of Generative AI in different contexts.

The methodology follows three stages:

- **Gesture design and selection:** The capture gesture was defined by forming a frame using the index fingers and thumbs of both hands, delimiting a rectangular area in space, as illustrated in Fig. 1.
- **Prototype development:** A functional prototype was designed based on hand tracking and the capture of the area defined by the gesture. The technical requirements of the system were established in this stage, including criteria for gesture recognition, capture parameters, and tool usability.
- **Experimental evaluation:** Analysis of the prototype with users to collect feedback.

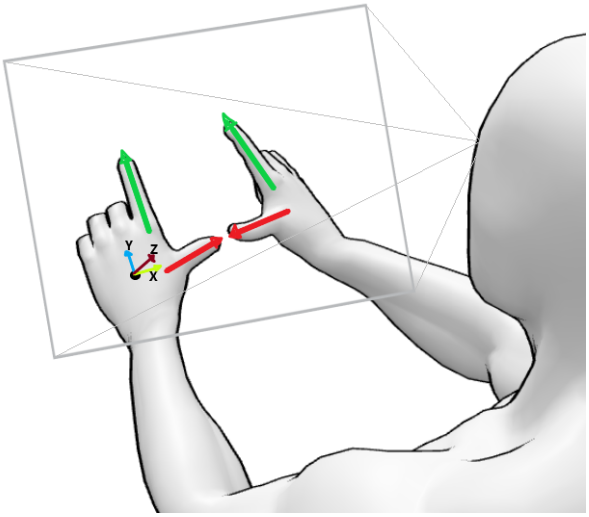


Fig. 1: Snapshot gesture

IV. System Architecture

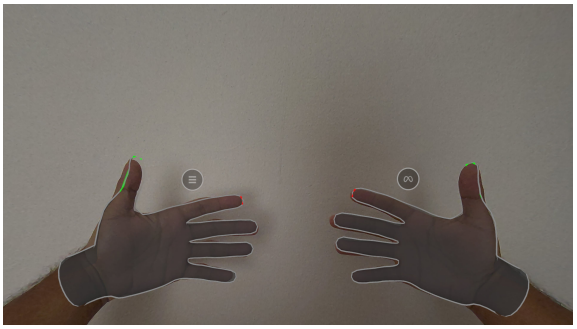
The prototype was developed using the Unity 6 game engine, with the Meta Quest 3 headset as the testing platform. The implementation relied on the main components of the Meta XR ecosystem, including:

- Meta XR Core SDK for hand tracking;
- Meta Passthrough API for real-world environment visualization through passthrough;
- Meta Building Blocks for rapid integration of functionalities and prefabs;
- and the WebCamTextureManager available in the GitHub project *Unity-PassthroughCameraApiSamples*.

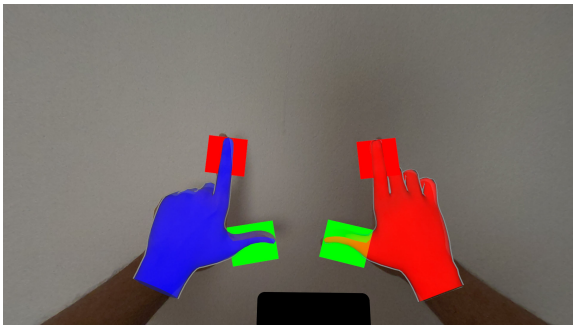
A. Architecture

The tool architecture is made up of four modules:

- **Gesture Recognition Module:** Responsible for identifying the gesture defined in the study. Detection is based on reading the positions of hand bones using the OVRSkeleton. The following conditions are checked: (I) the back of both hands must have their normals facing the reference camera; (II) the vectors formed by the index finger bones must be angled close to the Y-axis (pointing upward); (III) the vectors formed by the thumb bones must be close to the horizontal plane and in opposite directions; (IV) the thumb tips must be close to each other. All these checks are performed using coordinate conversions relative to the camera (rather than the world space), as exemplified in Fig. 2a (gesture not detected) and Fig. 2b (gesture successfully detected).



(a) Unrecognized gesture



(b) Recognized gesture

Fig. 2: Gesture recognition results: (a) gesture not detected, (b) gesture successfully detected.

- **Cropping Area Generation Module:** Once the gesture is recognized, the framing is computed using the centroid between both index fingertips as the center of the

capture plane. The plane orientation is defined by the normal vector derived from the positions of the index finger bones, ensuring that the frame always faces the user. The size of the capture area is proportional to the distance between the index fingers, dynamically adjusting the scale of the *RawImage* object. Once positioned, the system retrieves the real-world passthrough feed via the *WebCamTextureManager* API and maps it onto the *RawImage*, providing a real-time preview of the framed area before capture.

- **Frame Capture and Management:** The capture is triggered after holding the gesture for two seconds, at which point a 3D virtual frame is generated and can be manipulated using mid-air selection gestures. Based on the study by Dube *et al.* [6], mid-air gestures such as *push* and *pinch* perform better when combined with haptic feedback. Accordingly, virtual frame control was defined as follows: the frame can be moved or rotated (*pinch*), clicked (*push* with haptic feedback), and upon clicking, a menu becomes visible allowing the user to: download the image to the device, delete the frame, save its position, or retrieve details. This behavior is illustrated in Fig. 3, which shows the mid-air *push* gesture used to trigger the contextual menu on a captured frame. Positioning and images are stored on the device using the Meta SDK's *spatial anchors*. This allows virtual frames to be assembled and saved anywhere in the environment.

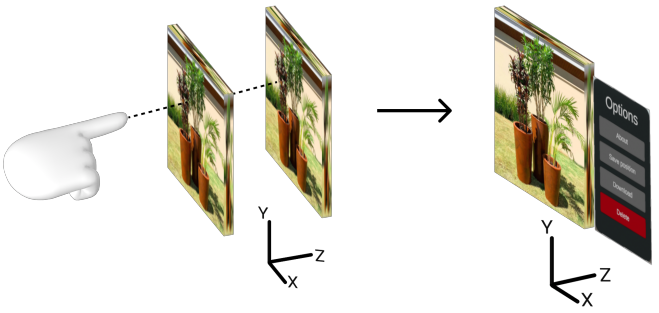


Fig. 3: Click

- **Generative AI:** When the *"details"* option is selected on a virtual frame, the system triggers the module that connects to *Google Gemini*, where a customized prompt is sent along with the captured image. The textual result generated by the LLM is then converted into audio using *Google TTS AI*. After processing, the audio is saved to the frame and can be played or paused at any time.

Figure 4 illustrates a snapshot of the current state of the application. Multiple captures from different environments are arranged as a virtual exhibition. Additionally, the selected frame displays a floating menu, where the sound icon indicates that an audio description associated with the image is currently playing.

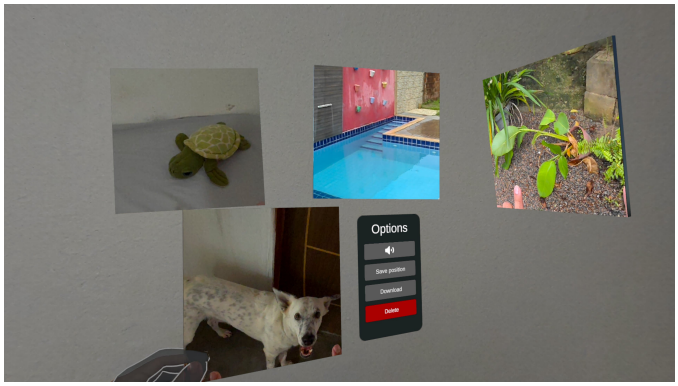


Fig. 4: Multiple captured frames

V. Future Work

This work presented the development of a gesture-based capture tool in Augmented Reality, integrating a custom gesture, spatial rendering, and generative AI-based scene description. Although the system is already functional and allows image capture and manipulation using natural gestures, user evaluation will be conducted in a future stage.

As next steps, we plan to assess the usability and effectiveness of the proposed gesture, as well as the application's potential in supporting narrative creation with the aid of Generative AI in various contexts. To this end, a user study will be conducted in a controlled environment, with emphasis on non-expert users. Furthermore, we intend to:

- Refine gesture detection criteria to make it more robust under different lighting conditions and viewing angles;
- Investigate improved strategies for communication between users and the Generative AI;
- Explore additional features such as object cropping, collaborative scenarios, or real-time narrative support.

These improvements will contribute to validating and enhancing the proposed solution, paving the way for broader applications in immersive experiences with a focus on natural interaction and personalized visual storytelling.

VI. Conclusion

We introduced AirSnip, a gesture-based AR tool that enables real-world scene capture, spatial rendering, and Generative AI integration on the Meta Quest 3. The system demonstrates the feasibility of using a custom "snipping" gesture for intuitive, controller-free interaction, creating interactive virtual frames enriched with AI-generated descriptions. As a work in progress, the prototype provides a foundation for future usability studies and feature enhancements, aiming to extend its applications in education, storytelling, and other immersive scenarios.

References

[1] Arena, F.; Collotta, M.; Pau, G.; Termine, F. An Overview of Augmented Reality. Computers 2022, 11, 28. https://doi.org/10.3390/computers11020028

[2] R. Phursule, K. Sirpor, P. Virmalwar, S. Zadbuke and P. Avachat, ”Augmented Reality Snipping Tool,” 2023 4th International Conference for Emerging Technology (INCET), Belgaum, India, 2023, pp. 1-4, doi: 10.1109/INCET57972.2023.10170651. keywords: Graphics;Three-dimensional displays;Machine learning;Real-time systems;Workstations;Augmented reality;Augmented Reality(AR);Virtual Reality(VR);Snipping Tool

[3] Lucas S. Figueiredo, Mariana Pinheiro, Edvar Vilar Neto, Thiago Chaves, Veronica Teichrieb. Sci-Fi Gestures Catalog. 15th Human-Computer Interaction (INTERACT), Sep 2015, Bamberg, Germany. pp.395-411, ff10.1007/978-3-319-22668-2_30ff. ffhal-01599861f

[4] M. Chen, A. Monroy-Hernández and M. Sra, ”SceneAR: Scene-based Micro Narratives for Sharing and Remixing in Augmented Reality,” 2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Bari, Italy, 2021, pp. 294-303, doi: 10.1109/ISMAR52148.2021.00045. keywords: Three-dimensional displays;Visual communication;Social networking (online);Multimedia Web sites;Lighting;Media;Reliability engineering;Human-centered computing;Visualization;Visualization techniques;Treemaps;Visualization design and evaluation methods,

[5] S. Lyu, “The Application of Generative AI in Virtual Reality and Augmented Reality”, Journal of Industrial Engineering; Applied Science, vol. 2, no. 6, pp. 1–9, Dec. 2024.

[6] Tafadzwa Joseph Dube, Yuan Ren, Hannah Limerick, I. Scott MacKenzie, and Ahmed Sabbir Arif. 2022. Push, Tap, Dwell, and Pinch: Evaluation of Four Mid-air Selection Methods Augmented with Ultrasonic Haptic Feedback. Proc. ACM Hum.-Comput. Interact. 6, ISS, Article 565 (December 2022), 19 pages. https://doi.org/10.1145/3567718