

CHAPTER 1. INTRODUCTION

1.1 Project Background

In many disciplines, especially education, research, and the creative industries, 3D modeling is becoming an integral component of workflows. It is particularly important in education and research for visualizing spatial relationships and helping learners understand complex concepts such as design and geometry. The 3D creative industry uses 3D modeling to generate digital assets for animation, games, simulations, and product design. 3D modeling is important in the early design phases to create the initial shape of a design to build a structure for the rest of the model(Harlan et al., n.d.; Sosna et al., 2025).

Three dimensional (3D) spatial modeling has a strategically significant role to play, yet 3D modeling systems continue to use 2D interaction technologies (input devices like the mouse or keyboard). This continues to widen the chasm between a user's natural way of interacting with 3D space and the way a 3D modeling system functions. As a result, 3D spatial modeling systems tend to be counterintuitive and have a steep learning curve for novice users new to 3D spatial modeling. Several studies have assessed reconfigurable input devices as no longer meeting the requirements of 3D interactions and as an imperative to design new devices (Cárdenas-Sainz et al., 2022; Uba & Jurewicz, 2024)

In order to close this gap, several studies have begun exploring hand-based interaction methods. This line of research has the ability and potential for the real-time capture of users' hand movements on camera to serve as a direct input method for 3D modeling. More intuitively, users can now know and understand how to interact with virtual objects the same way that humans manipulate physical objects. The greatest challenge for gesture-based approaches, however, is how to transform arbitrary, free, and imprecise movements of the hand into 3D geometry with structure and meaning. AI in this sense acts as an AI component in the hand gesture 3D modeling system. AI is meant to help with hand motion tracking and assist with the surface reconstruction of 3D objects based on user gestures. In the AI-based surface reconstruction system, the system can

define a 3D form (mesh) in real time, despite the user being rough or providing incomplete user gesture. Recent gesture-based and human-AI collaborative modeling systems have shown that AI can improve the robustness and interaction of the system while maintaining the control of the user. For instance, HotGestures shows no-bound hand gestures as a shortcut to rapid tool use and selection in a 3D modeling system (Song et al., 2023) , and 3Description's portrayal of AI is as a collaborative tool in the aid of the user's verbal and gesture interpretation, in a way that preserves human imagination and creativity and is not a substitute for it (Cai, 2023). Thus, AI acts as a tool to assist the user in the translation of their gestures to 3D geometric modeling.

Within the national context, the hand movement AI 3D modeling systems have strong relevance to education and research. In the Indonesian context, access to professional 3D modeling hardware and professional modeling software acts as a blockage for learning 3D modeling. Therefore, an AI-based 3D modeling system using simple and intuitive devices like built-in cameras offers an alternative that is accessible and easy to learn for beginner users to learn the basic concepts of 3D modeling.

Although this study primarily concerns education and research, the developed method may also be applicable to initial ideation within the creative sector later on.

1.2 Problem Statement

Progress in hand tracking and 3D surface reconstruction continues at a rapid pace. However, these advancements are seldom integrated into a viable, interaction-centered method of early-stage 3D content generation. Most past studies have tackled these problems separately, which results in little to no evidence that simple webcam-based hand tracking provides enough interaction to support modeling workflows for ideation and blocking within popular DCC tools and game engines.

Specifically, previous studies have largely failed to provide a zero-hardware-cost solution for complete 3D surface generation. Existing hand-based

modeling systems, such as HandPainter and BuildingSketch, heavily rely on expensive Virtual Reality (VR) or Mixed Reality (MR) headsets and specialized depth sensors to function. Consequently, they fail to address the accessibility gap for novice users who only possess standard computing hardware. Furthermore, studies that do utilize standard video input (e.g., NeuralRecon) primarily focus on scene rendering or implicit representations, failing to extract the clean, exportable polygon meshes required for standard 3D modeling pipelines. This leaves a significant gap for an accessible, webcam-only system that can seamlessly translate natural gestures into interoperable, production-ready 3D assets.

From the HCI standpoint, webcam-based tracking, such as MediaPipe Hands, can offer a sufficiently reliable stream of hand tracking data and can be used for air drawing and gesture modeling. Thus, the main concern is whether these interaction methods can be designed to the point where users can easily and reliably draw shapes in air, meaning that air drawing can be a practical input method of 3D modeling in its early stages).

From a technical standpoint, to create a 3D surface within a volume, the combination of Temporal Signed Distance Fields and mesh extraction can be used, but these techniques rely on particular input requirements. Because user input is in the form of sparse and possibly noisy strokes, the main challenge is to design a framework that can effectively capture dynamic hand gestures in the form of stable 3D models with the least possible latency.

Various artificial intelligence technologies such as trained denoisers or even just simple shape priors, can be used as auxiliary tools to help reduce noise or help with small gaps in geometry. However, the focus of this research is not about innovating more AI technologies. This project focuses on designing Human-Computer Interaction (HCI) techniques (gesture vocabularies, mappings, types of feedback, previews, etc.) and integrating off-the-shelf, pragmatic machine learning technologies. This is done within the boundaries that the technologies improve the usability of the interaction and are not the leading focus of the research.

Based on these points, the study addresses three focused problems:

- a. To what extent can webcam-captured hand movements be designed and encoded (gesture vocabulary, stroke mapping) so users can express 3D shapes naturally and reliably on typical laptop hardware?
- b. How should a pipeline be constructed, mapping gestures to geometric samples, incremental TSDF fusion, lightweight stabilization/post-processing, and interactive preview, so it best supports expressive, low-friction air-drawing workflows while keeping responsiveness acceptable for ideation
- c. When and for whom are the exported glTF/GLB assets sufficiently usable in Blender/Unity for early-stage modelling, and what degree of manual cleanup is acceptable for the target user groups (students, indie developers, 3D artists during blocking)?

1.3 Project Objectives

This research investigates the human-computer interaction method (the combination of air-drawing and hand-tracking) for rapid 3D and 2D ideation. This method attempts to tackle the three issues stated above. In this application, AI and ML components (stand-alone, integrated, or lightweight modules) are considered tools that assist the main research, and will not be the primary components of the research.

- a. Design and evaluate a gesture vocabulary and encoding scheme for webcam-based hand input that enables users to express simple 3D shapes through air-drawing. *Evaluation criteria:* expressiveness (can users create target shapes), learnability (time to learn gestures), and robustness under ordinary indoor lighting and typical laptop webcams.
- b. Implement an interaction-to-geometry pipeline that maps gestures to geometric samples, performs incremental TSDF fusion with lightweight stabilization/post-processing, and provides real-time visual feedback for ideation. Off-the-shelf ML modules may be integrated only to improve feedback/stability. *Evaluation criteria:* responsiveness (perceived lag), reconstruction coherence for interactive preview, and user-reported satisfaction during ideation tasks.

- c. Produce exportable glTF/GLB assets and evaluate their usefulness for early-stage modelling workflows. *Evaluation criteria:* import compatibility in Blender/Unity, typical manual cleanup required (qualitative/quantitative), and target-user suitability identified via a small user study (participants from 3D artists, students, or indie developers).

1.4 Project Scope

Research is directed and focused with the inclusion of certain limitations. The research limitations have been categorized into two types: limitations resulting from the user side and limitations from the system being developed:

1.4.1 User Limitations

- a. Device and Environment :

To use this system, an RGB camera device, such as a webcam, and a sufficient quality lighting setup, are required. For better hand tracking, the lighting and backgrounds should be less complex. This system is not meant to operate in darker or high visual noise environments.

- b. User Coverage :

The system can only support one user at a time. So, one user means one pair of hands for input. Tracking multiple people during a session or collaborative multi-user system input is not included in this study.

- c. Gestures Limited to Bare Hands :

User interaction is limited to the use of bare hands (there are no tools like controllers or physical objects). Users have to perform supported hand movements by the system, hence there is a recognized set of gestures. Also, movements beyond that may not get supported. Additionally, the system provides for one or two hand use (again, no aids like controllers or other objects) during the interaction.

1.4.2 System Limitations

- a. **Limited Model Complexity:**

The developed modeling system is designed to generate simple to moderately

sized 3D objects (e.g., primitives or single objects). The system is not designed to handle highly complex 3D scenes or 3D scenes with multiple objects. Models created with the system are made for representing static objects. The system also does not support the animation of characters or the deformation of objects.

b. Surface Reconstruction and Detail:

In this system, the surface reconstruction AI focuses on building smooth surfaces based on hand movement. Demands for very high geometric detail and complexity pose the system limitations; for instance, complex surface textures or fine grained details like very complex surface textures or ornaments may be difficult to reconstruct. Furthermore, this study does not deal with the generation of color and/or textures, and the output of the system is a geometric mesh with no material and/or texture constituents.

c. Platform and Performance:

In this study, the prototype was developed and tested on a laptop with specific CPU, RAM, and integrated GPU configurations. System performance, especially with respect to the analysis of real-time hand tracking, the application of GPU-based AI, and the 3D surface reconstruction of objects, is highly dependent on the laptop's configuration. Therefore, performance will differ on laptops with lower configurations.

This study does not consider performance optimizations across different platforms (i.e. between laptops and desktop computers, or between laptops and mobile computers). Additionally, using the system on computers with lower-than-standard-configured systems will most likely result in a lower frame rate and will create an experience of interactive delays, and thus will not achieve a desirable real-time interaction effect. This study also does not consider support for computers other than laptops (i.e. mobile computing and embedded systems).

The restrictions mentioned define the boundaries of the study. This includes the primary operations of the system in accordance with the goals. Other related aspects are seen as beyond the present boundaries of the study. These aspects may be considered future work.

1.5 Project Significance

This project holds significant value across multiple domains, specifically bridging the gap between natural human-computer interaction and accessible 3D content creation. The contributions are categorized as follows:

This research adds to Human-Computer Interaction (HCI) literature by examining mid-air gestures as an alternative input technique for 3D spatial interactions. Usability data was collected, and novice users were observed in an experiment, air-drawing via a webcam. This provided data and findings regarding the learnability and expressiveness of the modality, as well as the ergonomic trade-offs of air-drawing compared to mouse and keyboard interactions.

a. Technical Contribution:

The project showcases a new implementation pipeline that combines readily available computer vision (MediaPipe) and volumetric surface reconstruction (TSDF and Marching Cubes) running exclusively on CPU/GPU of regular consumer laptops. It shows that an elaborate, incremental 3D geometry pipeline can be deployed and maximized without depth sensors and can easily output .glb meshes.

b. Industrial Contribution:

For creative industries and software companies, this system provides a low-friction, rapid prototyping tool during the early ideation and “blocking” phases of 3D asset creation. The system outputs standard .glb files. As such, the system integrates easily into a user’s professional Digital Content Creation (DCC) workflows, such as Blender and Unity, and quickens conceptual design work for indie developers and 3D artists.

c. Educational Contribution:

This system removes the expensive issues with VR headsets and drawing tablets, so access to learning 3D modeling becomes greater. It is an easy, free-learning tool for students, and even hobbyists. It allows users to learn the basics of spatial and geometric concepts by simply moving their hands.

1.6 Chapter Summary

This chapter introduced the motivation for exploring hand-tracking-based interaction as an alternative approach to early-stage 3D modelling, highlighting the limitations of conventional input devices and the fragmented nature of existing hand tracking and reconstruction solutions. It discussed the challenges of using webcam-captured hand movements as expressive and reliable modelling input, as well as the difficulties of converting sparse, noisy gesture data into stable geometric representations suitable for export. Based on these considerations, the chapter formulated three focused problem areas related to input design, interaction-to-geometry processing, and output usability, and derived corresponding project objectives that emphasize human-computer interaction rather than the development of new AI models. The chapter also outlined the significance of the project in terms of HCI research, system integration, accessibility, and early-stage modelling workflows, thereby establishing a clear foundation and scope for the subsequent chapters of the study.