Learning Low-dimensional Latent Spaces for Hand-pose Interaction Systems

Songpei Xu

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Computing Science
College of Science and Engineering
University of Glasgow

December 2024

Abstract

In recent years, mid-air gesture recognition and control have emerged as touch-free interaction mechanisms with significant potential, but human-centered design challenges persist, particularly due to the complexity of user hand movements, user-learnability of interaction system, and usability across varied contexts. This research addresses limitations in current mid-air gesture systems by introducing strategies that reduce the latency of real-time interaction and enhance hand pose interpretability. We introduce a hand pose recognition approach using only two video frames, optimizing for speed and accuracy while minimizing time dependencies, thereby allowing users to experience a more responsive system. Building on these insights, the Hand-pose Embedding Interactive System (HpEIS) employs a Variational Autoencoder to map gestures to a two-dimensional space, introducing visualized feedback mechanisms that improve user experience. At the same time, interaction stability improved through smoothing and anti-jitter methods. While this approach improves robustness in dynamic movements, further challenges remain in adaptability and flexible user control. To expand flexibility, the HandSolo model introduces a disentangled hand pose embedding space, supporting multi-dimensional control with independent degrees of freedom, thus enabling interactions adaptable across devices and contexts. Coupled with a Visual Interaction Evaluation Strategy (VIEs), HandSolo provides a guidance for system designers to align model capabilities with user preferences. The experimentation underscores the effectiveness of these systems. This integrated research establishes a framework for mid-air hand pose control, advancing usability, flexibility and extensibility in the interaction design of mid-air hand pose with high dimensional input.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Professor Roderick Murray-Smith, for his continuous guidance and support throughout my PhD. He has helped me in many ways, from shaping my research direction to improving my daily work and study habits. In particular, I am immensely grateful for his patience in teaching me how to write research papers, a skill that will benefit me in my future academic pursuits. His advice on research direction, critical thinking, and thesis writing has been indispensable. Beyond academic guidance, I truly appreciate his efforts in fostering a supportive research environment, organizing group meetings, and arranging social gatherings that brought the team closer together.

I am also very grateful to my second supervisor, Dr.Chaitanya Kaul, for his encouragement and help during my research. He has provided me with valuable ideas and helped me explore different directions. His support has reduced my stress and given me confidence in my work. I also appreciate his patience in explaining difficult concepts and revising my thesis. His help has made my research process much smoother.

I would like to thank my lab colleagues, including Mr.Andrew Ramsay, Dr.Sebastian Stein, and many others, whose support has significantly contributed to my research. Their technical assistance, insightful discussions, and valuable feedback have helped me improve my work. Their willingness to share knowledge and provide engineering support has been essential to the completion of my PhD.

I am deeply grateful to Moodagent, a pioneering music streaming service company. The company's generous financial support, provision of data, and engineering assistance have been invaluable to my research. Their proprietary music emotion analysis technology has provided a foundation for my work. I sincerely appreciate their collaboration and the resources they have made available to me.

A special thank you goes to my boyfriend, Dr. Xuri Ge, now an Assistant Professor at Shandong University. Throughout my PhD, he has been a constant source of encouragement, both academically and personally. We have worked on related research topics, attended conferences together, and traveled to different countries. These experiences have not only enriched my academic journey but also helped me manage stress and stay motivated.

I am also deeply grateful to my family for their unwavering support, both financially and emotionally. They have stood by my side, encouraging me to pursue the life I desire, even though it meant being far away from home. Their belief in me has given me the strength to persevere through difficult times and complete my PhD.

Lastly, I would like to extend my appreciation to my home in Glasgow, which has provided me with a place to rest, eat, and recharge. Special thanks to Morrison's Grocery near my home and M&S Grocery near the university for making sure I never went hungry during these four years.

This PhD journey has been full of challenges, but with the support of so many people, I have been able to complete it. I am truly grateful to everyone who has helped me along the way.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This dissertation is the result of my own work, under the supervision of Professor Roderick Murray-Smith.

Chapter 1

Introduction

This chapter introduces the necessity of touch-free interaction and focuses on the demand for mid-air hand pose interaction and its challenges. In particular, ***i)*** *we analyze the limitations of current interaction methods, emphasizing the importance of device-free technologies in enhancing user experience.* ***ii)*** *we explore the benefits of mid-air hand pose interaction while highlighting challenges related to accuracy, user adaptability, and environmental factors. Finally, the chapter summarizes the key contributions of this thesis and provides an overview of the subsequent chapters, establishing the background and technical framework for this research.*

1.1 The Need for Touch-Free Interaction

As technological change increases across all industries brought about by the development of artificial intelligence, enthusiasm for new technologies and sciences is growing. But the question that follows is whether the new technologies give users a comfortable experience in practical applications. In other words, whether the way the user interacts with the smart device or smart object can meet the user's expectations or requirements. The rapid development of artificial intelligence is not only reflected in specific technological or laboratory environments, but also affects areas such as healthcare, entertainment, and industry. Whether it is widely used smart devices such as mobile phones, virtual environment based VR, AR, or still in the developmental stage of embodied intelligence based robots, it is applied in various fields we are familiar with. And interacting with intelligences still faces many challenges in the research of AI applications.

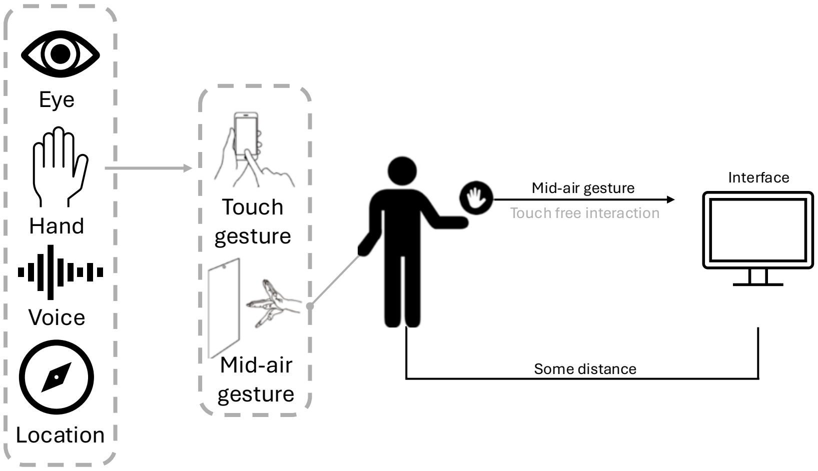


Figure 1.1: The touch free interaction and mid-air hand pose interaction concept figure(Guo & Pan 2023).

The Figure 1.1 shows the concept of touch-free interaction. For sanitary and health concerns, how to make more efficient, understandable touch-free interactions has attracted the interest of a growing number of researchers (Jalaliniya et al. 2013). Such sanitary concerns were particularly serious during the COVID-19 outbreak and in the post-outbreak era (Huang et al. 2020, Pearson et al. 2022). Even outside of outbreaks, public interaction interfaces also increase the risk of transmitting bacteria and viruses. So replacing traditional contact interaction with touch-free interaction is a natural approach for designers to consider. Based on the requirement of touch-free interaction, some mature and simple interactions have already been applied in daily life. For example, some simple payments using voiceprint (Alver 2007) or face recognition (Nasution et al. 2020). Compared with the traditional password input, the contactless voice or face recognition provides a cleaner and safer interaction for users.

However, while researchers are designing safe methods of interaction, they must not affect the convenience of the user interacting with the system. If the user has to learn a new mode of interaction all over again in order to interact with the system, the user might be annoyed with the interaction and use it less (Weise et al. 2020). In addition, in some specific scenarios, considerations based on convenience and naturalness of interaction may provide a more convenient interaction experience for the user. The ability to control a device using gestures, voice, or proximity sensors makes daily activities more seamless. For example, swiping the hand to skip songs or adjust the volume is faster than picking up the device to navigate the menu (Xu et al. 2023). In scenarios where both hands are busy, such as cooking in the kitchen or carrying groceries, voice commands or mid-air gestures enable users to interact with their smart devices simply and quickly without interrupting the current activity of their hands.

Similarly, a touch-free interaction system provides a much easier, less stressful interaction for people with disabilities or mobility issues. A smart home system that can be controlled by voice is obviously more user-friendly for those who cannot use physical buttons, or touch a screen due to mobility issues (Jeet et al. 2015). At the same time, in scenarios where direct contact is inconvenient, non-contact interaction offers more accessible possibilities. For example, wearing gloves in cold environments or working in sterile laboratories require a contactless interaction (Pearson et al. 2022). As an alternative to traditional contact interaction, touch-free interaction increases the accessibility of the system and reduces the interaction limitations in some specific scenarios. Especially in situations that may be dangerous or contaminated, a touch-free interaction is necessary and efficient. For example, when surgeons perform surgery (O'Hara et al. 2014), gesture-based controls enable them to view and manipulate digital imaging tools while maintaining sterility. Workers handling toxic substances or working in extreme heat (Dang & Cheffena 2024) can operate safely and efficiently without physical contact with the machine through voice-activated commands or proximity sensors.

With advances in artificial intelligence and sensor technology, contactless interaction interfaces are becoming more accurate, flexible and diverse. The demand for contactless interactions is increasing, both for the ordinary user as well as for designers or industry. On the one hand, touch-free interactions, which are supported by different new technologies and devices, bring a more immersive interaction experience to the user. For example, in the application of virtual and augmented reality (Kim et al. 2024, Wu et al. 2019, Zhang et al. 2017), touch-free controls allow users to operate virtual objects in a natural and immersive way. Using gestures to pick up, move, or throw objects provides a much deeper engagement than interactions limited to those based on other physical agents, such as the mouse. This experience gives users a more immersive experience both in gaming and in daily operations. On the other hand, as mentioned before, touch-free interactions are attracting more users due to the futuristic technology of the interactions and novel interaction experiences, meanwhile, the consumer products that employ touch-free interactions offer companies benefits in terms of reducing costs and increasing product flexibility. For example, smart TVs are equipped with gesture or voice controls now, which reduces the demand for traditional remote controls by users, so a traditional remote control (Dezfuli et al. 2012) with many functions is no longer a necessary accessory for TVs. Unlike the usual physical controls, touch-free systems can be adapted to different users and environments. For example, voice assistants (Hoy 2018) can recognise a personal voice and provide customised responses. Gesture systems (Kim et al. 2015) can be adapted to a user's range of movement or personal preferences. It certainly provides greater competitiveness to the product.

1.2 The Need for Mid-Air Hand Pose Interaction

In recent years, mid-air gesture recognition and control methods have garnered considerable attention as a touch-free interaction mechanism across various applications. Mid-air gestures provide a more intuitive control experience than voice commands or proximity-based systems. Realistic expressions of physical movement can be combined with multiple forms of feedback to give the user a more realistic interaction process.

While in some scenarios, such as voice-control-based smart home systems (Jeet et al. 2015), touch-free interactions based on similar voice recognition offer considerable convenience, there are still many environments where interactions such as voice control or voiceprint recognition are not appropriate (Seaborn et al. 2021). For example, in a noisy office or a meeting room in the process of a meeting, users are unable to give commands loudly and clearly, which means that it may be difficult for the interaction system to recognise voice commands clearly and accurately. In addition, in some public places, mid-air gestures, especially minor micro-gestures, can provide a more silent and unobtrusive interaction solution and reduce the disruption of the user's interactions to the surrounding area, whether based on privacy considerations or on the user's own mental need (Sharif & Tenbergen 2020).

In addition, with the development of various sensor technologies as well as extended reality (Bhardwaj et al. 2021, Dube & Arif 2023), users are placing more and more emphasis on the sense of participation in interactions in various scenarios. In other words, interaction modes that provide a higher sense of participation and interaction, whether in games, work, or daily activities, tend to provide higher user satisfaction. The interaction mode of mid-air gesture has an exceptional performance in such considerations. Specifically, mid-air gesture-based interaction processes provide more direct methods of manipulation in extended reality-based games (Du et al. 2022). Pointing a finger directly to select an object or swiping the palm of a hand to browse a menu reflects the intuitive and habitual actions we perform in the real world, and gestures are easier to learn and remember and respond faster than abstract voice commands. In museums, galleries, and some facilities used for exhibits or storytelling (Baraldi et al. 2015, Shapiro et al. 2017), the exploration by hand movement creates a visually more immersive experience that attracts and maintains the user's attention, which increases the utility of display facilities that are designed to output information.

In summary, mid-air gesture interaction provides a more flexible and wider range of application scenarios than other touch-free interaction methods such as voice control. It gives the user a more engaging and experiential interaction process, which has motivated more researchers to research mid-air gesture interaction.

1.3 Difficulties of Mid-Air Hand Pose Interaction

Many researchers are actively seeking to optimize interactions between mid-air hand poses and smart systems; however, designing such interactions while adhering to a human-centered approach remains challenging. Firstly, the shape and size of hands are different for different users, for example, the size of a child's small hand and an adult's hand, and the differences in these unchangeable characteristics contribute to the stability and accuracy of the effect of the interaction model. The encoding of hand information, including size and shape, will invariably result in differences. However, in many interaction scenarios, the size of the hand is not a primary concern; instead, the focus is on the diverse interaction outcomes that can be achieved through hand gestures. Consequently, addressing the potential bias introduced by different feature information can enhance the stability and accuracy of the interaction.

The latency of the interaction also affects the user's interaction experience (Liu & Heer 2014). If the users wait too long from the time they make a hand movement to the time they get the response from the interaction system, it will affect the consistency of the interaction. Sometimes, even a small delay can be unbearable. Like the annoyance when the user wants to play or pause, but has to wait a second. This is why many applications or smart devices want to ensure the interaction is fluent. Reducing latency during interaction is an important issue for interaction designers to consider.

Many existing gesture recognition or classification models are supervised learning (Ge et al. 2024, Hayashi et al. 2021, Liu et al. 2021, Qi et al. 2024, Stančić et al. 2017), which means that a large amount of data is required to achieve accurate gesture recognition. In addition, many gesture recognition methods based on supervised learning are preset with some gesture categories (Liu et al. 2020, Nguyen et al. 2023). And in real life, many users find it difficult to complete some gestures when interacting with systems. Meanwhile, some dynamic gestures, such as rotation, stretching, etc., are likely to not get the desired interaction results because the gestures are not performed in a proper manner. For users who are not used to precise gestures, preset gestures may not be easily completed. The challenge of how to address the accuracy requirements of hand movements during interaction remains an important research direction.

In addition, when focusing on human-centred interactions, an important issue that researchers should also consider is the system learning cost of the user (Novack & Goldin-

Meadow 2015, Paik et al. 2015). Not all users are able to quickly understand and master the interaction meanings of different gestures (Pukari et al. 2023), especially some complex hand movements. In some cases, it may take a long time for users to become familiar with a new interaction mode. For example, if the interaction system expects the user to drag a progress bar with fingers, it may take several times for the user to get a clear idea of how the finger position corresponds to the progress bar. On the other hand, gesture controls with static feedback may be easy to understand, such as finger taps to control start and pause, etc. However, for dynamic controls without fixed hand movement states, it may be more difficult for the user to understand how to perform the interaction or what kind of interaction results will be triggered by the change of hand movements. Therefore, ways to reduce the user's learning time and improve the user's understanding of how the interactive system works in a superficial way will help develop a more user-friendly and practical interactive system.

For interactive system designers, a well-designed interactive system should facilitate seamless migration across devices and environments to ensure that it is effective and adaptable (Hosseini et al. 2023, Myers et al. 2000). Many existing mid-air gesture interaction systems have been developed based on specific usage scenarios, such as the contactless interaction of mobile phones, which receives hand movement signals through the internal sensors of the mobile phone, and then controls the flipping of a video or an e-book (Kallio et al. 2003, Lu et al. 2014). However, it is still difficult to have a general processing method and processing logic used for interaction for different scenarios and different types of high-dimensional sensor inputs. This means that similar research that has already been performed is likely to require the development of entirely new models and interaction logic due to differences in hardware such as sensors. Therefore, a more unified and migratable approach to mid-air hand gesture or pose interaction needs to be investigated.

These considerations are crucial in the broader scope of user-smart-device and smart-interaction system design, as they influence the system's overall practicality, usability, and applicability across different scenarios. This research builds upon these foundational insights by addressing several limitations present in existing methods, particularly in handling long-time sequence dependencies in models and enhancing the interpretability and Extensibility of mid-air gesture interactions.

1.4 Overview of Thesis and Contributions

The main objective of this thesis is to propose a generic framework that can be used for mid-air hand pose interaction with high-dimensional inputs. **The framework will be able to be**

used for the dimensionality reduction and visualisation of high-dimensional dynamic mid-air hand pose and provides an extensible interaction design process and ideas for real-time, smoothness, and flexibility of the interaction system. Based on the objectives, the central research question of this thesis is: *How can machine learning be used to design mid-air hand pose interaction methods that improve efficiency, robustness, and user experience in touch-free interaction?*

In Chapter 2, we will introduce existing gesture interaction methods, devices, machine learning foundations for gesture recognition, and evaluation methods for interactive systems from three aspects: mid-air gesture interaction, machine learning in interaction, and evaluation of interaction, and briefly describe the problems and advantages.

Current mid-air gesture methods face persistent challenges, particularly in terms of model dependency on long-time sequences and limited interpretability in gesture interactions. Long sequence dependencies hinder the speed and accuracy of gesture recognition and response, while a lack of interpretability limits intuitive understanding and usability. To address these challenges, we propose a continuous interaction strategy in Chapter 3 that integrates visual feedback of hand poses and recognizes gestures using only two video frames. This approach minimizes time dependencies, enhances response speed, and provides a clearly structured gesture embedding space, allowing users to interact with a more interpretable and responsive system.

Specifically, our approach utilizes frame-based hand pose features from MediaPipe Hands, containing 21 landmarks, embedded into a two-dimensional pose space via an autoencoder, thereby compressing high-dimensional hand pose data while retaining essential interaction cues. A PointNet-based model is then applied to classify gestures, enabling device interaction and exploration control. By jointly optimizing the autoencoder with the classifier, we have developed a gesture-discriminative embedding space, improving classification accuracy and processing speed. Experimental results indicate that our embedding space achieves superior performance in gesture classification accuracy (75.2%) and interaction processing delay (2.4ms) compared to existing methods, providing users with a more interpretable and efficient interaction experience.

Despite these advances, challenges such as jitter, instability, and non-smoothness in dynamic movements remain prevalent, especially when diverse hand conditions are considered. Many existing interaction systems rely on black-box gesture recognition models that deliver rapid, direct feedback but often lack interpretability and adaptability for complex interaction tasks. To tackle these issues, we designed a Hand-pose Embedding Interactive System (***HpEIS***) as a virtual sensor in Chapter 4, which maps users' flexible hand poses to a two-dimensional visual space using a Variational Autoencoder (VAE) trained on a large variety of hand poses, with only a camera required for hand pose acquisition. To improve stability and smoothness, we introduced several processing measures, including quaternion processing, hand-pose data augmentation, an anti-jitter regularization term added to the loss function, stabilizing post-processing for movement turning points, and smoothing post-processing based on One Euro Filters. These enhancements were validated in target selection tasks, where the system achieved significantly improved task completion times with the gesture guidance window (approximately 10s) compared to without it (approximately 20s). Questionnaire responses further indicated that HpEIS provided an engaging, stable, and smooth interaction experience, with all participants agreeing that hand-pose interactions offered greater novelty and appeal than traditional touch- or voice-based systems.

To further expand the applicability of hand-pose interactions across diverse scenarios, in Chapter 5, we developed an adjustable hand-pose space disentanglement approach for a learnable VAE-based high-to-low dimensional embedding model, ***HandSolo***. This model decomposes the latent embedding space into multiple independent one- or two-dimensional spaces, enabling multi-DOF (degrees of freedom) control. HandSolo introduces a flexible, extensible interaction system paradigm, supporting multi-dimensional configurations and multiple DOF combinations. To maximize model performance and user comfort, we also proposed a Visual Interaction Evaluation Strategy (VIEs) to help system designers understand model capabilities and users' preferences. Experimental studies demonstrated the effectiveness of our embedding disentanglement designs through a discovery Experiment I for VIEs, an inspiration Experiment II for approach extensibility, and an exploration Experiment III for the virtual interaction system.

In summary, we present an integrated mid-air hand pose interaction system that combines visual feedback and low-dimensional embedding learning to address key limitations of previous models in hand-pose control. This system provides flexible, stable, and interpretable touch-free interactions, adaptable to multiple devices and contexts. Our experiments show that this system offers a robust, engaging user experience and can serve as an alternative for mid-air hand pose control. This work thus contributes to the field by building on foundational insights and addressing critical aspects of adaptability, usability, and scalability in mid-air hand pose interaction design.

This research has made contributions in the following areas:

1. We developed a hand pose recognition and control strategy based on two video frames, achieving fast and high-precision hand pose classification through the joint optimization of an autoencoder-based and a PointNet-based model.

2. We introduced the Hand-pose Embedding Interactive System (HpEIS), which utilizes a Variational Autoencoder to map user hand poses to a two-dimensional visual space, providing visualized interactive feedback.

3. We developed a hand pose space disentanglement model (HandSolo) that supports flexible interactions across multiple dimensions and degrees of freedom.

4. Through systematic experimental design, we evaluated the effectiveness and practicality of the proposed methods, offering references for future research and practical applications.

Most of the thesis generalizes and builds on the following publications accepted by various international conferences and journals, as follows:

1. Xu, S., Kaul, C., Ge, X. and Murray-Smith, R., 2023, June. Continuous interaction with a smart speaker via low-dimensional embeddings of dynamic hand pose. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE. (Chapter 3)

2. Xu, S., Ge, X., Kaul, C. and Murray-Smith, R., 2024, July. HpEIS: Learning Hand Pose Embeddings for Multimedia Interactive Systems. In 2024 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE. (Chapter 4)

3. Xu, S., Ge, X., Kaul, C. and Murray-Smith, R., Hand Solo A Mid-Air Hand Pose Interaction Method Based on Disentangled Degrees-of-Hand-Freedom. Accepted by ACM Multimedia (ACM MM) 2025. (Chapter 5)

All experiments in this research has been accepted by Ethics Committees, application number 300230025, 300210276.

Chapter 2

Background and Related Work

This chapter reviews key research in the field of mid-air hand pose interaction, focusing on three main aspects. ***i)*** *We explore broader studies on touch-free interaction methods, followed by a more detailed introduction to the devices and techniques related to mid-air hand pose interaction. We also reference studies on both dynamic hand poses and static gestures, providing insights into real-world applications and key considerations in gesture-based interaction.* ***ii)*** *We examine the role of machine learning in interaction, which contributes to the development of our interaction models. Our research is inspired by autoencoders, with a particular emphasis on variational autoencoders and their underlying theories. A wide range of probabilistic models and supervised learning techniques serve as valuable references for gesture recognition and classification in mid-air interaction.* ***iii)*** *We discuss basic evaluation methods for interactive systems, highlighting the widespread application of Fitts' law and closed-loop feedback in assisting designers in refining and iterating interaction systems to enhance usability and efficiency. By integrating these areas, this chapter aims to provide an overview of advancements in mid-air hand pose interaction and their relevance to the research presented in this work.*

2.1 Mid-air gesture interaction

2.1.1 Touch-Free Interaction

In recent years, touch-free interaction technology has attracted much attention, and its applications have been extended to smart home (Geeng & Roesner 2019, Kühnel et al. 2011, Wu & Fu 2011), healthcare (Da Gama et al. 2015, Meng et al. 2013), and other fields. In some scenarios, proximity sensing (Cheung & Lumelsky 1989, Hsiao et al. 2009) provides a simple interaction solution. Proximity sensors typically utilise technologies such as infrared and ultrasonic signals (Cheung & Lumelsky 1989) to detect the presence or proximity of a user, thereby triggering predefined actions without the need for physical contact (Wu et al. 2024). While proximity sensing is effective for basic applications such as remote monitoring, safety prevention and healthcare systems (Fu et al. 2023, Wu et al. 2024), it lacks the flexibility required for complex interactions because it cannot respond to specific commands or gestures. This limitation makes proximity sensing unsuitable for applications that require more complicated interactions.

Voice control is one of the widely adopted touch-free interaction methods in some interaction scenarios that require more functionality, especially in smart home (Mittal et al. 2015) and personal assistant applications (López et al. 2018, O'Brien et al. 2020). Voice assistant systems such as Amazon's Alexa (Hoy 2018, Lopatovska et al. 2019) and Google Assistant (López et al. 2018) allow users to speak commands and then provide corresponding results or interfaces. The touch-free, eye-free feature of voice control may be beneficial for users who may have barriers to the use of digital technologies (Jakob et al. 2021), or who have physical ailments (Derboven et al. 2014). However, voice control systems often encounter problems such as noise interference (Martinek et al. 2020) and dysarthria (Ballati et al. 2018), which can hinder the accuracy and consistency of commands (Gong & Poellabauer 2018). Additionally, privacy concerns remain, as these systems often rely on continuous listening, which can lead to user concerns about the security of personal data (Cheng & Roedig 2022).

With the maturity of eye-tracking devices nowadays, eye-tracking has become a more immersive way of interaction (Carter & Luke 2020, Krafka et al. 2016), especially in virtual reality (VR) (Clay et al. 2019), augmented reality (AR) (Koulieris et al. 2019) and assistive technologies (Majaranta 2011). On the one hand, eye-tracking systems can detect and respond to the user's gaze, enabling contactless control by recognising the user's gaze point on the interface. In VR environments, this allows for seamless control of in-game objects (Ren et al. 2011) or navigation through virtual space (Andersen et al. 2012). On the other hand, eye tracking allows for interaction in special situations, such as users with physical ailments (Wästlund et al. 2015). However, for users in real-life scenarios, the high cost of eye-tracking hardware and the problem of eye fatigue during a long period of use pose a challenge to its widespread adoption (Valtakari et al. 2021).

In contrast, gesture recognition offers a flexible and natural (Panwar & Mehra 2011) approach to contactless interaction that is effective in a range of applications from utilities (Ban-

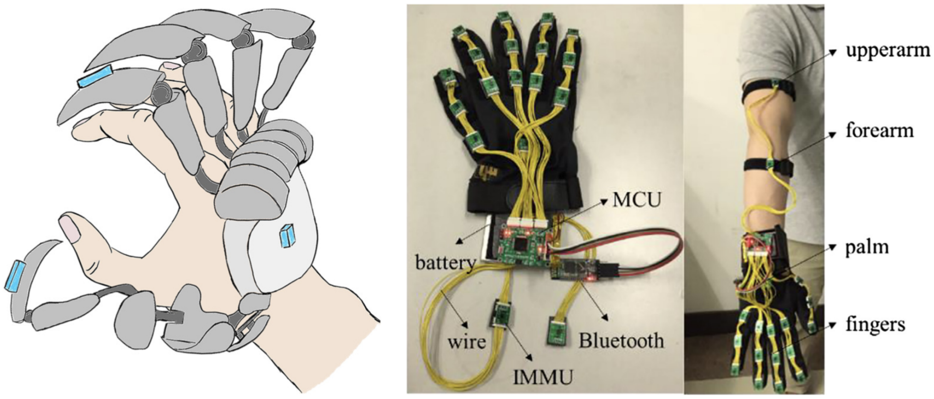


Figure 2.1: Data gloves. Collects data through positioning points on the glove, including coordinates, velocity, acceleration, flexion (Rusu et al. 2021).

garu et al. 2020) to smart home systems (Kühnel et al. 2011). Gesture recognition systems allow users to interact in a simpler way. For example, Microsoft Kinect (Mousavi Hondori & Khademi 2014, Panger 2012, Yang et al. 2019). provide gesture-based control for VR and AR, allowing users to more accurately manipulate virtual objects without physical contact (Da Gama et al. 2015, Meng et al. 2013). Due to the flexibility (Panwar & Mehra 2011) of the human hand, gesture recognition can provide more varied interactions (Panwar & Mehra 2011, Zhou et al. 2023) than just recognising the presence of the user, and allows users to interact silently and with small movements (Xu et al. 2024, 2023) in a space that requires silence or confinement. Additionally, unlike eye tracking, many gesture recognition methods do not require heavy equipment support and are therefore less likely to cause fatigue after prolonged use (Kim et al. 2015). Although some camera-based gesture recognition systems may face challenges related to lighting conditions (Ren et al. 2011), machine learning and the support of different sensor types and accuracies significantly improve the accuracy of gesture interactions (Ren et al. 2011, Yan et al. 2023).

2.1.2 Device

At present, common interactive technologies mainly focus on voice, touch screen and other fields, but the technology of direct gesture without visual attention and direct recognition in the air is still in the development stage. Gesture recognition has the advantage of silent, visual-attention-free messaging over traditional, widely-used voice and touch-screen interactions. (Ahmed et al. 2021, Guo et al. 2021, Sun et al. 2021, Yeo et al. 2015) Specifically, speech recognition interaction is more suitable for use in home environments and personal

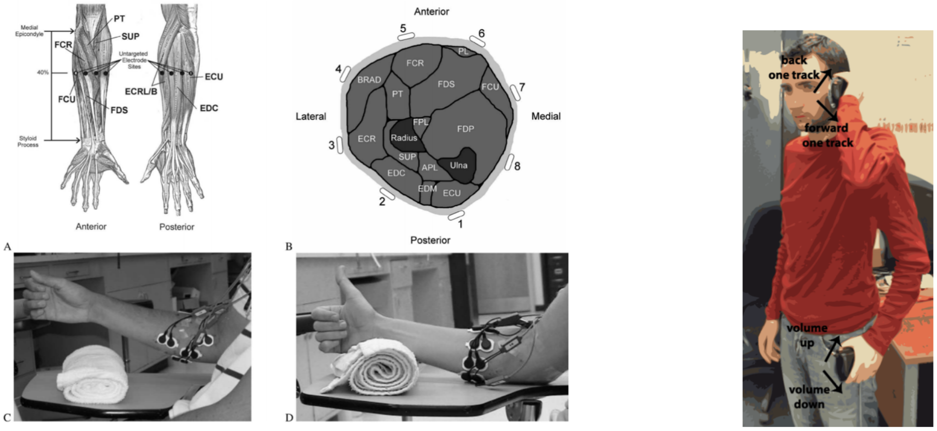


Figure 2.2: (left) Surface EMG (Farrell et al. 2008). The human skeleton and muscles used in EMG and their cross-sectional drawings, as well as the EMG test equipment. (right) Inertial sensing (Strachan et al. 2007). Different positions of the phone relative to the body trigger different responses. Different tilt of the phone in different directions triggers different responses.

space, because once entering public space, the sound emitted by interaction may become unacceptable. For example, when listening to music in the library, the quiet environment does not allow users to interact loudly with the system (Kimura et al. 2019). The most obvious downside of touch-screen interaction is that it is often accompanied by visual attention, which means it can be inconvenient for users in situations where they cannot be distracted, such as when they want to hang up a phone call that comes in unexpectedly while playing a computer game (Yeo et al. 2015). Interaction through gestures can be silent and rapid.

Gesture, as one of the most important ways to express ideas and convey information, is flexible, convenient and less restricted. Therefore, the research on interaction design based on gesture recognition has attracted great attention in recent years. (Guo et al. 2021) In recent years, popular gesture recognition sensing technologies include data gloves (Figure. 2.1) (Dipietro et al. 2008, Fang et al. 2018, Rusu et al. 2021), vision (Lin & Ding 2013, Sun et al. 2018), surface EMG (Figure. 2.2 left) (Farrell et al. 2008), ultrasound (Yang et al. 2018), pressure sensing (Dementyev & Paradiso 2014), inertial sensing (Figure. 2.2 right) (Fang et al. 2018, Strachan et al. 2007), motion-sensing (Wen et al. 2016, Xu, Pathak & Mohapatra 2015) and so on. However, most of the devices involved in these methods are inconvenient and limited and have limited ability to recognize subtle hand gestures. (Guo et al. 2021, Sun et al. 2021)

- Depth Cameras

To develop a prototype system that can be used to interact with the music system, subtle

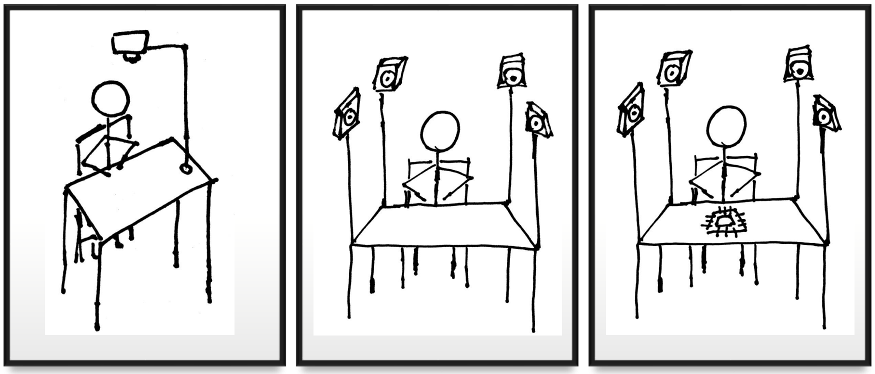


Figure 2.3: (left) camera above the user (middle) cameras are around and above the user or table in a room (right) cameras are around the user and sensor is on the table.

gestures, such as common music application functions, play, pause, previous, next, volume control, *etc*, need to be recognized. So researchers trying to use depth cameras (Figure 2.3) (Kim et al. 2015, Ren et al. 2011, Wilson & Benko 2010). One of the benefits of using a depth camera is that it can be positioned in the environment, avoiding the bulkiness of large wearable devices such as headsets or hand-held (Kim et al. 2015, Wilson & Benko 2010). We can simulate scenarios with multiple sensors for better performance, or more immediately commercially realistic ones with a single sensor associated with the smart speaker.

- Soli Radar

To support fast recognition response, camera-based gesture recognition usually causes user privacy problems. A highly efficient, low-energy miniature gesture sensor developed by Google based on the physics of millimetre wave radiofrequency radiation—Soli Radar (Figure 2.3, Figure 2.4). The sensor can recognize micro gestures well and has a range of 10-15 metres, so can also track room context. (Figure 2.5) (Lien et al. 2016)

In the existing research on small wearable devices, the research on bracelets, watches and rings accounts for a large proportion, because similar jewellery or accessories wearable devices are more lightweight and natural. (Kim et al. 2010, Sun et al. 2021, Vatavu & Bilius 2021, Yin et al. 2021) However, a lot of research focuses on discrete gesture recognition (Sun et al. 2021), so the establishment of a wearable device that can be used for both discrete and continuous gesture recognition and tracking is a promising direction. Wearable devices also allow immediate identification of the user, and the ability to personalise the content to them. Besides providing personalized services for each user, it can recognize the user's information

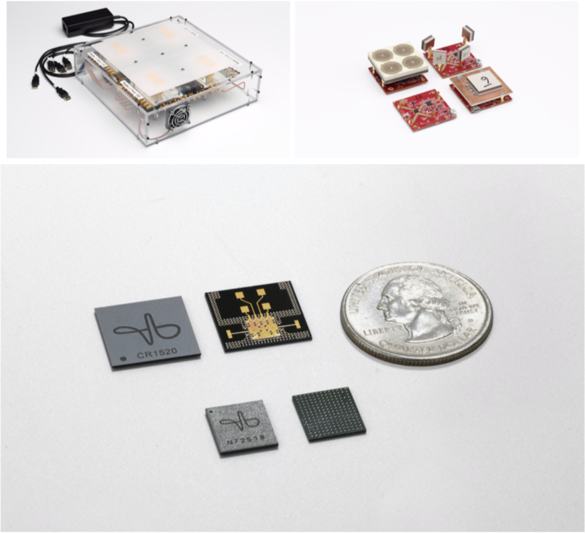


Figure 2.4: A highly efficient, low-energy miniature gesture sensor developed by Google based on the physics of millimetre wave radiofrequency radiation—Soli Radar (Lien et al. 2016).

when other users enter the room and then provide personalized services to the master user based on the information of other users, (Kim et al. 2010, Vatavu & Bilius 2021, Yin et al. 2021) such as recommending playlists suitable for multiple people. This is a good implementation for tasks that want to identify the current environment. If there are many people in the room, we can make a decision on an appropriate recommendation after obtaining information about each person's basic personal interests.

2.1.3 Gesture and Hand Pose

Among the existing research on the use of gestures for interaction and control (Anthony et al. 2012, Lien et al. 2016, Mlakar et al. 2021, Potts et al. 2022, Shakeri et al. 2017, Sun et al. 2021), the two main directions are contact gesture control and mid-air gesture control. Contact gesture control means that the hand comes in contact with some medium and makes a series of gestures to interact (Anthony et al. 2012, Chamunorwa et al. 2022, Potts et al. 2022). Contact gestures have been used in many studies for smart furniture design as well as for the design of gesture exploration tools. Embedding interaction control interfaces on a set of objects and furniture that users are already familiar with allows users to interact more explicitly in the context in a way that is more familiar to them in their familiar environment (Chamunorwa et al. 2022, Parilusyan et al. 2022, Potts et al. 2022). The obvious benefit of this

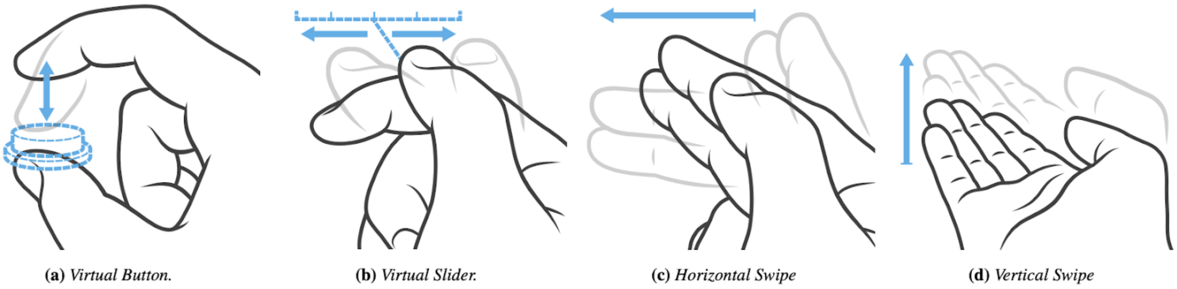

(a) *Virtual Button.* (b) *Virtual Slider.* (c) *Horizontal Swipe* (d) *Vertical Swipe*

Figure 2.5: The Soli radar can recognize multiple micro gestures at a range of 10-15 metres (Lien et al. 2016).

approach to interaction through media is that it avoids the application of additional devices such as cameras (Chakraborty et al. 2018, Ren et al. 2011), sensors (Lien et al. 2016), or data gloves (Dipietro et al. 2008, Fang et al. 2018, Rusu et al. 2021). In addition, due to the diversity of mediums, the same gesture on different mediums can be given different meanings (Chamunorwa et al. 2022). However, at the same time, the way and range of gestures can be accomplished is limited.

On the contrary, in the existing studies on mid-air gesture control, a large number of devices are used for the detection of gestures, such as EMG (Farrell et al. 2008), data gloves (Dipietro et al. 2008, Fang et al. 2018, Rusu et al. 2021), soli radar (Lien et al. 2016), sensors embedded in wearable devices (Kim et al. 2010, Sun et al. 2021, Vatavu & Bilius 2021, Yin et al. 2021). Cameras and depth cameras are used in many studies of gesture recognition (Chakraborty et al. 2018, Ren et al. 2011). In our study, we also used the depth camera as the beginning of the research to perform mid-air gesture interaction and control.

Based on this, we are also inspired by some well-established methods for gesture and pose recognition, such as Mediapipe (Lugaresi et al. 2019), a tool that can apply cameras for recognition, segmentation and tracking of different objects, as well as for the detection and marking of keypoints. Mediapipe hands is a hand and finger tracking method. It can predict 21 3D landmarks of a hand from a single frame only using machine learning (ML) (Lugaresi et al. 2019, Zhang et al. 2020). It does a good prediction and marking of hand keypoints even when the hand is partially occluded. Compared to embedding the images of video frames directly, representing each frame with 21 3D- point coordinates largely reduces the complexity of the model and the prediction time.

In the existing research on gestural interaction, the design of interactive gestures has been broadly classified into two categories, one for discrete gestures and one for continuous gestures. Discrete gestures mean that feedback is obtained after the full gesture has been triggered, while continuous gestures obtain the real-time feedback while the gesture is in progress. Compared to the methods (Kim et al. 2010, Liao et al. 2021, Vogiatzidakis & Koutsabasis 2021) that focus only on discrete gesture interaction, our approach handles both discrete and continuous gesture-based interaction scenarios.

2.1.4 Reality and Interaction

Thanks to the development of the Internet of Things, ubiquitous computing (ubicomp) is increasingly being mentioned in the field of modern interaction. The main function of ubicomp is to help users perform various operations and tasks more efficiently. This means that context-awareness (CA) is an important part of the research in the field of ubicomp. (do Nascimento et al. 2021, Marquardt & Greenberg 2015) Context is the information that can represent entities and influence the interaction between users and the system. (Abowd et al. 1999, Prekop & Burnett 2003)It includes the identification of characters (including identity, interests and hobbies, number.), item identification (name and location), environmental recognition (environment information such as an object, sound, temperature.), people and entities position (the position relationships between humans, position relationships entities, the relationship between humans and entities, and the humans' entities' orientation.) recognition. (Marquardt & Greenberg 2015)

2.2 Machine Learning in Interaction

2.2.1 Autoencoder

Low-dimensional embedding methods have been widely used to support real-time hand gesture control while minimising the complexity of models deployed on user devices. Traditional linear mappings, such as Principal Component Analysis (PCA) (Zebari et al. 2020) and Factor Analysis (FA) (Gisbrecht et al. 2015), have been used to project high-dimensional data into lower-dimensional spaces. However, the linear nature of PCA limits its ability to identify complex, non-linear relationships within the data, potentially leading to a loss of information during dimensionality reduction.

To address these limitations, non-linear dimensionality reduction techniques like t-distributed Stochastic Neighbour Embedding (t-SNE) (Gisbrecht et al. 2015) and Uniform Manifold Approximation and Projection (UMAP) (McInnes et al. 2018) have been introduced. These methods effectively represent complex structures in the data by preserving local and global relationships respectively. t-SNE focuses on keeping local neighbourhood relationships and is particularly effective for visualising high-dimensional data in two or three dimensions. However, it has scalability issues as the computational cost increases quadratically with the number of data points (Linderman & Steinerberger 2019). UMAP, on the other hand, provides a more efficient approach to reduce dimensionality while preserving both global and local structures. Despite these advantages, both t-SNE and UMAP lack the ability to reconstruct the original data from the embedded representation, which limits their utility in tasks requiring reversible transformations.

In contrast, autoencoders provide a robust framework for learning non-linear embeddings that allow both dimensionality reduction and data reconstruction (Hinton & Salakhutdinov 2006). An autoencoder consists of an encoder, which maps the input data into a latent space representation, and a decoder, which reconstructs the original data from this representation. The latent space learned by autoencoders can represent meaningful structures, making them highly suitable for real-time gesture control applications. In addition, recent advances such as convolutional autoencoders (Masci et al. 2011) have demonstrated their ability to extract spatial features, further improving their effectiveness in processing gesture data. Autoencoders also allow fine-tuning of hyperparameters, such as the number of layers and units, to achieve an optimal trade-off between representation quality and computational efficiency.

An extension of autoencoders is the Adversarial Autoencoder (AAE), which uses the principles of Generative Adversarial Networks (GANs) to enforce specific priors on the latent space (Makhzani et al. 2015). By combining the reconstruction capabilities of autoencoders with the adversarial training process of GANs, AAEs can improve the quality of latent representations and provide better control over the generative process. This makes AAEs particularly attractive for tasks requiring structured latent spaces. However, AAEs generally require complex training procedures due to the challenge of balancing adversarial losses and reconstruction goals. While AAEs have been successfully applied in various generative tasks, they are less commonly used in gesture-related systems, where the probabilistic modelling and generative capabilities of Variational Autoencoders (VAE) offer a more direct benefit. VAE models extend the traditional autoencoder architecture by introducing a probabilistic approach to learning latent representations (Kingma 2013). Unlike standard autoencoders, which map data into fixed latent vectors, VAEs encode input data into a probability distribution, typically Gaussian, over the latent space. This allows sampling and generation of new data, enabling VAEs to be powerful generative models. For gesture control systems, this property is particularly useful for data augmentation, since synthetic samples can improve the performance of downstream classifiers (Pu et al. 2016). In addition, VAEs promote disentangled representations, separating independent factors within latent space. However, this comes at the cost of additional complexity, such as latent variables cannot represent meaningful information effectively (Razavi et al. 2019). Solutions such as Beta-VAE have been proposed to reduce these problems by balancing reconstruction quality and latent space disentanglement (Higgins et al. 2017). Despite these challenges, VAEs remain a prime choice due to their flexibility and ability to generate meaningful latent representations, making them highly adaptable for gesture-related applications.

2.2.2 Probabilistic Model

In gesture recognition, probabilistic models are useful for capturing variations in human hand movements. Hidden Markov Models (HMMs) are often used to classify continuous gestures and (Starner et al. 1998, Vogler & Metaxas 1999) used an HMM to recognise American Sign Language (ASL) in real-time, achieving highly accurate gesture recognition results by training on long sequences of different gestures from different users. The effectiveness of HMM in modelling complex, multi-featured gestures was demonstrated. On the other hand, Bayesian networks can even achieve more than 99% accuracy in the dynamic recognition of two-handed continuous gestures (Suk et al. 2010). These models perform well in distinguishing gestures, making them effective in hand pose recognition or classification.

Furthermore, probabilistic models not only improve gesture recognition accuracy, but also support applications ranging from assistive technologies to immersive virtual environments. Specifically, in the field of VR- and AR-based extended reality interactions, HMMs can be used to use data obtained based on eye-tracking for the prediction of users' visual behaviours, leading to a smoother and more immersive museum visit experience (Pierdicca et al. 2018). HMMs have been integrated into automotive control systems for recognising driver gestures for touch-free interaction with in-vehicle entertainment systems to improve safety and comfort (Deo et al. 2016).

2.2.3 Supervised Learning

Gesture Recognition

Mid-air gesture recognition and control has recently attracted increasing research attention in multimedia applications. Early works such as Kinect in the Kitchen (Panger 2012) has explored mid-air gestural control and feedback using a Kinect in cooking scenarios, where common devices have limited displays and touch is less appropriate when cooking. Recently, many studies (Qian et al. 2020, Shakeri et al. 2017) have proposed mid-air interaction methods for driving based on mid-air gesture recognition and control, which can prevent driver distraction caused by conventional physical handling or touch. However, conventional gesture interaction methods require a physical device as support and focus on solving physical controls and interactions, such as simple music control, device selection. Hence, in recent years deep learning based gesture recognition methods (Ahmed et al. 2022, Ur Rehman et al. 2021) have been studied to enhance mid-air gesture control and interaction for many applications. For instance, (Ur Rehman et al. 2021) proposed a deep learning architecture based on the combination of a 3D Convolutional Neural Network (3D-CNN) and a Long Short-Term Memory (LSTM) network, which takes advantages of spatial-temporal information from 30-frame video sequences. To address these limitations, the Video Transformer Network (VTN) (Neimark et al. 2021) introduced a transformer-based architecture, allowing for more effective global temporal modeling while reducing redundancy. Though they achieve significant improvements, these methods remain unsatisfactory due to long time sequences dependencies and high-dimensional feature inputs of models. In addition, the interpretability of the user gesture interaction process is not addressed by most current methods, which is widely regarded as a critical component in real-world gesture control.

Gesture Classification

With the application of depth cameras, in addition to the traditional application of RGB images for object recognition, motion recognition, path tracking and other tasks, the use of depth images for video-related tasks is increasingly being studied. Depth image is an image that takes the distance (depth) from the image collector to each point in the scene as a pixel value, which directly reflects the geometry of the visible surface of the scene. The depth image can be calculated as point cloud data after coordinate transformation, and point cloud data with rules and necessary information can also be back-calculated as depth image data (Song & Xiao 2014). As show in Figure 2.6, each pixel point in the image frame provided by the depth data stream represents the distance from the object at that particular (x, y) coordinate to the closest object to the camera plane in the field of view of the depth sensor. Currently, depth images are acquired by LIDAR depth imaging method (Frueh et al. 2005), computer stereo vision imaging (Woodfill & Von Herzen 1997), coordinate measuring machine method (Liao et al. 1999). Compared with the traditional use of RGB images for computer vision related work, depth images can still maintain good picture recognition in low light or exposure state (Frueh et al. 2005, Song & Xiao 2014). In addition, in terms of practical applications,

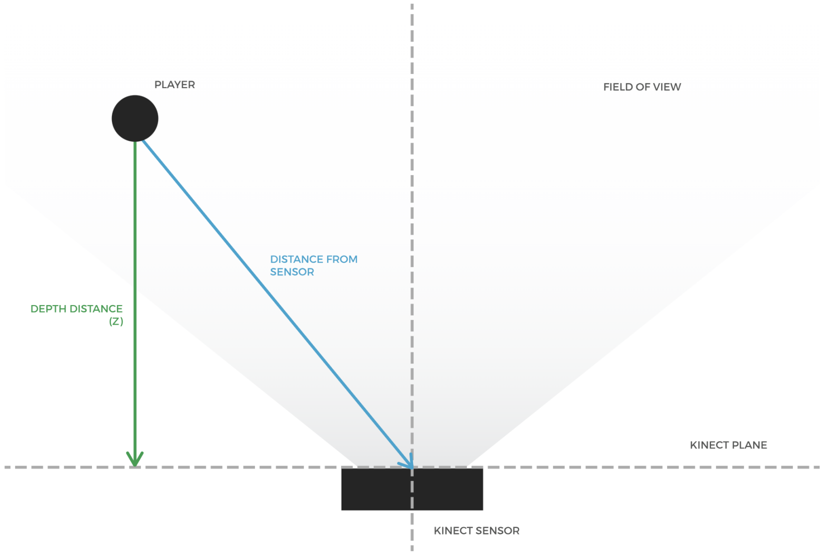


Figure 2.6: Depth distance values are the distances between the points and the sensor plane (Pterneas 2023).

images that do not contain specific details but only depth information can effectively protect the privacy of users (Elezovikj et al. 2013).

In many existing video classification studies, depth images are often used in conjunction with other techniques to achieve good results. For example, deep images and radar data are applied for target recognition, detection (Chen et al. 2022, Han et al. 2022) and classification (Yue-Hei Ng et al. 2015). The extraction of depth information and other information and temporal information of video frames by CNN, self attention and other models are used to perform related tasks (Chen et al. 2022, Han et al. 2022, Yue-Hei Ng et al. 2015).

In addition, because of its ability to detect moving objects and motion trends, the optical flow method is also used for video-related classification (Mahmoodi & Salajeghe 2019, Yu et al. 2014) and detection (Mase 1991) tasks. Optical flow is a method that uses the change of pixels in an image sequence in the time domain and the correlation between neighboring frames to find the correspondence that exists between the previous frame and the current frame, so as to calculate the motion information of objects between neighboring frames(Beauchemin & Barron 1995, Murray-Smith 2017). In general, optical flow is generated by the movement of the foreground target itself in the scene, the motion of the camera, or the joint motion of both (Beauchemin & Barron 1995). By calculating the optical flow between the previous and the current frames, we can obtain the optical flow field as shown in the Figure 2.7, with the direction of the arrow representing the direction of the movement of that pixel point (Murray-Smith 2017). The optical flow carries not only the motion informa-

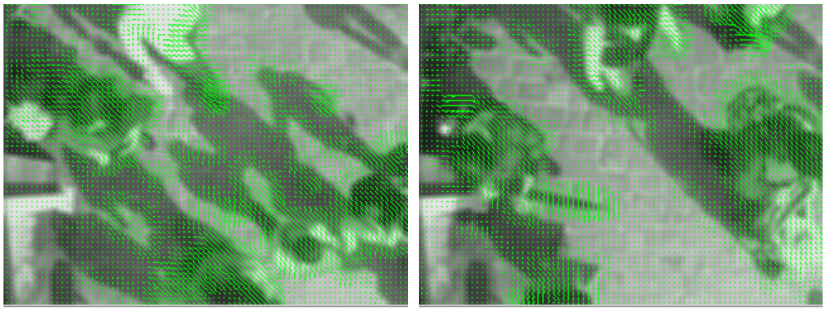


Figure 2.7: Visual optical flow field of realistic scene. The direction of the optical flow is shown by the line segments (Baltes et al. 2015).

tion of the moving object, but also rich information about the three-dimensional structure of the scene, and it is able to detect the moving object without knowing any information about the scene. However, the traditional optical flow method is not applicable when the object to be detected is moving too fast (Beauchemin & Barron 1995).

With the development of 3D technology, point cloud classification of data is an important task in processing 3D data, which is widely used in automatic navigation, robot navigation, 3D reconstruction and other fields. Researchers initially used traditional methods to classify point clouds. For example, multi-view projection converts a point cloud from multiple perspectives into a 2D image, thus allowing the application of 2D CNNs (Su et al. 2015). Although these methods are effective, they lose important 3D geometric information, thus limiting its ability to represent complex spatial relationships. PointNet (Qi et al. 2017) addresses these limitations by directly processing raw, unordered point clouds. It uses shared MLPs for point-wise feature extraction and a max-pooling operation to achieve permutation invariance. This innovation eliminates the need for projection while maintaining the raw structure of the point cloud. However, PointNet does not model local neighbourhoods, making it less effective at detecting local geometric relationships. To improve local feature learning, DGCNN (Wang et al. 2019) introduces a dynamic graph representation in which local neighbourhoods are constructed and updated during training. This method provides richer geometric information compared to PointNet, while maintaining the global context. However, the dynamic graph computation increases the general computational cost, which can be problematic for large datasets.

2.3 Interaction evaluation

2.3.1 Evaluation Methods

Research on touch-free and mid-air interaction has explored different ways of evaluation. Objective measures such as accuracy, response time, and throughput have been widely used, to compare systems under controlled conditions (Hincapié-Ramos et al. 2014, Soukoreff & MacKenzie 2004). These studies also highlight the problem of fatigue in longed mid-air use, with Consumed Endurance (CE) proposed as a metric to capture physical demand. On the contrary, subjective evaluations complement these measures by focusing on user perception. Scales such as SUS, NASA-TLX, and UEQ are commonly applied to assess usability, workload, and overall experience (Brooke et al. 1996, Hart & Staveland 1988, Laugwitz et al. 2008). Gesture interaction research has further emphasized naturalness, intuitiveness, and memorability (Wobbrock et al. 2009).

Most evaluations have been carried out in controlled laboratory settings, where fixed tasks and simple backgrounds reduce noise and ensure comparability. Such tests provide strong internal validity but may not capture the complexity of real-world use (Hincapié-Ramos et al. 2014). Many works (Hart & Staveland 1988, Wobbrock et al. 2009)suggest that combining objective and subjective methods in controlled environments remains the dominant approach, but it also points to the need for more diverse and realistic evaluation scenarios.

2.3.2 Fitts' law

In order to measure the good design of HCI interfaces and to improve the user experience, Fitts' law provides an important way of evaluating the design of interactive systems. By using the following equation:

$$MT = a + b \cdot \log_2\left(\frac{D}{W} + 1\right), \tag{2.1}$$

where MT represents the movement time, D is the distance to the target, W is the width of the target, and a and b are constants to compute the relationship between the movement time, the distance to the target, and the size of the target when the user is completing the interaction task, Fitts's law gives a quantifiable assessment index. Specifically, the further away the target is, the more difficult it is to reach. The smaller the target, the harder it is to hit it (MacKenzie 1992).

Fitts' law is widely used in guiding the design of human-computer interaction interfaces, for example, the bottom bar (dock) is placed at the bottom of the screen by default in mac os, and the start menu is at the bottom left corner of the screen in windows, where the W of the bottom edge and corners are infinitely large, and so the target is infinitely selectable (Tidwell 2010). Moreover, in systems that interact with a mouse, Fitts' law suggests targeted design references in the interaction design of other input types such as visual, haptic. For example, in order to help people with physical and speech impairments to interact effectively, the feasibility of gaze interaction is confirmed in (Rajanna & Hammond 2022) by Fitts' law. In addition, the comparison between mouse and gesture input in (Sambrooks & Wilkinson 2013) confirms that the difficulty of gesture interaction is due to the user's unfamiliarity with gesture-sensing devices and inaccuracy of gestures, which also informs our research, i.e., to consider what can be used to allow the user to become quickly familiar with gestures and to use gestures for interaction. In more complex interactions, such as robotic arm design and interaction design in virtual reality, the use of the Fitts' Law to compare the effects of tasks performed with a robotic arm (Guo 2022) or in virtual environments (Shi et al. 2023) with the effects of tasks performed by human beings in real life will help designers to increase the effectiveness of their designs and improve the user experience.

2.3.3 Closed-Loop Feedback

A closed-loop system is a dynamic loop system that includes perception, feedback, and action (Crossman & Goodeve 1983). To adapt the interaction system to different users' needs and behaviour changes in HCI research, designers often consider using a closed-loop system to adjust the response or design based on the user's behaviour, which may go through several loops of such adjustment (Fischer et al. 2022). Compared with open-loop systems, the dynamic feedback and adjustment of closed-loop systems brings more adaptability to the design of interactive systems, especially in the optimisation process of user experience, and the closed-loop feedback can help designers to adjust the system behaviour to improve the interaction (Renaud & Cooper 2000, Zhongcheng et al. 2005). For example, (Cockburn et al. 2007) explores dynamically changing menu layouts based on real-time input adjustments to accommodate user behaviour. In addition, a system that adjusts the content and difficulty of a lesson based on real-time student performance was introduced in (Luckin et al. 1999) to meet individual learning needs and improve student engagement and learning outcomes.

2.4 Conclusion

In this chapter, we review basic and widely used touch-free interaction methods as well as commonly used interaction methods and devices. Based on this, we highlight work related to hand interaction and provide a brief analysis of existing research on gestures as well as hand poses, including not only contact interaction but also static and dynamic mid-air gesture interaction. To further illustrate the need for research on gesture interaction, we also investigate research related to mid-air gesture interaction in real-world applications. These works provide support for the necessity and justification of our research on mid-air hand pose interaction.

In order to provide better processing of mid-air hand interaction poses for more stable and usable interaction models, we review machine learning and deep learning based related work in the study of gestures and hand poses. These include relevant probabilistic models for gesture recognition and classification, and supervised learning models. As well we highlight the autoencoder based approach which can be used not only for dimensionality reduction but also for generation. In our study of hand pose interaction, the above models provide the theoretical basis for dimensionality reduction visualisation and gesture recognition and generation.

In addition, to better evaluate and improve our interaction system, we discuss methods commonly used for HCI evaluation, including Fitts' law and closed-loop feedback. These methods provide theoretical support for evaluating our mid-air hand pose interaction system, allowing us to make improvements to our system and further refine our evaluation strategies to provide comfortable interaction experiences and interaction design strategies for users as well as interaction designers.

While previous studies have contributed significantly to understanding and designing a user-friendly mid-air hand poses and gestures interaction system, there are still some limitations that have not been addressed. Many studies focus on predefined gestures interaction, but they fail to address dynamic hand poses interaction problems. This limits their interaction applicability in a flexible real scenario. In addition, the dominant gesture classifiaction and recognition approaches, such as deep-learning- and machine-learning-based methods, suffer from high latency when predicting and low interpretability when user interact with the system. These challenges indicate a need for a low latency stable model to predict dynamic hand pose, which provides an interpretable real-time interaction process. Additionally, the relevant literature has not sufficiently explored whether it is possible to interact effectively through the flexible use of different hand movement states, which is the key to designing an interaction system that can be more flexibly extended to more scenarios. Therefore, further research is necessary in order to design an mid-air hand interaction system that can be used more flexibly. This will lead to the development of an interaction design framework that can be extended to multiple interaction scenarios. The aim of the research is to make the interaction process more flexible by introducing the disentanglement of hand poses, and to make the interaction system extensible to more scenarios, as well as to provide an interaction system evaluation scheme for interaction system designers. Although previous research has provided a foundation for the study of mid-air hand-pose interactions, there are still many problems. This thesis will provide a thought for flexible and extensible mid-air hand-pose interaction by proposing a Learning Low-dimensional Latent Spaces for Hand-pose Interaction Systems approach, which will be described in detail in the following chapters.

Chapter 3

Visualization of High-dimensional Hand Pose in Low-dimensional Space

Current mid-air gesture recognition methods struggle with processing long-sequence frames and lack interaction interpretability, limiting real-world usability. This chapter introduces a continuous interaction strategy that integrates visual hand-pose feedback with gesture recognition and control. Our approach leverages frame-based hand pose features extracted from MediaPipe Hands. These features are embedded into a two-dimensional pose space using an autoencoder, followed by a PointNet-based model for gesture classification. The recognized gestures are then used for device control and interaction exploration. ***i)*** *By jointly optimizing the autoencoder with the classifier, we learn a discriminative embedding space for gesture recognition.* ***ii)*** *Through accuracy evaluation and interaction processing latency analysis, we demonstrate that our method achieves a clear embedding space which implements the embedding fast.* ***iii)*** *We validate the system's effectiveness with experienced users exploring various parts of the gesture space by adjusting their hand poses.*

3.1 Introduction

With the increasing use of virtual reality (VR) and augmented reality (AR) technologies and the increased need for direct control with the eyes away from the screen (Khundam 2015, Yousefi et al. 2016), the direct recognition and control of mid-air gestures is now an important area of interaction and artificial intelligence research. Mid-air gesture recognition and control (MG-RC) has recently attracted ever-increasing research attention in multimedia, due to its widely multimedia interactions and applications (Xu et al. 2024), e.g., interactive control of different applications by the mid-air gestures for electronic devices. However, the current MG-RC methods remain unsatisfactory due to the processing of long-sequence frames in the model and a lack of interpretability of the process of mid-air gesture interactions, which are widely regarded as important cues in real-world gesture control. This chapter presents a new continuous interaction strategy with visual feedback of hand pose and mid-air gesture recognition and control for a smart music speaker, which utilizes only 2 video frames to recognize gestures. Frame-based hand pose features from MediaPipe Hands (Lugaresi et al. 2019), containing 21 landmarks, are embedded into a 2 dimensional pose space by an autoencoder. The corresponding space for interaction with the music content is created by embedding high-dimensional music track profiles to a compatible two-dimensional embedding. A PointNet-based model (Qi et al. 2017) is then applied to classify gestures which are used to control the device interaction or explore music spaces. By jointly optimising the autoencoder with the classifier, we manage to learn an embedding space for discriminating gestures. We demonstrate the functionality of the system with experienced users selecting different musical moods by varying their hand pose.

We propose a simple PointNet-based classification network to recognize the predefined discrete gestures and continuous hand poses by fewer frames (2 frames) in low-dimensional inputs (2 dimensions) from an autoencoder. Our approach handles both discrete and continuous gesture-based interaction scenarios. Finally, we defined corresponding functions for the different recognized gestures, which include discrete gesture control for music start/stop and continuous hand pose control for real-time musical space exploration. Our proposed pipeline overcomes to some extent the disadvantages of high-dimensional feature input and long sequence frames, and implements a continuous hand pose to explore the music space. The low frame dependency of the gesture recognition stage gives our model the advantage of low latency. Visible user interaction based on the autoencoder encoding gives the user more freedom of choice and exploration, which is not exploited in the literature.

3.2 Background

3.2.1 Dimensionality reduction

Since the gesture data collected by various types of sensors have high dimensionality and may even reach thousands of dimensions, dimensionality reduction of the data is necessary for better processing of the data and better visualization. There are two main methods of dimensionality reduction, 1) retaining only the most relevant variables in the original dataset (feature selection). 2) finding a set of smaller new variables, each of which is a combination of the input variables and contains essentially the same information as the input variables (feature extraction). For the former, the main methods include high correlation filter, random forest (Reddy et al. 2020, Zebari et al. 2020). For the latter, the main methods include, principal component analysis (PCA) (Cao et al. 2003, Partridge & Calvo 1998, Zebari et al. 2020), independent component analysis (ICA) (Cao et al. 2003, Zebari et al. 2020), t-SNE (Gisbrecht et al. 2015), UMAP (McInnes et al. 2018). A pair of variables with high correlation increases the multicollinearity in the data set, so it is necessary to delete one of them with high correlation filter. Random forest is one of the most commonly used methods for dimensionality reduction, which will explicitly count the importance of each feature in the dataset (Zebari et al. 2020). PCA is one of the most widely used techniques to deal with linear data (Cao et al. 2003, Reddy et al. 2020). We can use ICA to transform the data into independent components, using fewer components to describe the data (Cao et al. 2003, Zebari et al. 2020). Conversely, t-SNE is suitable for nonlinear data processing, and the visualization of this method is more straightforward than other methods (Ge et al. 2024, Gisbrecht et al. 2015). UMAP is suitable for high-dimensional data, and this method is faster compared to t-SNE (Becht et al. 2018, McInnes et al. 2018).

3.2.2 User Feedback and Visualization

When the user is interacting with the device, a feedback system that is easy for users to understand will give the user a good interaction experience (Aula & Surakka 2002, Ryoo & Aggarwal 2007). For example, the tik-tik sound when rotating the dial, the vibration of the smart bracelet when clicking the button. As a form of user feedback, visual feedback is one of the forms we consider in our work, especially when users interact with mid-air gestures, visual gesture space can give users a more visualized interaction experience than only using sound or vibration as a feedback method (Pavlovic et al. 1997). At the same time, visual gesture interface has an active role in exploring the types of gestures that are more user-friendly, easier to use, and more easily recognized by the model (Aloba et al. 2020, Dang & Buschek 2021). In addition to the aforementioned visualization tools such as t-SNE (Gisbrecht et al. 2015), UMAP (McInnes et al. 2018), clustering algorithms (Jang et al. 2014) and autoencoder (Dang & Buschek 2021, Rusu et al. 2022) also perform well in visualization. Take autoencoder as an example, when the output dimension of encoder in autoencoder is two-dimensional, then it is a coordinate that can be visualized in two-dimensional space. A well-performing gesture space can map different gestures to different locations in the gesture space (Dang & Buschek 2021, Rusu et al. 2022). For continuous gesture segments, a smooth line in the gesture space will represent the hand movement.

3.2.3 Music and Interaction

Interacting with multimedia content has become a widespread part of daily entertainment, including watching videos, playing games, and listening to music. However, technical and environmental limitations can affect the user experience. Factors such as device deployment and interaction needs across different scenarios influence how users interact with smart multimedia for entertainment when they have entertainment needs. Taking the interaction of smart speakers as an example, the need to listen to music may occur at any occasion and time, including driving, cooking, etc. Unlike watching movies or playing games, which require prolonged screen attention, listening to music does not demand continuous visual focus. Music can be enjoyed without direct hand operation or eye contact. Therefore, smart speaker-based exploration is an easy start to conduct research on the interaction of mid-air hand pose with smart multimedia systems, as the impact of other modal interactions can be reduced, such as unnecessary voice control and visual feedback, and focused on hand poses.

Currently, touchscreen-based interaction with smart music players is very common. Users can easily browse playlists, adjust volume, and select tracks. However, traditional touch-sensitive surfaces lack adaptability to various environments, which may limit user experience (Modaberi 2024). When cooking in the kitchen, touchscreens become impractical due to dirty hands. When the user is driving, touchscreen interactions can distract drivers and pose safety risks (Mitsopoulos-Rubens et al. 2011). Additionally, while voice control is well-developed in many applications, it may not be suitable for individuals with speech impairments (Harish & Poonguzhali 2015). An alternative interaction solution will meet their needs for interaction with smart speakers in multiple scenarios (Kendon 2014).

Unlike touch and voice control, gesture interaction offers a silent way to control music playback, volume, and track selection while reducing the need for visual attention and surface contact (Carfì & Mastrogiovanni 2021). By using hand's or finger's natural movements, users can control music playback without physical contact or verbal commands. This flexible and convenient interaction solution meets the diverse interaction needs of different people, such as a driver who is driving (Qian et al. 2020, Shakeri et al. 2017) or someone in a quiet public place (Kimura et al. 2019).

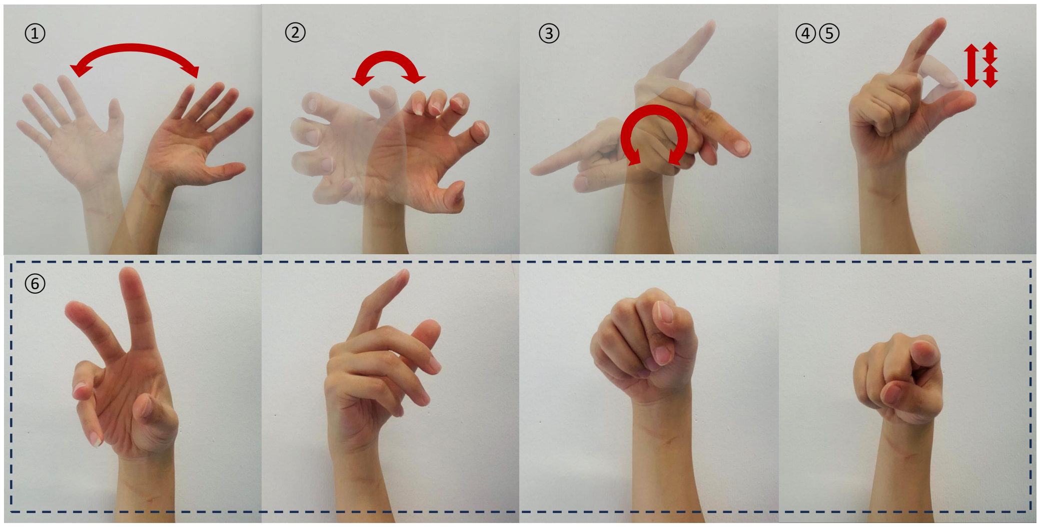


Figure 3.1: Five functional hand-pose schematics and some generic movements. The top row shows the collection of five functional hand pose schematics, with pinch and double pinch at the end. The bottom row shows examples of generic movements.

3.3 Methodology

3.3.1 Dataset

We investigate and design interactive control gestures that conform to human habits for the smart music speaker, including 'continuous palm sweep (right)' (the arm and hand away from body), 'continuous palm sweep (left), 'hand dial rotation (right)', 'hand dial rotation (left)', index finger drawing 'circle clockwise', drawing 'circle counterclockwise', 'pinch' and 'double-pinch'. In fact, since these gestures are in pairs and opposite directions, we collect 8 gestures in total. In addition, we incorporate generic hand pose movements, i.e., the free variation and movement of the hand posture, into the mid-air hand-pose embedding space to enhance the flexibility of the system interaction. Specifically, using the Intel$^{\circledR}$ RealSense$^{\text{TM}}$ LiDAR Camera L515 depth camera (frame rate is 30 fps), we collected 25 clip videos for each gesture for 7 volunteers. Each video duration varies from 1 to 3 seconds. To avoid the influence of the background, we choose a white wall about one meter away from the camera as the background and the collected gestures are 40-50 cm away from the camera. After that, we extract about 90, 000 frames containing gestures from the collected clip videos. In Table 3.1, we provide the detailed information of our collected gesture dataset.

Figure 3.1 presents the collected hand poses used for interaction, including five functional hand poses, called palm sweep, hand dial rotation, index finger and thumb pinch, and index finger and thumb double pinch. These hand poses encompass multiple angular rela-

Table 3.1: Overview information of our collected hand-pose frames corresponding to Figure 3.1.

No.	Hand Pose	Hand Pose Split	Frames	Total Frames
①	palm sweep	palm sweep (right) palm sweep (left)	11002 10807	21809
②	hand dial rotation	hand dial rotation (clockwise) hand dial rotation (counterclockwise)	9526 9386	18912
③	index finger circle	index finger circle (clockwise) index finger circle (counterclockwise)	10520 10460	20980
④	pinch	-	-	9181
⑤	double pinch	-	-	8980
⑥	generic movement	-	-	9781
Total				89643

tionships between the thumb and index finger, as well as the relationship between different inter-articular angles and flexion changes of a single finger or two fingers. The collected gesture data will continue to be used in subsequent research and experiments.

3.3.2 Music Space

Our music data is provided by our industrial partner, Moodagent, including about 55,000 music tracks. Each track is represented by 34 features, including predicted subjective scores for 6 emotion types, 14 genre types, and 14 style types, with scores ranging from 1 to 7. In addition, we do not perform augmentation operations on the music space embedding. These features are derived from the track's audio signal using a convolutional neural network to predict human subjective classifications. The music features are embedded by UMAP down to a 2-dimensional music space for human interaction. In this work, we focus on the exploration of a music space with different emotions, including sadness, joy, fear, erotic, anger and tenderness, which the user can interact with via continuous hand pose changes. We colour the music space with scores for different emotions.

3.3.3 Mediapipe

MediaPipe Hands (Lugaresi et al. 2019, Zhang et al. 2020), a real-time hand landmarks detection model, is used to extract 3-dimensional (3D) coordinate information of 21 hand land-

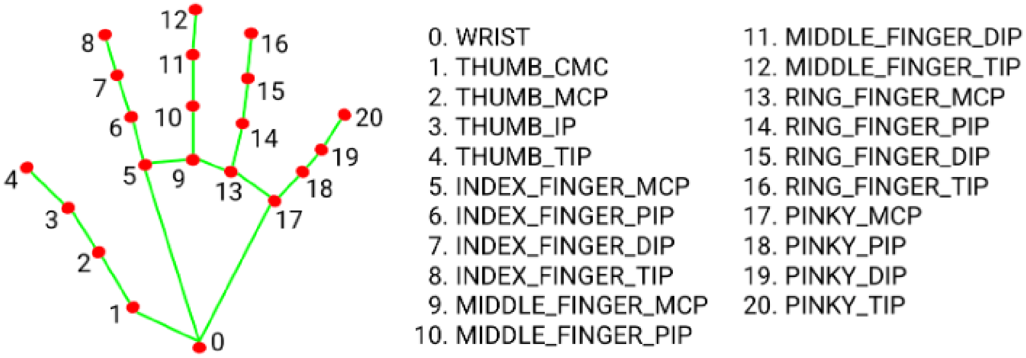


Figure 3.2: Mediapipe 21 hand landmarks (Lugaresi et al. 2019). In this experiment, we will process the 3D coordinates of the hand landmarks and perform dimensionality reduction on the vector composed of the coordinates. Where the vector is the sequential concatenation of x,y,z coordinate values of the landmark numbers in the figure.

marks as shown in Figure 3.2, which saves a significant model space compared to directly using pixel-level image (image size is 480 × 620). Then, an autoencoder based on fully connected layers (Rusu et al. 2022) is designed to reduce the dimensionality of the 3D coordinates of the 21 hand landmarks. The encoder and decoder in autoencoder contain four fully-connected layers respectively, and the layers use LeakyReLU (Xu, Wang, Chen & Li 2015) for non-linear activation.

3.3.4 Low-dimensional Embedding

In this work, the main purpose of the low-dimensional embedding is that the space distribution of low-dimensional hand pose features can be visualized. In this way, gesture interactions are more understandable and more controllable when interacting with a music space. Our work explores the impact of different data inputs on subsequent classification and visualization. Specifically, we acquired RGB color images, depth images, and keypoint coordinates for gesture keypoint detection using Intel® RealSense™ LiDAR Camera L515 depth camera. We configured different layers and cells of the encoder and decoder for the different inputs.

3.3.5 Gesture Classification

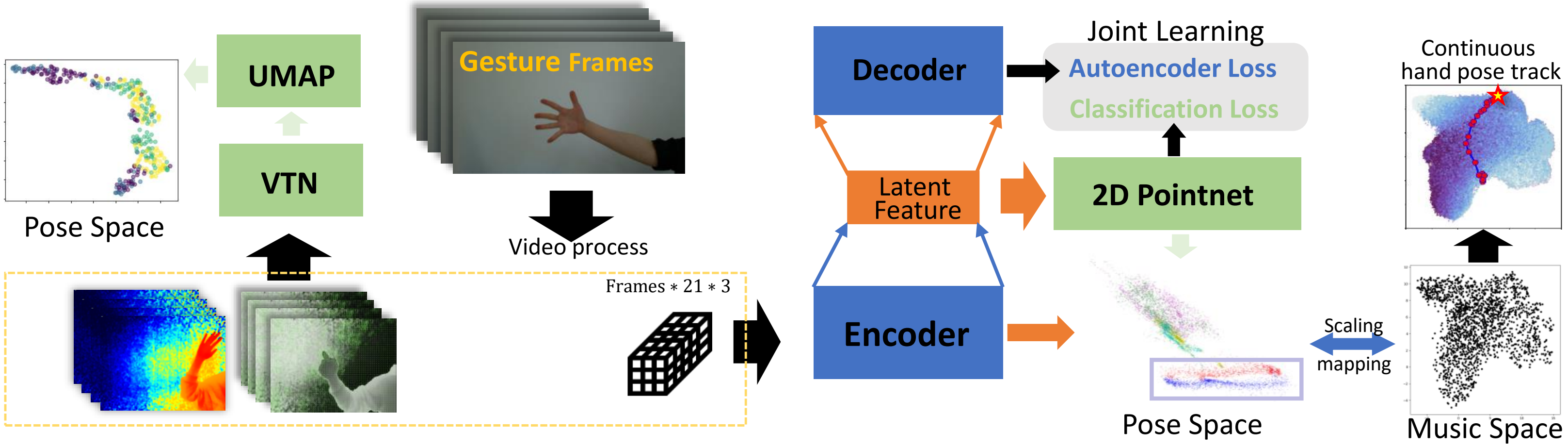


Figure 3.3: Pipeline of mid-air gesture control. The figure shows the dimensionality reduction process of UMAP and VTN for mid-air gestures as well as the dimensionality reduction process of our model. The inputs are image frames and the outputs are the 2D mid-air gesture space after dimensionality reduction. Further, we show an example of connecting the 2D gesture space of our model to the music space and performing a hand pose tracking.

In order to use different gestures to control different functions in gesture interaction, we need to implement a classification of gestures. Although our low-dimensional embedding model can achieve gesture clustering to some extent and has a gesture clustering visualization result, it still cannot classify gestures. Therefore, we decided to add an auxiliary classification model based on the low-dimensional embedding model. The main roles of this model are (1) to assist in training the low-dimensional embedding model, so that the low-dimensional embedding model can get more discriminative gesture clustering visualization results. (2) Classify the gestures so that the classification results can be used for different control and interaction operations.

Video Transformer Network for Depth Images and Optical Flow

In our study, we first focus on the Intel$^{\circledR}$ RealSense$^{\text{TM}}$ LiDAR Camera L515 depth camera. The depth image collected by this camera is an image with a resolution of 480×620. We process this image into a pseudo-RGB image, which is shown in the lower left corner of Figure 3.3. The RGB color image collected by this camera is a three-channel image with a resolution of 480×620. We applied the optical flow method to the original RGB color image and obtained the optical flow field and the corresponding optical flow matrix for the second image in the lower left corner of Figure 3.3.

In order to process the frames information, we applied the video classification model Video Transformer Network (VTN) (Neimark et al. 2021) which can extract temporal and spatial information in the transformer to classify the video frame sequence. In addition to outputting the gesture category during gesture classification, we also output the input of the fully connected classification layer as the encoding of each frame. In this embedding phase, the depth images are embedded into a 256-dimensional vector, while the optical flow is embedded into a 192-dimensional vector. In order to reduce the dimensionality of the frames into a two-dimensional vector that can be visualized, we directly applied UMAP (McInnes et al. 2018) to visualize the clustering of the VTN output. We used UMAP clustering with Euclidean distance, minimum neighbors of 20, and minimum distance of 0.5 units.

The Video Transformer Network (VTN) (Neimark et al. 2021) model is a novel model for video processing and classification. The application of splitting each frame into 16 small patches and transformer extracts both temporal and spatial information between video frames. For our task, due to the simple content and objects of our gesture depth images, too-deep transformer layers are not conducive to classification, so we chose to adjust the depth of the transformer layer to 2. To fit the model, we process the original depth image (480×620)

into a pseudo-RGB map with three channels ($3 \times 480 \times 620$) first and then resized the image to 224. The depth information provides 3D coordinates of the hand, rather than only 2D projections, which improves the robustness of gesture recognition. In this way, the VTN can learn both the spatial hand structure and the temporal dynamics across frames. On the other hand, we use OpenCV to process the depth image directly to get the optical flow matrix, which has a size of $2 \times 480 \times 620$. To apply VTN to the optical flow, we modify the 'in channel' of the spatial transformer of VTN to 2. For the VTN model, the loss function we choose is cross entropy.

To compare with the VTN model, we also try the ResNet-34 (He et al. 2016). The difference between the depth image and optical flow as input is that the convolution layer of the first layer of ResNet-34 is set to 3 and 2 respectively. The loss function we use is cross entropy.

PointNet for 2D Points

We employ a highly efficient and effective PointNet (Qi et al. 2017) that directly consumes point information, based on multiple linear layers, to classify the predefined gestures. PointNet-based classification network reduces the model size and training time by using low-dimensional inputs from the autoencoder, as well as assisting the pose space to obtain better distinguishability in the autoencoder by their jointly learning. Specifically, the input to our PointNet-based classification model is the output of the encoder, i.e., the latent embedding. In contrast to the original PointNet, the channel of the input of our classification model depends on the dimensions of the latent embedding, and not a fixed three dimensions. In our models in this chapter, we use channel=2, and in subsequent research, we will explore cases such as channel=4 and channel=5 to adapt to our research. Moreover, considering that our outputs are classification results for gestures with fewer classes, and in order to obtain results with higher accuracy, the output dimensions of each fully connected layer of our PointNet-based classification model are set to 512, 192, and 6. Furthermore, inspired by the popular sequence-based methods (Neimark et al. 2021, Ur Rehman et al. 2021), we also explore and take advantage of the sequence information from the gesture video. Specifically, the frame-based sequence features are encoded by the autoencoder as new inputs of the classification network. Then, we get the corresponding predicted gesture categories from the optimized classification network. In this work, we explore single-, 2- and 8-frame sequence inputs for the PointNet-based classification, and we chose a 2-frame sequence as our final inputs, based on trading-off classification performance with minimizing time latency.

Our goal is to jointly learn the parameters of the proposed fully-connected layer based autoencoder and PointNet-based classification by minimizing a loss function over the training set, which employs mean square loss (Sara et al. 2019) and cross-entropy loss (Zhang & Sabuncu 2018), respectively. We employ Adam (Kingma & Ba 2014) with 0.001 weight decay and 0.001 learning rate as the optimizer of our joint model.

3.3.6 Interaction

We propose a novel, user-friendly strategy for control interaction and exploration in a visible music space, where the continuous hand pose is used for continuous exploration.

To connect the pose space to the music space, we use a physical mapping, *i.e.*, first scaling the music space and the pose space to the same range and computing the cluster centres for each type of music and then computing the distance to the coordinates of the real-time pose in the two-dimensional pose space. The music category with the closest distance to the coordinates of the real-time pose will be highlighted.

We mark different emotions in different colours in the music space. The colours range from light to dark, indicating light to heavy emotional expressions of music. In this way, users have more freedom and controllability to explore the music space with an entire music database by the continuous mid-air hand pose movement, as shown in Figure 3.3.

3.4 Discussion

In this section, we report results on dimensionality reduction and classification of gestures over different models and data, and visualize and interact with gestures in air. We designed gestures according to the desired interaction function and performed gesture data collection. Based on this, we compare the results of depth image, optical flow and hand landmarks with different models for reducing dimension and classifying. Finally, we visualize the gesture data in the gesture space and interact with the music space.

We first present an empirical finding that highlights the effectiveness of a well-selected fully-connected autoencoder in the proposed pipeline, which will get low-dimensional pose spaces for interactions. Then the effectiveness of well-designed VTN-based, ResNet34-based and PointNet-based classifiers will be proved.

Table 3.2: Classification results for different models with depth images and optical flow

Model	Data	Frames	Classes	Precision (%)	mAP (%)	Time (ms)
VTN	Depth Image	4	6	34	29	130
	Depth Image	8	6	46	40	260
	Optical Flow	4	6	29	27	120
	Optical Flow	8	6	29	26	290
ResNet 34	Depth Image	4	6	29	26	120
	Depth Image	8	6	31	29	240
	Optical Flow	4	6	33	27	110
	Optical Flow	8	6	33	29	250
VTN+ VTN	Depth Image + Optical Flow	4	6	50	43	270
	Depth Image + Optical Flow	8	6	50	43	550
VTN+ ResNet 34	Depth Image + Optical Flow	4	6	45	40	200
	Depth Image + Optical Flow	8	6	55	49	480

3.4.1 Effects of Autoencoder

As shown in Figure 3.4, the 2D outputs of the different gestures from the encoder in the autoencoder are plotted in a 2D space on the display. Compared with the widely used UMAP (indicated by (a), (a') and (a")), which has significant effects on dimensionality reduction (McInnes et al. 2018) and clustering (Becht et al. 2018), the proposed fully-connected layer based autoencoder (indicated by (b), (b') and (b")) can better distinguish the distribution of different gestures with lower model complexity. This allows users to see the positions of different gestures in the pose space and the relationship between different gestures. When hand pose points in the pose space are sufficiently dispersed, subtle hand pose changes will be clearly tracked. Furthermore, we explore the distributional effects of using different frame sequences (2-frame and 8-frame) on the encoding of the proposed fully-connected autoencoder. By comparing the visualization results with the first column of Figure 3.4 with single-frame inputs, using multi-frame information can significantly improve inter-class gesture clustering and intra-class dispersion. As low latency is very important for user interaction, we focus on 2 frames in our subsequent study, which can avoid long time latency caused by longer frame sequences dependencies in other methods (Ur Rehman et al. 2021).

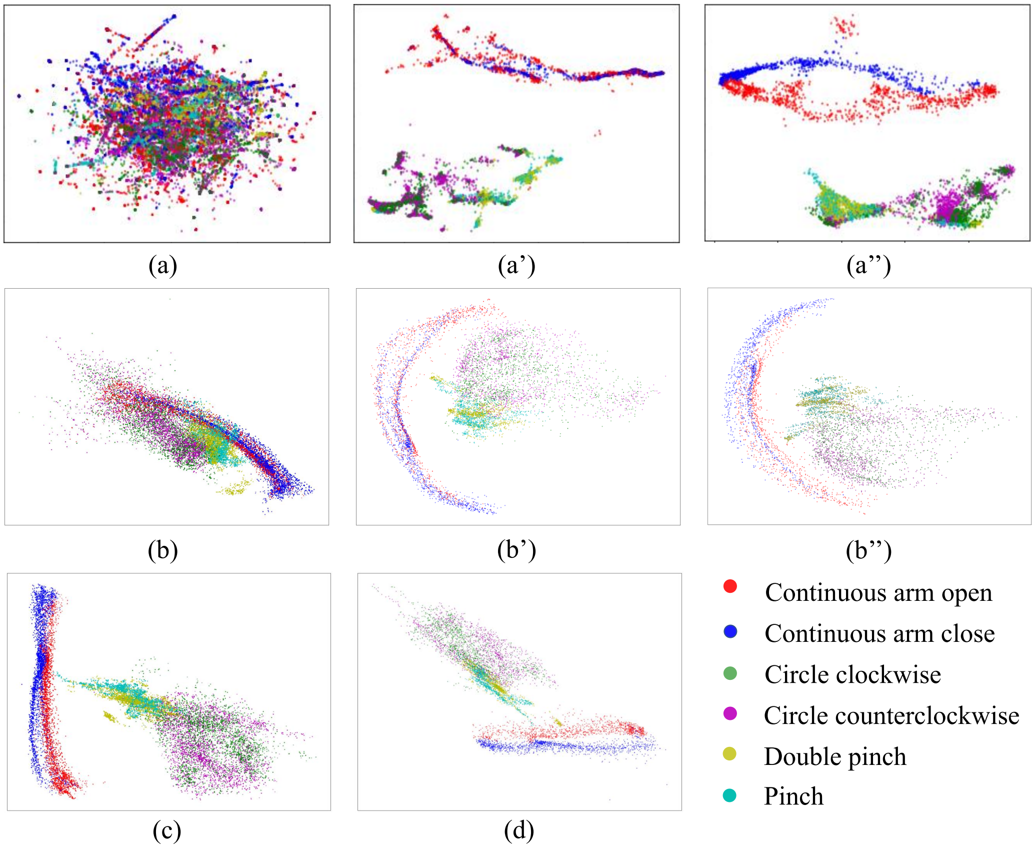


Figure 3.4: Visualization of different embedding methods of mid-air gestures. (a), (a’) and (a") use single-, 2- and 8-frame gesture sequence inputs of UMAP (McInnes et al. 2018), respectively; (b), (b’) and (b”) use single-, 2- and 8-frame gesture sequence inputs of our fully-connected autoencoder without classification; (c) and (d) are joint autoencoder and PointNet-based classifier on single- and 2-frame gesture sequences, respectively.

3.4.2 Effects of Classification

VTN- and ResNet-based classification

At the beginning of our experiments, we first consider transferring original video frames to depth images (480 × 620) and use the depth images with different frame sequence lengths as inputs to ResNet- and VTN-based models (He et al. 2016, Neimark et al. 2021) and outputs as gesture categories. We considered 4- and 8-based frame sequences, respectively. In the Table. 3.2, we give the gesture classification results. It contains the frame sequence classification accuracy and the category classification accuracy. In addition, we also calculate the time of the prediction of the model. From the Table 3.2, we can see that the use of depth images does not work well (the maximum precision is 46%) for gesture recognition and classification.

Based on this, we consider whether we can add dynamic information to the frame-sequence input. So we used the optical flow ($2 \times 480 \times 620$) of RGB frames as input to the model and applied it to ResNet- and VTN-based models. The results are also shown in Table. 3.2. We also tested combining the two data types and performing joint learning on the model, but the results were still not satisfactory, with the highest classification accuracy of 55% (using Depth image and Optical flow on VTN + ResNet34).

PointNet-based classification

Different from previous approaches, *e.g.* (Rusu et al. 2022), where only clustering is used for gesture reduction and visualization, we utilize a PointNet-based classification network to recognize the mid-air gestures for interactions and to further guide the visualization of the distribution of pose space. As shown in (c) and (d) of Figure 3.4, by joint learning with the classification network, the proposed autoencoder can better distinguish the space distribution of inter- and intra-class mid-air gestures. Although this increases the training time, there is no additional cost in the inference process of gesture feature dimensionality reduction and clustering. In addition, in Table 3.3, we provide the detailed classification results for different frame-based gestural sequence features from the autoencoder. We also provide inference time for the entire mid-air hand pose interaction process of each sequence. Compared to the 12.6 ms required by the original PointNet without autoencoder, the recognition and interactions with autoencoder take 2.3 ms, 2.4 ms and 3 ms on 1-, 2- and 8-frame based sequences, respectively. It demonstrates that the joint learning of the fully-connected autoencoder and PointNet-based classifier can improve the clustering effect of the autoencoder while also keeping the accuracy of the classification. Meanwhile, compared to using only the encoder output of Autoencoder for classification, the 2-frame-based PointNet improves the precision by 3.9% and the mAP by 6.3%. Finally, by balancing classification effectiveness and low latency requirements, we finally choose 2-frame sequences as inputs. Combining our classification results as well as the latency results, our PointNet-based model will be used as an auxiliary model in our interaction model to help the model accurately discriminate different gestures with a small latency.

3.4.3 Visualization of Latent Space

In order to provide a better experience and understanding when using the smart music player, we display the selected music position and use the dynamic hand pose continuous control process to map to locations in the music space. As shown in the top row of Figure 3.5,

Table 3.3: Comparisons of Classification models with hand landmarks

Method	Autoencoder	Frames	Precision(%)	mAP(%)	time (ms)
–	√	1	72.4	58.8	2.3
PointNet based	–	1	70.6	62.1	12.6
	√	1	71.3	61.9	2.3
	√	2	75.2	62.5	2.4
	√	8	77.2	64.9	3.0

when we specify different target music positions in music spaces, they can all be reached by continuous movement and exploration of mid-air hand pose. We provided the detailed music space information in Appendix A. For the Figure 3.5 (d), (e), we measure the time for an experienced user to reach the specified target point when exploring the pose space for two consecutive times, 4.4 s and 2.1 s, respectively. This demonstrates that users can learn to explore one point in the music space by continuous dynamic mid-air hand pose control to enhance their understanding of interaction with the music space, and thus reach the goals faster. The three different tracks of hand pose in the Figure 3.5 (f) shows that users can reach the same target music position with continuous control by different dynamic hand pose, and that the starting position and pose of the hand does not affect the exploration of the target music. Notably, the latency of the interaction process of a frame-based gesture sequence (2 frames) is about 2.4 ms, as shown in Table 3.3, including the inference of autoencoder and classifier and drawing.

3.4.4 Limitation

Method Limitation

While improving response speed and interaction accuracy, our current mid-air hand gestures recognition and interaction methods often struggle to provide stable and smooth user experiences when faced with diverse hand gestures and movements. The reliance on model-driven gesture embeddings, although effective for classification, does not fully account for variations in users' hand physiology or unpredictable motion patterns.

Additionally, many existing systems lack interpretability in gesture recognition and control processes, which limits users' understanding and undermines the system's user-learnability. This lack of real-time feedback and interpretability makes it difficult for users to grasp how their gestures are processed and recognized, leading to potential frustration and inefficiency

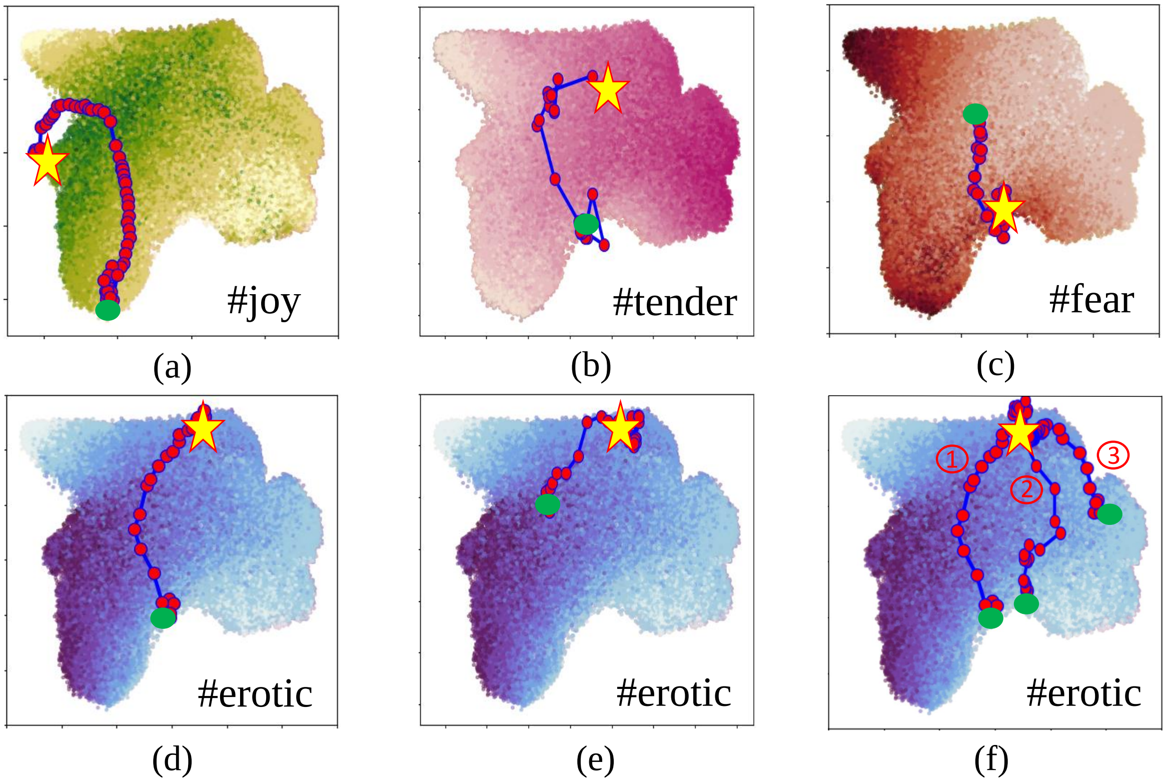


Figure 3.5: Visualization of hand-pose-controlled track selection in the music spaces of 55,000 music tracks. The yellow ☆ indicates the target music position point, the green ● indicates the hand pose starting point, and hand-pose-selected movement tracks are in blue. Different colours of points in the space represent music with different emotions, and the darker the colour, the higher the emotional value.

in practical applications. Moreover, without clear visual or sensory feedback, users may find it challenging to adjust their gestures accurately, reducing the overall usability and comfort of mid-air interactions. Addressing these limitations requires the development of models that not only classify gestures effectively but also provide meaningful, interpretable feedback that enhances users' understanding and control within the interaction system.

Study Limitation

This study has some clear limitations. The dataset was collected from only seven participants, which is a small number, so it cannot fully show the variation of hand poses in a larger population. All participants were young adults. The system was not tested with children, older adults, or people with different backgrounds. This reduces the diversity of the study and makes the results less general. And all gesture recordings were made with the right hand only. The system may work differently with the left hand or with users who use both hands in daily life. Because of these limits, the findings may not represent all users. Future work should include more participants, cover different age groups, and collect data from both hands. This would improve the reliability of the results and give a more complete evaluation of the hand pose interaction method.

3.5 Conclusion

In this chapter, we study the problem of mid-air hand pose control and visible interaction for a smart music speaker and propose a hand pose space encoding and visualization model by a fully-connected autoencoder joined with a PointNet-based gesture classification network. Specifically, a new mid-air gesture dataset is collected to train and evaluate the proposed mid-air gesture recognition and control method. The proposed autoencoder embeds gestures into low-dimensional spaces suitable for visualisation and interaction, which helps to unify the pose space with the music space for interactions. In addition, the auxiliary PointNet-based classification network further improves the clustering of gestures in the visualized pose space and maintains better classification performance (75.2%) than just using Autoencoder (72.4%). Moreover, the proposed interaction strategy requires only a few gesture frames (2-frame sequence) of input to get a continuous control that the user can explore, which helps to maintain a low interaction latency (2.4ms). The chapter also provides an exploratory demonstration of the ability to control and select different areas of the user space via continuous hand pose changes by experienced users. However, ensuring the stability and smoothness of hand movements within the low-dimensional gesture space remains a challenge that must be addressed to provide users with a more fluid and natural interactive experience. Additionally, enhancing the interpretability of the interaction can further reduce the learning curve for users and improve overall efficiency. Therefore, in the next chapter, we will focus on tackling these issues to refine and optimize the interaction process.

Chapter 4

Stable and Smooth Interaction System with Hand Pose Estimation

Mid-air gesture and hand-pose interaction provide an alternative to touchscreens, especially for diverse hand conditions. The previous chapter introduced a low-dimensional space for hand poses but faced challenges like jitter and instability. Many gesture methods rely on black-box models, limiting interpretability for visualization tasks. To address this, we propose a VAE-based Hand-pose Embedding Interactive System (HpEIS), mapping hand poses to a 2D visual space for guided exploration. Specifically, ***i)*** *we introduce improvements, including stability processing and smoothness processing.* ***ii)*** *we evaluate HpEIS based on target selection task completion time and final distance to the target point, comparing performance with (10s) and without (20s) the gesture guidance window condition.* ***iii)*** *experimental results and user feedback indicate that HpEIS enables a user learnable, movement stable and smooth mid-air hand movement interaction experience, which can be flexibly started from any position.*

4.1 Introduction

In this chapter, we focus on the more challenging topic of mid-air hand pose embedding and interaction. Challenges in adopting this potentially more intuitive and convenient approach to multimedia control are due to the complexity and challenges of hand-pose embedding. The complexity is mainly due to the inherent flexibility of human hands, resulting in a wide variety of complex postures. The flexibility of hand postures is a fundamentally crucial consideration in designing interactions. Unlike virtual reconstruction of hand poses directly using external handheld devices or radar, encoding hand poses and visualizing their positions in low-dimensional space can be challenging due to their low dimensionality but high complexity. Moreover, interacting in a visualized embedding space using hand-pose movement is subject to many uncontrolled factors, such as unavoidable physiological jitter.

We present a Hand-pose Embedding Interactive System (HpEIS) as a virtual sensor, which maps users' flexible hand poses to a two-dimensional visual space using a Variational Autoencoder (VAE) model trained on a large variety of hand poses. HpEIS enables visually interpretable and guidable support for user explorations in multimedia collections, using only a camera as an external hand pose acquisition device. We identify general usability issues associated with system stability and smoothing requirements through pilot experiments with expert and inexperienced users. We then design stability and smoothing improvements, including hand-pose data augmentation, an anti-jitter regularisation term added to the loss function, stabilising post-processing for movement turning points and smoothing post-processing based on One Euro Filters. In target selection experiments (n=12), we evaluate HpEIS by measures of task completion time and the final distance to the target point, with and without the gesture guidance window condition. Experimental results and questionnaire responses indicate that HpEIS provides users with a learnable, flexible, stable and smooth mid-air hand movement interaction experience.

1) We introduce a virtual sensor HpEIS, a user-learnable hand-pose embedding interactive system, to provide new inspiration for interacting with multiple multimedia collections with flexible hand movement. 2) We proposed an augmented VAE model to encode the hand pose into a visualized latent space for interaction, which contains an innovative augmentation strategy with an anti-jitter regularisation term based on a pilot experiment. 3) We introduced stabilization post-processing and smoothing post-processing to specifically deal with instability due to physiological jitter and system sensitivity. 4) We designed a real-time user guidance window based on the hand pose reconstruction in the hand pose interaction system. This greatly improves the user interaction experience. 5) Our user study experiments demonstrate substantial advancements in the user experience of our system in terms of stability and smoothness, while maintaining interaction flexibility.

4.2 Background

We will introduce the background of using mid-air-hand-pose interaction, such as hand pose categories and common devices, interaction flexibility and interpretability. Moreover, the related works of these tasks are also described in detail.

4.2.1 Interaction Flexibility

In existing research on hand-pose interaction, various sensors have been applied to different interaction tasks (Hayashi et al. 2021, Liu et al. 2021, Qin et al. 2021, Wilhelm et al. 2020). Thus, different interaction methods are used to meet the needs of people interacting in different scenarios. However, as mentioned in (Vogiatzidakis & Koutsabasis 2022, Xu et al. 2023), among these typical interaction methods (Groenewald et al. 2016, Kong et al. 2020, Nacenta et al. 2013), the dominant one is still the static hand-pose interaction. This means that the user needs to give a specified static and non-continuous hand pose and then the system gives the corresponding feedback after acquiring and recognizing the hand representation. These studies were widely used in multimedia single-control scenarios, such as volume and top/bottom selection for multimedia music. It leads to less flexibility in the system interaction. In addition, this kind of static hand-pose control usually uses black-box operation, i.e., direct feedback without interpretable interaction.

4.2.2 Interaction Interpretability

As artificial intelligence (AI) models become more powerful, the user interaction requirements also become more complex (Silva et al. 2019). While users do not need to understand exactly how the model works internally, there is a growing demand for users to seek explainable interactions with intelligent systems. It will help less experienced users understand and use AI systems faster (Došilović et al. 2018, Taka et al. 2022), allowing them to experience the rapid advances in AI more quickly and more realistically. Therefore, it is especially important to break the black box between the user and the system (Došilović et al. 2018), increase the user's trust in the AI system, and improve the transparency and interpretability of AI (XAI) (Banovic et al. 2023, Gilpin et al. 2018). On one hand, to improve the XAI, researchers proposed new XAI tools (Sharma et al. 2019), e.g., head-mounted VR/AR glasses (Du et al. 2022, Masurovsky et al. 2020, Tseng et al. 2023) and handheld sensing devices., to reconstruct virtual hands in virtual environments for users to understand virtual world interactions. However, expensive and uncomfortable bulky equipment limits the access of many users. On the other hand, reducing high-dimensional representations and interacting on their visualized low-dimensional latent space has the potential to be an intuitive and interpretable interaction strategy (Dang & Buschek 2021, Rusu et al. 2022, Strachan et al. 2007, Xu et al. 2023). For instance, (Xu et al. 2023) introduced a fixed gesture-based music space exploration scheme that enforced a physical mapping of the hand pose latent space to a music latent space. This is an interpretable study of gesture interaction but reduces the flexibility of the interaction due to their requirements for fixation and recognition of gesture categories. Additionally, the additional mapping requirement of hand-embedding space and multimedia latent space also reduces the migratability of their interaction systems.

4.2.3 Quaternion

Quaternion-based representations are frequently applied in 3D transformations within human-computer interaction (HCI), where accurately and stably tracking hand poses in 3D space is vital. Unlike conventional rotation representation methods, such as Euler angles or rotation matrices, quaternions can effectively mitigate issues related to gimbal lock. Gimbal lock occurs when two of the three rotational axes align, leading to a loss of one degree of freedom and resulting in erratic behavior during complex movements. This issue can disrupt the continuity and smoothness of hand pose tracking, a critical requirement for mid-air gesture interactions that need to operate seamlessly across a range of orientations (Kuipers 1999). A quaternion q is defined as

$$q = w + xi + yj + zk, \tag{4.1}$$

where w, x, y, z are real numbers, and

$$i^2 = j^2 = k^2 = ijk = -1. \tag{4.2}$$

This four-dimensional representation encapsulates rotation in a way that provides continuous, singularity-free results, which is crucial for maintaining consistent tracking during dynamic and multi-axis movements.

In the context of mid-air hand gesture interactions, maintaining consistency despite variations in hand size, distance from the camera, and orientation changes remains a key challenge (Xu et al. 2023). Recent research has explored the use of quaternions to create more resilient systems that can compensate for these variations (Arsenault 2014, Elouariachi et al. 2020, Patil et al. 2019). By encoding hand rotations in quaternion space, interaction models can avoid the pitfalls of orientation-based noise and misalignment, ensuring that gestures are accurately interpreted regardless of user-specific or environmental factors. This characteristic is particularly useful for enabling gesture control in variable lighting conditions and diverse backgrounds, where traditional 2D or non-robust 3D representations may fail (Xu et al. 2023).

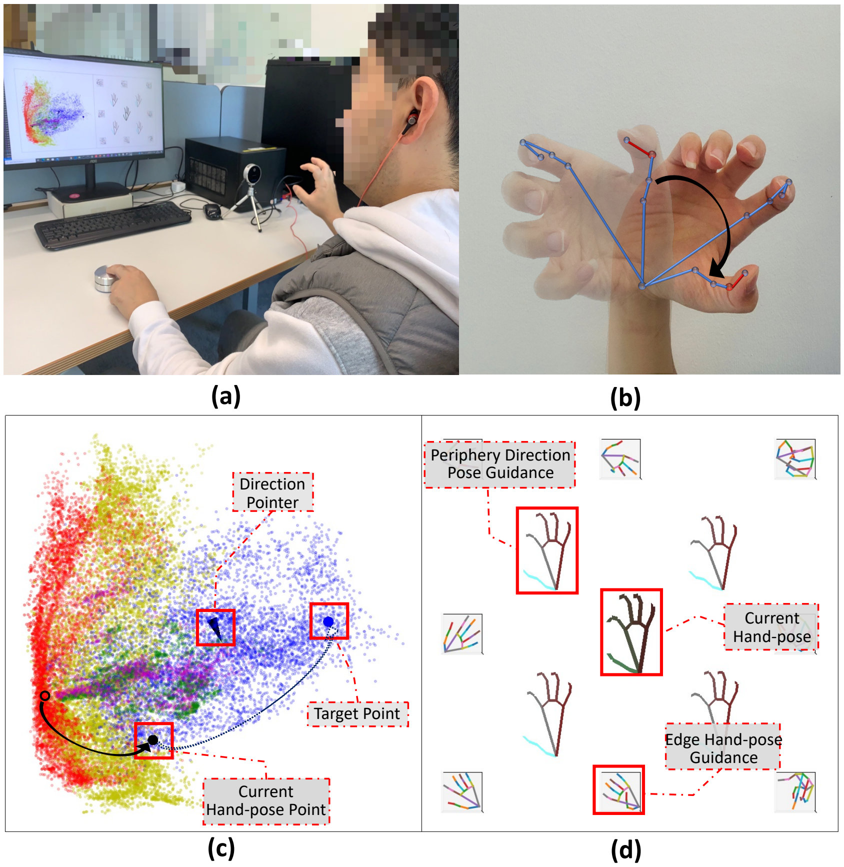


Figure 4.1: Example scenes of our designed **H**and-**p**ose **E**mbedding **I**nteractive **S**ystem (termed **HpEIS**). (a) A user is exploring the hand-pose embedding space and interactively finding target points that can be given new multimedia meanings. (b) Partial mid-air hand-pose movement. (c) The hand-pose embedding space visualization. (d) User Guidance window with hand-pose reconstruction.

4.3 Overview of System

The pipeline of our model is in Figure. 4.2. We will give preliminaries of our work in Section 4.3.1 and a detailed introduction to obtaining the mid-air hand-pose embedding space in Section 4.3.2. And then, the design of visualization guidance in the system, and the design of interaction flexibility, stability and smoothness will be introduced in Section 4.3.3, Section 4.3.4, Section 4.3.5 and Section 4.3.6, respectively.

4.3.1 Data Processing

As a new multimedia adapter, our mid-air hand-pose movement interaction system aims to control multimedia through a flexible-stable-smooth hand pose interaction strategy in a visu-

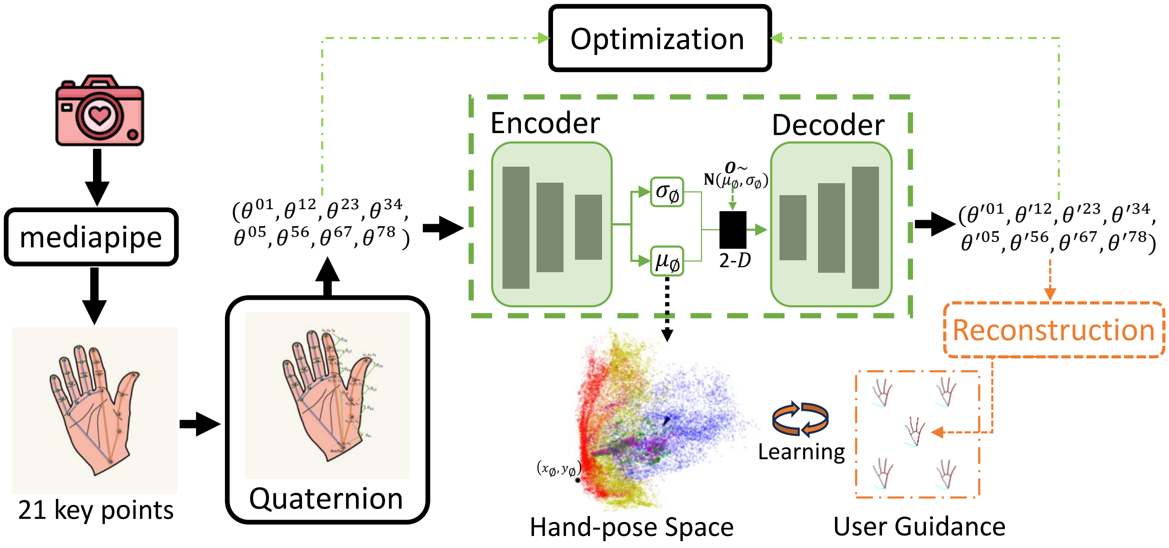
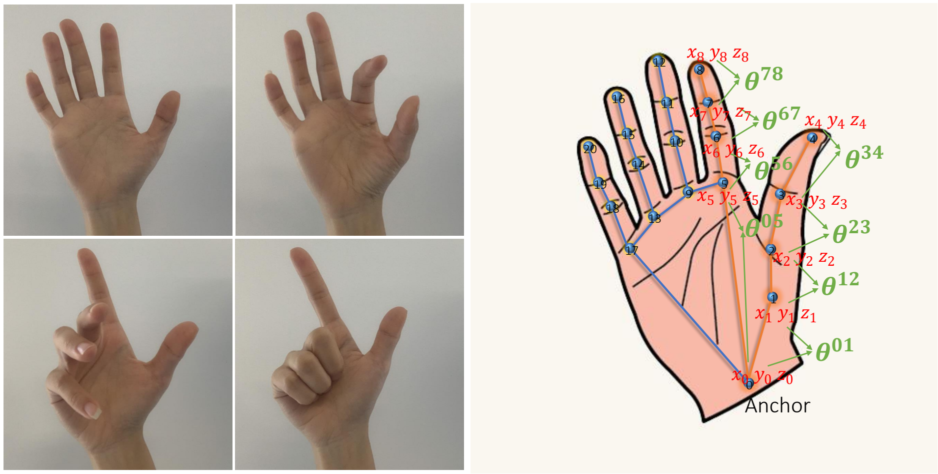

Optimization
Encoder
Decoder
mediapipe
$(\theta^{01},\theta^{12},\theta^{23},\theta^{34},$
$\theta^{05},\theta^{56},\theta^{67},\theta^{78})$
$\sigma_{\emptyset}$
$\mu_{\emptyset}$
$O\sim N(\mu_{\emptyset},\sigma_{\emptyset})$
2-D
$(\theta'^{01},\theta'^{12},\theta'^{23},\theta'^{34},$
$\theta'^{05},\theta'^{56},\theta'^{67},\theta'^{78})$
Reconstruction
21 key points
Quaternion
$(x_{\emptyset},y_{\emptyset})$
Learning
Hand-pose Space
User Guidance

Figure 4.2: The pipeline of our mid-air hand-pose embedding space construction and user guidance reconstruction.

alized embedding space, which can be mapped with any multimedia space. In this work, as shown in Figure 4.3 (a), we use the thumb and index finger to form a hand pose for interaction and consider the relative angle information of both fingers to the wrist simultaneously. We argue that simple hand poses are easier to detect and embed for the model and speed up the interaction for the system. In addition, simple and straightforward hand poses make them easier to remember and accomplish. Note that, different poses of the remaining fingers do not affect the mid-air hand-pose definition, thus maintaining a certain degree of hand flexibility and pose robustness. To get the hand-pose embedding space, we use an external camera device to capture the hand inputs $\mathscr{G} = \{g_1, \dots, g_n\}$, where n is the length of hand sequence of each user, and employ MediaPipe (Lugaresi et al. 2019) to detect the 21 key points $\mathscr{P} = \{p_1, \dots, p_{21}\}$ for each hand frame, where each point has three-dimensional coordinates $p_i = (\hat{x}_i, \hat{y}_i, \hat{z}_i)$. We find that different experimental users with different hand sizes at different distances from the camera device usually produced different interaction results using mid-air hand poses with identical meanings in the embedding space. To address this issue, we employ quaternion (Rieger & Van Vliet 2004) to transfer the hand key landmarks, avoiding the effect of hand size, hand position in the sensor field and distance from the sensor. Specifically, as shown in Figure 4.3 (b), we choose the 9 landmarks from the thumb and index finger, containing an anchor point at the wrist and 8 landmarks at the joints of two fingers, to calculate the quaternion number of adjacent landmarks. After transferring, we obtain new quaternion-based rotation angle hand-pose movement representations $\mathscr{Q} = \{q_1, \dots, q_n\}$



(a) (b)

Figure 4.3: Example of MediaPipe extraction and quaternion conversion. (a) means that we only consider the thumb and index fingers and their angles with the wrist, and the posture of other fingers does not affect the hand-pose representation. (b) means the quaternion conversion for 9 detected landmarks.

based on the detected landmarks, where $q_i = [\theta_i^{01}, \theta_i^{12}, \theta_i^{23}, \theta_i^{34}, \theta_i^{05}, \theta_i^{56}, \theta_i^{67}, \theta_i^{78}]$ for each hand frame. Afterwards, we use Variational Autoencoder (VAE) (Kingma & Welling 2014) to encode and decode all hand pose representations and use the intermediate 2-dimensional latent representations as coordinates to compose the embedding space $\mathscr{S} = \{s_1, \ldots, s_m\}$ for interaction and visualization, where $s_i = (x_i, y_i)$. Finally, we can obtain a 2-dimensional hand-pose embedding space $\mathscr{S}$, which allows users to explore in all directions with different hand postures. Once it is assigned to a specific multimedia space, such as music space and book space, then we can pair free or purposeful explorations of the corresponding multimedia space. Additionally, we provide new users with visualized guidelines based on the decoder in trained VAE for hand pose movement, while also reducing user discomfort once the embedding space is updated.

4.3.2 Dispersed Latent Space with Continuity Movement

Figure 4.2 shows the pipeline of our mid-air hand-pose embedding space construction. For each hand frame g_i, we use a Variational Autoencoder (VAE) (Kingma & Welling 2014) to obtain the latent 2-D embedding as the hand pose coordinate $s_i = (x_i, y_i)$, which conforms to a learnable Gaussian distribution with mean μ and standard deviation σ, with the quaternion-based rotation angle representation $q_i \in \mathrm{R}^8$ as input. It contains a Multi-Layer Perceptron (MLP) based encoder and an MLP-based decoder. Specifically, each MLP contains 4 fully connected layers, where the neuron numbers in the encoder are 128, 96, 64 and 2, respectively and the reverse in the decoder. The 2-D latent embeddings are normalized between 0 and 1.

Our goal is to make the sample representations before encoding and after decoding as similar as possible by optimization, and the latent space used for decoding obeys the Gaussian distribution. In other words, the Mean Squared Error (MSE) between input $q_i = [\theta_i^{01}, \theta_i^{12}, \theta_i^{23}, \theta_i^{34}, \theta_i^{05}, \theta_i^{56}, \theta_i^{67}, \theta_i^{78}]$ and output $q_i' = [\theta_i'^{01}, \theta_i'^{12}, \theta_i'^{23}, \theta_i'^{34}, \theta_i'^{05}, \theta_i'^{56}, \theta_i'^{67}, \theta_i'^{78}]$ and the Kullback Leibler(KL) divergence of the distribution of low-dimensional data and the standard normal distribution are as small as possible, which are as follows:

$$\mathscr{L}_{MSE} = \frac{1}{8}\sum_{j=1}^{8}\left(\theta_i^{(j)} - \theta_i^{'(j)}\right)^2, \tag{4.3}$$

$$\begin{aligned}\mathscr{L}_{KL}\left(P_\phi(O|q_i)||P\varphi(O)\right) &= \mathscr{L}_{KL}\big(\mathscr{N}(\mu_\phi, \sigma_\phi), \mathscr{N}(0,1)\big)\\ &= -\frac{1}{2}\sum_{j=1}^{2}\left(1+\log\left(\left(\sigma_\phi^{(j)}\right)^2\right) - \left(\sigma_\phi^{(j)}\right)^2 - \left(\mu_\phi^{(j)}\right)^2\right),\end{aligned} \tag{4.4}$$

where P_ϕ is a binary independent Gaussian distribution with learnable mean μ_i and standard deviation σ_i, P_φ is a binary standard Gaussian distribution $\mathscr{N}(0,1)$. O is a randomly re-sampled 2-D latent feature from P_ϕ to decode. Finally, we use binary mean μ_ϕ as the coordinate (x_i, y_i) to present the hand-pose position in the embedding space.

4.3.3 User Guidance Window

To bring users, especially novice users, a learnable exploration and interaction experience, we provide an innovative method called user guidance window in Figure 4.2, which uses a trained decoder to obtain hand-pose decoding representations in four directions around the current hand pose and inversely uses quaternion to reconstruct the image expression of these neighbour hand poses. Specifically, given the current mid-air hand pose g_i, we use the trained encoder to obtain the means (x_μ, y_μ) of latent representation as the current position in the hand-pose space. Then, we sample 20 latent representations with the same distribution as the current hand pose as neighbour hand poses, and take 4 latent representations in four directions as the final hand-pose guidance inputs to the trained decoder for decoding and reconstruction. During the reconstruction process, we invert the quaternion and recalculate the positional relationship of the thumb, index finger and wrist, according to the decoded quaternion-based neighbour hand-pose representations $\{q_1', q_2', q_3', q_4'\}$. In addition, as shown in Figure 4.1 (d), we selected eight orientations at the edge of the entire space for hand posture reconstruction as fixed orientation guidance. Note that, since we focus on reconstructing the thumb and index finger, the other finger postures are the same as the current gesture. The main purpose of our user guidance window is to provide visual guidance for novice users when exploring a new embedding space, which can greatly reduce the unfamiliarity of the space. Alternatively, the user guidance window can be closed once the user has mastered the hand-pose habit of exploration.

4.3.4 Design for Flexibility

We believe that system flexibility is crucial to user interaction experience. Different from the current gesture interaction systems (Rusu et al. 2022, Xu et al. 2023), we mainly consider the flexibility design in two aspects:

(1) Data flexibility: Fixed hand-pose data is often limited and their spatial distribution is often local due to hand-pose similarity, which will lead to many other irregular hand poses not being positioned or positioned incorrectly. To this end, our HpEIS considers embedding space global exploration. In order to generate the entire hand-pose embedding space, we train our VAE using arbitrary hand poses, which contain specified hand poses, but also more generic movement, which ensures the distributional diversity of the space.

(2) Interaction flexibility: On one hand, we consider interacting with only two fingers, while other fingers do not affect the hand-pose embedding space. As mentioned in (Lang & Schieber 2004), mechanical coupling limits the independence of the index, middle, and ring fingers to the greatest degree and neuromuscular control primarily limits the independence of the ring, and little fingers during large-arc movements. Only considering the thumb and index finger reduces the influence of the non-independence of other fingers, allowing for hand flexibility during user hand movements. On the other hand, due to the inclusion of arbitrary mid-air hand poses in the training data, our system allows users to take any mid-air hand pose to start exploring the embedding space. Thus, this makes our system interaction flexible.

4.3.5 Design for Stability

As a user system, stability is a crucial part of enhancing the user experience. To improve the interactive stability of the VAE-based hand-pose embedding space, we first conduct pilot experiments (details in Section 4.4.3) on the current interactive system. However, we found three intuitive instability issues during our pilot experiments, which are (1) different hand sizes and the distance from hand to camera will influence the embedding place in the latent embedding space, (2) hand-pose embedding space is overly sensitive to subtle hand changes caused by physiological jitter and (3) space track exhibits jump due to error detection of hand key landmarks from intrinsic defects of the hand detector.

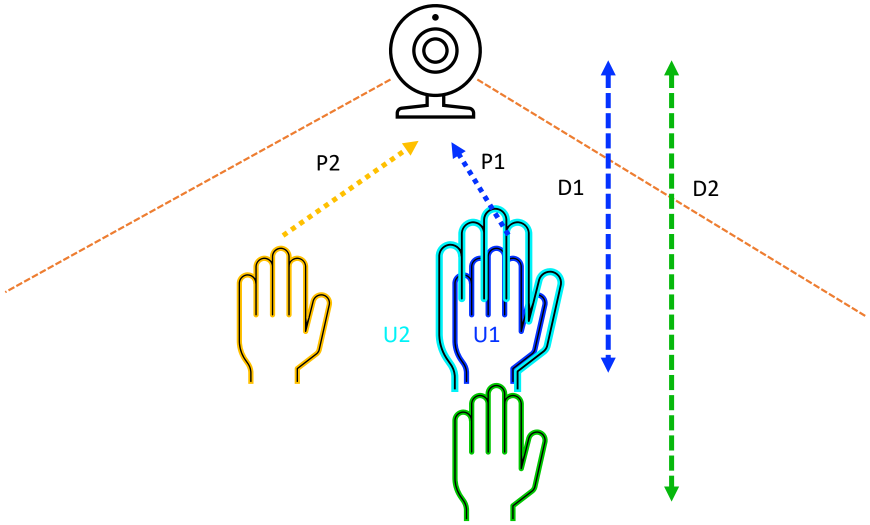


Figure 4.4: Quaternion experimental set up. U1 and U2 are different users and represent different hand sizes. P1 and P2 represent different startingpositions for hand movement. D1 and D2 represent different distances to the camera.

Quaternion Design

Compared to simple 3D coordinates, relative coordinates (Feng et al. 2003) and quaternion (Shoemake 1985) can well avoid the effects of different hand sizes, different distances from the camera, and different positions in the screen. Specifically, relative coordinates and quaternion arrays are only related to the selection of anchor (Feng et al. 2003, Shoemake 1985), so they can reduce the bias caused by distance and position. In contrast to relative coordinates, the quaternion array only cares about the rotation angle between different points and anchor points, which avoids the position deviation of the same gesture in the gesture space caused by the hand size (Hsu et al. 2018, Shoemake 1985). Therefore when we want to have a more general system, using the quaternion array will give more stable results than relative coordinates. Fig. 4.4 shows an illustration of our different setup for experiments. Where U1 and U2 are different users and represent different hand sizes. P1 and P2 represent different starting positions for hand movement. D1 and D2 represent different distances to the camera.

Model Design

To address the mentioned issues more effectively with the algorithm, inspired by (Sinha & Dieng 2021), as shown in Figure 4.5, we design a new data augmentation strategy based on pilot experiments on the trained VAE to mitigate the position oscillations in embedding space

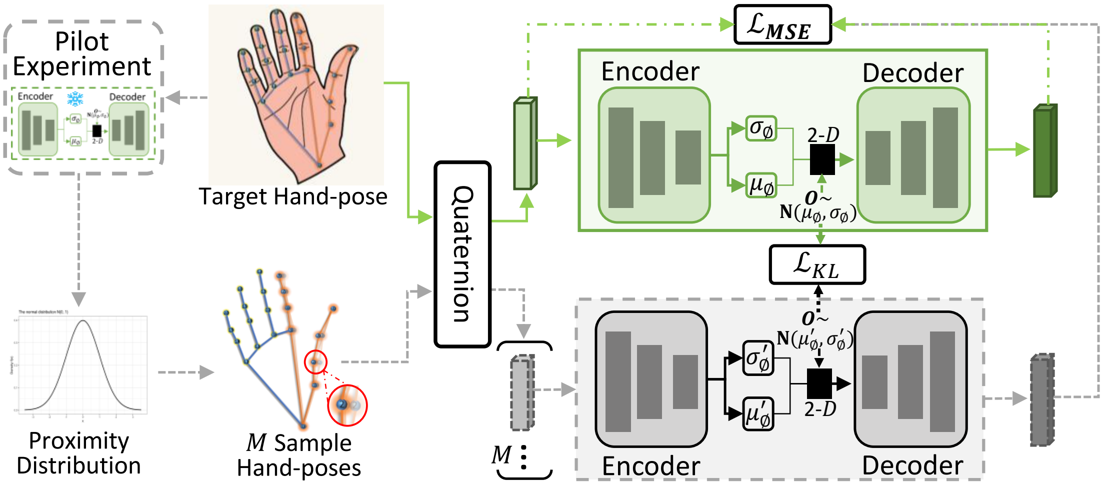


Figure 4.5: The data augmentation strategy and the VAE re-training process with stability constraints.

due to hand jitter and add a new optimization objective function for this purpose to refine and retrain the VAE. Specifically, we first trained a pilot model on the collected hand poses, then we designed a pilot experiment (details in Section 4.4.3) to further collect the real-interaction data, which contains 10 hand poses closest to each target hand-pose location during the pilot experiments as the stable proximity hand poses. After that, we can get the proximity distribution for each landmark of the target hand pose based on the collected proximity hand poses and then we sample M (M=100) new proximity hand poses as the augmented hand poses, where the mean is from the target feature and the standard deviation is averaged from proximity hand poses. Note that, because we only consider two fingers, we only have 9 landmarks. Next, we use quaternion to convert the target and augmented hand-poses landmarks and input them into the new VAE. Our goal is to overcome the system sensitivity caused by slight movement or jitter by optimizing the minimum distance between the jittered hand poses and the target hand pose and the space distribution of both. The objective functions between the target hand-pose representation q_i and each augmented representation q_m can be described as:

$$\mathscr{L}_{MSE} = \frac{1}{8}\sum_{j=1}^{8}\left(\theta_i^{(j)} - \theta_m^{(j)}\right)^2 \tag{4.5}$$

$$\mathscr{L}_{KL}\left(P_{\phi'}(O'|q_m)||P_{\phi}(O|q_i)\right) = -\frac{1}{2}\sum_{j=1}^{2}\left(1 - \log\left(\frac{\sigma_{\phi}^{(j)}}{\sigma_{\phi'}^{(j)}}\right)^2 - \frac{\left(\sigma_{\phi'}^{(j)}\right)^2 + \left(\mu_{\phi'}^{(j)} - \mu_{\phi}^{(j)}\right)^2}{\left(\sigma_{\phi}^{(j)}\right)^2}\right). \tag{4.6}$$

Where the $\mathscr{L}_{KL}$ is similar to (Sinha & Dieng 2021) for minimizing the difference in distribution between training hand poses and proximity hand poses. Additionally, the new $\mathscr{L}_{MSE}$ between the proximity hand poses and the target hand pose serves as a regularization term to further limit the difference between the jittered hand pose and the target hand pose in reality. Notably, we finally minimize the objective functions containing the loss computation of VAE itself.

Post-processing Design

In order to solve the error detection issue caused by MediaPipe, for example, inaccurate landmark detection occurs when the finger is perpendicular to the camera (last image in the bottom row of Figure 3.1) or when the finger is obscured (second image in the bottom row of Figure 3.1), we design a post-processing measure to further improve the stability of system interaction. Here we consider the Euclidean distance change of the coordinates in the embedding space to be greater than 0.1 as the mutation coordinates, i.e. unstable phenomenon, and there may exist error detection. We neutralize the unstable point s_i with the following operation:

$$s_i = \begin{cases} \frac{1}{2}(s_{i-1} + s_i), & \text{if} \sqrt{(s_{i-1} - s_i)^2} > 0.1 \\ s_i, & \text{if} \sqrt{(s_{i-1} - s_i)^2} < 0.1. \end{cases} \tag{4.7}$$

4.3.6 Design for Smoothness

During the pilot experiments, the smoothness of system interaction is required to be improved. The main problem is that there are jagged changes in the interaction path during the user's interactive exploration of the embedding space. Additionally, when users approach the target point, the decrease in movement speed leads to more pronounced space path non-smoothness around the target point due to hand jitter. To further improve the smooth experience during user interaction, we follow One Euro Filter (Casiez et al. 2012) to balance denoising and preservation of path dynamics. One Euro Filter is an effective method to filter out noise while maintaining important features of the signal with two main parameters, cutoff frequency CF and slope threshold ST. A higher cutoff frequency will reduce the response of the filter and is more suitable for processing fast-changing signals. A lower slope threshold will increase the intensity of the filter to suppress the noise and is more suitable for processing low-changing signals.

In our system, we use One Euro Filters with different parameters, i.e. cutoff frequency and slope threshold, to handle different hand-pose interaction scenarios separately. Specifically, we divide user interaction scenarios into two types, fast-moving interaction and slow-moving interaction. The former tends to occur during the initial stage of the interaction, where the user's hand moves faster, so the hand moves a greater distance between frames. Conversely, the latter tends to occur at the end of the interaction near the target point, where the user's hand moves slowly and the movement distance is small. Furthermore, we argue that users value the speed of exploring to the target point at the initial stage of interaction, but value the accuracy of reaching the target point at the end of the interaction stage. To this end, according to the distance between two neighbouring hand-pose points in embedding space, we choose two different group parameters for One Euro Filters, as follows:

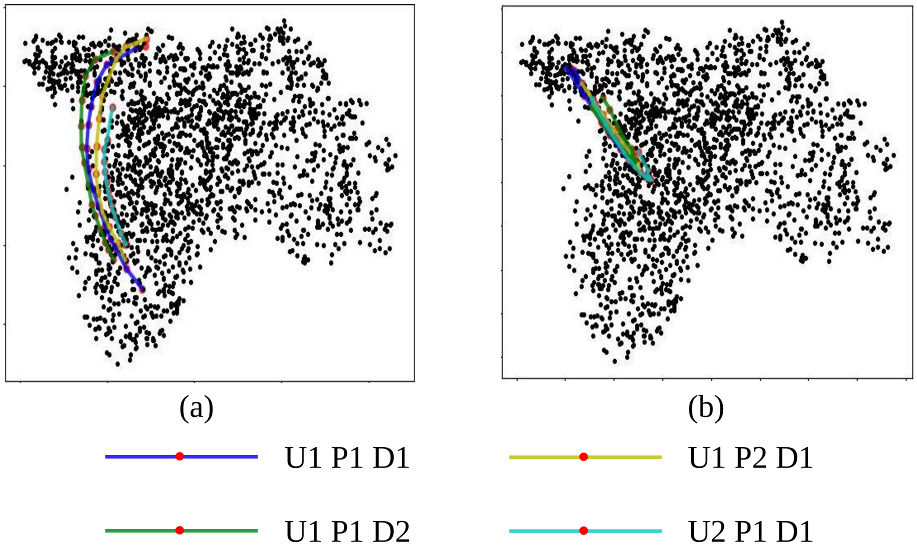
$$(CF, ST) = \begin{cases} (0.04, 0.85), & \text{if } \frac{1}{9}\sum_{j=1}^{9} L > 0.015 \\ (0.005, 0.75), & \text{if } \frac{1}{9}\sum_{j=1}^{9} L < 0.015, \end{cases} \tag{4.8}$$

$$where \;\; L = \sqrt{(\hat{x}_{i-1}^{(j)} - \hat{x}_i^{(j)})^2 + (\hat{y}_{i-1}^{(j)} - \hat{y}_i^{(j)})^2 + (\hat{z}_{i-1}^{(j)} - \hat{z}_i^{(j)})^2}. \tag{4.9}$$

Note that, due to the consideration that real-world hand movements are less sensitive than corresponding movements in embedding space, we use the movement distance between two neighbouring hand poses, g_{i-1} and g_i, calculated from 9 landmarks of two fingers in the real world as the parameter setting condition.

4.4 Experiments

In this section, we will first introduce our comparative results based on the 3D spatial coordinates and the transformation of quaternions. On this basis, we will elaborate on our pilot experimental design for the interactive system and its results, thereby refining the research designs and constructing the final HpEIS system. Finally, we will validate our HpEIS system and the effectiveness of the user guidance window through a user study.


(a)

(b)

Figure 4.6: The figure (a) shows the low-dimensional embedding of ’continuous arm open’ when quaternion is not used, and the figure (b) shows the low-dimensional embedding of ’continuous arm open’ when quaternion is used. In contrast to figure (a), the use of quaternions in figure (b) brings more stability to the model, because the trajectories are more dispersed in a) than b). For greater visibility, the background Music Space using a subset (2,000 music tracks) of the music dataset is shown in black.

4.4.1 Quaternion Test

During our designing process, we found that different experienced users with different distances from the camera usually produce different control results on the interaction with the music space for the same hand pose. To have a more stable interaction with the music space, we explore the use of quaternions (Saxena et al. 2009) to make the system invariant to hand size, hand position in the camera field, and distance from the camera on gesture embedding.

Fig. 4.6 compares two experienced users (indicated U1 and U2) of different palm sizes (palm width and palm length) without and with quaternion conversion (as (a) and (b)), respectively, to control the music space at different locations (start position and distance) from the camera using the same gesture. Specifically, the palm width and length for U1 are 7 cm and 15 cm respectively, and for U2 are 10 cm and 17 cm, where palm width is the distance from the widest part of the palm and palm length is the distance from the root of the palm to the tip of the middle finger. P1 and P2 indicate different starting positions, with a difference of 20 cm. D1 and D2 indicate different distances from the camera, which are 45 cm and 100 cm, respectively. Compared with (a) and (b) in Fig. 4.6, it suggests that the use of quaternion effectively reduces the effects of hand size, hand position and distance from the camera on the low-dimensional embedding of the gestures, thus allowing the interactive system to work

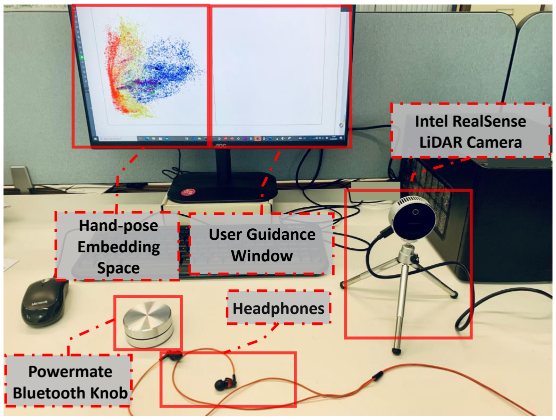


Figure 4.7: Detailed display of experimental equipment.

more stably.

4.4.2 Experimental Setup

We introduce pilot experiments with two users, one expert user and the other a less experienced user. 12 participants (7 men and 5 women) aged between 22 and 38 took part in our User Study. Additionally, 12 participants were divided equally into two groups. The first group completed the experiment without a user guidance window, and the other group was guided with a user guidance window.

We conduct pilot experiments to identify problems, refine research designs and assess feasibility. Specifically, the pilot experiments contain 4 experiments, which are *the baseline experiment* in Section 4.4.3, *the augmentation experiment* in Section 4.4.3, *the stability experiment* in Section 4.4.3, and *the smoothness experiment* in Section 4.4.3. The User Study in Section 4.4.4 is for assessing our interaction system and interface. In all experiments, as shown in Figure 4.7, the camera used to capture hands was Intel® RealSense™ LiDAR Camera L515 camera (frame rate is 30 fps). The hand movement plane was specified to be within 30-60 centimetres from the camera. The hand starting position and starting pose were not fixed. In addition, when the user finds the target point, a Bluetooth knob (Griffin PowerMate) was used as a labelling tool and then headphones produce audio feedback.

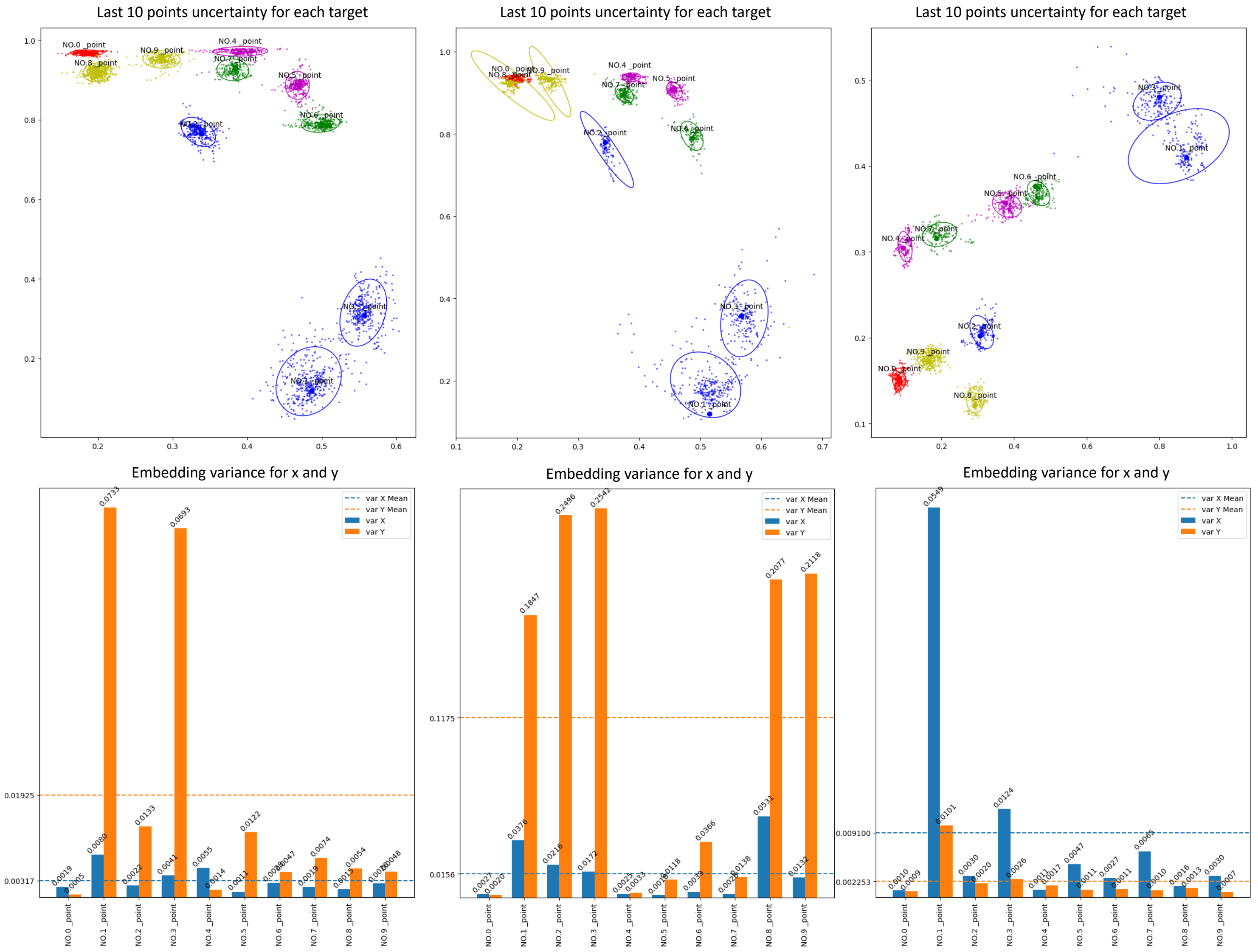
4.4.3 Pilot Experiment

Tasks

Two experimenters, including an expert user and an inexperienced user, conducted an exploration of the hand-pose embedding space. Our goal was to verify whether the user could find the target point in the 2D embedding space and to detect the interaction problems in the process of finding the point. Firstly, we identify 10 randomly selected points to avoid hand-pose inertia in the trained embedding space as the target points for pilot users. The pilot users were presented with a black circular pointer in the embedding space representing the current hand-pose position in the space. Target points for different hand poses are marked by different coloured circles, and different coloured points encoded in the embedding space mean the collected hand-pose embedding points. Then, we asked pilot users to press and hold the knob when they intuitively thought they had found a point to mark the point found, and the system would display the next target point. We evaluate the interactive usability of our system by the distance (between the final point and the goal point) and the time taken (between the start and the stop) when users thought they had reached the target point within a limited time. After the pilot experiments, we computed and showed the uncertainty of the last ten frames of each attempt around different target points.

Baseline Experiment

As mentioned in Section 4.3.2, we trained our VAE on the collected hand poses as the pilot model. Figure 4.8 (a) showed the results of the last ten frames to target points in the embedding space from the basic VAE and Figure 4.10 (a) showed the distance of each hand pose from the target points and the time-consuming. We found two obvious problems: (i) It is relatively easy for the users to approach the area of target points, but it is harder to anchor the hand pose to the target point or an acceptable target point, as shown in Figure 4.8 (a); (ii) The instability and non-smoothness of the user's interaction, i.e. the oscillation of the distance from the target points in Figure. 4.10 (a). As shown in Figure 4.9, by directly analyzing the corresponding detected hand landmarks, we argue that both problems are mostly caused by physiological jitter and system sensitivity. Specifically, we randomly chose a target point in the embedding space and then took 10 frames after reaching the target point where the hand stopped moving. We define these 10 frames as hand-stabilized, but it is clear from the coordinates of two landmarks at the top of Figure 4.9 that the last 10 frames also have physiological jitter. Moreover, by combining the uncertainty variance of the 3D coordinates of the hand landmarks before encoding (at the end of Figure 4.9) and the uncertainty variance of the 2D coordinates after encoding (at the end of Figure 4.8), we can clearly see that the variance in the system is larger, thus indicating that there are system sensitivities that lead to interactive instability.



(a) Basic VAE

(b) StdAug VAE

(c) AvgStdAug VAE

Figure 4.8: The uncertainty of each target point and the variance for embedded coordinates (x, y) for Basic VAE (a), with standard deviation augmentation (StdAug) VAE (b) and with average standard deviation augmentation (AvgStdAug) VAE (c).

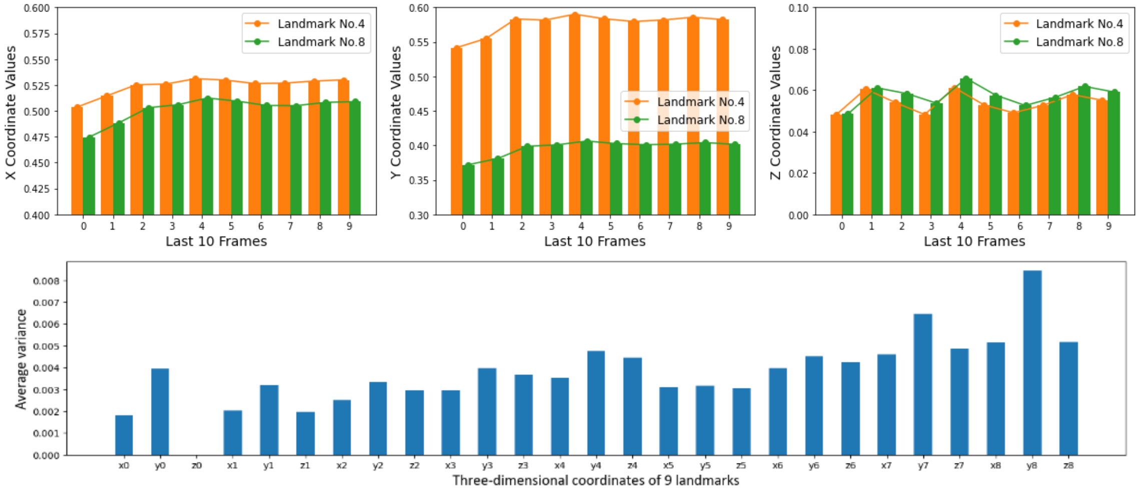


Figure 4.9: Analysis of Hand physiological jitter. The three-dimensional coordinates of 9 detected landmarks and their average variances from the last 10 frames when the user thinks they are close to the random target point and stops moving their hand.

Augmentation Experiment

To improve the stability of the system interaction process, especially to address the instability and sensitivity when approaching the target points, we proposed the data augmentation strategies in Section 4.3.5 and retrained the VAE model on the augmented hand poses with an objective function. In particular, we explored two augmentation strategies. (i) **StdAug VAE**: For the last ten frames of 200 explorations (target points are not fixed), we calculate the standard deviation of each coordinate of each point on their original landmarks, and finally average the standard deviation of each target point and then sample. (ii) **AvgStdAug VAE**: Different from StdAug VAE, we first calculate the standard deviations of the last 10 frames of each exploration and then calculate the average of the standard deviations of the 200 explorations.

By comparing (a), (b) and (c) in Figure 4.8, we observed that StdAug VAE made the distribution of hand poses more uncertain but AvgStdAug VAE was better. Specifically, for AvgStdAug VAE, the both principal axis (0.0091) and secondary axis (0.0023) of the mean-variance value of the binary distribution are smaller than that of basic VAE (0.0193, 0.0032). StdAug VAE shows the worst results (0.1175, 0.0156). Among all the target points, the point (No.1 point) with the largest uncertainty in AvgStdAug VAE also has a smaller variance (0.0549) than that (0.0733) of the same point in basic VAE. For the other target points, the variance figures of Figure 4.8 show that AvgStdAug VAE has smaller variances. Moreover, as shown in the uncertainty figures with 95% confidence intervals, AvgStdAug VAE allows

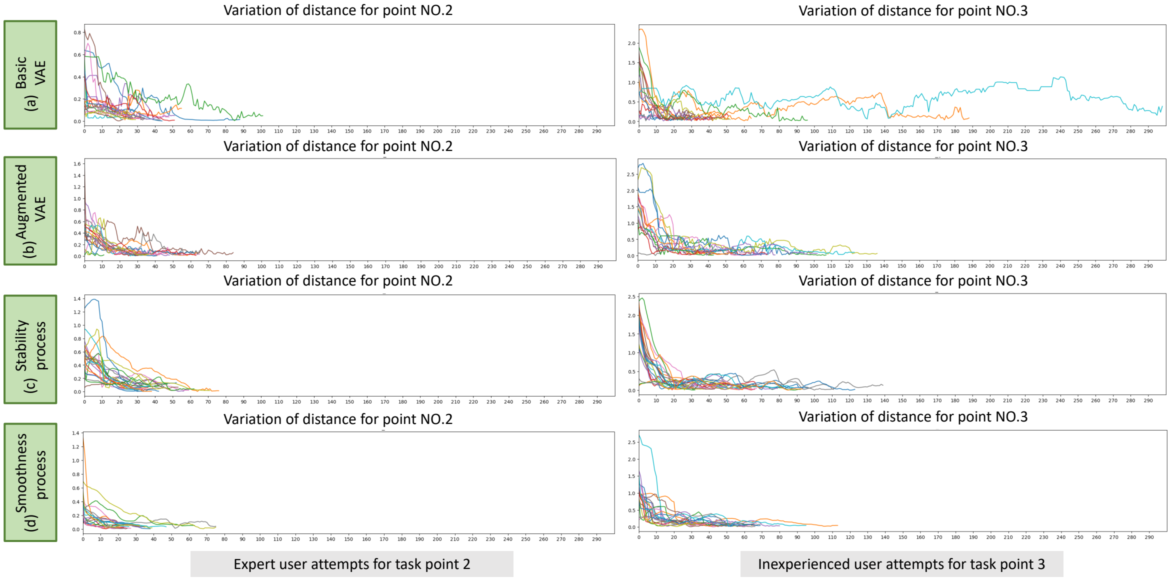


Figure 4.10: Distance changes for each attempt of the two users in the pilot experiments in relation to time on two random target points. The horizontal axis represents time (s/10) and the vertical axis represents Euclidean distance with regularization. The length of the curve shows how long it takes the user to complete a task. The convex and concave cusps of each curve represent a situation where the hand movement is unstable, in other words, fewer convex and concave cusps means less jitter.

most points to contract near the target point, somewhat reducing the constant jumping that occurs when the hand poses close to the target point. We believe that this owes to AvgStdAug VAE can mitigate the inaccuracy of the explorations due to multiple bad cases, and thus more accurately obtain the distribution of the hand poses from each exploration to target points.

Additionally, we also displayed the distance of each interaction from the target points and the time from the start points in Figure 4.10 (b). Compared with basic VAE (Figure 4.10 (a)), our AvgStdAug VAE with anti-jitter KL loss can better improve the stability of interactions when approaching target points. Note that, the user may have bad cases due to fatigue. Although we have improved the stability when approaching the target points, there is still instability and non-smoothness during the interaction. Next, we further propose two improvements.

Stability Experiment

To further improve the stability during interactions, we further introduced a stable post-processing design in Section 4.3.5. By comparing Figure 4.10 (b) and Figure 4.10 (c), we can find that the stable post-processing design can better improve the stability during interactions. In particular, Figure 4.10 (b) shows many sudden increases or decreases of distance variation, i.e., some convex and concave cusps in (b), which are mostly higher than 0.1. After our stability post-process, a large proportion of the cusps have been replaced by more gentle curves in Figure 4.10 (c). These changes are particularly evident in the distance variation curves after 6s for both types of users. It reflects that our proposed stable post-processing has an excellent capability to improve the stability of system interaction.

Smoothness Experiment

We proposed the smoothness experiment to address the phenomenon of non-smoothness according to Section 4.3.6. By implementing smooth post-processing, the distance-time curves between the gesture-embedding position and the corresponding target point in the space have been significantly improved, as shown in Figure 4.10 (d). Due to the staged filtering effect of the One Euro Filter with different parameters, subtle wave peaks during the interaction process were significantly reduced, resulting in a smoother trajectory and stability near the target point. Moreover, the different filter settings in the first and second stages of interactions allow for a better balance between speed considerations in the first stages of interactions and accurate considerations at the end of interaction.

4.4.4 User Study: System and Interface Assessment

In this Section, we conduct user studies to evaluate the proposed interactive system on the designed hand pose embedding space, as well as the interface.

Experiment Design

After the pilot experiments, we solved a series of stability and smoothness issues and completed the final hand pose movement interactive system. Then, we conducted a user study to evaluate the performance of the proposed HpEIS. Before introducing the study experiments, we first introduce the configuration of the two versions of the proposed HpEIS. The main difference between the two versions is that we propose an innovative user guidance window based on hand pose reconstructions from the decoder of the VAE in version 2 (as shown in Fig. 4.2). We believe that providing users with guidance on hand posture movement is beneficial to the overall space exploration and interaction experience, especially for novice users.

Then, we conducted user study experiments. 12 participants were evenly divided into two

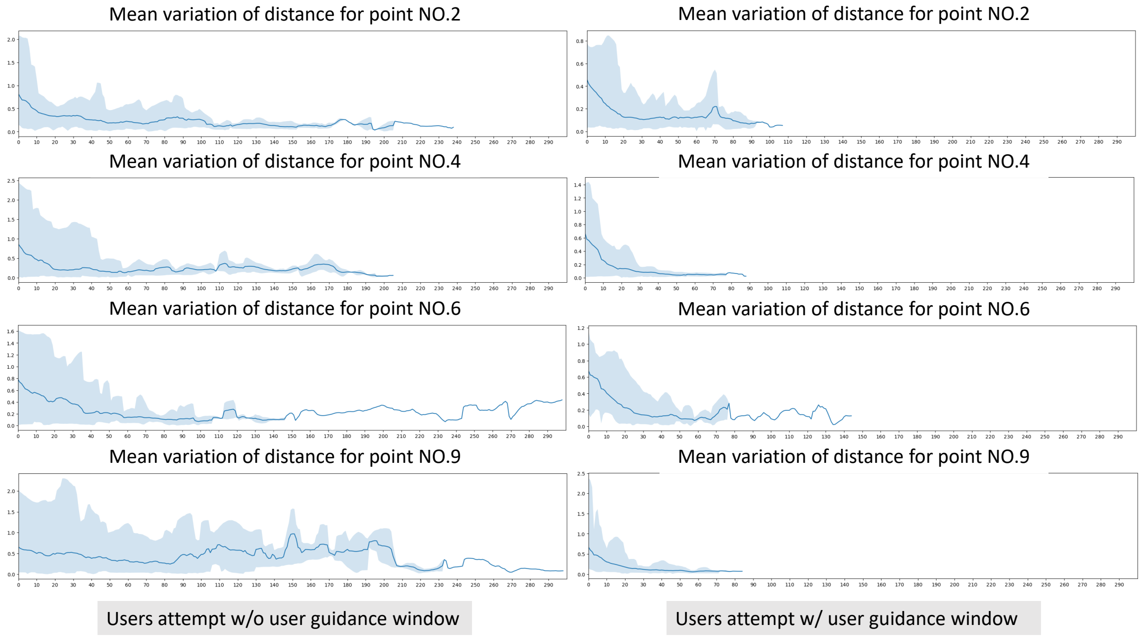


Figure 4.11: The average time-distance curves for the two groups of participants. The horizontal axis represents time (s/10) and the vertical axis represents Euclidean distance with regularization. Using the user guidance window can reduce the user's approaching time to the target point, which is 20s and 10s for the mean time, respectively. Fewer convex and concave cusps and a smaller light blue background area mean the overall interaction process is more stable and smoother when using the guidance window. It suggests that the user guidance window has a superior capability to guide users, especially new users.

non-overlapping groups to prevent the influence of experience factors. The first group completed the experiment without a user guidance window, and the other group was guided with a user guidance window. Participants were granted a 10-minute period for initial acclimatization before the formal experiment commenced to facilitate familiarity with the system. The staff took the initiative to elucidate the requisite hand pose (collected hand poses) that participants could use and stipulated that hand poses could be free. Note that, the second group was individually introduced to the use of the user guidance window. Next, participants in both groups began a user test in which each participant conducted a total of 100 interactive explorations and each time the target point was a random one of 10 specific target points. At the end of the experiment, participants were asked to complete a questionnaire to help us improve the system.

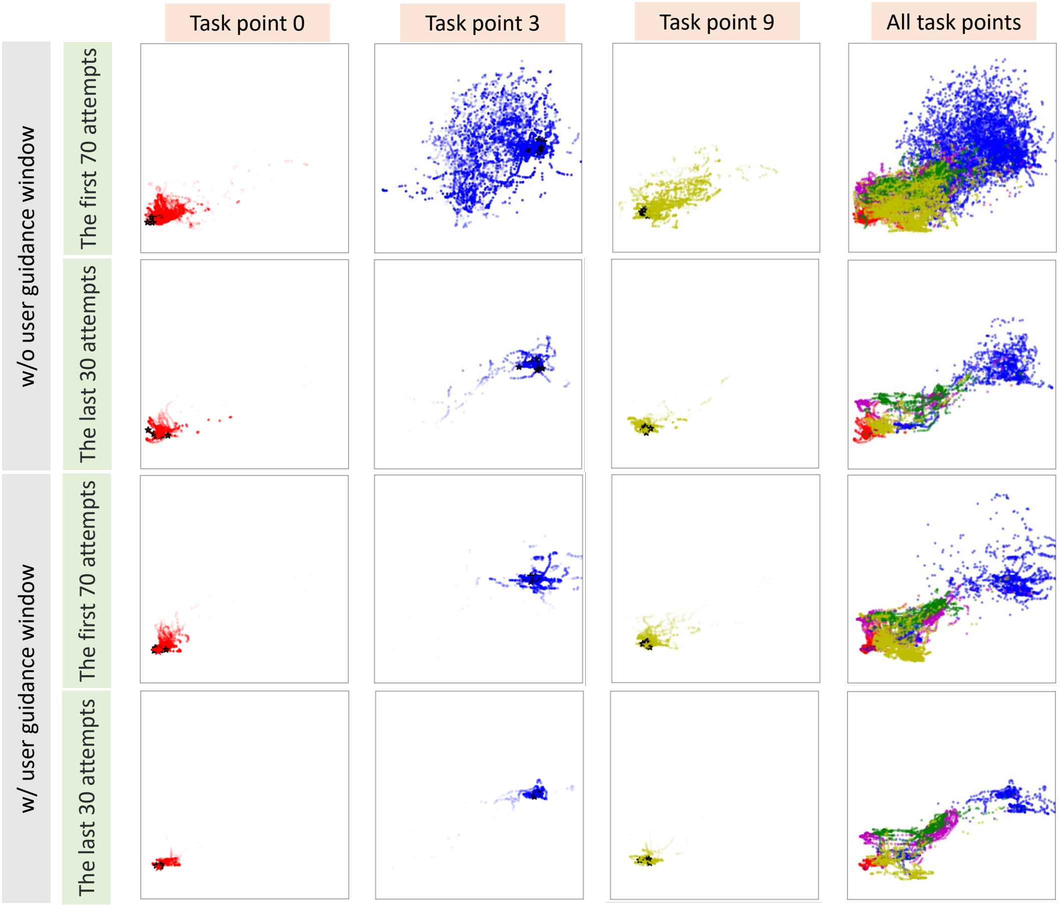


Figure 4.12: Two random participants' hand-pose points were visualized, one used the user guidance window and one did not. The black stars are stopping points of the final hand poses from a random participant. In particular, we split the first 70 and last 30 attempts of hand movements for each participant. The area covered by the gestures points that the participants need to explore when trying to find the target becomes smaller as the number of attempts increases, which demonstrates the user-learnability of our system interactions.

Tasks

The task is the same as our pilot experiment, i.e. finding specific target points. The goal is to swiftly and accurately navigate towards the designated target point by employing varied hand movements or distinct hand poses. Participants are entrusted with the intuitive determination of their proximity or arrival at a given target point. Once the target point is believed to be reached, the participant stops the interaction by performing a long press on the knob, which triggers an audio feedback. We set a maximum threshold of 30 seconds for exploration for each target point.

Data and Analysis

We evaluate the developed HpEIS from two aspects, the duration of interactive exploration and the final distance from the target point. We argue that these two aspects of interaction metrics can reflect the user interaction experience. In addition, the real-time distance curves (Figure 4.11) between the user's current position and the target point can also reflect the stability and smoothness of the interaction during user exploration.

Figure 4.11 shows the average time-distance curves for the two groups of participants. In addition, we believe that the computationally averaged area (i.e., the light blue background area can somewhat indicate the stability and smoothness of the interaction. Due to space limitations, we randomly selected 4 target points as examples for display. By comparing the results of the two groups of experiments, it is observed that using the user guidance window can reduce the user's approaching time to the target point, which is 20s and 10s for the mean time, respectively. Fewer convex and concave cusps and a smaller light blue background area mean the overall interaction process is more stable and smoother when using the guidance window. It suggests that the user guidance window has a superior capability to guide users, especially new users.

Moreover, as shown in Figure 4.12, we further visualized the location of the stopping point of the final hand pose for two random participants, one used the user guidance window and one did not. By comparison, it is clear that the participant with the guidance of the user window had more accurate interaction results. This again demonstrates the effectiveness of the user guidance window. Additionally, we analyzed the first 70 attempts and the last 30 attempts of each participant, and it can be seen that there is a certain unfamiliarity when the user first experiences our HpEIS. However, as the interaction progresses, users can learn the regularity of the hand poses in the embedding space, which makes it possible to become more familiar with the interaction of our system. This further demonstrates the user-learnability and acceptability of interactions on our hand pose embedding space.

Questionnaire Analysis

After the user experiment, 12 participants (6 with the user guidance window and 6 without it) filled out a questionnaire. It contained 23 questions, of which 22 were mandatory multiple-choice questions and 1 was a short-answer question. Three of the questions were specific to participants using the guidance window. Our questionnaire used a 10-point Likert scale and included the NASA Task Load Index (Hart & Staveland 1988) for workload assessment. The details of the questionnaire and the results can be found in the **Appendix D**. Specifi-

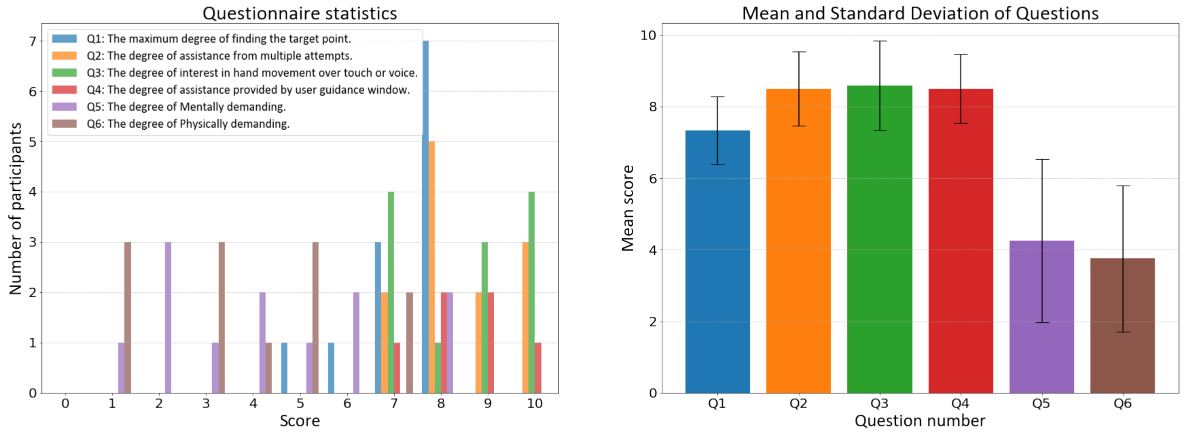


Figure 4.13: Statistical analysis of questions related to interactive hand poses. The graph on the left counts the number of participants with different scores for each question. The graph on the right shows the mean score and standard deviation for each question. Where the error bars are the standard deviation. On the left graph, we provide the questions.

cally, as shown in Figure 4.13, the figure on the left presents the responses to each question, which provides a clear picture of participants' overall evaluation of the system. The figure on the right offers a more precise view through statistical analysis based on these ratings. We found that more than half (58.3%) of the participants felt that they were able to find the target point accurately (Mean (*M*)=7.33, Standard Deviation (*SD*)=0.78). 10 participants (83.4%) largely felt that after a few more attempts they could memorize some information about the hand-pose space and target points could be found more easily (*M*=8.5, *SD*=0.92) in subsequent tasks. Among 6 participants who used the user guidance window, half always used it and the other half sometimes. 5 of them agreed that the guidance window helped them to a great extent to perform the task better (*M*=8.5, *SD*=0.83). Some participants stated that our HpEIS requires learning to find the correct way to use it, which requires a certain mental demand (*M*=4.25, *SD*=1.96) and physical demand (*M*=3.75, *SD*=1.75). Overall, all participants agreed that the proposed hand-pose interactions are more engaging and novel than traditional touch-based or voice-based interactions (*M*=8.58, *SD*=1.15).

HpEIS connects multimedia space

We believe the proposed HpEIS is a flexible multimedia adapter used to connect multiple multimedia spaces and directly interact without additional operations, such as song selection in the music space and movie selection in the movie space, etc. Here we apply it to a music-embedding space, which provides a new inspiration for research on connecting multimedia.

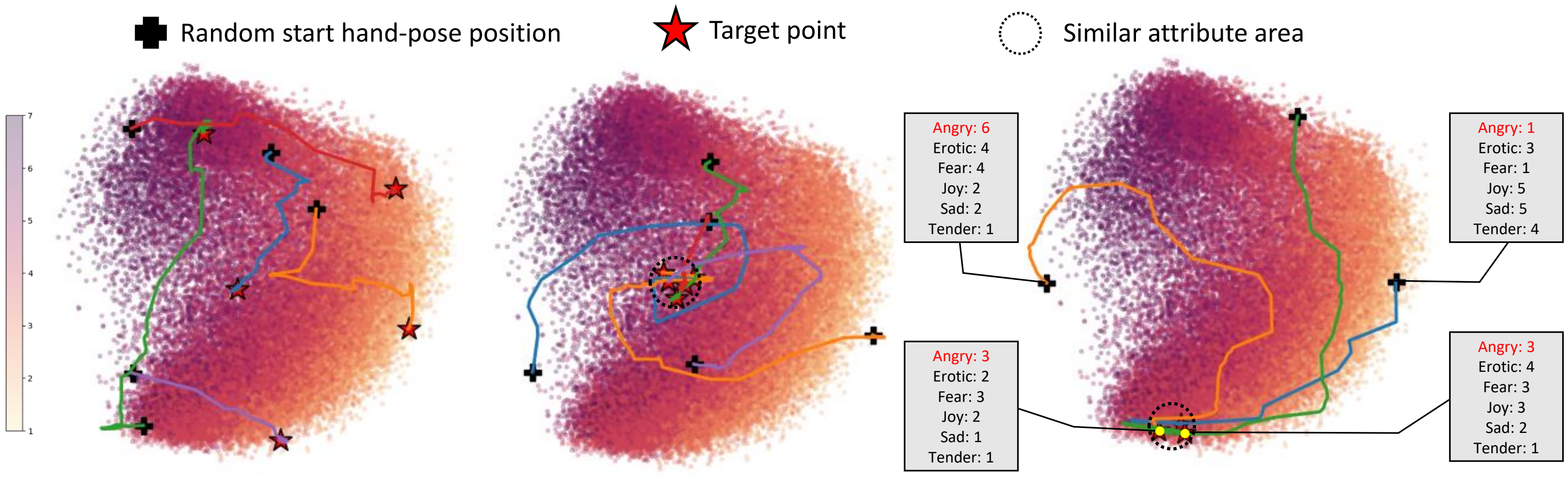


Figure 4.14: A simple interactive application of the proposed HpEIS on a music multimedia space of expressive angry music (The darker color, the more intense the 'Anger' expressed in the song.). The first one is the exploration and interaction of music with different levels of anger. The latter two respectively represent multiple explorations and interactions corresponding to music with a similar attribute. Specifically, two yellow points are on a circle of similar attributes with radius 0.1 in the music space, and the real distance in the high dimension are 1.41 and 2.24 respectively. The distance between the left start point and end points in higher dimensions is about 3.16, and the distance between the right start point and end points in higher dimensions is about 6.00, but only 0.47 and 0.65 in the music space. All distances are Euclidean distances.

In particular, given a music-embedding space embedded from a similar VAE model, we normalize its extent to be the same as the hand-pose embedding space extent. After that, we can directly interact with the music-embedding space via hand movement. As shown in Figure 4.14, any hand pose can intervene in the space for interaction, thus ensuring interaction flexibility. For exploring music areas with similar attributes, we can use different starting hand poses to explore the target areas from different directions and paths, thereby maintaining the user's sense of novelty in the music system. Notable, the multimedia space is not limited to music space, any other multimedia space interaction can use the same HpEIS system. This improves the transferability of HpEIS and reduces repetitive work, which is not available in other interaction studies (Xu et al. 2023).

4.5 Limitation

4.5.1 Method Limitation

Although we have been able to visualise high-dimensional hand movement poses in low-dimensional space and allow different hand poses to be distributed in different locations in the potential space by implementing classification of the hand poses. Based on this, our stability design and smoothness design allow us to build a more stable and smooth interaction system, which can be used to explore some 2D-based data spaces. However, the current system still has the disadvantage of having many hand poses in one region of the latent space, as shown in Figure 4.1 (c) and more details will be illustrated in the next chapter, so there is still a lack of good control or disentanglement of individual hand poses. In other words, the system has less control of how the hand poses could relate to different aspects of latent space control.

In addition, due to the limitations of the dataset, the current systems can only have good interaction control for the categories of hand poses that have been collected. This property makes our control of hand movements less flexible. We would like to be able to control different controls by different hand movement states, whereas existing interaction systems are difficult to achieve separate controls corresponding to different movement states without more extensive datasets and a large overlapping area due to the above two limitations.

4.5.2 Study Limitation

This chapter also has some study limitations. The number of participants in the experiment was still small, as already discussed in Chapter 3, and this remains a constraint. More importantly, all experiments were conducted in a controlled environment with a plain white background. This condition simplified the recognition task by removing visual noise and distractions, but it does not reflect the complexity of real-world scenarios. In everyday settings, users may interact in front of varied backgrounds, in different rooms, or even outdoors, where the colors, textures, and lighting conditions are much less stable. Such changes may influence the performance of the hand pose interaction method, and the current results may therefore overestimate the robustness of the system. Another limitation is that the controlled setting restricted the evaluation of how the system adapts to diverse and dynamic conditions. For example, sudden changes in lighting, cluttered scenes, or moving objects in the background were not considered in this study. These factors can affect both the accuracy of hand detection and the stability of the interaction. Addressing them will be necessary to make the method practical for real-world use.

4.6 Conclusion

In this chapter, we developed a HpEIS to adapt multimedia collections via mid-air hand movement using only a camera with MediaPipe software. HpEIS engages in (i) providing user-learnable and interpretable visible space exploration experiences, and (ii) providing flexible, stable, and smooth hand-pose interactions. To this end, we proposed the augmented VAE model encoding to obtain a 2D normalized hand-pose embedding space, which can be visualized for multimedia interaction via hand movement. A series of stability and smoothness post-processing operations have improved the overall user experience of the interactive system. We evaluated HpEIS in the task of finding the target points in the hand-pose embedding space with and without a user guidance window, as well as mapping music embedding space exploration. The hand movement curves of our user experiments show that our HpEIS provides a flexible, stable and smooth hand-pose interaction method, and the application of our HpEIS in a music space provides an idea for 2D-space interaction of mid-air hand poses with smart multimedia systems. It guides a stable, smooth and easy-to-learn approach for visualising mid-air hand poses for interaction with smart multimedia. However, we still want to further enhance hand interaction by making it more natural and flexible. Rather than restricting interactive hand poses to predefined gestures within a dataset, we seek a method that allows control based on the continuous state of hand movements. At the same time, we aim to minimize the need for constantly collecting new gesture data to accommodate evolving interaction demands. Therefore, in the next chapter, we will explore these challenges and expectations through the use of a disentangled low-dimensional hand pose space.

Chapter 5

A Mid-Air Hand Pose Interaction Method Based on Disentangled Degrees-of-Hand-Freedom

Previous research has shown that high-dimensional hand poses can be mapped to a low-dimensional space, enabling stable and smooth visualization of mid-air movements. However, these methods are limited by predefined poses and a 2D space, restricting their functionality. Furthermore, distinguishing between overlapping hand-pose intervals within one representation space remains a significant challenge. To address these limitations, we proposing a generalizable strategy for handling high-dimensional inputs. In this chapter, we address three key challenges in hand-pose-based interaction research: ***i)*** *disentangling the hand-pose embedding space to ensure independent interactions of different hand poses within a unified representation,* ***ii)*** *enhancing the extensibility of hand-pose encoders for broader interactions, and* ***iii)*** *developing an optimisation evaluation framework to assess interaction performance and adaptability. Our goal is to improve the practicality and extensibility of mid-air hand movement interactions for real-world applications.*

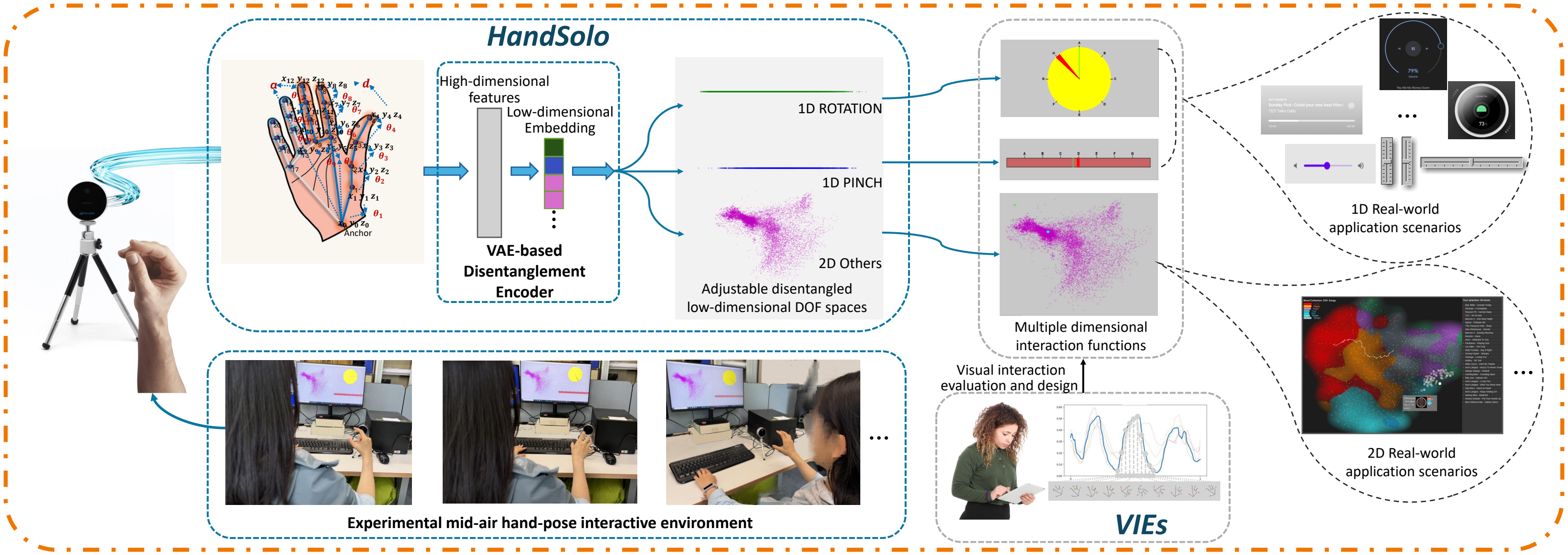

HandSolo
Anchor
High-dimensional features
Low-dimensional Embedding
VAE-based Disentanglement Encoder
1D ROTATION
1D PINCH
2D Others
Adjustable disentangled low-dimensional DOF spaces
Multiple dimensional interaction functions
1D Real-world application scenarios
2D Real-world application scenarios
Visual interaction evaluation and design
VIEs
Experimental mid-air hand-pose interactive environment
...

Figure 5.1: Our ***HandSolo*** mid-air hand pose interaction method can disentangle the high-dimensional hand poses captured by a camera into low-dimensional degrees of hand freedom (DOF) embedding spaces. It can independently apply different hand poses, such as PINCH, ROTATION and others, to one-dimensional (1D) or two-dimensional (2D) interactive control. The two 1D disentangled spaces and one 2D disentangled space in the example shown above can be mapped to a variety of practical uses, such as dial control of music volume, dragging a slider of progress bars, selecting tracks in a streaming media space (Vad et al. 2015). The visualization analysis tool, *VIEs*, we proposed to help designers design and combine DOFs to find reasonable mid-air interactive hand poses.

5.1 Introduction

Mid-air hand-pose interaction aims to use different hand poses and movements for contactless system control. In recent years, mid-air hand-pose interaction has made significant progress in the field of intelligence with the development of smart devices, e.g. contactless hand pose interactions in intelligent driving systems (D'Eusanio et al. 2020, Qian et al. 2020) help reduce distractions while driving, facilitates control of furniture such as curtains with gestures in smart homes (Liu et al. 2020), exploration of medical images through contactless hand poses in surgery (Alonazi et al. 2023, Lee et al. 2020, Mahmoud et al. 2021, Zhao et al. 2023). The main challenge is to implement various mid-air hand-pose interactions using a unified easy-to-adjust approach.

Our previous research has demonstrated that we can perform dimensionality reduction of high-dimensional hand poses, thus allowing changes in mid-air hand poses to be visualised in low-dimensional space, and making mid-air hand movement more stable and smooth in low-dimensional space. Our approach can effectively reduce the disadvantages of bulky, expensive, and uncomfortable devices based on large devices such as VR and AR. However, the former still faces huge challenges, such as processing of the high-dimensional characteristics of hand-pose signal features and the constraints in gesture interaction based on fixed hand-pose classification.

Specifically, due to the widespread use of cameras on different devices, such as mobile phones, homes, and even cars, camera-based mid-air hand pose interaction applications have been further extensively studied. However, they still face the two challenges, i.e. high-dimensional hand-pose signal features and user-friendly interaction design based on a powerful hand-pose encoder. Our work in Chapter 3 introduce a continuous interaction strategy with visual feedback of hand pose and mid-air hand pose recognition and control for a smart music speaker, relying on an autoencoder joint training framework for embedding fixed hand-pose categories. Chapter 4 further propose a new hand-pose embedding interactive system with stabilising and smoothing post-processing operations, mapping a user's flexible hand poses to a two-dimensional visual space using a Variational Autoencoder (VAE) trained on different hand poses. The methods mapped mid-air hand-pose movements onto a continuous two-dimensional platform for a stable and smooth interaction design. Although these interaction methods have some amount of flexibility and are easy to visualize, they are only adapted to fixed hand-poses and can only be used in a two-dimensional continuous space, which results in a limited functionality. Furthermore, based on the latent encoding obtained via the training of different mid-air hand-poses, although the different hand-poses can be

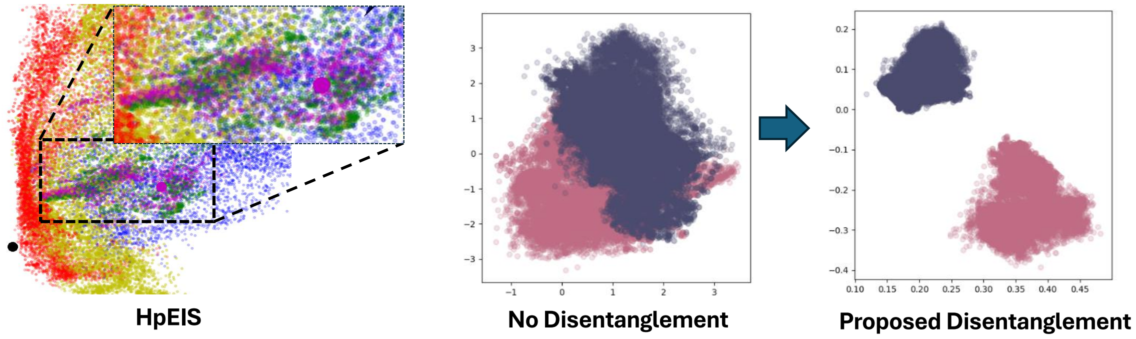


Figure 5.2: Comparisons of traditional hand-pose embedding space from HpEIS in Chapter 4, and ours without and with disentanglement. Note that the embedding figure comes directly from the Chapter 4. The mixture of multiple hand poses allows for continuous interaction of different hand-pose movements, but it is difficult to separate the different hand-pose spaces independently, resulting in ambiguity in space representation. Compared with HpEIS and no disentanglement VAE model, our disentanglement VAE model can better separate the unified mixed embedding spaces into two independent embedding spaces which can be used for independent interactions.

mapped to a unified representation space, there is still an underlying challenge that different hand-pose intervals in the unified hand-pose space are difficult to split, especially in areas where hand-poses overlap, as shown in the left of Figure 5.2.

In order to achieve interaction independence of different mid-air hand poses, the flexible mid-air hand movement interaction is then extended to more real-world interaction scenarios, thus providing a generalisable application strategy for interaction with more high-dimensional inputs. In this chapter, we simultaneously highlight and address three important issues in current hand-pose based interaction research for real-world application scenarios, including (i) separating the joint hand-pose embedding space of all poses in a representative dataset such that different hand-poses from the same encoder can be allowed to interact independently leading to a disentangled representation of the different poses, (ii) enhancing the extensibility of hand-poses from said hand-pose encoders, and (iii) designing an interaction optimisation evaluation scheme.

To address the above mentioned issues, we propose a new paradigm for mid-air hand-pose interactive system design, focusing on separating the hand-pose spaces from a trainable VAE-based encoder to obtain a more flexible representation of degrees of freedom (DOF) of hands and perform virtual control of different functions. The key challenges lie in (i) how to separate hand-poses and keep their interaction independent, and (ii) how to make system designers better at improving interaction design based on interaction feedback. To address the first issue, mainstream research (D'Eusanio et al. 2020) tends to use different classifiers to obtain the corresponding categories of hand-poses and map them to their respective embedding spaces, which are used for the corresponding hand-pose interaction control. However, this requires independent encoding of different hand-poses and the technique tends to be less robust to new hand-poses. The latter is often achieved using user questionnaires to obtain a rough design but without a reasonably detailed design paradigm. Thus, this remains an open research problem.

In this chapter, we propose a adjustable hand-pose space disentanglement approach for a VAE-based hand-pose embedding model (named ***HandSolo***) to obtain optional dimensional and independent hand-pose embedding spaces for interactions based on different degrees of freedom (DOFs) of hands. In addition, we further introduce a new visual interaction evaluation strategy (*VIEs*) to help system designers understand the optimal interaction scheme during user hand-pose interaction. In particular, different from our previous work, we incorporate a disentanglement penalty term into a trainable VAE model to separate different embedding spaces of hand-pose movements with different DOFs. It plays a crucial role in the separation and control of hand poses, as well as in the study of high-dimensional signal interaction. On one hand, the disentangled hand-pose DOF representation from the low-dimensional embedding space of high-dimensional hand poses allows for different functional hand interaction designs in corresponding independent distinct low-dimensional spaces. On the other hand, we can disentangle arbitrarily required hand poses with different DOFs from a hand space coming from the same encoder, resulting in low-cost controllable interaction systems. Furthermore, to better design a human-computer interaction system, we further propose the visual interaction evaluation strategy for system designers to choose the best disentangled embedding space of hand-pose DOFs for interactions. The whole new paradigm of mid-air hand-pose interactive system design we proposed is shown in Figure 5.1.

Contributions:

- We propose a method, *HandSolo*, to realize multiple interactive functions with adjustable mid-air hand-pose movements by disentangling the embedded low-dimensional latent space from a learnable VAE model into different independent hand pose subspaces with different DOFs.

- We explore new hand information representations (named new hand features) to better enhance the spatial extent of the disentangled embedding for specific hand poses, thus providing better interaction design options for different interaction functions.

- We design a new visual interaction evaluation strategy (VIEs) to help the system designer design and improve optimal interaction systems based on visual analysis of user feedback. This is a new effort that improves on the traditional questionnaire-based human-computer interaction strategies.

- We design a simple mid-air hand-pose system with the proposed new approaches and use it to demonstrate the effectiveness of our proposed methods. We further demonstrate the applicability of our HCI system in real-world scenarios through user studies.
- We demonstrate the extensibility of our proposed HandSolo model for more hand-pose interactions with more DOF latent embedding spaces.

5.2 Background

5.2.1 Mid-air Hand-pose Interaction

Directly using hands to complete different control functions is a convenient way of interaction in the real world, because hands are a medium and a powerful tool to interact with the surrounding environment. By changing different hand poses, such as pinching, rotating, we can trigger different interactive functions and further realize interactive control. For example, we only need two fingers to control the TV volume, which enables fun new styles of interaction.

Some studies (Khundam 2015, Yousefi et al. 2016, Zhang et al. 2017) have attempted to model a virtual hand in virtual systems and perform hand interactions in the virtual environment based on virtual reality (VR) and augmented reality (AR) devices. For example, (Kim et al. 2024) designed a skin stretch display to simulate hand interaction in a VR environment that features four independently controlled stretching units surrounding the forearm. Although these approaches may work well to simulate hand interactions in virtual environments, they are difficult to implement in the real world and rely on external physical devices, such as head-mounted displays.

Detecting hand pose signals in the real world and parsing the signal content to implement corresponding interactive functions has recently attracted ever-increasing research attention in the HCI community. There are two main methods: one is based on radar hand signal representation (Hajika et al. 2024, Lee et al. 2024, Sluÿters et al. 2023), and the other is based on camera hand signal representation (Qi et al. 2024, Zahra et al. 2023). Hand interaction systems based on radar signals are often only applied to coarse-grained gesture recognition interactions due to the inherent crudeness of signal representation, such as different gestures corresponding to different interaction feedback. For instance, (Sluÿters et al. 2023) used microwave radar sensors to recognize gestures for interactions, which only filters the raw radar signals and reduces them to two physically meaningful features, i.e. the hand-radar distance and the effective permittivity of the hand. Although gesture interaction based on radar signals has certain efficiency advantages, i.e., the feature dimension of radar signals is not high, the gesture interaction it can do has obvious limitations, such as many continuous hand movements cannot be recognized, the popularity of radar equipment is not high, and the distance between the hand and the radar cannot be too far. However, camera-based feature extraction for hand has attracted more and more attention because it is widely configured. For example, (Zahra et al. 2023) designed a camera-based hand gesture recognition and interaction system based on a traditional genetic algorithm. They achieved the same recognition performance and interaction functions as using radar equipment. Moreover, (Xu et al. 2023) proposed an autoencoder with a PointNet-based classifier, which encodes high-dimensional hand features in a low-dimensional embedding space. It allows multiple hand-pose movements to interact continuously in the embedding space, which can achieve more efficient embedding than radar signals. These studies have given us great inspiration, as they demonstrate the effectiveness and convenience of mid-air hand interaction. However, current research still has defects in the independence and coherence of hand-pose movement interaction, so we further address these issues in this study.

5.2.2 Hand-pose Embedding

In this section, we will introduce some existing hand-pose embedding methods (Lee et al. 2024), based on the camera sensor. Some studies (Islam et al. 2020, Nguyen et al. 2023, Wang et al. 2020) employ the Convolutional Neural Network (CNN) to extract the hand features and recognise them to realize different interactive functions. However, their high-dimensional properties make them difficult to apply in real scenarios. Dimensionality reduction techniques (Gisbrecht et al. 2015, Rusu et al. 2022) are introduced into the hand-pose embedding, transforming and creating a low-dimensional representation of high-dimensional hand signals. For example, (Rusu et al. 2022) used a data glove equipped with accelerometers to record high-dimensional hand movement data that are thereafter reduced to 2D embeddings using autoencoders. However, its dependence on devices makes it not extend well to multiple high-dimensional sensors and more real-world scenarios, such as dirty hands when cooking. Chapter 3 and Chapter 4 proposed a VAE model joint learning with a hand pose classifier under the real-world camera, which can better embed the hand movement into a 2D embedding space and allow the user to interact in this space. However, these methods encode all the collected fixed gestures or hand poses into a unified 2D embedding space, which makes the interaction single-functional and less independent among different hand poses. Disentanglement methods (Fu et al. 2024, Higgins et al. 2017, Locatello et al. 2020, Tran et al. 2021) are popular and effective methods to achieve interpretable application of model to different tasks, and are mainly used to disentangle feature representations from a unified model into independent feature representations with different attributes and thus achieve independent adaptation to different tasks. Inspired by this, in this study, we would like to introduce the disentanglement strategy into the high- to low-dimensional VAE to allow high-dimensional hand signals to be mapped into hand-pose embedding spaces with different DOFs using a unified encoder and achieve different interaction functions.

5.2.3 Disentanglement Metrics

To assess the ability to disentangle models in representation learning, (Kumar et al. 2017) developed the SAP score. This score is used to measure the difference in predictive ability between the two most predictive latent variables in each generative factor. This assessment metric is simple but assumes that the relationship between the variables is linear, which is less applicable in more complex datasets. Similarly, the Beta-VAE score (Higgins et al. 2017) uses a linear classifier to predict generative factors from latent variables. The higher the score, the better the separation. However, it relies on the assumption that the relationship between factors and latent variables is linearly separable, which limits its flexibility. To improve the limitations of the Beta-VAE score, (Kim & Mnih 2018) proposed the FactorVAE score. It measures the independence between latent variables by using discriminators. This approach obtains more complex dependencies, but it increases the computational cost and the choice of hyperparameters is a challenge.

In addition, Mutual Information Gap (MIG) (Chen et al. 2018) calculates the gap between the highest mutual information and the second highest mutual information between latent variables and generative factors. It penalises information overlap between latent dimensions and encourages more unique representations. However, for high-dimensional data, MIG can be very expensive to compute.

To better assess the ability of latent variables to match generative factors, the DCI (Eastwood & Williams 2018) uses three metrics to assess representations: disentanglement, completeness, and informativeness. However, the classes of generative factors are not always known, especially for some unsupervised or weakly supervised learning, so DCI does not work in such cases.

5.3 Designing of Mid-air Hand-pose Disentanglement Model

In this section, we present the details of our new mid-air hand-pose disentanglement model, including the fundamental hand-pose embedding model. and our model's extensible exploration. Our main innovation is an extensible hand-pose space disentanglement approach based on VAE (HandSolo).

5.3.1 Adjustable Hand-pose Space Disentanglement VAE (*HandSolo*)

As shown in Figure 5.3, we propose a *HandSolo*, a disentanglement approach for the widely used VAE model (Kingma & Welling 2014, Rezende et al. 2014, Xu et al. 2024, 2023), to disentangle the unified hand-pose embedding space into different independent sub-space for meaningful hand poses with different Degrees-of-Freedoms (DOFs). Please note that, in our study, the DOF is different from the traditional "hand DOF" (Hollerbach 1985, Muceli & Farina 2011), such as bones and joints. We set different movement patterns and characteristics of the hand as different DOFs, because we found that different dimensions of embeddings can represent different hand poses, e.g. joint rotation, fingertip distance, different finger movement speeds. On the one hand, our HandSolo can reduce various complex high-dimensional hand-pose representations to a specified low-dimensional space and disentangle it into desired hand poses to control the system. This is an efficient fundamental model with low hardware requirements. On the other hand, we introduce more detailed hand features to better represent the corresponding hand movements with different DOFs and improve the robustness by increasing specific hand-pose movement embedding capability and reducing the influence of physical factors such as camera distance and different individual hands. In addition, we further explore the potential of the proposed HandSolo to extend more new hand DOFs to represent new hand poses, thus reducing the design cost of more new interaction requirements.

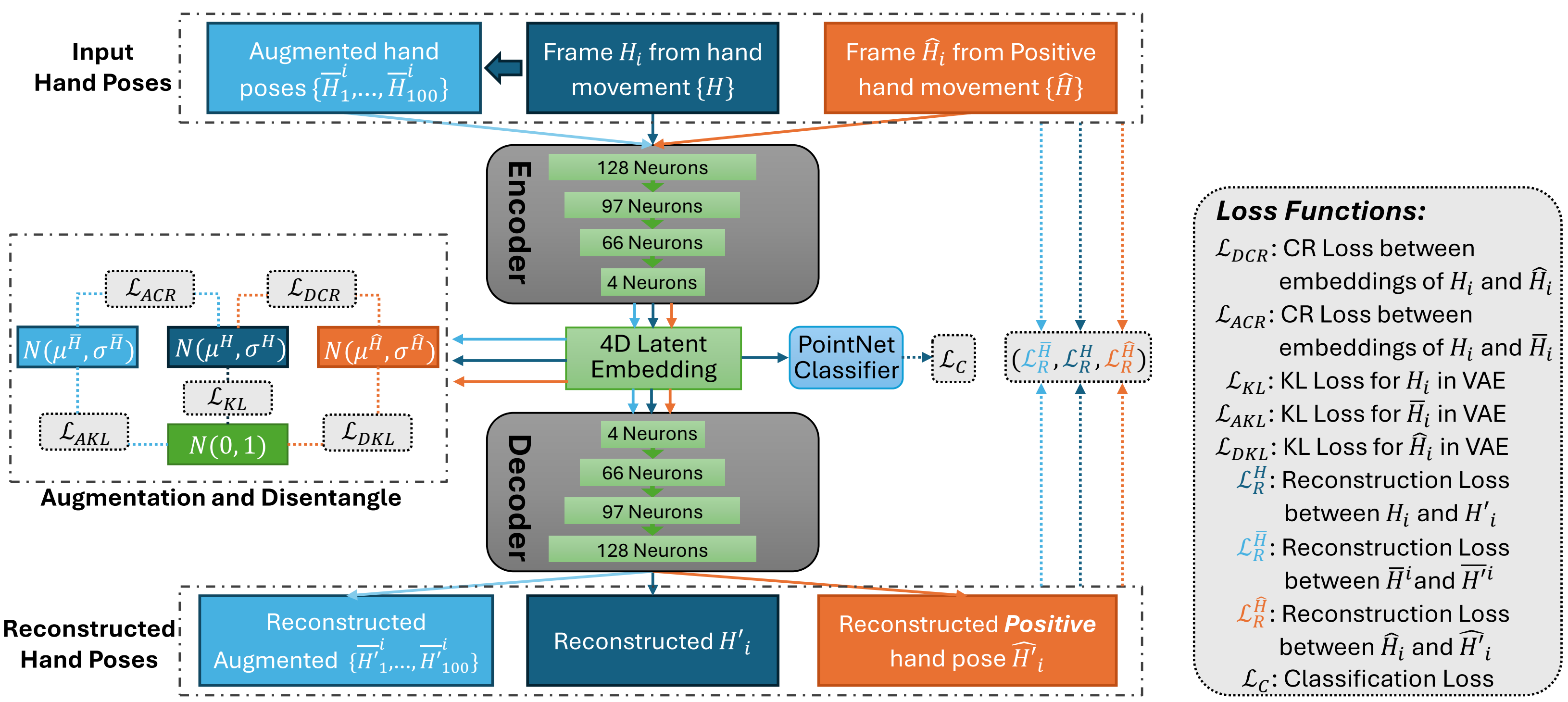

Input Hand Poses
Augmented hand poses $\{\overline{H}^i_1,\ldots,\overline{H}^i_{100}\}$
Frame H_i from hand movement $\{H\}$
Frame $\widehat{H}_i$ from Positive hand movement $\{\widehat{H}\}$
Encoder
128 Neurons
97 Neurons
66 Neurons
4 Neurons
$\mathcal{L}_{ACR}$
$\mathcal{L}_{DCR}$
$N(\mu^{\overline{H}},\sigma^{\overline{H}})$
$N(\mu^{H},\sigma^{H})$
$N(\mu^{\widehat{H}},\sigma^{\widehat{H}})$
$\mathcal{L}_{KL}$
$\mathcal{L}_{AKL}$
$N(0,1)$
$\mathcal{L}_{DKL}$
Augmentation and Disentangle
4D Latent Embedding
PointNet Classifier
$\mathcal{L}_C$
$(\mathcal{L}_R^{\overline{H}},\mathcal{L}_R^{H},\mathcal{L}_R^{\widehat{H}})$
Decoder
4 Neurons
66 Neurons
97 Neurons
128 Neurons
Reconstructed Hand Poses
Reconstructed Augmented $\{\overline{H'}^i_1,\ldots,\overline{H'}^i_{100}\}$
Reconstructed H'_i
Reconstructed *Positive* hand pose $\widehat{H'}_i$
Loss Functions:
$\mathcal{L}_{DCR}$: CR Loss between embeddings of H_i and $\widehat{H}_i$
$\mathcal{L}_{ACR}$: CR Loss between embeddings of H_i and $\overline{H}_i$
$\mathcal{L}_{KL}$: KL Loss for H_i in VAE
$\mathcal{L}_{AKL}$: KL Loss for $\overline{H}_i$ in VAE
$\mathcal{L}_{DKL}$: KL Loss for $\widehat{H}_i$ in VAE
$\mathcal{L}_R^{H}$: Reconstruction Loss between H_i and H'_i
$\mathcal{L}_R^{\overline{H}}$: Reconstruction Loss between $\overline{H}^i$and $\overline{H'}^i$
$\mathcal{L}_R^{\widehat{H}}$: Reconstruction Loss between $\widehat{H}_i$ and $\widehat{H'}_i$
$\mathcal{L}_C$: Classification Loss

Figure 5.3: The framework of the proposed adjustable hand-pose space disentanglement VAE for **HandSolo**. The model principal is a basic Encoder-Decoder VAE, containing four linear layers respectively. It uses three inputs, namely target hand-pose frame H_i, augmented hand poses $\{\overline{H}^i\}_{[1,100]}$ and positive sample hand-pose frame $\hat{H}_i$. We optimise the three hand-pose inputs using the regular VAE loss functions, containing reconstruction loss and KL loss in VAE. Besides, we introduce a PointNet-based classification loss to constrain the regular classes of hand pose. Finally, we construct a disentanglement loss and augmentation loss based on Consistency Regularization (Sinha & Dieng 2021) (CR) loss with the positive sample and augmented samples, allowing the same hand-pose embedding to be further approximated and maintaining interaction stability.

Basic VAE Model

Variational Autoencoder (VAE) (Kingma & Welling 2014) is a widely used model for high-dimensional to low-dimensional mapping learning due to its efficiency. Inspired by Chapter 4, we employ the same augmented VAE model to learn a unified hand pose embedding latent space. This is due to its effective stability and smooth designs that can provide powerful interaction experience guarantees. In detail, as shown in Figure 5.3, the augmented VAE model with an encoder containing four fully connected layers with 128, 97, 66, and 4 neurons and the opposite number of neurons in the decoder. Our VAE model also contains a joint learning classification model based on the PointNet (Qi et al. 2017, Xu et al. 2023) to further improve the distinguishability of different hand-pose embeddings. This classifier accepts a low-dimensional latent space embedding as input, and the classifier loss $\mathscr{L}_C$ is joint with VAE loss. The VAE loss contains a reconstruction loss function $\mathscr{L}_R^H$, i.e. Mean Squared Error (MSE), for target hand pose H_i from the hand movement $\{H_i\}$ and a Kullback Leibler (KL) divergence $\mathscr{L}_{KL}$ of the distribution of low-dimensional data and the standard normal distribution.

New Disentanglement Approach

To eliminate semantic ambiguities between different hand-pose embeddings for multiple independent interaction requirements, we propose a HandSolo approach to represent different DOFs in terms of independent low-dimensional sub-spaces learnt from a VAE-based disentanglement model. Here, *we refer to the learned low-dimensional latent mappings as different degrees of freedom (DOFs)*, e.g., in Figure 5.3 we have 4-dimensional (4D) DOFs and we can flexibly set or combine different DOFs to form the corresponding hand pose. In this way, we can assign independent disentangled embedding spaces with fixed dimensions to corresponding hand-pose movements with different DOFs for different interaction requirements. For example, we utilize the disentangled one-dimensional space to independently represent “PINCH”, where the DOF values represent the distance of two fingertips. The independent embedding space allows “PINCH” interaction to be focused and uninterrupted by other degrees-of-freedom hand poses.

To realize the disentanglement ability in the VAE model, we introduce the ***Positive*** hand movement, which has the same interactive meaning as target hand movement (either multiple repetitions by the same user or attempts by different users). The motivation of our HandSolo is to narrow down the embeddings of hand poses (Here they are composed of predefined DOFs) from different hand movements with the same interaction meaning (Here we call them as ***Positive*** hand movements $\{\hat{H}\}$). As shown in Figure 5.3, we first train the VAE model for the Positive hand pose $\hat{H}_i$ for target hand pose H_i from ***Positive*** hand movement by a reconstruction loss $\mathscr{L}_R^{\hat{H}}$ and KL divergence $\mathscr{L}_{DKL}$. To narrow the embeddings between the target hand pose and the positive hand pose, we employ a Consistency Regularization (Sinha & Dieng 2021) (CR) (Here we called **disentanglement penalty term**) to make the embeddings of H_i and $\hat{H}$ consistently. It is formatted as $\mathscr{L}_{DCR}$ in Figure 5.3.

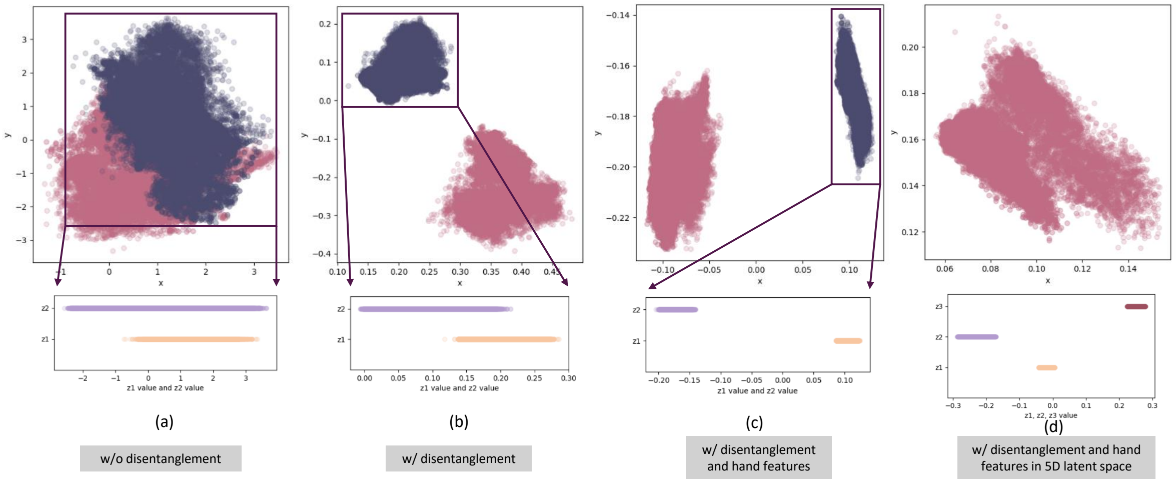
In addition, due to the augmentation strategy, we also consider the reconstruction loss $\mathscr{L}_R^{\overline{H}}$ and KL divergence $\mathscr{L}_{AKL}$ for augmented samples $\{\overline{H}^i\}_{[1,100]}$. To improve the stability of hand poses in corresponding embedding space, we also employ a Consistency Regularization $\mathscr{L}_{ACR}$ to make the embeddings of augmented neighboring hand poses $\{\overline{H}^i\}_{[1,100]}$ (the main cause of the system jitter phenomenon) consistent with target hand pose H_i. We sample 100 augmentations for each hand pose sample and the augmentation details follow with Chapter 4. Finally, our whole disentanglement VAE model (HandSolo) is optimized as follows:

$$\mathscr{L}_{VAE} = (\mathscr{L}_R^H + \mathscr{L}_{KL} + \mathscr{L}_C) + (\mathscr{L}_R^{\hat{H}} + \mathscr{L}_{DKL} + \mathscr{L}_{DCR}) + (\mathscr{L}_R^{\overline{H}} + \mathscr{L}_{AKL} + \mathscr{L}_{ACR}). \tag{5.1}$$

As shown in Figure 5.4, we visualize the ability to represent different hand poses with different combinations of DOFs from the 4D latent embedding space, based on different variants of the VAE model. Specifically, we regard two of the 4 dimensions as two-dimensional point coordinates (x, y), and we further use the left two disentangled dimensions as two independent one-dimensional coordinate movements, termed z_1 and z_2. In this way, we can fully demonstrate the multi-scenario utility of our HandSolo. Comparing the basic VAE (Figure 5.4 (a)) and our disentanglement VAE (Figure 5.4 (b)), when we consider the proposed disentanglement approach in the VAE model, the overlapping embedding spaces with different DOFs, which can be treated as three hand poses for interactions, one for 2D interaction and others for 1D interactions, are separated clearly. However, when we carefully disentangle the two disentangled 1D sub-spaces, there is still a certain overlapping interval. This indicates that two hand poses with the same DOF range may be activated in one interaction, which is problematic. To address this problem, we further introduce some new features for hand-pose representation, detailed in the next section.

New Feature Enhancement

In this section, we will introduce the inputs of our disentanglement VAE (HandSolo). On one hand, we follow Chapter 4 to represent the i-th hand pose by 21 key landmarks detected by MediaPipe (Lugaresi et al. 2019). Different from Chapter 4, to better represent the complex hand poses, we consider more landmarks (total 13 landmarks $\{(x_k, y_k, z_k)\}$, $k \in [1, 13]$) from



(a)

w/o disentanglement

(b)

w/ disentanglement

(c)

w/ disentanglement and hand features

(d)

w/ disentanglement and hand features in 5D latent space

Figure 5.4: Comparisons of different disentangled embedding spaces from different VAE variants. When considering disentanglement (b), we can obtain more independent sub-spaces for different degrees of freedom (DOFs) of hand poses. When further introducing new hand features (c), the distinguishability of the embedding spaces for different DOFs is more obvious, further facilitating interaction independence. Additionally, we can set higher dimensions of embedding space (d) to represent more DOF of hand poses and the still excellent discriminability reflects the cost-effectiveness and robustness of our model. We also provide more comparative results with different DOF combinations and settings in **Appendix C**.

three main fingers, not only two. Specifically, we represent i-th hand pose from hand movement as $q_i = [\theta_i^1, \theta_i^2, \theta_i^3, \ldots, \theta_i^{11}, \theta_i^{12}]$ after the quaternion conversion (Rieger & Van Vliet 2004) in Chapter 4. The quaternion is calculated based on two neighbour landmarks as shown in Figure 5.5 and can avoid the effect of hand size, hand position in the sensor field and distance from the sensor. However, there is still a certain overlapping interval as shown in Figure 5.4 (b), when we only consider the quaternion features. On the other hand, to improve the robustness and further reduce subspace ambiguity, we introduce new features to better represent the specific hand poses based on the detected landmarks referring to their physiological properties. As shown in Figure 5.4 (b) and (c), two 1D DOF sub-spaces can represent ROTATION (z_1) and PINCH (z_2). When we add two new features, i.e. the angle α between Landmark 12 and Anchor 0 and the physical distance d between Landmark 4 and Landmark 8, two 1D DOF sub-spaces can be disentangled better and no overlapping interval compared with Figure 5.4 (b) and (c).

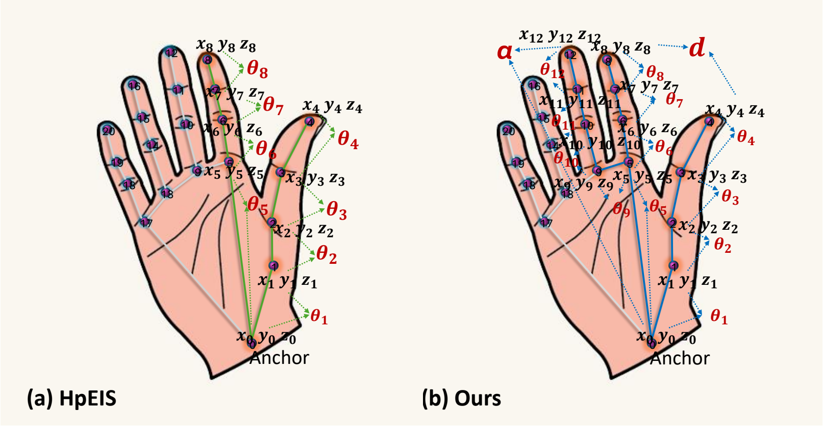


Figure 5.5: Different from the existing HpEIS in Chapter 4, we consider more additional features to represent the complex hand poses. By introducing new features, such as the angle between two landmarks and distance between two fingertips, the mapping correspondence of each hand-pose movement in the corresponding disentangled embedding space becomes larger and maintains linear correspondence for better interactions. The detailed comparsions are shown in Figure 5.6.

Extensibility Setup

In addition, we further analyse and visualize the importance of new features on Figure 5.6, to demonstrate that the embeddings can be disentangled into exclusive hand DOF without being affected by other factors. We use z_2, the fourth dimension in the 4D embedding space, as an example (Here we observe that PINCH is the best hand pose, details are given in Section 5.4.1.). When we consider the angle α and physical distance d (Figure 5.6 (b)), the PINCH interaction can perform more linearly without being affected by other factors, such as wrist rotation and hand moving from the sensor, compared with Figure 5.6 (a). This is because we only consider key hand movements that affect PINCH, and take other potential factors into account in hand features, thus ensuring that they are ignored as the PINCH subject movement occurs. Specifically, wrist rotation induces a smaller change in the DOF dimension of the fingertip distance, highlighting a more effective disentanglement of these degrees of freedom in the latent space. It also demonstrates that the disentanglement between wrist rotation and fingertip distance (PINCH) is improved.

In this way, as shown in Figure 5.6 (b), when PINCH occurs in the real world, our HandSolo can achieve a linear movement in the disentangled 1D embedding space. This can be employed in many interactions, such as controlling the progress bar of music, and controllable opening/closing of curtains or lights in a smart home.

As shown in Figure 5.7, our HandSolo has the extensibility to scale multiple DOFs to

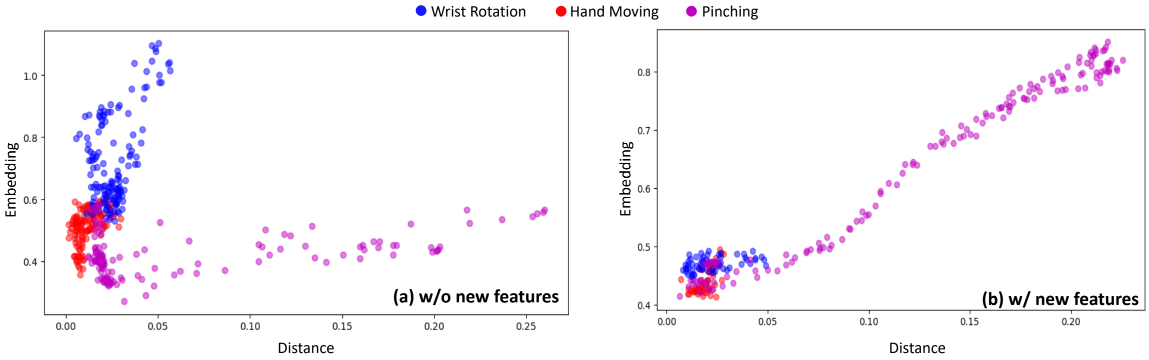


Figure 5.6: Mapping correspondence between the distance between the thumb's tip and index finger's tip and the embedding of PINCH hand pose in disentangled low-dimensional space. (a) for without new features, (b) for with two hand features. With the additional new features, the DOF for expressing the PINCH hand pose show a clearer linear correspondence with the variation of distance between the two fingertips, which is shown in magenta. The wrist rotation (blue), which was originally more sensitive, becomes less noticeable in this DOF. As a comparison, we also give the smaller impact of moving the hand back and forth relative to the camera (red).

accommodate new hand pose requirements without complex setups. This has great potential for real-world applications and lays the foundation for realizing low-cost and easy-to-use mid-air hand-pose interactive systems. In particular, in our disentanglement VAE model, if we want to add more hand poses for more interactive function controls, we only need to change the dimension of the latent embedding space and leave the rest unchanged. As shown in Figure 5.4 (d), we add another DOF in our disentanglement VAE by setting the dimension to 5. Compared with the 4D embedding space, in Figure 5.4 (c), our model can maintain good discriminability. We provide more details experiments to discuss this and demonstrate its effectiveness in Section 5.4.3.

5.3.2 Experiment I Disentanglement Test

Based on our previous description, in this section we compare the effects of the models. Inspired by Group-MIG (Tran et al. 2021), we designed a metric to assess the performance

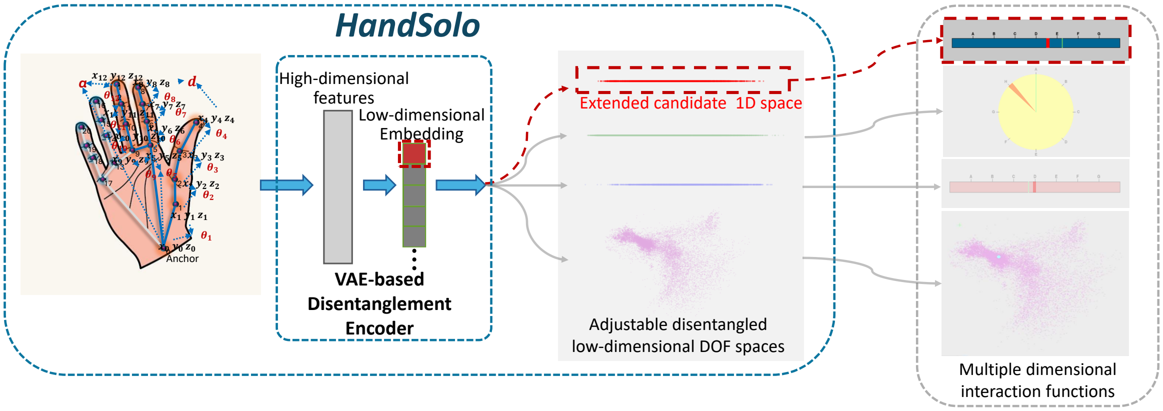

HandSolo
Anchor
High-dimensional features
Low-dimensional Embedding
VAE-based Disentanglement Encoder
Extended candidate 1D space
Adjustable disentangled low-dimensional DOF spaces
Multiple dimensional interaction functions

Figure 5.7: Extensibility Setup. Our HandSolo can directly extend a new disentangled candidate space to accommodate new hand pose requirements without complex setups. For example, we can add an additional dimension of VAE embeddings as a new DOF to disentangle a new 1D hand pose for another independent interaction.

of our model and named it Latent-MIG L:

$$L = \frac{1}{d}\sum_{i=1}^{d} M(z_i), \tag{5.2}$$

$$M(z_i) = \frac{H(z_i) - I_{\max}(z_i)}{H(z_i)}, \tag{5.3}$$

$$I_{\max}(z_i) = \max_{j\neq i} I(z_i, z_j), \tag{5.4}$$

$$I(z_i, z_j) = H(z_i) - H(z_i \mid z_j), \tag{5.5}$$

$$= \int p(z_i, z_j) \log \frac{p(z_i \mid z_j)}{p(z_i)} dz_i dz_j, \tag{5.6}$$

$$H(z_i) = -\int p(z_i) \log p(z_i)\, dz_i, \tag{5.7}$$

in which z_i is the value of dimension i of latent embedding, $H(z_i)$ is the entropy of the random variable z_i, $p(z_i)$ is its probability density function. $I(z_i, z_j)$ denotes the mutual information between z_i and z_j, while $H(z_i \mid z_j)$ is the conditional entropy of z_i given z_j. The joint probability density function is given by $p(z_i, z_j)$, with $p(z_i \mid z_j)$ as the conditional probability density. $I_{\max}(z_i)$ represents the maximum mutual information between z_i and any other latent variable. $M(z_i)$ is the mutual information gap for z_i. The density function $p(z)$ is estimated using kernel density estimation (KDE),

$$p(z) = \frac{1}{nh} \sum_{k=1}^{n} K\left(\frac{z - z_k}{h}\right), \tag{5.8}$$

where h is the bandwidth, $K(\cdot)$ is the kernel function, and z_k are the sample points.

The Latent-MIG can be used to assess the degree to which each latent variable is separated from the others in our learned representation of mid-air hand pose. Latent-MIG relies on entropy and mutual information. Higher values of mutual information indicate greater dependence between two variables. To assess the disentanglement ability of our model, Latent-MIG calculates the maximum mutual information between each latent dimension and any other dimension. However, traditional MIG calculation methods require discrete data, and to accommodate the continuous data of our mid-air hand movements, probability density estimation is required to compute entropy and mutual information. Kernel density estimation (KDE) approximates probability distributions by smoothing the observed data points using a kernel function (e.g., Gaussian).KDE provides a flexible method for estimating single-variable distributions and joint distributions. However, incorrect choice of bandwidth can lead to under- or over-smoothing, which can affect entropy calculations. So to prevent numerical errors, we truncate the estimated density to a small positive value before taking the logarithm. Unlike Group-MIG, Latent-MIG focuses only on the independence between latent dimensions and does not assess how well latent variables or groups of latent variables represent known data factors. This allows it to be used for unsupervised learning without predefined factors, thus enabling training without predefined gestures. Furthermore, the reason that we did not consider optimisation with the same metrics during training is because optimising mutual information (MI) during training is difficult as it requires the estimation of unknown distributions $p(z)$ and $p(z_i \mid z_j)$, which is computationally expensive to estimate using methods such as KDE. In addition, the probability estimates computed by KDEs are non-parametric methods that are non-differentiable and therefore not suitable for gradient-based optimisation. MI also involves hard-to-solve integrals that require numerical approximations with high variance, further increasing the computational difficulty.

We obtain the results as shown in Table 5.1. In the table, we show the Latent-MIG L of the four models. In this case, the VAE (Xu et al. 2024) is derived from our previous study. Also we compare Beta-VAE(Higgins et al. 2017), our base HandSolo, and our HandSolo with the addition of two hand features. For a further exploration, we also give the results of our extension HandSolo, i.e., 5-dimension latent space. In this experiment, besides considering the accuracy of the hand pose, we also consider the Latent-MIG, which is mainly used to

Table 5.1: HandSolo disentanglement performance results

Score	VAE	Beta-VAE	HandSolo	HandSolo Two extra features	Extension HandSolo
Acc(%)	97.16	96.87	**98.55**	98.46	97.28
Latent-MIG	0.4784	0.6726	0.7475	**0.7840**	0.6913

compare the degree of disentanglement of the data, i.e., the more disentangled the data is, the larger the Latent-MIG is, and vice versa. In this case, we emphasised larger Latent-MIG results considering that we wanted greater numerical differentiation of the different dimensions to be used to interact more flexibly and accurately with different DOFs.

Table 5.1 shows that the addition of the disentanglement module can increase the degree of discretisation of the dimensions in the latent space. Meanwhile, the Latent-MIG score of our HandSolo (0.7475) is 0.0749 higher than that of Beta-Vae (0.6726), while adding more hand features gives the highest Latent-MIG score of 0.7840. Compared with Beta-Vae which has the disentanglement effect, the Latent-MIG scores of our model are improved by 2.78%- 16.56%. While all the disentanglement models have higher Latent-MIG scores than that of VAE (0.4784), which verifies that our disentanglement model obtains a more discriminative latent space than that of the non-disentangled VAE. The accuracies of our proposed models are all higher than Beta-Vae (96.87%). Although the accuracy of adding hand features (98.46%) is not higher than that of the basic HandSolo (98.55%), there is not a significant difference between them. With higher Latent-MIG scores, the HandSolo with added hand features is more effective in implementing flexible interaction control through mid-air hand DOFs. Although our extension's setting results in lower Latent-MIG scores (0.6913) for the 5D HandSolo model, it is still a significant improvement over the VAE model without disentanglement or Beta-Vae. The possible reason for the reduced Latent-MIG scores for the extension's setting is that disentanglement of higher dimensions is more difficult, more complex, and more demanding on the model. We believe that more gesture data will help the disentanglement of higher dimensions.

5.4 Designing of Mid-air Hand-pose Interaction System

In this section, we present the details of our new mid-air hand-pose interaction system, including the prototype of our interaction system. Our main innovation is a new visual interaction evaluation strategy (VIEs) for system design. We provide the details below.

5.4.1 Interaction System

In this section, we introduce an example interaction system in Figure 5.8, which can employ our proposed HandSolo to control different interactive functions by the disentangled hand poses. In this way, we can evaluate our proposed HandSolo for mid-air hand pose interactions. We take 4D DOF latent embedding space from the trained disentanglement based VAE as an example. We also introduce examples of the application of hand poses with different DOF components. Finally, we provide a visual interaction evaluation strategy (VIEs), which can be used by system designers to choose the best embedding range and the most usable hand movements range for user interaction.

System Scenario

In the designed interaction system, we set the application scenario to interact with smart multimedia. When interacting with smart multimedia, e.g., smart music speaker, the most frequently used functions include volume level control, progress bar dragging, and searching in the media library according to preferences. To this end, based on our hand pose characteristics, we decompose the four-dimensional latent space of the trained disentangled VAE into three different interactive hand poses with different DOFs. That is, one two-dimensional DOFs for 2D interaction and two 1D DOFs for independent interactions, corresponding to the three interactive scenarios, such as searching music in a 2D media library (Vad et al. 2015), volume level control, and progress bar dragging. In this way, we can demonstrate the power of our model for different hand poses. To better understand these interactive processes, as shown in Figure 5.8, we provide a visualization interactive function interface that can be directly applied to the physical device. The dial and slider can simulate user interactive functions such as adjusting the volume and dragging the progress bar, respectively. Additionally, in a two-dimensional space, other hand poses control the cursor to select points that represent music tracks in a media library. Specifically, 2D interactive area be practically applied to multimedia exploration applications, such as virtual music tracking, where hand poses are mapped to corresponding music tracks. In this space, the user's hand movement is visually represented by a moving green cross pointer, the blue point represents the target point. In the dial area at the right side's upper part, we have marked letters at each $\frac{\pi}{4}$ interval around a circular dial. Numbers can alternatively replace these letters to indicate volume levels. The dial's starting position is set at the top of the circle, with the letters arranged clockwise. The user's hand rotation movement will guide a green linear pointer that moves around the circle from the center of the circle, allowing the user to control the rotation angle. The red sector is

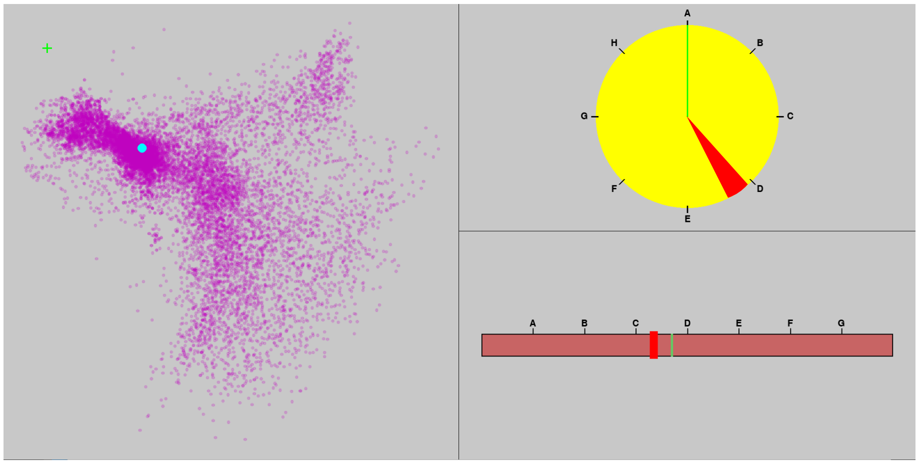


Figure 5.8: A sample virtual interaction system that integrates various virtual interaction objects such as a 2D interactive area, a dial and a bar with a slider.

the target point. In addition, on the lower right side, we design a horizontal bar with multiple marker points. The number and positioning of these markers can be adjusted to suit various functions and UI designs. Here we adapt hand pinch to control a vertical green linear pointer as a slider that moves horizontally. The red rectangular area on the bar is the target point. Each target point in all tasks is shown in only one of the areas and not shown in the other two. For our experiment, we set three sizes of different target points in each interaction, which randomly appeared. The details are further provided in **Appendix B**. Due to our excellent disentanglement capability, our system enables users to switch between various functions via different hand poses, facilitating volume adjustment, progress bar manipulation, exploration and selection of music tracks. These three designs illustrate one possible application of hands. And it is versatile and can be adapted for various functions and smart devices or applications. In the next, we will look for the best hand pose with different DOF choosing to implement each interaction separately.

DOF Setting of Interaction

To find a better single 1D DOF or combined multi-dimensional DOFs in the embedding space to match the hand pose interactive requirements, we provide an example to balance different settings (More examples with different DOF settings are provided in Figure C.3 of **Appendix C**). This is a selective process and requires the designer to make a trade-off. For example, we employ the trained 4D embedding space to represent four DOFs, called O_0, O_1, O_2, O_3. As shown in the Figure 5.9, we compare single 1D DOF from O_3 (a) and

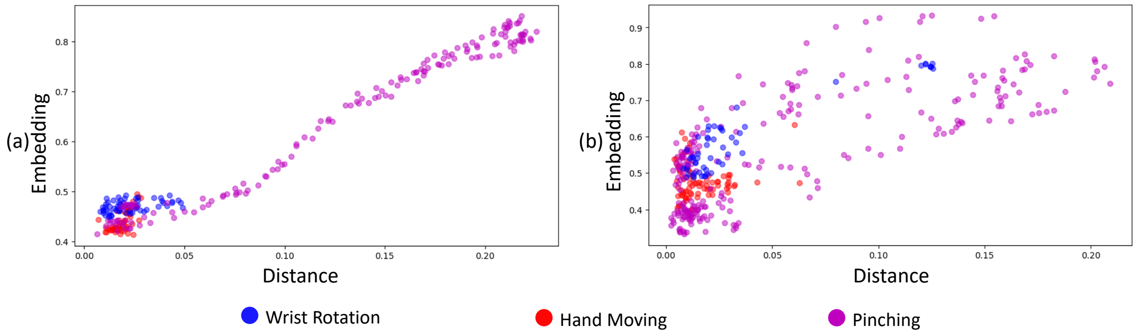


Figure 5.9: (a) and (b) is one of the examples to compare different DOF choices of PINCH (Slider) representation from the trained disentanglement VAE, where the DOF of (a) is O_3 and (b) is O_2.

O_2 (b), respectively. For each DOF, we try three hand poses to experiment with feedback in the corresponding disentangled space. Through the comparisons, we can observe that it is smooth and maintains linear growth when pinching. Therefore, we set DOF O_3 as the PINCH hand pose and treat it as the final interaction of bar dragging. Similarly, we use O_2 to control the dial rotation because it perfectly expresses hand rotation, which fits perfectly to adjust the volume of music by dial rotation. And the other DOFs (O_0, O_1) form a two-dimensional space that is used to match the music track space for interaction by other hand poses.

With the above description, we have constructed a mid-air hand-pose interaction system based on our proposed HandSolo for musical multimedia. ROTATION is used to control volume, PINCH is for dragging the slider on the progress bar and others are for music searching.

Visual Interaction Evaluation Strategy (VIEs)

We introduce a new visual interaction evaluation strategy (VIEs) to help the system designer design and improve optimal interaction systems based on visual analysis of user feedback, like refining the range for the selected gestures. This is a valuable new effort that improves on the time-consuming of traditional questionnaire-based human-computer-interaction design strategies. Specifically, as shown in Figure 5.10, we visualize the user interaction process into a unified view and present a reference line of skilled user hand movements. Here, we define the standard movement completed by our experimental designers as ‘skilled user movement’. These standard movements are baseline movements made by our designers after several attempts of our hand movement poses. These movements can be regarded as a benchmark for the hand interaction movements initially designed by our designers for the interaction system. The view is two-dimensional, with the horizontal coordinate representing the movement time and the vertical coordinate representing the value of the corresponding hand pose in the corresponding disentangled embedding space. The hand pose analysis curves of multiple users will be presented in the same view simultaneously. Through visual comparison, the best user hand pose range and comfortable hand poses can be clearly obtained. Then the embedding interval corresponding to the best hand movement range can be mapped to the corresponding interactive device. In this way, the system designer can provide users with a comfortable and reasonable interactive system through our proposed VIEs.

5.4.2 Experiment I Discovery of VIEs

We conducted three experimental studies to introduce the interaction system's design and illustrate our proposed model's effectiveness. In Experiment I, we experimentally showed **how the VIEs could be used to help system designers discover optimal design choices of our HandSolo for comfortable and easy-to-use interactions**. In this experiment, we will help the designer to discover more comfortable and reasonable hand poses and ranges of movement by comparing multiple hand pose movement curves of different users with those of our experienced designers, so as to achieve a more optimised interaction system.

Note that, the experimental setup of all experiments included an RGB camera of Intel® RealSense™ LiDAR Camera L515 (frame rate is 30 fps) and a monitor displaying the interactive interface as shown in Figure 5.8. At the beginning of each experiment, participants were briefed on the specific requirements of the experiments. The researchers would demonstrate to them the hand poses they needed to accomplish. All participants were instructed to sit within one meter of the camera, with hand movements to be performed at a distance of 30 cm from the camera. Participants were assured that their faces and other personal information would not be recorded, with only their hands and arms being captured.

Experiment Design

The purpose of this experiment was to investigate whether interactive system designers could design a comfortable and easy-to-use interactive system without resorting to user feedback in the form of questionnaires or interviews. The experiment was based on the disentangled four-dimensional embedding space from our trained disentanglement VAE. Before the experiment, as mentioned in section 5.3.1, we chosen the best three hand poses with different DOFs based on the trained VAE, including ROTATION from 1D DOF O_2, PINCH from 1D DOF O_3 and SWIPE from 2D DOFs (O_0, O_1). In this case, as shown in Figure 5.8, ROTATION was used for the interaction of Dial, PINCH was for slider adjustment and SWIPE was for 2D interaction.

Four participants (3 female and 1 male) took part in this experiment. All participants were right-handed, aged between 25 and 35 years. Here we simulate a system designer making system design and interaction choices. The researcher gave an example of each hand pose before user experimentation.

Task

In this task, participants were asked to complete the experiment as much as possible with the hand movement they felt comfortable with. When the experiment started, the participants were required to perform three complete hand interaction poses in the Dial interaction area, Slider interaction area, and 2D exploration interaction area on the interactive screen of Figure 5.8. The participants completed the complete interaction with the most comfortable starting and ending hand pose. After completing a hand pose, the participants clicked the space on the keyboard to proceed to the next pose. These processes simulate the user's optimal hand pose interaction process.

Experiment Data

As shown in Figure 5.10, the recordings of "skilled user movement" for three hand poses were shown as the guidelines in blue lines. To avoid incomplete entries or possible inaccurate recognition of one single hand movement, each hand pose was repeated three times. After this, taking the system designer's point of view, we could visualise the user tests into the same view for evaluation and selection of optimal interaction intervals. The Line of "skilled user movement" presented the changes within the corresponding hand-pose embedding space during the uniform hand movement of a skilled user. The other user test lines, which represent user interaction operations in real situations, reflected user interaction habits. Compared with them, the designers could design a comfortable and easy-to-use interaction based on the optimal comfortable interactive control interval of the user's curve response. VIEs showed whether the corresponding hand-pose embedding space could be used for comfortable easy-to-use interaction design.

Findings

As shown in Figure 5.10, we compared the movement results of three different hand poses of four participants in the disentangled low-dimensional embedding space with the correspond-

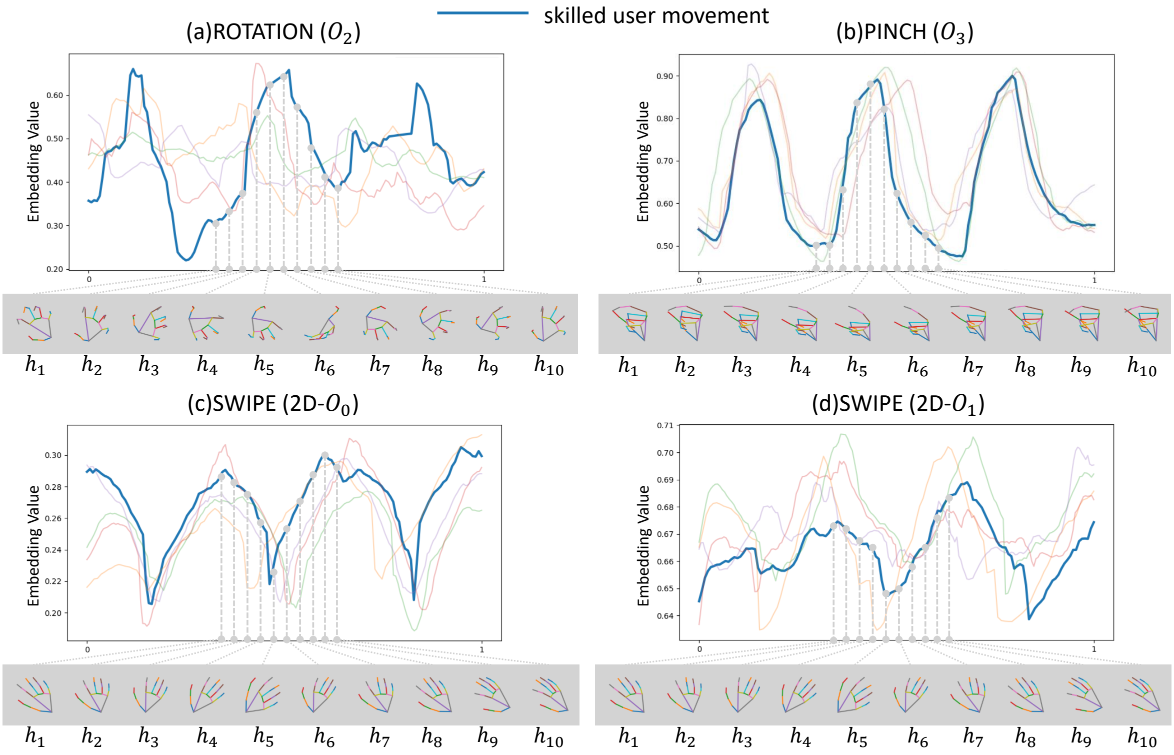


Figure 5.10: Visual Interaction Evaluation Strategy (VIEs) of three hand pose, including ROTATION, PINCH and SWIPE. ROTATION and PINCH interactive in the corresponding 1D embedding spaces and SWIPE in the 2D space. To allow designers to understand the interaction details more clearly, we expanded the 2D embedding space into 1D values for visualization. The blue line is "skilled user movement", presenting the embedding changes within the corresponding hand-pose embedding space during uniform hand movement for guideline. The horizontal axis represents the regularized time, and the vertical axis represents the embedding value.

ing "skilled user movement" blue guideline curves. For easy comparison, we normalized the time of all movements to 0-1.

Specifically, the PINCH movement curve shown in Figure 5.10 (b) has a high degree of overlap with the "skilled user movement" curve, which means that the user's DOFs of hand poses have a good disentanglement effect. It means all PINCH movement ranges could be well mapped to the corresponding embedding space for the corresponding comfortable easy-to-use system interactions. The embedding values from 0.50 to 1.0 could be chosen and remapped for PINCH interaction in real-world systems, such as progress bars.

However, the ROTATION movement in Figure 5.10 (a) was roughly similar to the "skilled user movement" in terms of movement trend, but there were still some undesirable situations. For example, the hand pose skeleton h_6 in Figure 5.10 (a) was not easy for all users to do, so most users could not reach the maximum value of the ideal situation shown in “skilled user movement”. Similarly, from the hand movement curves of the four users, it could be seen that the comfortable hand-starting pose is near the hand pose shown in h_3. Therefore, the designer could consider reducing the interaction area corresponding to a smaller movement range when designing the interactive system, i.e., renormalizing the area of 0.3-0.6 of the hand pose to adapt to the hand movement mode that makes the user feel comfortable and remapped to the Dial.

Additionally, Figures (c) and (d) respectively showed the movement curves of the x and y coordinates of the SWIPE hand movement in two-dimensional space. The x-curve showed a better degree of disentanglement of the DOFs, and the overlap of the hand pose movement curves was higher. Although the y-curve did not achieve a better disentanglement effect to a certain extent, the SWIPE hand pose alone had little impact on the y-disentanglement-space, and its fluctuation range was only about 0.04, which was in line with the physiological law of SWIPE hand movement. So we could still consider using these two dimensions to express the two-dimensional disentanglement space for hand pose exploration.

In conclusion, the designer could design a comfortable and easy-to-use interaction system that meets the user’s comfortable interactions based on the proposed HandSolo by only using our VIEs.

5.4.3 Experiment II Extensibility Discovery

In this experiment, we gave a further extensibility experiment of our proposed disentanglement VAE to provide evidence for the further interaction extension. This experiment aimed to **evaluate the simple extensibility of our HandSolo to provide a new method for the mid-air hand-pose-based HCI research community.** Unlike the 4D spaces in Experiment I, in order to verify the extensibility of our interaction model, we will use 5D latent space as an example in this experiment to demonstrate that our model is also highly capable in the disentanglement of the other-dimensions DOFs’ latent space. Effective interaction control of hand pose can still be performed with different disentangled hand DOFs.

Experiment Design

The purpose of this experiment was to explore the different possibilities of our HandSolo system in the process of exploring more hand DOFs from the proposed disentanglement VAE. In this experiment, we invited participants who completed Experiment I to complete Experiment II at the same time. They completed the similar task as Experiment I. We first trained a disentanglement VAE model with a five-dimensional latent embedding space, in which we disentangled three 1D spaces and one 2D space. *Notably, we have developed one more new 1D DOF space than Experiment I.* We selected four hand poses for control, namely ROTATION, PINCH, SWIPE, and WRIST. The first three were the same as Experiment I. The WRIST hand pose was to lightly squeeze the fist and rotate the wrist. The starting position was the direction of the fist facing the camera, and then rotated inward to the maximum angle. Participants saw the interactive interface shown in Figure 5.11 (f). The blue slider was for the WRIST pose. The researchers still recorded the "skilled user movement" first. In the experiment, the researchers demonstrated the four hand poses to the participants and asked them to make each hand pose three times respectively.

Task

The participants were asked to complete the same task as Experiment I for Experiment II with the only difference being that we added a new Slider function to simulate the linear interaction of the WRIST hand pose.

Findings

Similar to Experiment I, we utilized the proposed VIEs to visualize the relationship between the disentangled hand-pose embedding spaces of participants and the normalized time in Figure 5.11. When comparing Figure 5.10 and Figure 5.11, we found that more DOFs in the disentanglement VAE could be used to extend new hand-pose embedding space for additional interactive functions.

On one hand, after we extend more DOFs in our disentanglement VAE, the original three hand poses in Experiment I still maintained stable disentanglement capabilities, with some of the previous poses performing even better. Specifically, ROTATION obtains clearer disentanglement results, compared with Figure 5.10 (a), in which different participants performed similar hand pose movement curves in the results shown in Figure 5.11 (a). When the hand moved, a more overlapping hand pose movement curve could be obtained, which showed that the disentanglement result of 1D DOF O_2 was more stable and robust. The disentanglement results of PINCH were still excellent and did not require major adjustment or remapping in terms of DOFs for disentanglement, i.e., the range of y in Figure 5.11 (b). Regarding the DOFs disentanglement of hand poses in 2D space, we found that, based on the same hand poses, although 2D hand pose interactions could be realised as interactive functions, the stability was still challenged due to the multi-DOF combinatorial dexterity.

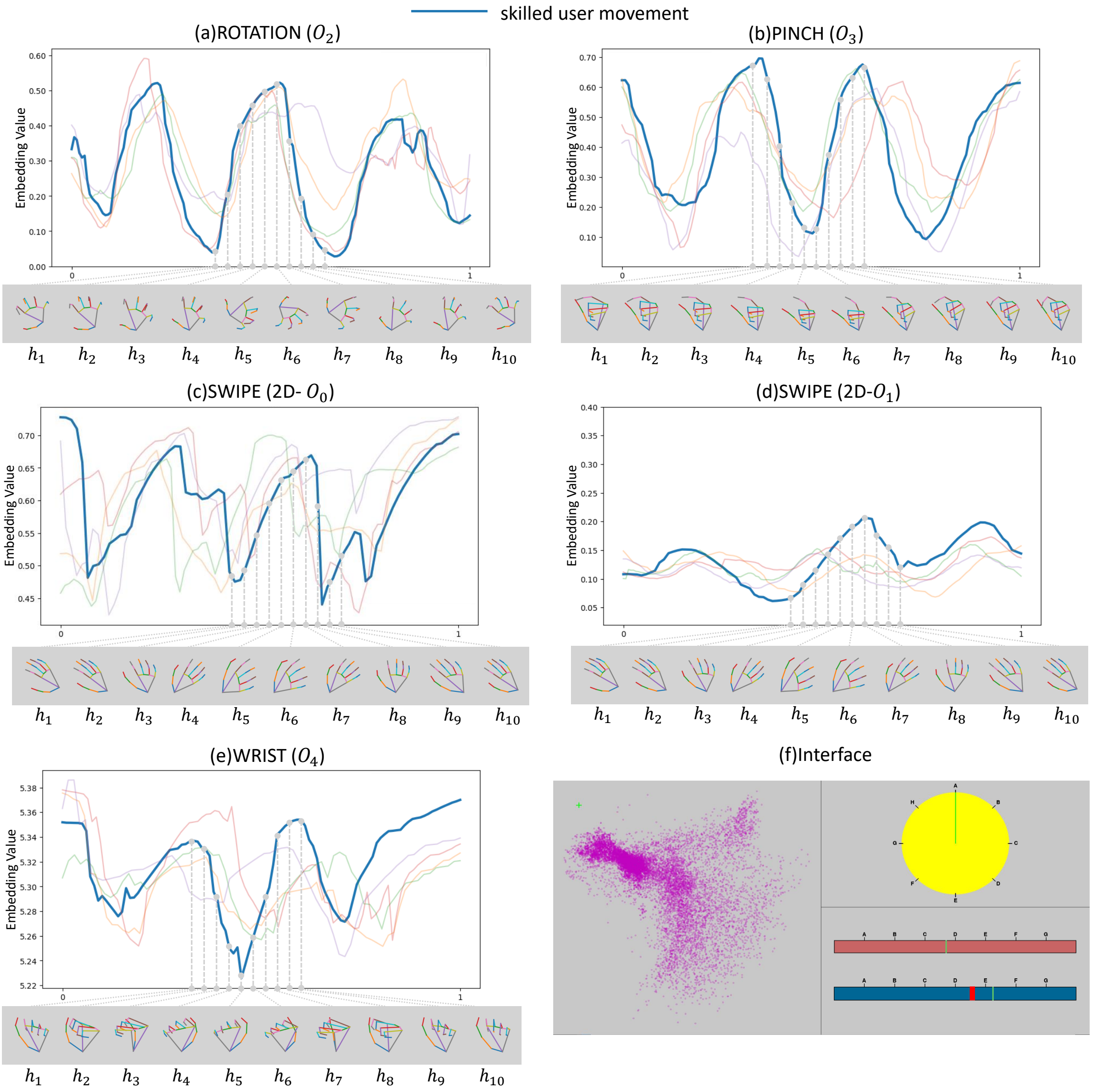


Figure 5.11: Similar with Figure 5.10, (a) to (e) showed the VIEs of four hand-pose movements. It included existing three hand poses, i.e. ROTATION, PINCH, SWIPE and a new additional hand pose, WRIST. ROTATION, PINCH and WRIST interactived in corresponding 1D embedding spaces and SWIPE in the 2D space. WRIST was the hand pose that we choosed to represent the extended DOF (O_4) spaces. This indicated the extensibility of our HandSolo. (f) was the extended interactive interface the participants see when they did Experiment III. The blue progress bar in the lower right was the interactive object of WRIST, and the rest objects were the same as those in Experiment I.

On the other hand, for another disentangled one-dimensional DOF dimension O_4 that we extended in this experiment, we tried to use it to express the WRIST hand pose, because when we performed model testing (Figure 5.6), we found that the rotation of the wrist had a certain influence in the disentangled latent space. The experimental results shown in Figure 5.11 (e) verified the ability of our model to disentangle the new hand DOF to a certain extent. Specifically, the hand movement curves of the multiple participants in Figure 5.11 (e) trended approximately the same, reflecting the stability of the interaction. Besides, the embedding value range could still be used to set up the interactive function. However, the disentanglement ability in this dimension was not as good as in the remaining two 1D DOF disentanglement spaces, and the hand movement curves overlapped less than those in the remaining two 1D spaces. The main reason for the participants' large difference from "skilled user movement" was the flexibility of the participants' wrist movement; the back of the hand facing the camera pose as shown in h_5 was not achieved by all participants.

Overall, our disentanglement model (HandSolo) could be extended to the disentanglement space with more hand DOFs and maintained disentanglement consistency and interaction independence. It could help designers find more combinations of DOF and discover the rationality of hand poses. Our HandSolo provides designers with more inspiration in mid-air hand pose interaction design.

5.4.4 Experiment III Interaction System User Study

The Experiment III was used to **evaluate the user performance of the virtual interaction system based on our proposed HandSolo after VIEs optimisation**. In order to demonstrate the rationality, efficiency and usability of our interaction model, we invite more participants to conduct interaction experiments. In the experiments, we will collect data such as time to complete the experiment and distance moved., and evaluate our interaction system according to Fitts' law. A questionnaire was also conducted with the participants for qualitative analysis. At the end of the experiment we will give our evaluation results and discussion.

Experiment Design

We performed user experiments on interactive designs by performing target-points-finding in the virtual interaction system in Figure 5.8. 10 participants (4 females, 6 males) aged between 25 and 40 years old, all right-handed, took part. Each participant completed 40 target-point-finding tasks, which took about 30 minutes.

Specifically, the researchers first explained the experimental requirements and the interactive system interface as shown in Figure 5.8. Then they gave examples of different hand-pose interactive processes to the participants, including two 1D interactions, named ROTATE and PINCH, and a 2D interaction containing multiple such as SWIPE, CIRCLE. In our experiment, we defined the hand poses in two-dimensional interaction as OTHER. The participants were allowed to make some additional changes based on these displayed poses, such as wrist rotation, palm tilting forward and backward. Participants had 5 minutes to familiarize themselves with the experimental system and tasks before the experiment officially began. They were allowed to pause the experiment if they felt tired, and continued after a short break.

We defined 15 target points, with 5 target points in each interaction area. The target points appeared randomly in three areas. There were three sizes of target points in each interactive area (the visualize interface image is further given in **Appendix B**), and target points with different sizes were appear randomly. As shown in Figure 5.8, in the 2D interactive area, the target point was a light blue dot with a diameter of 10, 15, and 20 pixels, and the hand movement pointer in the area was a green crosshair. In the 1D Dial area, the target was an arc of 10, 15, and 20 pixels on the circle. The fan-shaped area corresponding to the calculated central angle was coloured red. The central angle was calculated by $\theta = \frac{S}{r}$, where θ was central angle, S was arc length, r was radius length. The hand movement pointer in the area was a green linear pointer with the radius length rotating around the center of the circle. In the 1D Slider area, the target point was a red rectangle with a width of 10, 15, and 20 pixels, and the hand movement pointer in the area was a green linear pointer that could slide on the progress bar. Each target point in all tasks was shown in only one of the areas and not shown in the other two.

Task

Each participant was asked to complete a task of finding target points in the different interaction areas of the virtual system in Figure 5.8. When the participant saw a target point in one interactive area, they needed to find the target point by changing their hand pose and moving their hand. They needed to get as close to the target point as possible. When participants thought the point of orientation on the hand was close enough or has entered the range of the target point area, they pressed the space key on the keyboard to record and continued to find the next target point. If the participant thoughts that they had tried hard but still couldnot reach the target point area range, they could still press the space key at the position where they thought their hands were closest to target point.

Experiment Data

In the experiment, we recorded the time each participant takes to find each target point, the size of the target point each time, and the position in the embedding space of the participant's hand pose when it is first marked as appearing in the interaction area, which could measure the distance to the target point. Based on the data recorded in the experiment, we evaluated the interactive system according to Fitts' Law (Fitts 1954, Rusu et al. 2022). The Index of Difficulty (*ID*) and Index of Performance (*IP*) values were calculated by

$$\begin{aligned} ID &= \log_2\left(\frac{D}{W}+1\right), \\ MT &= a+b\cdot\log_2\left(\frac{D}{W}+1\right), \\ IP &= \frac{ID}{MT}, \end{aligned}$$

where Movement Time (MT) was the time required to complete the task, D was the distance to the target point, W was the target size, and a and b were constants. The linear relationship between MT and ID for each type of hand pose, the statistical graph of IP, and the statistical graph of all hand poses were given respectively.

Findings

To evaluate the rationality of our hand-pose interactions, we visualized the Fitts' Law statistical analysis (Fitts 1954, Rusu et al. 2022) results of 10 participants (t=400) in Figure 5.12 and a histogram of the IP distributions for three hand poses was plotted in Figure 5.13 to show more details. As mentioned in section 5.4.4, we presented three interactive functions with three explored hand poses in their corresponding disentangled spaces from the unified 4D DOF embedding space. Note that the purpose of this study was not to generalise the results, but rather to illustrate the feasibility of our multiple disentanglement hand-pose interactions and to demonstrate how designers could further identify the hand poses that were most appropriate for their interaction prototypes.

By analyzing the linear relationship between ID and MT, we observed that for all three mid-air hand pose interactions, they all conformed to the Fitts' Law, that was, the more difficult the task, the more time they spent. It also demonstrated the feasibility of our three disentangled hand poses in their respective interaction sub-spaces. Specifically, comparing the IP box plots in Figure 5.12 with the IP histogram in Figure 5.13, the IP value of dragging the slider with the PINCH hand pose was higher than the other two interactions, and the

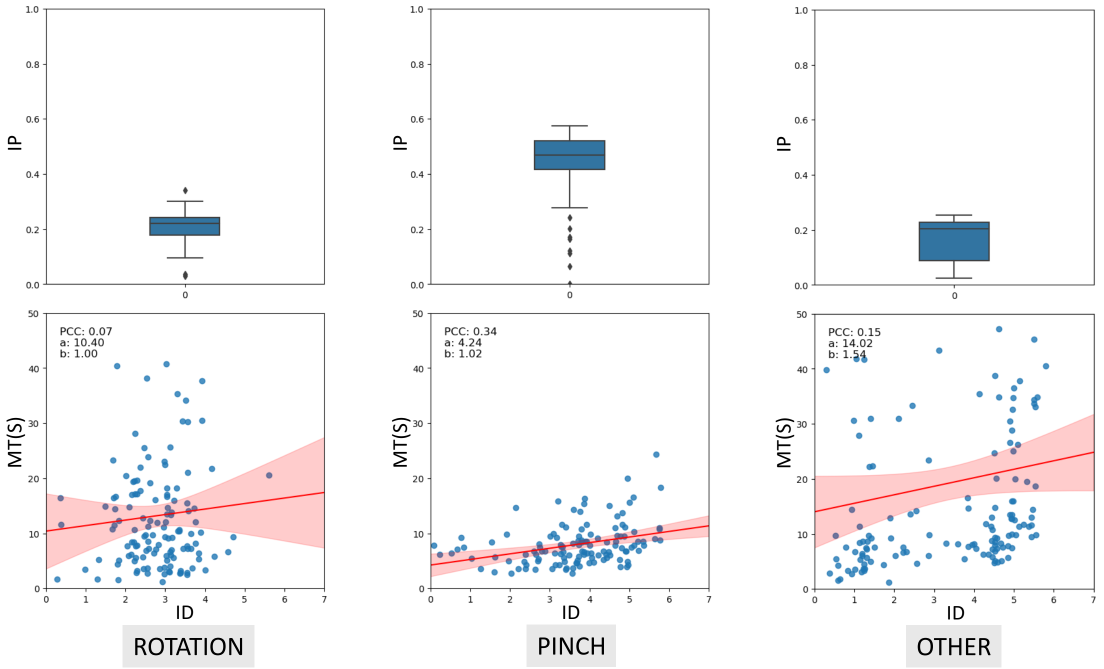


Figure 5.12: Box plots showed the IP of ROTATION, PINCH, and OTHER hand poses in the interaction experiment. The correlation diagrams showed the relations between ID and MT, and the PCC and fitted coefficients a and b were statistically calculated. Each blue point in the figure represented a task, the red line was the fitting line, and the red area is the 95% confidence interval.

majority of the PINCH hand pose tasks had higher IP values, i.e., the bins in orange in Figure 5.13. This indicated that the one-dimensional interaction of the PINCH hand pose is the easiest of the three interactions. Although the ROTATE hand pose for the one-dimensional interaction and the OTHER hand pose for the two-dimensional interaction shown in Figure 5.12 had approximate task difficulty, i.e., IP values, the fitting of IP is not very good on both small and large ID values, which may be related to the limits of experimental data. We are able to see from the more detailed histogram of the IP distributions, Figure 5.13, that the IP values of the ROTATION hand pose in blue are more in line with a Gaussian distribution, which is the type of distribution we expect to see. Whereas the green bins of OTHER hand poses show a bimodal distribution, where one of the centers is similar to the mean value of ROTATION hand pose, but the other center is located on the far left. In other words, there were still many 2D OTHER hand pose interaction tasks with a small IP. One possible reason for presenting this bimodal distribution is that the 2D interaction mode utilized the combination of two DOFs of hand. Without a clear expression of the DOFs of hand, it was difficult to accurately control the movement of two dimensions at the same time. This leads to two possible centers of distribution. Also, due to less experimental data, we did not

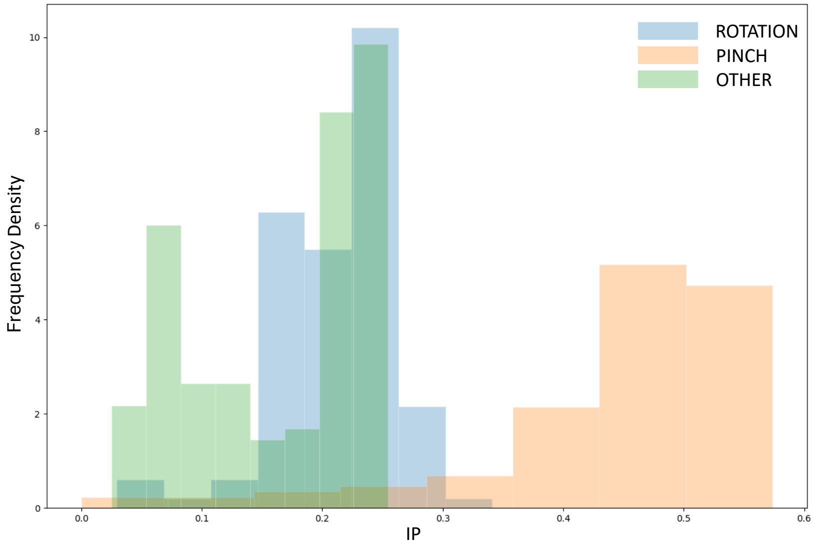
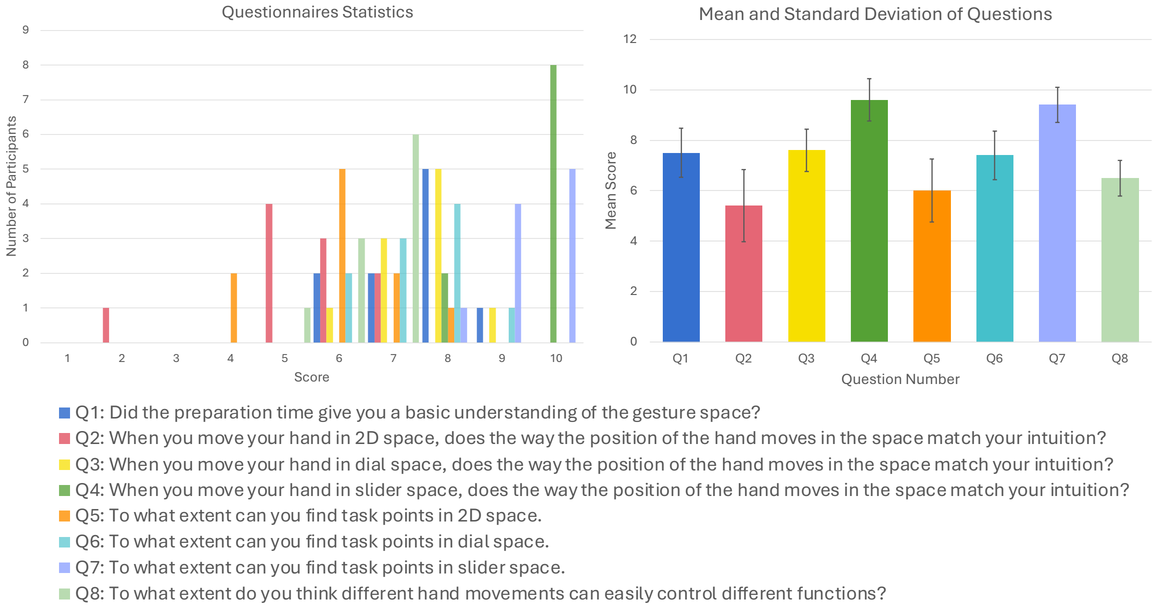


Figure 5.13: The histogram shows the IP distribution of the three hand poses. This figure gives more statistical details about Figure 5.12. Most of the IP values of the PINCH hand pose are larger than those of the other two poses, indicating that our disentangled PINCH hand pose is the easiest one among the three hand poses. The distribution of the ROTATE hand pose presents an approximate Gaussian distribution, while the OTHER hand pose presents an obvious bimodal distribution with a center located to the left of the mean of the ROTATION distribution, which indicates that in our disentanglement model, the one-dimensional ROTATION hand pose is easier than the two-dimensional hand pose.

obtain the Gaussian distributions for the PINCH hand pose and OTHER hand pose that we expected should comply with the law of large numbers. From the overall IP distribution, the interaction of the two one-dimensional hand poses was easier than the interaction of the two-dimensional hand poses. In addition, in terms of linear correlation shown by the Pearson correlation coefficient (PCC), the linear correlation shown by the PCC (0.07) of ROTATION was smaller, which to a certain extent showed that the task difficulty determined by the target point was not the only challenge faced by participants when completing task.

In general, our HandSolo could disentangle the unified hand-pose embedding space into different independent interactive sub-spaces for different DOF hand poses. Through further analysis of the results, we found that one-dimensional linear interaction was more suitable for designing comfortable and easy-to-use simple interactions than two-dimensional interaction, and two-dimensional interaction was suitable for difficult interactions with a certain interaction experience.



- 🟦 Q1: Did the preparation time give you a basic understanding of the gesture space?
- 🟥 Q2: When you move your hand in 2D space, does the way the position of the hand moves in the space match your intuition?
- 🟨 Q3: When you move your hand in dial space, does the way the position of the hand moves in the space match your intuition?
- 🟩 Q4: When you move your hand in slider space, does the way the position of the hand moves in the space match your intuition?
- 🟧 Q5: To what extent can you find task points in 2D space.
- 🟦 Q6: To what extent can you find task points in dial space.
- 🟪 Q7: To what extent can you find task points in slider space.
- 🟩 Q8: To what extent do you think different hand movements can easily control different functions?

Figure 5.14: Statistical analysis of questions related to interactive hand poses. The graph on the left counts the number of participants with different scores for each question. The graph on the right shows the mean score and standard deviation for each question. Where the error bars are the standard deviation. Below the graphs, we provide the questions.

Qualitative Analysis

Although one of the aims of our system design was to have a system evaluation method that did not require questionnaires, we still included questionnaires in this experiment in order to assess the effectiveness of our interactive system and designing methods. Each participant in the experiment was asked to complete a questionnaire after the experiment for approximately 5 minutes. It consisted of 22 mandatory multiple-choice questions and one optional short-answer question. Our questionnaire used a 10-point Likert scale and included the NASA Task Load Index (Hart & Staveland 1988) for workload assessment.

In Figure 5.14 we give statistical graphs for eight of the main questions related to the interaction process of different hand pose movements. We also give both the mean and standard deviation of the scores for each question. Here, Q2-Q4 and Q5-Q7 are two sets of questions related to different hand interaction movements, where Q2 and Q5 are related to interactions in 2D space. Q3 and Q6 are related to ROTATE hand movements, i.e., Dial space. and Q4 and Q7 are statistically related to PINCH hand movements, i.e., interaction feelings while controlling the Slider. The figure on the left shows that most of the ratings are concentrated in the 6-8 range, but there are still a lot of participants (8 participants for Q4 and 5 participants for Q7) who give a score of 10 for the interaction experience of controlling the slider

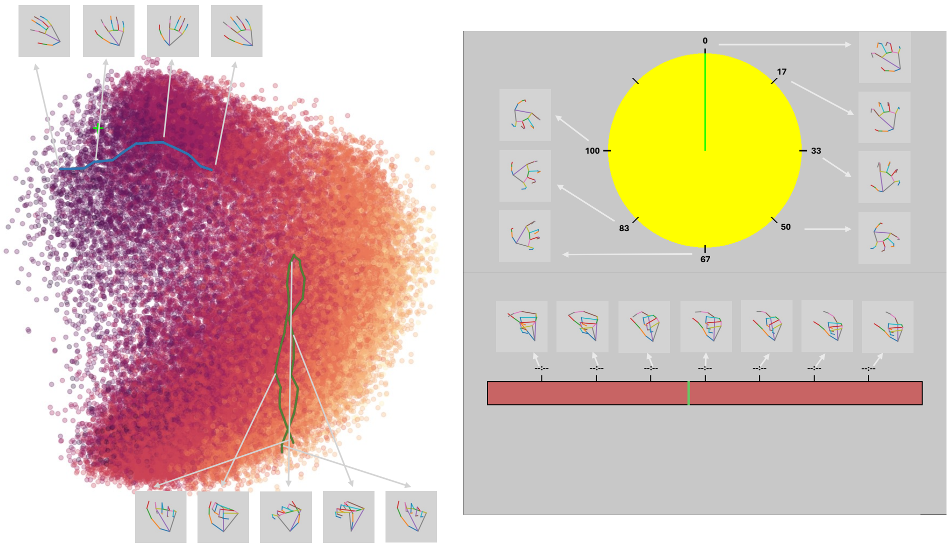


Figure 5.15: A virtual smart music player demonstrates our HandSolo application. The two-dimensional space on the left area is a music space that is freely explored by hand movements, with each point representing a music track. We have provided two different hand movement trajectories to show that we can control hand movement and explore the music space with different DOFs. The top right area shows the volume control dial, we also provide hand pose examples for the corresponding volume and map the hand poses on a $\frac{3}{4}$ circle to avoid uncomfortable poses. The lower area on the right shows the progress bar, controlled by PINCH, for which we have also provided hand pose examples.

with the PINCH. The lowest score of 2 was given to questions related to the exploration of two-dimensional space (1 participant for Q2). From the mean and standard deviation plots on the right, we can clearly see that most of the participants agreed that for controlling the slider with the PINCH was the most intuitive way (Mean (M)=9.6, Standard Deviation (SD)=0.84) of controlling and using PINCH to control was easier to find more target points (M=9.4, SD=0.70). In contrast, the exploration of 2D space presented the greatest challenge to participants, and there was greater divergence in terms of controllability. This is reflected not only in the fact that some of the participants felt that their mid-air hand movements in reality did not give them intuitive 2D-space feedback to a certain degree (M=5.4, SD=1.43), but also in the fact that some of them felt that the OTHER hand movements (M=6.0, SD=1.25) made it harder for them to find target points compared to the PINCH (M=9.4, SD=0.70) and the ROTATE (M=7.4, SD=0.97). The details of the questionnaire and more statistical results can be found in the **Appendix E**.

5.4.5 Virtual Application of HandSolo

To demonstrate the effectiveness and applicability of our interaction system, we designed a virtual smart music player interface for our interaction system as shown in Figure 5.15. This interaction interface can virtually connect our HandSolo interaction system with a smart music system, allowing the user to explore the two-dimensional music space by controlling it with hand DOFs, and to control the volume dials by ROTATE, and can also drag the progress bar through PINCH. In this application, our music data is processed and visualised as described in Section 4.4.4. Based on observations from our VIEs in 5.4.2, we designed the volume control area on a $\frac{3}{4}$ clockwise circle which is a comfortable user movement range, in order to reduce the uncomfortable hand movements that users use to control the dial rotation. At the same time, we used the basic four-dimensional latent space for the demonstration, i.e., one two-dimensional space and two one-dimensional spaces. It is worth noting that this application can still be extended to more dimensions, such as one two-dimensional space and three one-dimensional spaces, but more functions are needed to correspond to it.

In this simple demonstration, our skilled user achieves switching between different functions by changing hand movements freely, which indicates the applicability of our interaction system in a real scenario. We provide examples of hand poses in different positions in Figure 5.15. Where, in the 2D music space, we show two different hand movement trajectories which are controlled by different hand DOFs. Users can explore the 2D music space through more flexible hand poses besides PINCH and ROTATE. Using different Hand DOFs to control and explore demonstrates the flexibility of our interacting system. In addition, the hand movements represented by the blue and green trajectories in the 2D space reveal that our application of the four-dimensional hand DOFs latent space in this application can be extended with more independent hand DOFs, which allows to control more functions without having to re-collect the hand data and train a new interaction model.

5.4.6 Discussion

Through the above experimental setup, we demonstrate that the HandSolo method implements a extensible system in which multiple mid-air hand poses can independently interact with different functions. Moreover, VIEs provides interaction designers with a visual method for analysing the interaction of mid-air hand poses, which can be used to design more comfortable and easy-to-use interaction systems. However, there are still more possibilities to be explored in our design and certain limitations that need to be addressed in future work. On the one hand, when we are designing interactions with a combination of multiple DOFs, such as hand interactions in 2D-DOF embedding spaces, the multiple DOFs allow various hand poses to interact freely in the embedding space, which makes the interaction less stable although more flexible. In future work, we will further investigate the independence of different hand-pose movements in 2D embedding spaces. This can be easily solved by our HandSolo system, but needs to be carefully designed. On the other hand, we have shown that adding some customisation features to corresponding hand-pose interactions can further improve the interaction independence, but this still needs to be carefully designed by the designer. Therefore, how to accurately find new features to enhance the embedding of hand poses is still a development problem that needs to be further enhanced in our future work.

5.5 Limitation

5.5.1 Method Limitation

The current method still has limitations in 2D-space-exploring interactions. Although our method allows for more flexible mid-air control, the experimental results and user feedback show that fatigue is still a problem when users engage in interactions that require multiple hand DOFs. The increased physical and mental demand reduces the efficiency and comfort of long-term use. Compared with simpler 1D interactions, the 2D control still falls short of expectations in terms of ease of use and intuitiveness. This indicates that more work is needed to design interaction techniques that reduce user fatigue while preserving the expressive power of multi-DOF control.

5.5.2 Study Limitation

In this study, our user experiments were based on fixed tasks. Participants were asked to complete pre-defined interaction tasks under controlled conditions. While this design helped to ensure consistency and comparability across users, it restricted the scope of the evaluation. Real-world interactions are often open-ended and dynamic, requiring users to adapt gestures in more flexible ways. By focusing on fixed tasks, the study may not fully capture the variability and unpredictability of natural interaction scenarios. Future studies should therefore include more free-form tasks and open-ended usage contexts to better reflect realistic applications.

5.6 Conclusion

Elicitation of hand poses by designers can lead to poor performance during deployment if it does not take classification performance into account when selecting which combinations of hand poses to use with the given sensors. We used an approach which could gather hand data from proposed hand poses, as well as general random hand movements, and used machine learning to find combinations of useful trajectories which would not interfere with each other during classification. Novel hand-pose movement controls which were not in the original training data can be derived with this approach.

We describe the design and implementation of an adjustable mid-air hand-pose interaction method, named **HandSolo**, which can map high-dimensional data to controllable low-dimensional projections. It is based on a Variational AutoEncoder (VAE) deep neural network, with a disentanglement approach to separate out multiple low-dimensional degrees of hand-freedom. Further, we also propose a new visual interaction evaluation strategy (VIEs) to assist designers in analysing candidate hand poses and deciding between options.

We overcame two key challenges of the existing state of the art, (i) using machine learning techniques for disentangling independent multiple hand pose embedding spaces with different DOFs for different interaction functions through a unified VAE model, and (ii) visually analysing user interaction habits to adapt the interaction system to the trained model's capabilities, in particular to involve designers.

The user experiments provide three hand poses statistic results and the use case of different hand poses interact with a smart music speaker show that our designs not only provide multiple hand-pose interactions from a unified model, but also consider user interaction comfort, design consistency, and model classification capability. The extensibility experiment also provide visualized hand movement curves to show how the tools can support the development of new hand pose controls in multiple hand-pose interactions.

Chapter 6

Conclusion and Discussion

This chapter summarizes the key contributions of this thesis in developing a flexible, interpretable, and extensible mid-air hand pose interaction system. Our research improves interaction stability, usability, and learning efficiency by using low-dimensional representations and disentanglement strategies. Despite these advancements, challenges such as user fatigue and limited multi-modal feedback remain. Future work should focus on reducing interaction fatigue and enhancing accessibility through multi-modal feedback integration. This study can inspire more adaptive and immersive human-computer interaction systems.

6.1 Discussion

With the development of various methods and devices that are used to perform user interaction with smart systems, mid-air gestures for flexible and simple interaction with smart systems face many problems. The high-dimensional inputs of interaction devices and the complexity and variability of scenarios make it difficult to implement an easy-to-use and easy-to-understand interaction system. Among these, we develop a deep-learning-based interactive model that visualises high-dimensional changing hand pose input data in a low-dimensional space. In this way, we implement a flexible, interpretable, visualisation interaction tool. We address the difficulty of learning low-dimensional interactions with high-dimensional inputs while also providing a more flexible and controllable processing paradigm. In order to extend the interaction of mid-air hand poses with smart systems to a wider range of possible smart systems and scenarios, this thesis also proposes an extensible method for disentanglement in low-dimensional spaces. The method is shown to extend the deep-learning-based interactive model, which has been trained with a small amount of data, with different hand degrees of freedom, thus allowing for more flexible hand interactions, as well as providing an extensible interaction model that can be used in other high-dimensional gesture interaction scenarios. Furthermore, based on our proposed framework, our user studies demonstrate the controllability and easy-to-understand properties of our interaction system when users interact with our interaction system. Moreover, we also demonstrate that our framework suggests a visual evaluation approach that can be used as a reference by designers of interactive systems.

In order to address the problems still faced in real-time mid-air gesture interaction systems with respect to the processing of long-sequence-frame input and the interpretability of the interaction, in Chapter 3 **we propose a continuous interaction strategy with visual feedback.** The classification and processing time results in Section 3.4 demonstrate that this strategy has 75.2% of gesture recognition and the 2.4ms latency in response. Using only two video frames of mid-air gestures, Section 3.4.1 employ an autoencoder-based model to compress the high-dimensional hand data into a two-dimensional space. Meanwhile, our classification results in 3.4.2 shown that the auxiliary classification model also performs the classification of the gestures, which improves the processing speed (2.4ms) of the real-time interactions while maintaining the precision (75.2%) of the gesture recognition. In addition, the time and movement trajectories given by the visual interaction user study in Section 3.4.3 confirm that our proposed visual feedback strategy is equipped to improve the usability and interpretability of the interaction and enhance the understanding of user interaction. We provide the user with a visualized and structured feedback of the hand pose space by using the low-dimensional embedding space, which allows the user to understand the movement state of the hand movement in the hand pose space while interacting with the hand movement.

Although in terms of modelling, we can visualise the user's mid-air hand movements in a low-dimensional hand pose space, which improves the user's understanding and efficiency in using the interactive system. However, while making these advances, we found that users still face the problems of jittering, instability and non-smoothness during mid-air hand movements due to the physiological characteristics of the human body and the properties of the interaction device when interacting with the interaction system. At the same time, we found that the black-box models that traditional interaction systems rely on often make it difficult for users to learn how to control the interaction with different hand movements, thus reducing the applicability of the interaction model for complex tasks. Therefore, in Chapter 4, **we design a mid-air hand pose embedding system with stable and smooth design and visual guidance window.** Based on Chapter 3, we processed the data with quaternions in the model in Section 4.4.1 and introduced a regularisation term against jitter while we augmented the gesture data in Section 4.4.3, which reduce the physiological jitter and jitter during interaction caused by the sensor variability from mediapipe. In addition, to further make the interaction process more stable and smooth, we added the post-processing of the mid-air hand pose embedding during the interaction, including the post-processing of stabilising the movement inflection point and the One Euro Filter Casiez et al. (2012), which is the post-processing of smoothing the hand movement. Our users' hand movement curves in Section 4.4.4 show the time of complete tasks and the convex and concave cusps of the curves, which demonstrate the smoothness and ease of use of our stable and smooth design and guidewindow design. Meanwhile, we designed a hand pose guidance window to help users interact with the system. Our experiments' movement curves and hand-pose points in the space in Section 4.4.4 show that users can complete the task in about half the time than without the hand pose guidance window. In other words, users can learn how to interact dynamically in the low-dimensional space of mid-air hand poses more quickly by using the hand pose guidance window to prompt gestures in each direction. Section 4.4.4 provides a multimedia-based application of our HpEIS, which also demonstrates the ability to be used in a real scenario. Our hand movement trajectory curves in the music space demonstrate the flexibility and the ease-of-use ability of our system.

While designing the methodology for user interaction with the smart system, we have considered the aspects of comfort, stability, comprehensibility and ease of learning, but at the same time, we still want to extend our methodology so that we can make the mid-air hand pose interaction more flexible, including deploy and the hand movement, and increase the availability of our methodology in different scenarios. In addition, our current approach follows a traditional basic interaction design process, i.e., a closed-loop approach that involves user experimentation, user surveys, feedback, and system improvements. We wanted to find an easier way to reduce the workload of interaction designers. In Chapter 5, **we develop HandSolo, an extensible low-dimensional hand space disentanglement method, and VIEs, a designers' helping tool.** On the one hand, the method proposed in Section 5.3.1 allows flexible control of multiple hand DOFs through the introduction of a disentanglement penalty term that disentangles the low-dimensional representation in the latent embedding space of the high-dimensional mid-air hand pose input into multiple independent one- or two-dimensional spaces. The method can also be extended to different combinations of other degrees of freedom, as well as other high-dimensional inputs, improving the availability of the interaction system. This availability includes both different input devices and interactions in different environments. The accuracy and Latent-MIG results in 5.3.2 show the disentanglement performance of our model and demonstrate the extensibility of our system. Through the box plot and statistic plot of Fitts' law and statistic plot of questionnaires in 5.4.4, our user study confirms that our proposed interaction system complies with the basic rules of human-computer interaction and that different functions can be controlled by different hand DOFs in the low-dimensional hand pose space. However, there is still difficulty in controlling in 2D space. On the other hand, Section 5.4.1 also propose a Visual Interaction Evaluation strategy (VIEs). The Section 5.4.2 indicates that this strategy helps designers to directly understand user interaction preferences and interaction functions by visualising the difference between the mid-air hand movement curves of users and the experimental baseline curves, which can be used for further system design. The experiment in Section 5.4.3 not only further demonstrates the extensibility of our interactive system, but also demonstrates the stability of multi-dimensional disentanglement, which can provide a reference for designers to design and optimize interactive functions. The questionnaires results in Section 5.4.4 and our virtual application of the system in Section 5.4.5 demonstrate that our disentanglement method with VIEs provides users with a flexible and comfortable interaction experience, and at the same time provides interaction designers with an effective potential strategy for evaluating and optimizing the system that can be used in a real scenario.

6.2 Limitation and Future Work

Although our mid-air hand pose interaction system enables users to interact more flexibly and comfortably with smart systems in a controlled way, we still have some potential problems and limitations. As shown by the experimental results in Section 5.4.4 and its questionnaire in Section 5.4.4 and Appendix E, the problem of fatigue in interaction still exists. Current interaction systems still have high demands on the physical and mental capabilities of the user when faced with multi-hand-DOFs, i.e., the 2D-space-exploring interactions mentioned in the previous section. Moreover, while the ease of control in other 1D-interaction is higher than that of multi-hand- DOFs, still falls short of expectations.

On the other hand, in the process of interacting with a smart system, users usually need clear visual, auditory, or haptic feedback to confirm a successful interaction. Our current mid-air hand pose interaction system provides visual feedback, but still does not provide good accessibility to certain specific users and scenarios, like the interaction feedback for people with visual impairment and the interaction feedback needed by drivers when they are driving. It is obvious that simple visual feedback cannot satisfy the multiple interaction needs of users.

The experiments were conducted with only a small number of participants, all young adults, and only with the right hand. This reduces the diversity of the dataset and may limit the generalizability of the findings to broader populations. In addition, all tasks were performed in a controlled laboratory setting with a plain white background and stable lighting. These conditions reduce noise but fail to capture the complexity of real-world environments.

Addressing the problems and limitations mentioned in our current mid-air hand pose interaction system will provide some benefits in future research studies. Firstly, the most important research direction is how to reduce fatigue when users interact with smart systems, not only limited to physical fatigue, but also focusing on mental fatigue. In the process of interaction between users and smart systems, providing users with a good experience and feeling is the focus in HCI research. Therefore, more research can focus on more controllable hand DOFs disentanglement methods to provide more flexible and easy-to-understand hand movement control methods, especially two-dimensional and multi-dimensional hand DOFs movement interaction methods and the way to assess the systems.

In addition, considering the user's feedback needs, the combination of multiple feedback and response approaches is a way to enhance the user experience. By considering different users and different usage scenarios, multi-modal feedback, such as visual, sound, and haptic, enhances interaction clarity. Effective and appropriate feedback will bring a more immersive experience to the user and increase the accessibility of the interactive system, especially for smart systems applied in the entertainment and leisure domains.

Future studies should also expand the participant type to include users of different ages, left- and right-hand use, and larger sample sizes, in order to improve diversity and robustness. Testing should be carried out in multiple environments instead of just in the lab, which can better reflect real world use. These directions will make the system more generalizable and practical for daily applications.

6.3 Conclusion

In conclusion, this thesis explores a flexible approach for interaction of mid-air hand pose with smart systems, focussing on flexibility, interpretability, and extensibility of mid-air hand pose interactions. While there is still some user comfort and user feedback research to continue to explore, our research shows that the method based on interaction with smart systems proposed in this thesis can solve the problems of flexibility and stability that exist in mid-air hand pose interaction. At the same time, the user will learn how the hand space works in the interaction system through the interaction process, which will reduce the time required to learn and be familiar with the interaction system. This improves the interpretability and user-learnability of the interactive system. This work also highlights that in terms of the model of the interaction system, the flexible framework we propose can also be transferred to other scenarios and devices. This study demonstrates the development potential of our interaction system framework for transferring to more application scenarios and will provide design ideas for more flexible interactions with mid-air hand poses.

Appendix A

Stream Media Data

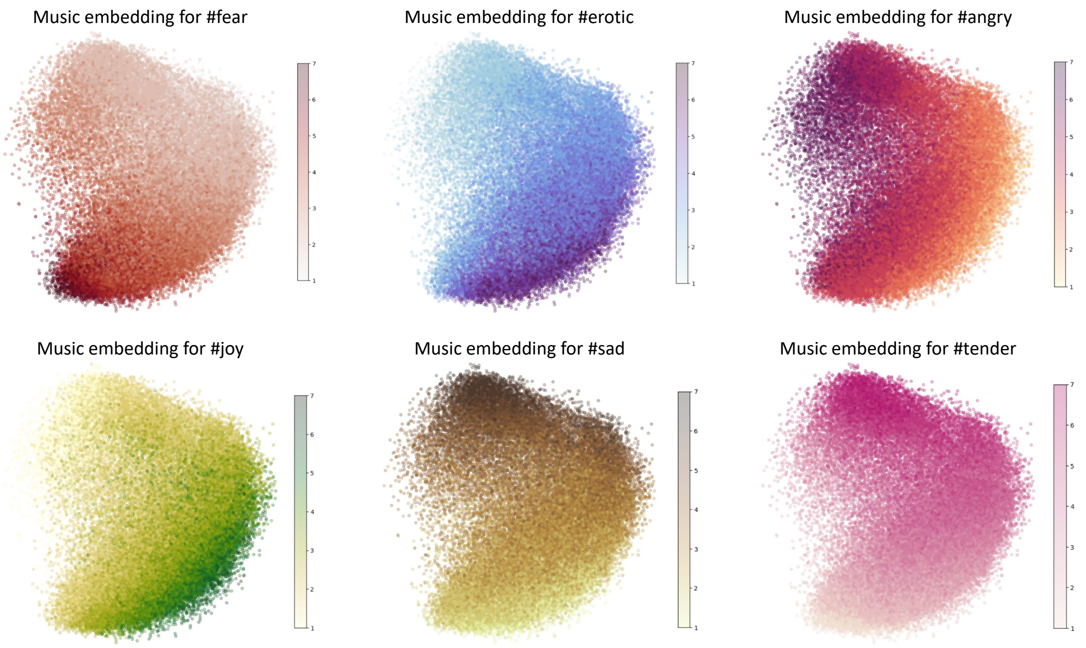


Figure A.1: The music embedding spaces for each emotion.

Figure A.1 shows our music embedding space. Similar to the VAE of the hand pose space, we use the same VAE setup, specifically, each MLP contains 4 fully connected layers, where the neuron numbers in the encoder are 128, 96, 64 and 2, respectively and the reverse in the decoder. The 2-D latent embeddings are normalized between 0 and 1. The only difference is that our original music features have 34 dimensions, and these features include scoring situations for style type, emotion type, etc., with scores ranging from 1 to 7. In addition, we do not perform augmentation operations on the music space embedding. In the figure, we colour the music space with scores for different emotions, and we consider 6 emotions, fear, erotic, angry, joy, sad, and tender.

Appendix B

Interaction Interface

In Experiment II, each time the participants do a target-point-finding task, they will see the target point appear in the corresponding area, as shown in Figure B.1. Each area shows the minimum size of one of the target points, and we display the three target point sizes in equal proportion next to it.

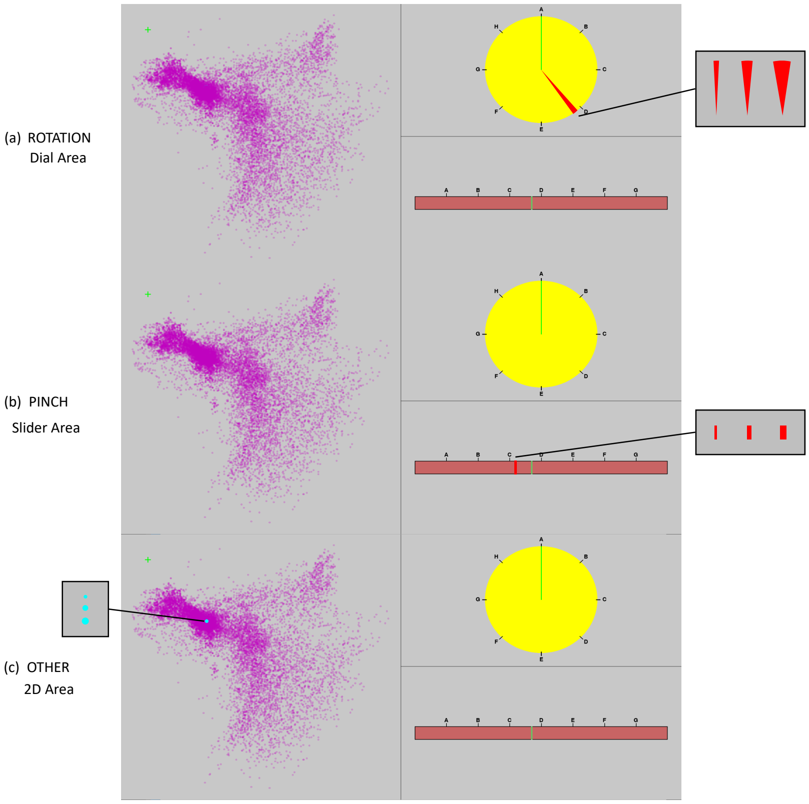


Figure B.1: Three different interfaces the participants saw when they did different target-point-finding tasks in different areas. Different target sizes for each area were also provided.

Appendix C

Disentangled DOF Spaces with Different Settings

We provide DOF-space figures for 4D latent space, including with and without disentanglement, and add extra hand features. And a disentangled DOF-space figure with extra hand features for 5D latent space.

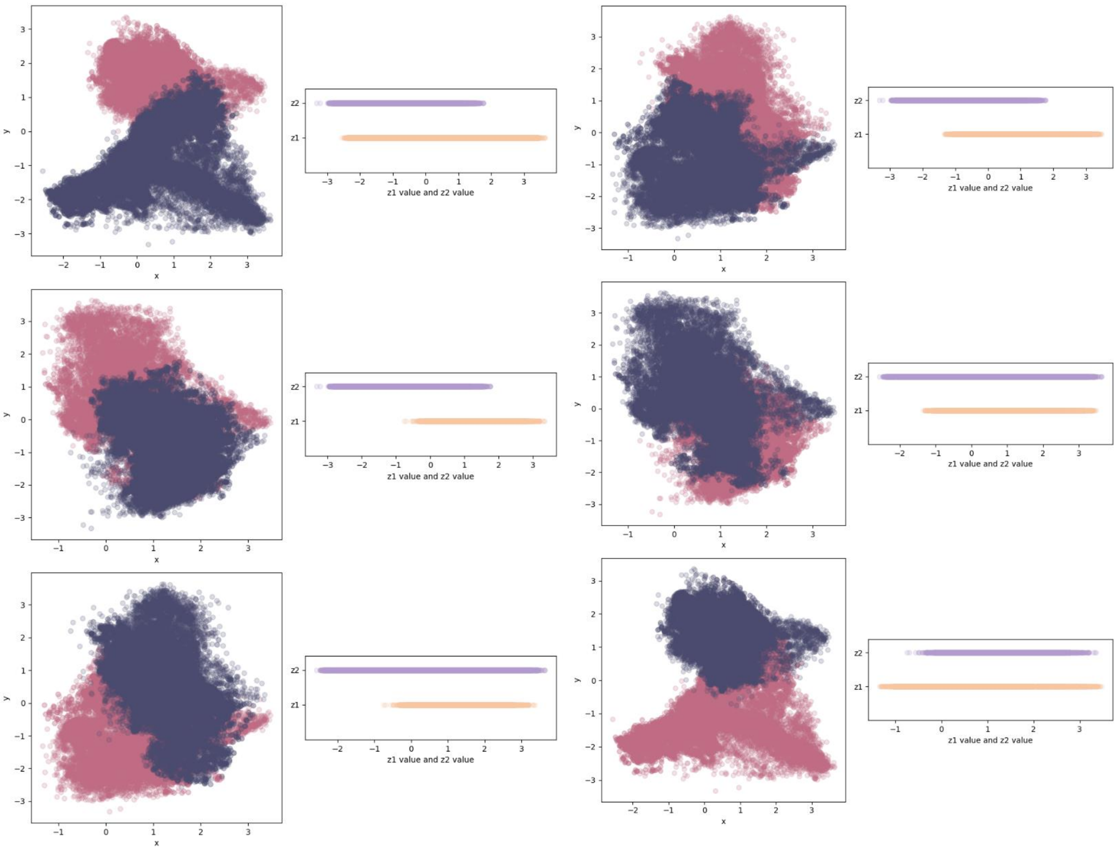


Figure C.1: 4D DOF spaces without disentanglement. The extracted 1D embeddings can be seen to have a high amount of overlap making it infeasible to utilize the hand poses for interactive control in practical use cases such as controlling a dial or slider.

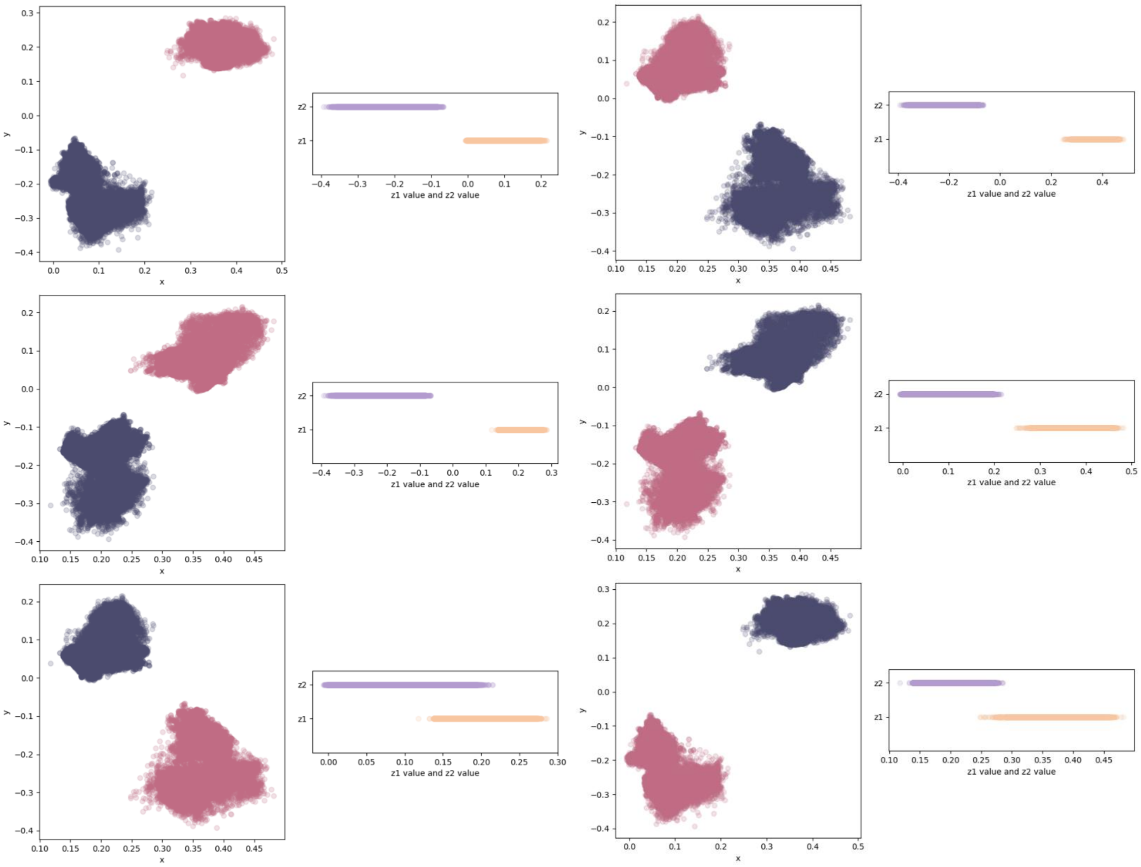


Figure C.2: 4D DOF spaces with disentanglement. The hand poses are disentangled in the latent space allowing for more fine-grained control of systems interfaced with our embeddings.

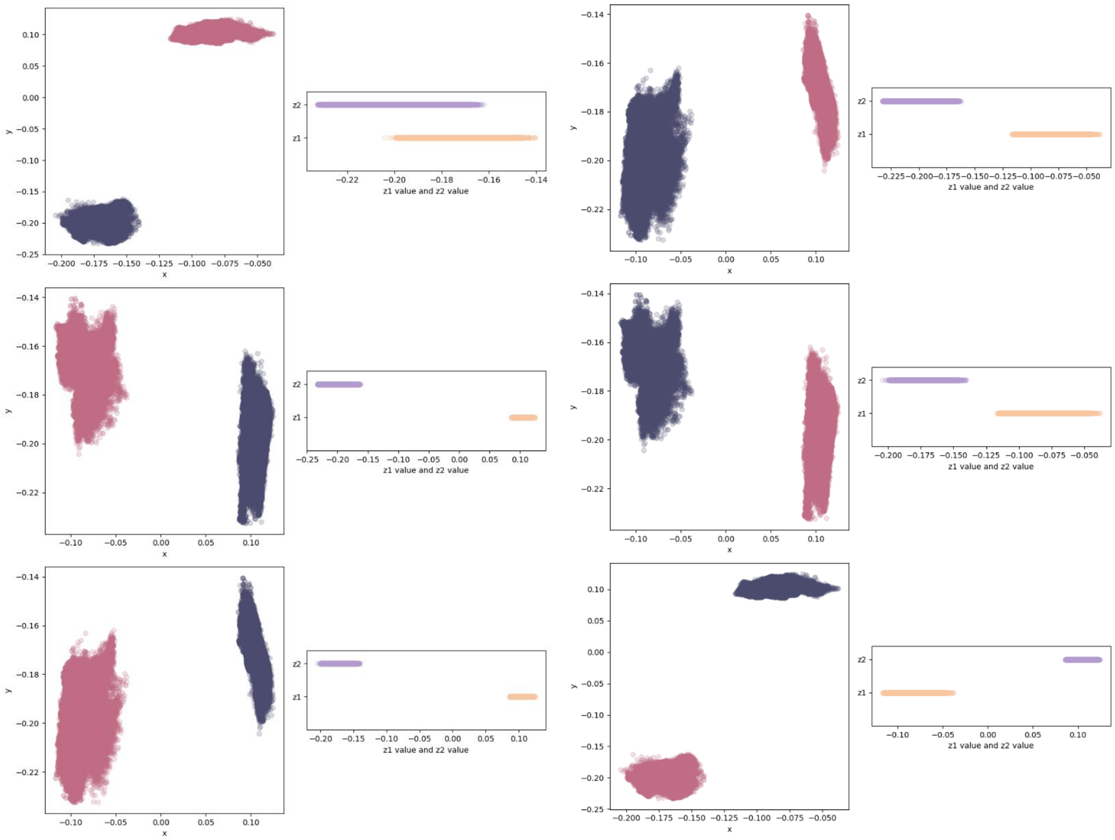


Figure C.3: 4D DOF spaces with disentanglement and extra hand features. The extra hand features are added to make the disentanglement results for specific hand DOFs clearer than in Figure C.2, thus making the corresponding interactions more robust and comfortable.

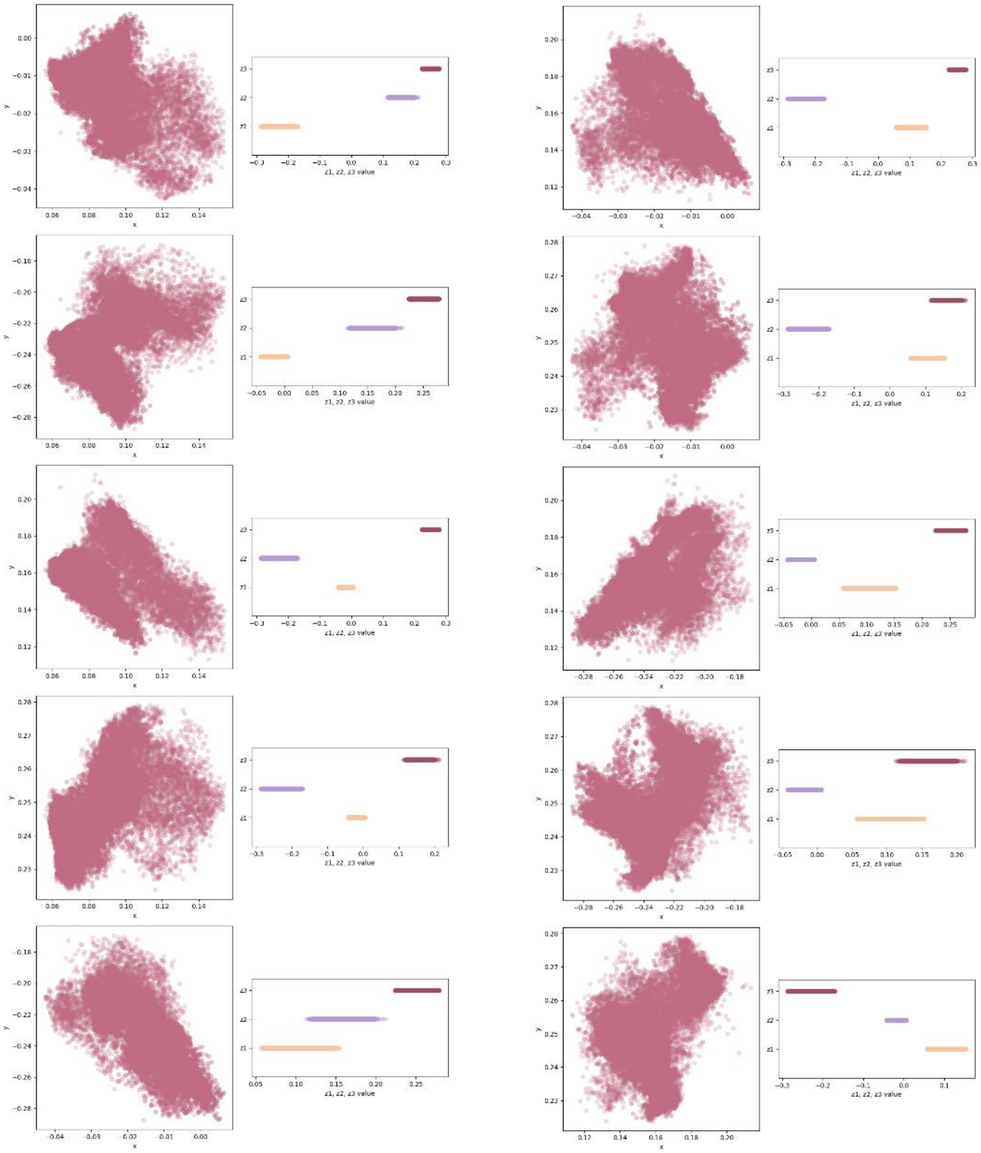
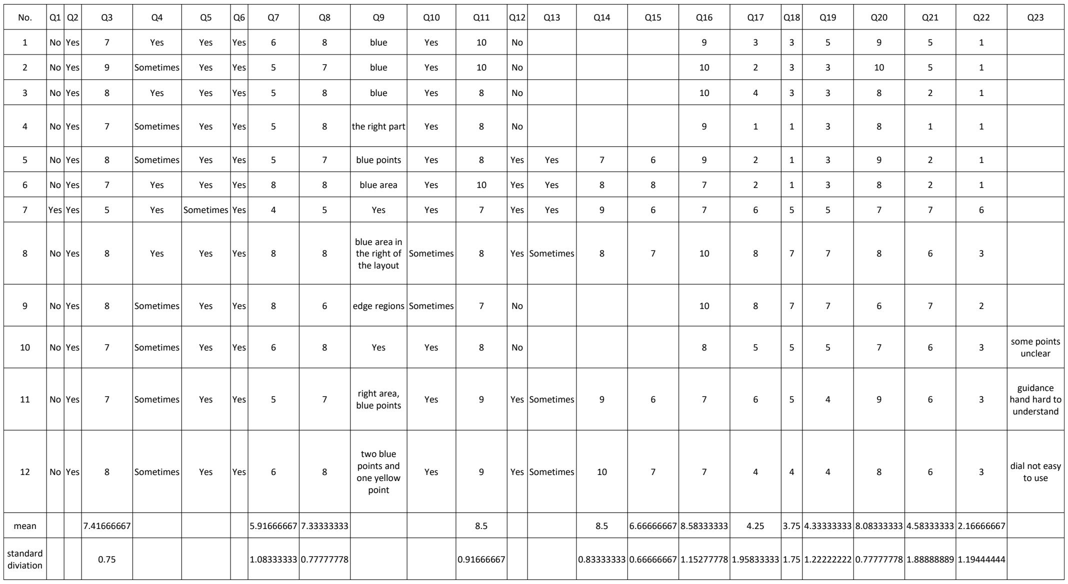


Figure C.4: 5D DOF spaces with disentanglement and extra hand features. Our HandSolo model is extensible and still maintains good disentanglement properties with similar hand DOFs as before.

Appendix D

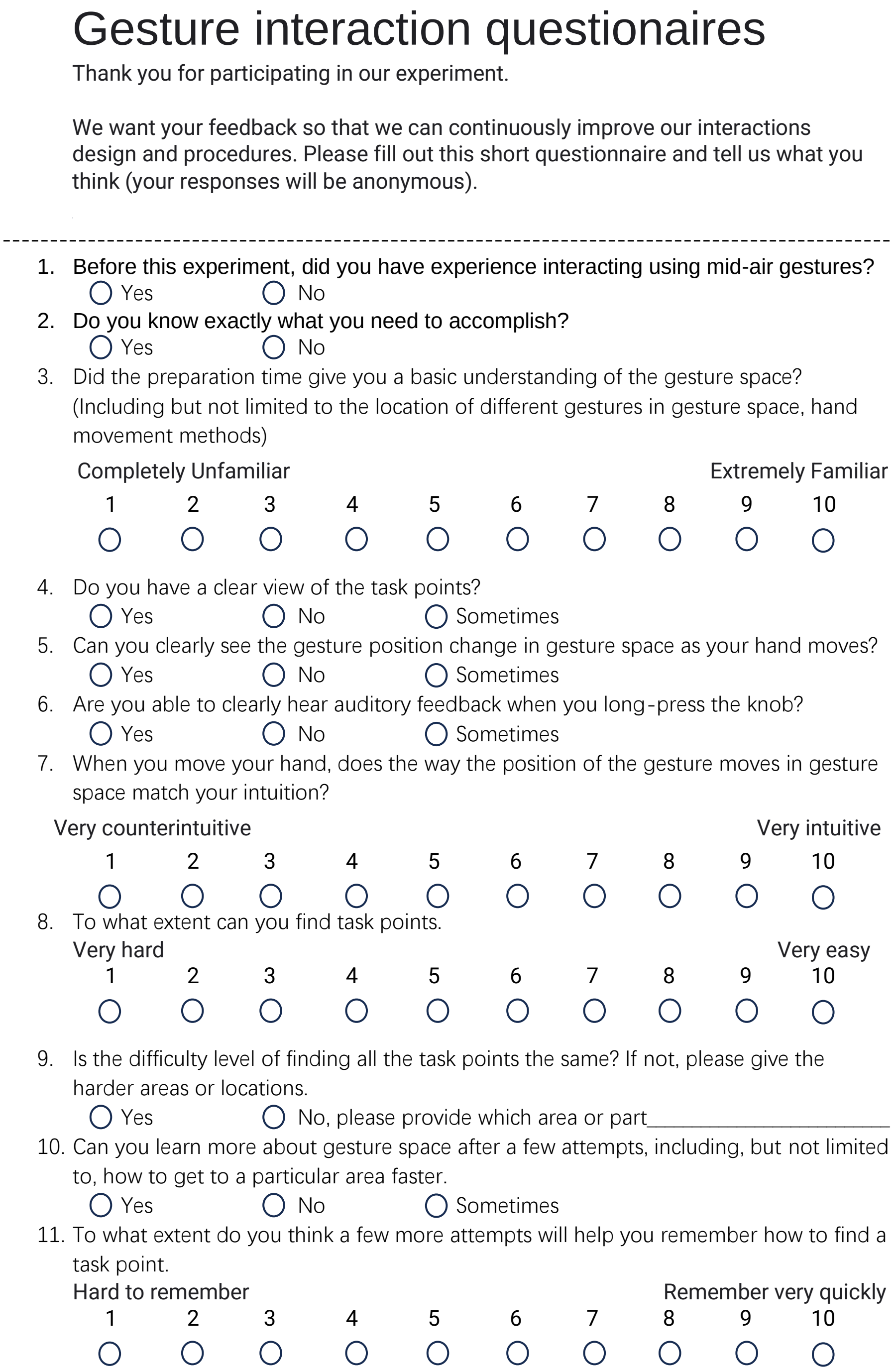
Questionnaire with more details for HpEIS

Fig. D.1 shows the statistical information sheet of the questionnaire, which contains all Likert linear scale questions and short answer questions. After this, we provide our questionnaires.



No.	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23
1	No	Yes	7	Yes	Yes	Yes	6	8	blue	Yes	10	No				9	3	3	5	9	5	1	
2	No	Yes	9	Sometimes	Yes	Yes	5	7	blue	Yes	10	No				10	2	3	3	10	5	1	
3	No	Yes	8	Yes	Yes	Yes	5	8	blue	Yes	8	No				10	4	3	3	8	2	1	
4	No	Yes	7	Sometimes	Yes	Yes	5	8	the right part	Yes	8	No				9	1	1	3	8	1	1	
5	No	Yes	8	Sometimes	Yes	Yes	5	7	blue points	Yes	8	Yes	Yes	7	6	9	2	1	3	9	2	1	
6	No	Yes	7	Yes	Yes	Yes	8	8	blue area	Yes	10	Yes	Yes	8	8	7	2	1	3	8	2	1	
7	Yes	Yes	5	Yes	Sometimes	Yes	4	5	Yes	Yes	7	Yes	Yes	9	6	7	6	5	5	7	7	6	
8	No	Yes	8	Yes	Yes	Yes	8	8	blue area in the right of the layout	Sometimes	8	Yes	Sometimes	8	7	10	8	7	7	8	6	3	
9	No	Yes	8	Sometimes	Yes	Yes	8	6	edge regions	Sometimes	7	No				10	8	7	7	6	7	2	
10	No	Yes	7	Sometimes	Yes	Yes	6	8	Yes	Yes	8	No				8	5	5	5	7	6	3	some points unclear
11	No	Yes	7	Sometimes	Yes	Yes	5	7	right area, blue points	Yes	9	Yes	Sometimes	9	6	7	6	5	4	9	6	3	guidance hand hard to understand
12	No	Yes	8	Sometimes	Yes	Yes	6	8	two blue points and one yellow point	Yes	9	Yes	Sometimes	10	7	7	4	4	4	8	6	3	dial not easy to use
mean			7.41666667				5.91666667	7.33333333			8.5			8.5	6.66666667	8.58333333	4.25	3.75	4.33333333	8.08333333	4.58333333	2.16666667	
standard diviation			0.75				1.08333333	0.77777778			0.91666667			0.83333333	0.66666667	1.15277778	1.95833333	1.75	1.22222222	0.77777778	1.88888889	1.19444444	

Figure D.1: The results table of questionnaires.



Gesture interaction questionaires

Thank you for participating in our experiment.

We want your feedback so that we can continuously improve our interactions design and procedures. Please fill out this short questionnaire and tell us what you think (your responses will be anonymous).

---

1. Before this experiment, did you have experience interacting using mid-air gestures?
   ○ Yes ○ No
2. Do you know exactly what you need to accomplish?
   ○ Yes ○ No
3. Did the preparation time give you a basic understanding of the gesture space? (Including but not limited to the location of different gestures in gesture space, hand movement methods)

Completely Unfamiliar Extremely Familiar

1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○

4. Do you have a clear view of the task points?
   ○ Yes ○ No ○ Sometimes
5. Can you clearly see the gesture position change in gesture space as your hand moves?
   ○ Yes ○ No ○ Sometimes
6. Are you able to clearly hear auditory feedback when you long-press the knob?
   ○ Yes ○ No ○ Sometimes
7. When you move your hand, does the way the position of the gesture moves in gesture space match your intuition?

Very counterintuitive Very intuitive

1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○

8. To what extent can you find task points.

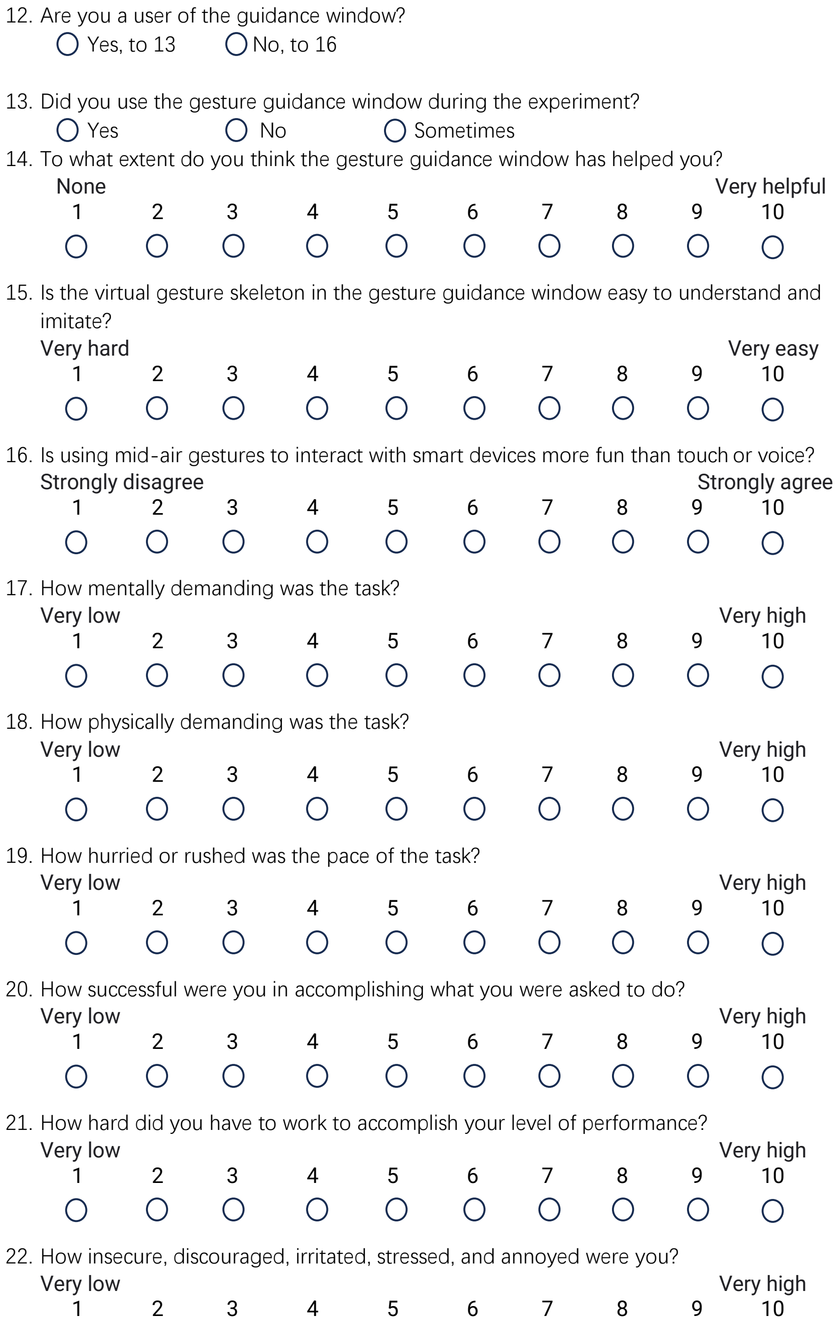
Very hard Very easy

1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○

9. Is the difficulty level of finding all the task points the same? If not, please give the harder areas or locations.
   ○ Yes ○ No, please provide which area or part________________________
10. Can you learn more about gesture space after a few attempts, including, but not limited to, how to get to a particular area faster.
   ○ Yes ○ No ○ Sometimes
11. To what extent do you think a few more attempts will help you remember how to find a task point.

Hard to remember Remember very quickly

1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○



12. Are you a user of the guidance window?

○ Yes, to 13 ○ No, to 16

13. Did you use the gesture guidance window during the experiment?

○ Yes ○ No ○ Sometimes

14. To what extent do you think the gesture guidance window has helped you?

None Very helpful

1 2 3 4 5 6 7 8 9 10

○ ○ ○ ○ ○ ○ ○ ○ ○ ○

15. Is the virtual gesture skeleton in the gesture guidance window easy to understand and imitate?

Very hard Very easy

1 2 3 4 5 6 7 8 9 10

○ ○ ○ ○ ○ ○ ○ ○ ○ ○

16. Is using mid-air gestures to interact with smart devices more fun than touch or voice?

Strongly disagree Strongly agree

1 2 3 4 5 6 7 8 9 10

○ ○ ○ ○ ○ ○ ○ ○ ○ ○

17. How mentally demanding was the task?

Very low Very high

1 2 3 4 5 6 7 8 9 10

○ ○ ○ ○ ○ ○ ○ ○ ○ ○

18. How physically demanding was the task?

Very low Very high

1 2 3 4 5 6 7 8 9 10

○ ○ ○ ○ ○ ○ ○ ○ ○ ○

19. How hurried or rushed was the pace of the task?

Very low Very high

1 2 3 4 5 6 7 8 9 10

○ ○ ○ ○ ○ ○ ○ ○ ○ ○

20. How successful were you in accomplishing what you were asked to do?

Very low Very high

1 2 3 4 5 6 7 8 9 10

○ ○ ○ ○ ○ ○ ○ ○ ○ ○

21. How hard did you have to work to accomplish your level of performance?

Very low Very high

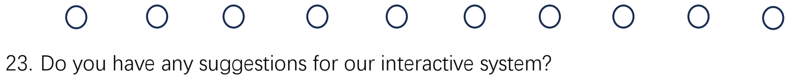
1 2 3 4 5 6 7 8 9 10

○ ○ ○ ○ ○ ○ ○ ○ ○ ○

22. How insecure, discouraged, irritated, stressed, and annoyed were you?

Very low Very high

1 2 3 4 5 6 7 8 9 10



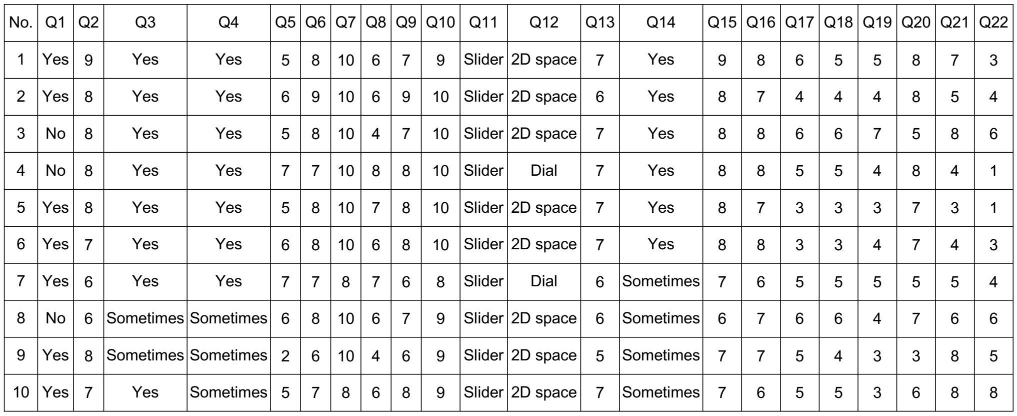
○ ○ ○ ○ ○ ○ ○ ○ ○ ○

23. Do you have any suggestions for our interactive system?

Appendix E

Questionnaire with more details for HandSolo

In this appendix, we give the full results of the questionnaire, as shown in Figure E.1. As well as a statistical chart for the NASA Task Load Index of the questionnaire as shown in Figure E.2. We have included our questionnaire at the end of this appendix. In Figure E.2, an average mental (M=4.8, SD=1.14) and physical (M=4.6, SD=1.07) requirements are shown. The speed requirements (M=4.2, SD=1.23) for our experiments shown in the graph are also not very high. Although the average level of completion as perceived by the participants was not low (M=6.4, SD=1.65), there were still participants who were dissatisfied with the level of task completion and most of them felt that they needed a lot of hard work (M=5.8, SD=1.87) to complete the task. Combined with the experimental results in the section 5.4.4, we believe that the main source of difficulty is the fact that the process of exploring the two-dimensional space can cause some distress to the participants.



No.	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22
1	Yes	9	Yes	Yes	5	8	10	6	7	9	Slider	2D space	7	Yes	9	8	6	5	5	8	7	3
2	Yes	8	Yes	Yes	6	9	10	6	9	10	Slider	2D space	6	Yes	8	7	4	4	4	8	5	4
3	No	8	Yes	Yes	5	8	10	4	7	10	Slider	2D space	7	Yes	8	8	6	6	7	5	8	6
4	No	8	Yes	Yes	7	7	10	8	8	10	Slider	Dial	7	Yes	8	8	5	5	4	8	4	1
5	Yes	8	Yes	Yes	5	8	10	7	8	10	Slider	2D space	7	Yes	8	7	3	3	3	7	3	1
6	Yes	7	Yes	Yes	6	8	10	6	8	10	Slider	2D space	7	Yes	8	8	3	3	4	7	4	3
7	Yes	6	Yes	Yes	7	7	8	7	6	8	Slider	Dial	6	Sometimes	7	6	5	5	5	5	5	4
8	No	6	Sometimes	Sometimes	6	8	10	6	7	9	Slider	2D space	6	Sometimes	6	7	6	6	4	7	6	6
9	Yes	8	Sometimes	Sometimes	2	6	10	4	6	9	Slider	2D space	5	Sometimes	7	7	5	4	3	3	8	5
10	Yes	7	Yes	Sometimes	5	7	8	6	8	9	Slider	2D space	7	Sometimes	7	6	5	5	3	6	8	8

Figure E.1: The results table of HandSolo questionnaires.

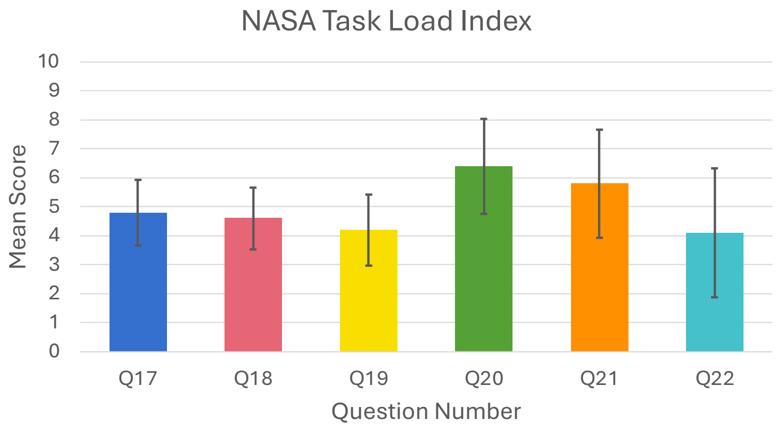
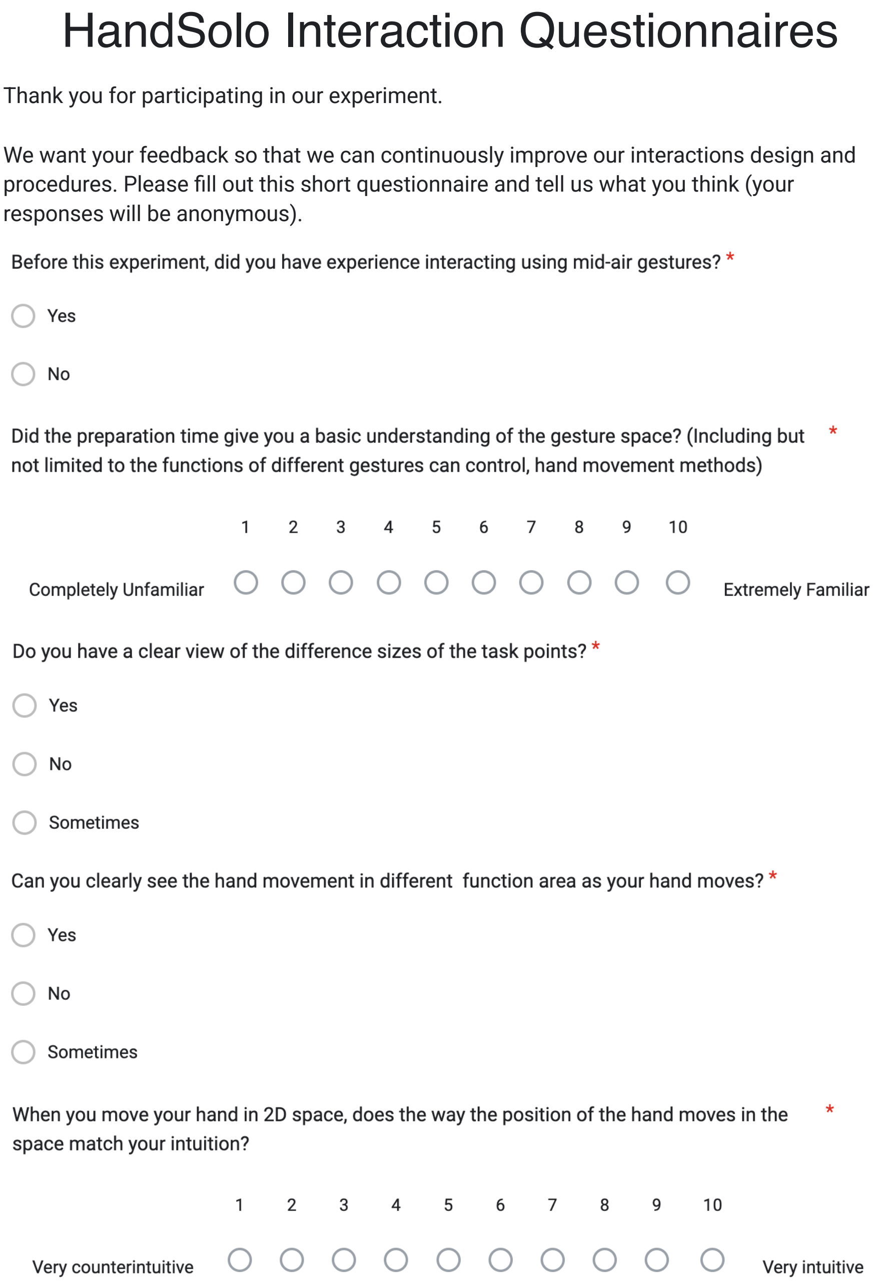


Figure E.2: The Nasa Task Load Index results mean and standard deviation.



HandSolo Interaction Questionnaires

Thank you for participating in our experiment.

We want your feedback so that we can continuously improve our interactions design and procedures. Please fill out this short questionnaire and tell us what you think (your responses will be anonymous).

Before this experiment, did you have experience interacting using mid-air gestures? *

○ Yes

○ No

Did the preparation time give you a basic understanding of the gesture space? (Including but not limited to the functions of different gestures can control, hand movement methods) *

	1	2	3	4	5	6	7	8	9	10	
Completely Unfamiliar	○	○	○	○	○	○	○	○	○	○	Extremely Familiar

Do you have a clear view of the difference sizes of the task points? *

○ Yes

○ No

○ Sometimes

Can you clearly see the hand movement in different function area as your hand moves? *

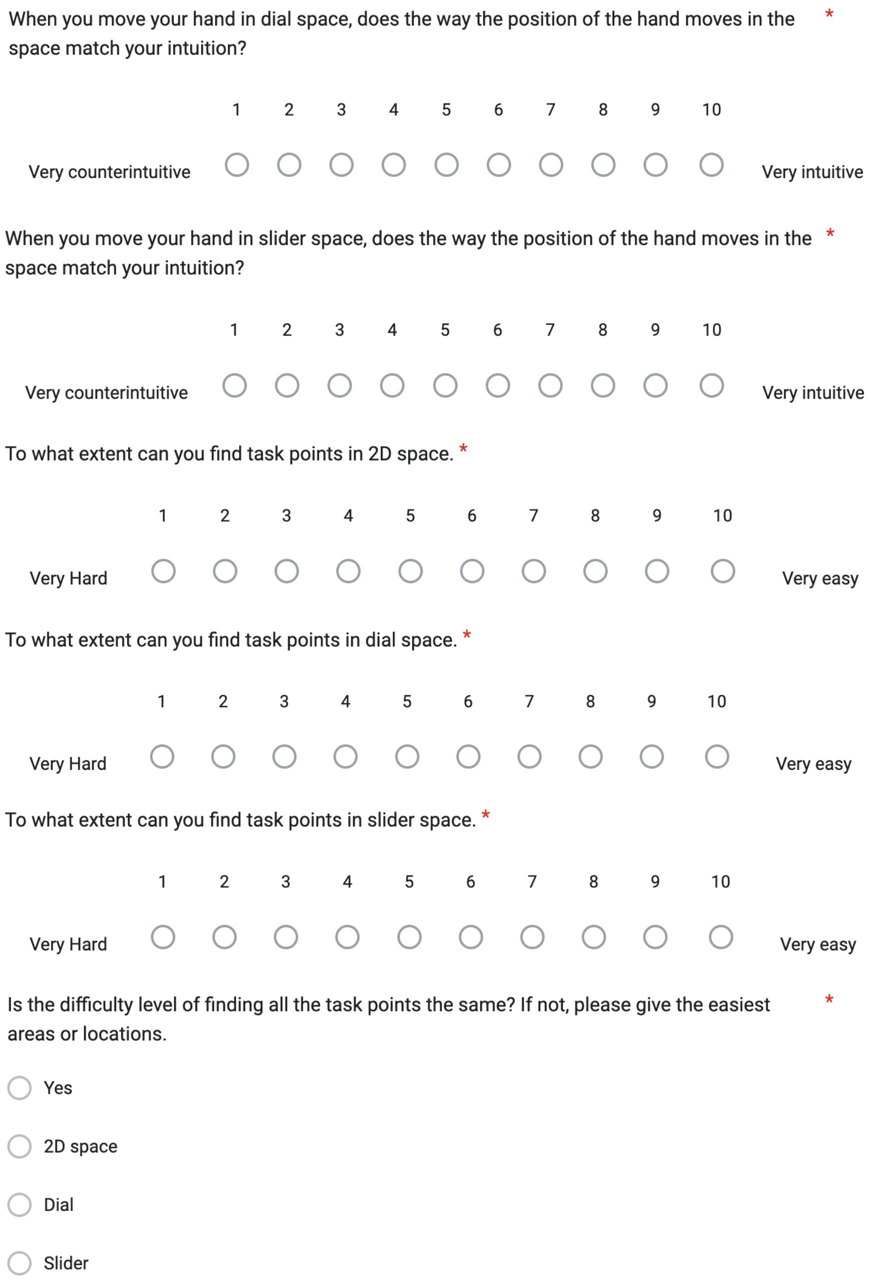
○ Yes

○ No

○ Sometimes

When you move your hand in 2D space, does the way the position of the hand moves in the space match your intuition? *

	1	2	3	4	5	6	7	8	9	10	
Very counterintuitive	○	○	○	○	○	○	○	○	○	○	Very intuitive



When you move your hand in dial space, does the way the position of the hand moves in the space match your intuition? *

	1	2	3	4	5	6	7	8	9	10	
Very counterintuitive	○	○	○	○	○	○	○	○	○	○	Very intuitive

When you move your hand in slider space, does the way the position of the hand moves in the space match your intuition? *

	1	2	3	4	5	6	7	8	9	10	
Very counterintuitive	○	○	○	○	○	○	○	○	○	○	Very intuitive

To what extent can you find task points in 2D space. *

	1	2	3	4	5	6	7	8	9	10	
Very Hard	○	○	○	○	○	○	○	○	○	○	Very easy

To what extent can you find task points in dial space. *

	1	2	3	4	5	6	7	8	9	10	
Very Hard	○	○	○	○	○	○	○	○	○	○	Very easy

To what extent can you find task points in slider space. *

	1	2	3	4	5	6	7	8	9	10	
Very Hard	○	○	○	○	○	○	○	○	○	○	Very easy

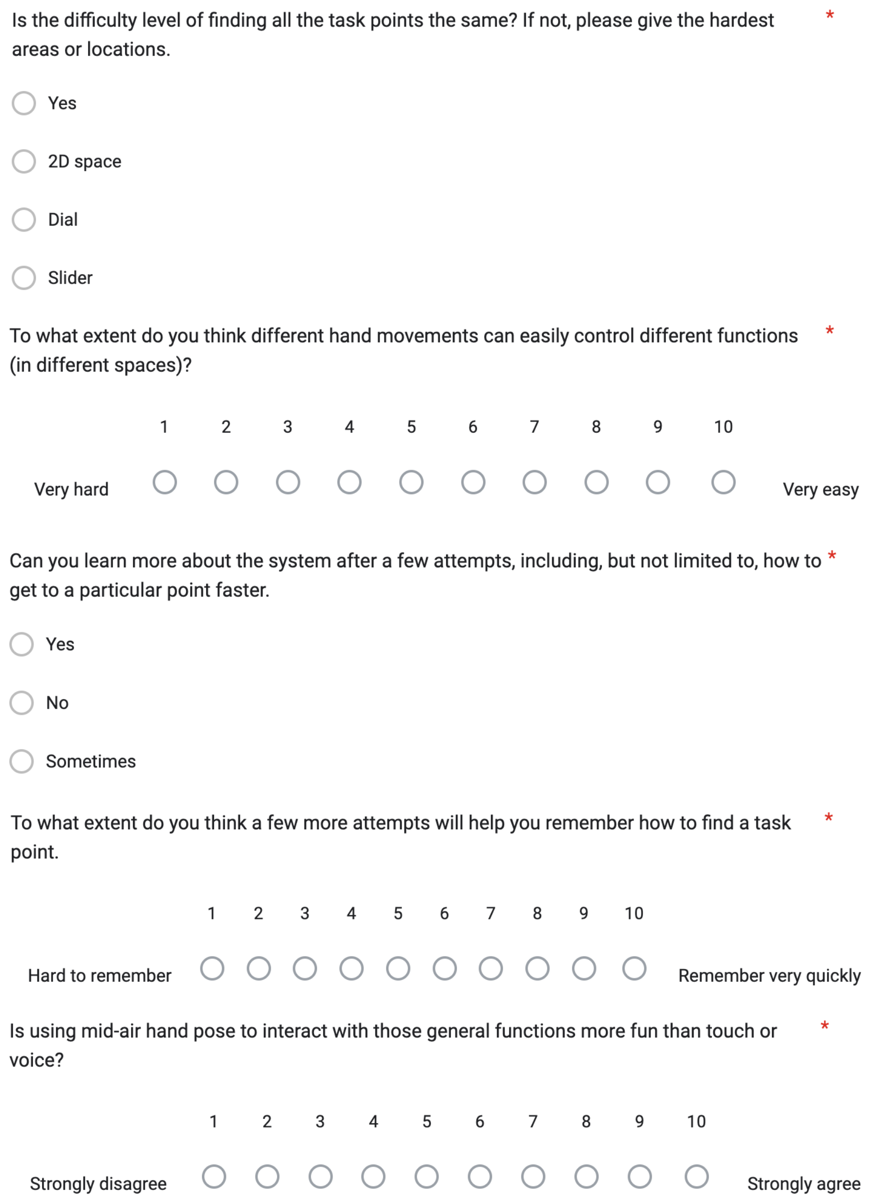
Is the difficulty level of finding all the task points the same? If not, please give the easiest areas or locations. *

○ Yes

○ 2D space

○ Dial

○ Slider



Is the difficulty level of finding all the task points the same? If not, please give the hardest areas or locations. *

○ Yes

○ 2D space

○ Dial

○ Slider

To what extent do you think different hand movements can easily control different functions (in different spaces)? *

	1	2	3	4	5	6	7	8	9	10	
Very hard	○	○	○	○	○	○	○	○	○	○	Very easy

Can you learn more about the system after a few attempts, including, but not limited to, how to get to a particular point faster. *

○ Yes

○ No

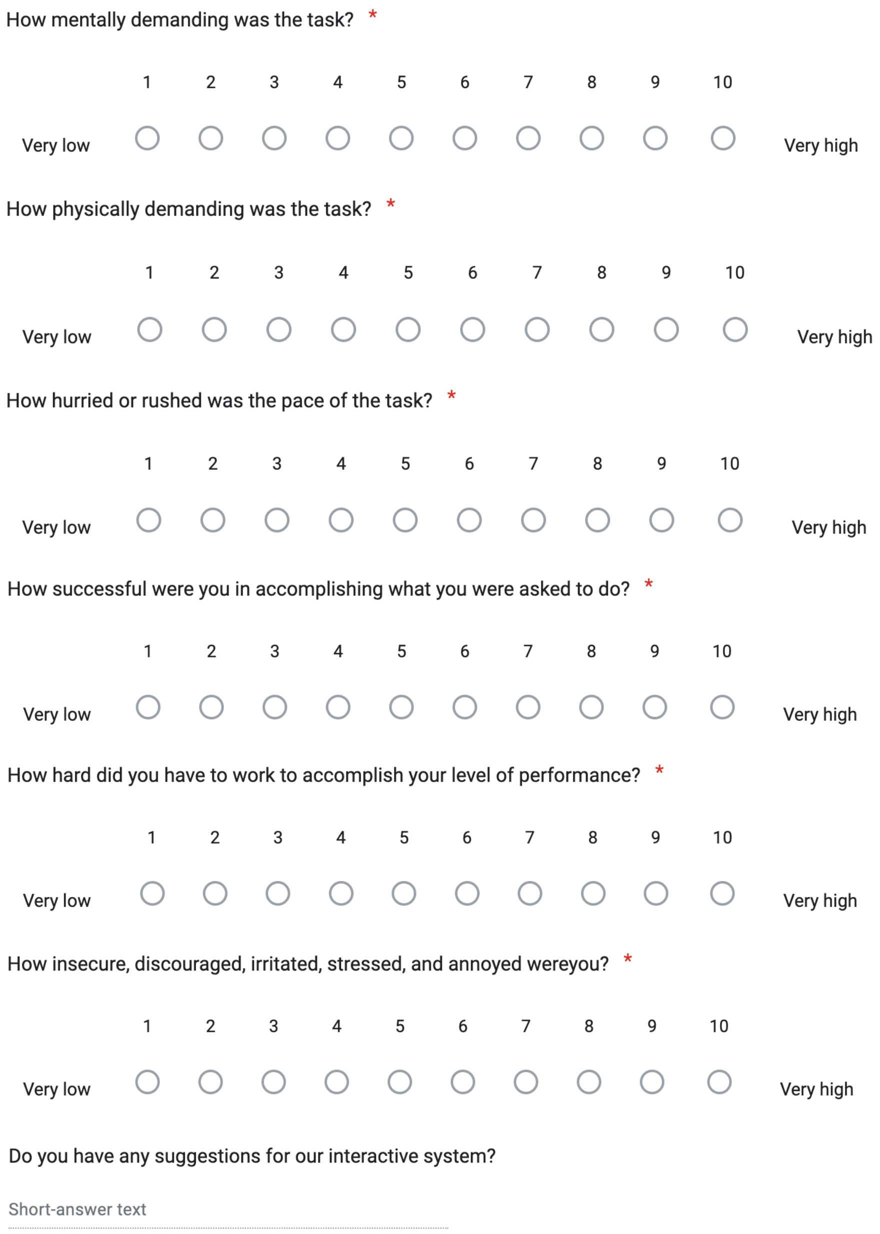
○ Sometimes

To what extent do you think a few more attempts will help you remember how to find a task point. *

	1	2	3	4	5	6	7	8	9	10	
Hard to remember	○	○	○	○	○	○	○	○	○	○	Remember very quickly

Is using mid-air hand pose to interact with those general functions more fun than touch or voice? *

	1	2	3	4	5	6	7	8	9	10	
Strongly disagree	○	○	○	○	○	○	○	○	○	○	Strongly agree



How mentally demanding was the task? *

	1	2	3	4	5	6	7	8	9	10	
Very low	○	○	○	○	○	○	○	○	○	○	Very high

How physically demanding was the task? *

	1	2	3	4	5	6	7	8	9	10	
Very low	○	○	○	○	○	○	○	○	○	○	Very high

How hurried or rushed was the pace of the task? *

	1	2	3	4	5	6	7	8	9	10	
Very low	○	○	○	○	○	○	○	○	○	○	Very high

How successful were you in accomplishing what you were asked to do? *

	1	2	3	4	5	6	7	8	9	10	
Very low	○	○	○	○	○	○	○	○	○	○	Very high

How hard did you have to work to accomplish your level of performance? *

	1	2	3	4	5	6	7	8	9	10	
Very low	○	○	○	○	○	○	○	○	○	○	Very high

How insecure, discouraged, irritated, stressed, and annoyed wereyou? *

	1	2	3	4	5	6	7	8	9	10	
Very low	○	○	○	○	○	○	○	○	○	○	Very high

Do you have any suggestions for our interactive system?

Short-answer text

Bibliography

Abowd, G. D., Dey, A. K., Brown, P. J., Davies, N., Smith, M. & Steggles, P. (1999), Towards a better understanding of context and context-awareness, *in* 'International symposium on handheld and ubiquitous computing', Springer, pp. 304–307.

Ahmed, S., Kallu, K. D., Ahmed, S. & Cho, S. H. (2021), 'Hand gestures recognition using radar sensors for human-computer-interaction: A review', *Remote Sensing* **13**(3), 527.

Ahmed, S., Kim, W., Park, J. & Cho, S. H. (2022), 'Radar based air-writing gesture recognition using a novel multi-stream CNN approach', *IEEE Internet of Things Journal* .

Aloba, A., Woodward, J. & Anthony, L. (2020), Filterjoint: Toward an understanding of whole-body gesture articulation, *in* 'Proceedings of the 2020 International Conference on Multimodal Interaction', pp. 213–221.

Alonazi, M., Ansar, H., Al Mudawi, N., Alotaibi, S. S., Almujally, N. A., Alazeb, A., Jalal, A., Kim, J. & Min, M. (2023), 'Smart healthcare hand gesture recognition using cnn-based detector and deep belief network', *IEEE Access* **11**, 84922–84933.

Alver, C. (2007), 'Voice biometrics in financial services', *Journal of Financial Services Technology, The* **1**(1), 75–81.

Andersen, N. E., Dahmani, L., Konishi, K. & Bohbot, V. D. (2012), 'Eye tracking, strategies, and sex differences in virtual navigation', *Neurobiology of learning and memory* **97**(1), 81–89.

Anthony, L., Brown, Q., Nias, J., Tate, B. & Mohan, S. (2012), Interaction and recognition challenges in interpreting children's touch and gesture input on mobile devices, *in* 'Proceedings of the 2012 ACM international conference on Interactive tabletops and surfaces', pp. 225–234.

Arsenault, D. (2014), A quaternion-based motion tracking and gesture recognition system using wireless inertial sensors, PhD thesis, Carleton University.

Aula, A. & Surakka, V. (2002), Auditory emotional feedback facilitates human-computer interaction, *in* ‘People and Computers XVI-Memorable Yet Invisible’, Springer, pp. 337–349.

Ballati, F., Corno, F. & De Russis, L. (2018), “hey siri, do you understand me?”: Virtual assistants and dysarthria, *in* ‘Intelligent Environments 2018’, IOS Press, pp. 557–566.

Baltes, J., Hosseinmemar, A., Jung, J., Sadeghnejad, S. & Anderson, J. (2015), Practical real-time system for object counting based on optical flow, *in* ‘Robot Intelligence Technology and Applications 3: Results from the 3rd International Conference on Robot Intelligence Technology and Applications’, Springer, pp. 299–306.

Bangaru, S. S., Wang, C., Zhou, X., Jeon, H. W. & Li, Y. (2020), ‘Gesture recognition–based smart training assistant system for construction worker earplug-wearing training’, *Journal of Construction Engineering and Management* **146**(12), 04020144.

Banovic, N., Yang, Z., Ramesh, A. & Liu, A. (2023), ‘Being trustworthy is not enough: How untrustworthy artificial intelligence (ai) can deceive the end-users and gain their trust’, *Proceedings of the ACM on Human-Computer Interaction* **7**(CSCW1), 1–17.

Baraldi, L., Paci, F., Serra, G., Benini, L. & Cucchiara, R. (2015), ‘Gesture recognition using wearable vision sensors to enhance visitors’ museum experiences’, *IEEE Sensors Journal* **15**(5), 2705–2714.

Beauchemin, S. S. & Barron, J. L. (1995), ‘The computation of optical flow’, *ACM computing surveys (CSUR)* **27**(3), 433–466.

Becht, E., Dutertre, C.-A., Kwok, I. W., Ng, L. G., Ginhoux, F. & Newell, E. W. (2018), ‘Evaluation of UMAP as an alternative to t-SNE for single-cell data’, *BioRxiv* p. 298430.

Bhardwaj, A., Chae, J., Noeske, R. H. & Kim, J. R. (2021), Tangibledata: Interactive data visualization with mid-air haptics, *in* ‘Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology’, pp. 1–11.

Brooke, J. et al. (1996), ‘Sus-a quick and dirty usability scale’, *Usability evaluation in industry* **189**(194), 4–7.

Cao, L., Chua, K. S., Chong, W., Lee, H. & Gu, Q. (2003), ‘A comparison of pca, kpca and ica for dimensionality reduction in support vector machine’, *Neurocomputing* **55**(1-2), 321–336.

Carfì, A. & Mastrogiovanni, F. (2021), ‘Gesture-based human–machine interaction: Taxonomy, problem definition, and analysis’, *IEEE Transactions on Cybernetics* **53**(1), 497–513.

Carter, B. T. & Luke, S. G. (2020), ‘Best practices in eye tracking research’, *International Journal of Psychophysiology* **155**, 49–62.

Casiez, G., Roussel, N. & Vogel, D. (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’, pp. 2527–2530.

Chakraborty, B. K., Sarma, D., Bhuyan, M. K. & MacDorman, K. F. (2018), ‘Review of constraints on vision-based gesture recognition for human–computer interaction’, *IET Computer Vision* **12**(1), 3–15.

Chamunorwa, M., Wozniak, M. P., Vöge, S., Müller, H. & Boll, S. C. (2022), ‘Interacting with rigid and soft surfaces for smart-home control’, *Proceedings of the ACM on Human-Computer Interaction* **6**(MHCI), 1–22.

Chen, O. T.-C., Chang, Y.-X., Jhao, Y.-W., Chung, C.-Y., Chang, Y.-L. & Huang, W.-H. (2022), 3d object detection of cars and pedestrians by deep neural networks from unit-sharing one-shot nas, *in* ‘2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)’, IEEE, pp. 1–8.

Chen, R. T., Li, X., Grosse, R. B. & Duvenaud, D. K. (2018), ‘Isolating sources of disentanglement in variational autoencoders’, *Advances in neural information processing systems* **31**.

Cheng, P. & Roedig, U. (2022), ‘Personal voice assistant security and privacy—a survey’, *Proceedings of the IEEE* **110**(4), 476–507.

Cheung, E. & Lumelsky, V. J. (1989), ‘Proximity sensing in robot manipulator motion planning: system and implementation issues’, *IEEE transactions on Robotics and Automation* **5**(6), 740–751.

Clay, V., König, P. & Koenig, S. (2019), ‘Eye tracking in virtual reality’, *Journal of eye movement research* **12**(1).

Cockburn, A., Gutwin, C. & Greenberg, S. (2007), A predictive model of menu performance, *in* 'Proceedings of the SIGCHI conference on Human factors in computing systems', pp. 627–636.

Crossman, E. R. F. & Goodeve, P. (1983), 'Feedback control of hand-movement and fitts' law', *The Quarterly Journal of Experimental Psychology* **35**(2), 251–278.

Da Gama, A., Fallavollita, P., Teichrieb, V. & Navab, N. (2015), 'Motor rehabilitation using kinect: a systematic review', *Games for health journal* **4**(2), 123–135.

Dang, H. & Buschek, D. (2021), Gesturemap: Supporting visual analytics and quantitative analysis of motion elicitation data by learning 2d embeddings, *in* 'Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems', pp. 1–12.

Dang, Y. & Cheffena, M. (2024), 'Multifunctional sensing platform based on single antenna for non-contact human-machine interaction and environment sensing', *IEEE Transactions on Antennas and Propagation* .

Dementyev, A. & Paradiso, J. A. (2014), Wristflex: low-power gesture input with wrist-worn pressure sensors, *in* 'Proceedings of the 27th annual ACM symposium on User interface software and technology', pp. 161–166.

Deo, N., Rangesh, A. & Trivedi, M. (2016), In-vehicle hand gesture recognition using hidden markov models, *in* '2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)', IEEE, pp. 2179–2184.

Derboven, J., Huyghe, J. & De Grooff, D. (2014), Designing voice interaction for people with physical and speech impairments, *in* 'Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational', pp. 217–226.

Dezfuli, N., Khalilbeigi, M., Huber, J., Müller, F. & Mühlhäuser, M. (2012), Palmrc: imaginary palm-based remote control for eyes-free television interaction, *in* 'Proceedings of the 10th European conference on Interactive tv and video', pp. 27–34.

Dipietro, L., Sabatini, A. M. & Dario, P. (2008), 'A survey of glove-based systems and their applications', *Ieee transactions on systems, man, and cybernetics, part c (applications and reviews)* **38**(4), 461–482.

do Nascimento, L. V., Machado, G. M., Maran, V. & de Oliveira, J. P. M. (2021), ‘Context recognition and ubiquitous computing in smart cities: a systematic mapping’, *Computing* **103**(5), 801–825.

Došilović, F. K., Brčić, M. & Hlupić, N. (2018), Explainable artificial intelligence: A survey, *in* ‘2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)’, IEEE, pp. 0210–0215.

Du, G., Guo, D., Su, K., Wang, X., Teng, S., Li, D. & Liu, P. X. (2022), ‘A mobile gesture interaction method for augmented reality games using hybrid filters’, *IEEE Transactions on Instrumentation and Measurement* **71**, 1–12.

Dube, T. J. & Arif, A. S. (2023), Ultrasonic keyboard: A mid-air virtual qwerty with ultrasonic feedback for virtual reality, *in* ‘Proceedings of the Seventeenth International Conference on Tangible, Embedded, and Embodied Interaction’, pp. 1–8.

D’Eusanio, A., Simoni, A., Pini, S., Borghi, G., Vezzani, R. & Cucchiara, R. (2020), Multimodal hand gesture classification for the human–car interaction, *in* ‘Informatics’, Vol. 7, MDPI, p. 31.

Eastwood, C. & Williams, C. K. (2018), A framework for the quantitative evaluation of disentangled representations, *in* ‘6th International Conference on Learning Representations’.

Elezovikj, S., Ling, H. & Chen, X. (2013), Foreground and scene structure preserved visual privacy protection using depth information, *in* ‘2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)’, IEEE, pp. 1–4.

Elouariachi, I., Benouini, R., Zenkouar, K. & Zarghili, A. (2020), ‘Robust hand gesture recognition system based on a new set of quaternion tchebichef moment invariants’, *Pattern Analysis and Applications* **23**(3), 1337–1353.

Fang, B., Sun, F., Liu, H. & Liu, C. (2018), ‘3d human gesture capturing and recognition by the immu-based data glove’, *Neurocomputing* **277**, 198–207.

Farrell, T. R. et al. (2008), ‘A comparison of the effects of electrode implantation and targeting on pattern classification accuracy for prosthesis control’, *IEEE Transactions on Biomedical Engineering* **55**(9), 2198–2211.

Feng, Z., Dalong, T. & Yingzi, W. (2003), Obstacle avoidance for mobile robots based on relative coordinates, *in* 'IEEE International Conference on Robotics, Intelligent Systems and Signal Processing, 2003. Proceedings. 2003', Vol. 1, IEEE, pp. 616–621.

Fischer, F., Fleig, A., Klar, M. & Müller, J. (2022), 'Optimal feedback control for modeling human–computer interaction', *ACM Transactions on Computer-Human Interaction* **29**(6), 1–70.

Fitts, P. M. (1954), 'The information capacity of the human motor system in controlling the amplitude of movement.', *Journal of experimental psychology* **47**(6), 381.

Frueh, C., Jain, S. & Zakhor, A. (2005), 'Data processing algorithms for generating textured 3d building facade meshes from laser scans and camera images', *International Journal of Computer Vision* **61**(2), 159–184.

Fu, J., Ge, X., Xin, X., Karatzoglou, A., Arapakis, I., Wang, J. & Jose, J. M. (2024), Iisan: Efficiently adapting multimodal representation for sequential recommendation with decoupled peft, *in* 'Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval', pp. 687–697.

Fu, X., Pan, X., Liu, Y., Li, J., Zhang, Z., Liu, H. & Gao, M. (2023), 'Non-contact triboelectric nanogenerator', *Advanced Functional Materials* **33**(52), 2306749.

Ge, X., Fu, J., Chen, F., An, S., Sebe, N. & Jose, J. M. (2024), Towards end-to-end explainable facial action unit recognition via vision-language joint learning, *in* 'Proceedings of the 32nd ACM International Conference on Multimedia', pp. 8189–8198.

Geeng, C. & Roesner, F. (2019), Who's in control? interactions in multi-user smart homes, *in* 'Proceedings of the 2019 CHI conference on human factors in computing systems', pp. 1–13.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M. & Kagal, L. (2018), Explaining explanations: An overview of interpretability of machine learning, *in* '2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)', IEEE, pp. 80–89.

Gisbrecht, A., Schulz, A. & Hammer, B. (2015), 'Parametric nonlinear dimensionality reduction using kernel t-SNE', *Neurocomputing* **147**, 71–82.

Gong, Y. & Poellabauer, C. (2018), 'An overview of vulnerabilities of voice controlled systems', *arXiv preprint arXiv:1803.09156* .

Groenewald, C., Anslow, C., Islam, J., Rooney, C., Passmore, P. J. & Wong, B. (2016), 'Understanding 3d mid-air hand gestures with interactive surfaces and displays: a systematic literature review'.

Guo, H. & Pan, Y. (2023), 'Interpretative structural modeling analyzes the hierarchical relationship between mid-air gestures and interaction satisfaction', *Applied Sciences* **13**(5), 3129.

Guo, L., Lu, Z. & Yao, L. (2021), 'Human-machine interaction sensing technology based on hand gesture recognition: A review', *IEEE Transactions on Human-Machine Systems* .

Guo, X. (2022), 'A fitts' law evaluation and comparison for human and manipulator on touch task', *Cognitive Computation and Systems* **4**(3), 265–272.

Hajika, R., Gunasekaran, T. S., Haigh, C. D. S. Y., Pai, Y. S., Hayashi, E., Lien, J., Lottridge, D. & Billinghurst, M. (2024), 'Radarhand: A wrist-worn radar for on-skin touch-based proprioceptive gestures', *ACM Transactions on Computer-Human Interaction* **31**(2), 1–36.

Han, X., Cong, P., Xu, L., Wang, J., Yu, J. & Ma, Y. (2022), 'Licamgait: Gait recognition in the wild by using lidar and camera multi-modal visual sensors', *arXiv preprint arXiv:2211.12371* .

Harish, N. & Poonguzhali, S. (2015), Design and development of hand gesture recognition system for speech impaired people, *in* '2015 International Conference on Industrial Instrumentation and Control (ICIC)', IEEE, pp. 1129–1133.

Hart, S. G. & Staveland, L. E. (1988), Development of nasa-tlx (task load index): Results of empirical and theoretical research, *in* 'Advances in psychology', Vol. 52, Elsevier, pp. 139–183.

Hayashi, E., Lien, J., Gillian, N., Giusti, L., Weber, D., Yamanaka, J., Bedal, L. & Poupyrev, I. (2021), Radarnet: Efficient gesture recognition technique utilizing a miniature radar sensor, *in* 'Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems', pp. 1–14.

He, K., Zhang, X., Ren, S. & Sun, J. (2016), Deep residual learning for image recognition, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 770–778.

Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S. & Lerchner, A. (2017), 'beta-vae: Learning basic visual concepts with a constrained variational framework.', *ICLR (Poster)* **3**.

Hincapié-Ramos, J. D., Guo, X., Moghadasian, P. & Irani, P. (2014), Consumed endurance: a metric to quantify arm fatigue of mid-air interactions, *in* 'Proceedings of the SIGCHI conference on human factors in computing systems', pp. 1063–1072.

Hinton, G. E. & Salakhutdinov, R. R. (2006), 'Reducing the dimensionality of data with neural networks', *science* **313**(5786), 504–507.

Hollerbach, J. M. (1985), Optimum kinematic design for a seven degree of freedom manipulator, *in* 'Robotics research: The second international symposium', Citeseer, pp. 215–222.

Hosseini, M., Ihmels, T., Chen, Z., Koelle, M., Müller, H. & Boll, S. (2023), Towards a consensus gesture set: A survey of mid-air gestures in hci for maximized agreement across domains, *in* 'Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems', pp. 1–24.

Hoy, M. B. (2018), 'Alexa, siri, cortana, and more: an introduction to voice assistants', *Medical reference services quarterly* **37**(1), 81–88.

Hsiao, K., Nangeroni, P., Huber, M., Saxena, A. & Ng, A. Y. (2009), Reactive grasping using optical proximity sensors, *in* '2009 IEEE International Conference on Robotics and Automation', IEEE, pp. 2098–2105.

Hsu, H.-W., Wu, T.-Y., Wan, S., Wong, W. H. & Lee, C.-Y. (2018), 'Quatnet: Quaternion-based head pose estimation with multiregression loss', *IEEE Transactions on Multimedia* **21**(4), 1035–1046.

Huang, S., Ranganathan, S. P. & Parsons, I. (2020), To touch or not to touch? comparing touch, mid-air gesture, mid-air haptics for public display in post covid-19 society, *in* 'SIGGRAPH Asia 2020 Posters', pp. 1–2.

Islam, M. M., Islam, M. R. & Islam, M. S. (2020), 'An efficient human computer interaction through hand gesture using deep convolutional neural network', *SN Computer Science* **1**(4), 211.

Jakob, D., Wilhelm, S., Gerl, A. & Ahrens, D. (2021), A quantitative study on awareness, usage and reservations of voice control interfaces by elderly people, *in* 'International Conference on Human-Computer Interaction', Springer, pp. 237–257.

Jalaliniya, S., Smith, J., Sousa, M., Büthe, L. & Pederson, T. (2013), Touch-less interaction with medical images using hand & foot gestures, *in* 'Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication', pp. 1265–1274.

Jang, S., Elmqvist, N. & Ramani, K. (2014), Gestureanalyzer: visual analytics for pattern analysis of mid-air hand gestures, *in* 'Proceedings of the 2nd ACM symposium on Spatial user interaction', pp. 30–39.

Jeet, V., Dhillon, H. S. & Bhatia, S. (2015), Radio frequency home appliance control based on head tracking and voice control for disabled person, *in* '2015 Fifth International Conference on Communication Systems and Network Technologies', IEEE, pp. 559–563.

Kallio, S., Kela, J. & Mantyjarvi, J. (2003), Online gesture recognition system for mobile interaction, *in* 'SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance (Cat. No. 03CH37483)', Vol. 3, IEEE, pp. 2070–2076.

Kendon, A. (2014), On gesture: Its complementary relationship with speech, *in* 'Nonverbal behavior and communication', Psychology Press, pp. 65–97.

Khundam, C. (2015), First person movement control with palm normal and hand gesture interaction in virtual reality, *in* '2015 12th International Joint Conference on Computer Science and Software Engineering (JCSSE)', IEEE, pp. 325–330.

Kim, H. & Mnih, A. (2018), Disentangling by factorising, *in* 'International conference on machine learning', PMLR, pp. 2649–2658.

Kim, K., Joo, D. & Lee, K.-P. (2010), Wearable-object-based interaction for a mobile audio device, *in* 'CHI'10 Extended Abstracts on Human Factors in Computing Systems', pp. 3865–3870.

Kim, K., Kim, J., Choi, J., Kim, J. & Lee, S. (2015), 'Depth camera-based 3d hand gesture controls with immersive tactile feedback for natural mid-air gesture interactions', *Sensors* **15**(1), 1022–1046.

Kim, T., Shim, Y. A., Kim, Y., Kim, S., Lee, J. & Lee, G. (2024), Quadstretcher: A forearm-worn skin stretch display for bare-hand interaction in ar/vr, *in* 'Proceedings of the CHI Conference on Human Factors in Computing Systems', pp. 1–15.

Kimura, N., Kono, M. & Rekimoto, J. (2019), Sottovoce: An ultrasound imaging-based silent speech interaction using deep neural networks, *in* 'Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems', pp. 1–11.

Kingma, D. P. (2013), 'Auto-encoding variational bayes', *arXiv preprint arXiv:1312.6114* .

Kingma, D. P. & Ba, J. (2014), 'Adam: A method for stochastic optimization', *arXiv preprint arXiv:1412.6980* .

Kingma, D. P. & Welling, M. (2014), 'Auto-encoding variational bayes', *the 2nd International Conference on Learning Representations* .

Kong, H., Lu, L., Yu, J., Chen, Y. & Tang, F. (2020), 'Continuous authentication through finger gesture interaction for smart homes using wifi', *IEEE Transactions on Mobile Computing* **20**(11), 3148–3162.

Koulieris, G. A., Akşit, K., Stengel, M., Mantiuk, R. K., Mania, K. & Richardt, C. (2019), Near-eye display and tracking technologies for virtual and augmented reality, *in* 'Computer Graphics Forum', Vol. 38, Wiley Online Library, pp. 493–519.

Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W. & Torralba, A. (2016), Eye tracking for everyone, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 2176–2184.

Kühnel, C., Westermann, T., Hemmert, F., Kratz, S., Müller, A. & Möller, S. (2011), 'I'm home: Defining and evaluating a gesture set for smart-home control', *International Journal of Human-Computer Studies* **69**(11), 693–704.

Kuipers, J. B. (1999), *Quaternions and rotation sequences: a primer with applications to orbits, aerospace, and virtual reality*, Princeton university press.

Kumar, A., Sattigeri, P. & Balakrishnan, A. (2017), 'Variational inference of disentangled latent concepts from unlabeled observations', *arXiv preprint arXiv:1711.00848* .

Lang, C. E. & Schieber, M. H. (2004), 'Human finger independence: limitations due to passive mechanical coupling versus active neuromuscular control', *Journal of neurophysiology* **92**(5), 2802–2810.

Laugwitz, B., Held, T. & Schrepp, M. (2008), Construction and evaluation of a user experience questionnaire, *in* 'Symposium of the Austrian HCI and usability engineering group', Springer, pp. 63–76.

Lee, A.-r., Cho, Y., Jin, S. & Kim, N. (2020), 'Enhancement of surgical hand gesture recognition using a capsule network for a contactless interface in the operating room', *Computer methods and programs in biomedicine* **190**, 105385.

Lee, C.-J., Zhang, R., Agarwal, D., Yu, T. C., Gunda, V., Lopez, O., Kim, J., Yin, S., Dong, B., Li, K. et al. (2024), Echowrist: Continuous hand pose tracking and hand-object interaction recognition using low-power active acoustic sensing on a wristband, *in* 'Proceedings of the CHI Conference on Human Factors in Computing Systems', pp. 1–21.

Liao, J., Wu, M. & Baines, R. (1999), 'A coordinate measuring machine vision system', *Computers in industry* **38**(3), 239–248.

Liao, Z., Luo, Z., Huang, Q., Zhang, L., Wu, F., Zhang, Q. & Wang, Y. (2021), SMART: screen-based gesture recognition on commodity mobile devices, *in* 'Proceedings of the 27th Annual International Conference on Mobile Computing and Networking', pp. 283–295.

Lien, J., Gillian, N., Karagozler, M. E., Amihood, P., Schwesig, C., Olson, E., Raja, H. & Poupyrev, I. (2016), 'Soli: Ubiquitous gesture sensing with millimeter wave radar', *ACM Transactions on Graphics (TOG)* **35**(4), 1–19.

Lin, J. & Ding, Y. (2013), 'A temporal hand gesture recognition system based on hog and motion trajectory', *Optik* **124**(24), 6795–6798.

Linderman, G. C. & Steinerberger, S. (2019), 'Clustering with t-sne, provably', *SIAM journal on mathematics of data science* **1**(2), 313–332.

Liu, H., Wang, Y., Zhou, A., He, H., Wang, W., Wang, K., Pan, P., Lu, Y., Liu, L. & Ma, H. (2020), 'Real-time arm gesture recognition in smart home scenarios via millimeter wave sensing', *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* **4**(4), 1–28.

Liu, H., Zhou, A., Dong, Z., Sun, Y., Zhang, J., Liu, L., Ma, H., Liu, J. & Yang, N. (2021), 'M-gesture: Person-independent real-time in-air gesture recognition using commodity millimeter wave radar', *IEEE Internet of Things Journal* .

Liu, Z. & Heer, J. (2014), 'The effects of interactive latency on exploratory visual analysis', *IEEE transactions on visualization and computer graphics* **20**(12), 2122–2131.

Locatello, F., Poole, B., Rätsch, G., Schölkopf, B., Bachem, O. & Tschannen, M. (2020), Weakly-supervised disentanglement without compromises, *in* 'International conference on machine learning', PMLR, pp. 6348–6359.

Lopatovska, I., Rink, K., Knight, I., Raines, K., Cosenza, K., Williams, H., Sorsche, P., Hirsch, D., Li, Q. & Martinez, A. (2019), 'Talk to me: Exploring user interactions with the amazon alexa', *Journal of Librarianship and Information Science* **51**(4), 984–997.

López, G., Quesada, L. & Guerrero, L. A. (2018), Alexa vs. siri vs. cortana vs. google assistant: a comparison of speech-based natural user interfaces, *in* 'Advances in Human Factors and Systems Interaction: Proceedings of the AHFE 2017 International Conference on Human Factors and Systems Interaction, July 17- 21, 2017, The Westin Bonaventure Hotel, Los Angeles, California, USA 8', Springer, pp. 241–250.

Lu, Z., Chen, X., Li, Q., Zhang, X. & Zhou, P. (2014), 'A hand gesture recognition framework and wearable gesture-based interaction prototype for mobile devices', *IEEE transactions on human-machine systems* **44**(2), 293–299.

Luckin, R., Du Boulay, B. et al. (1999), 'Ecolab: The development and evaluation of a vygotskian design framework', *International journal of artificial Intelligence in Education* **10**(2), 198–220.

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J. et al. (2019), 'Mediapipe: A framework for building perception pipelines', *arXiv preprint arXiv:1906.08172* .

MacKenzie, I. S. (1992), 'Fitts' law as a research and design tool in human-computer interaction', *Human-computer interaction* **7**(1), 91–139.

Mahmoodi, J. & Salajeghe, A. (2019), 'A classification method based on optical flow for violence detection', *Expert systems with applications* **127**, 121–127.

Mahmoud, N. M., Fouad, H. & Soliman, A. M. (2021), 'Smart healthcare solutions using the internet of medical things for hand gesture recognition system', *Complex & intelligent systems* **7**, 1253–1264.

Majaranta, P. (2011), *Gaze interaction and applications of eye tracking: Advances in assistive technologies: Advances in assistive technologies*, iGi Global.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I. & Frey, B. (2015), 'Adversarial autoencoders', *arXiv preprint arXiv:1511.05644* .

Marquardt, N. & Greenberg, S. (2015), 'Proxemic interactions: From theory to practice', *Synthesis Lectures on Human-Centered Informatics* **8**(1), 1–199.

Martinek, R., Vanus, J., Nedoma, J., Fridrich, M., Frnda, J. & Kawala-Sterniuk, A. (2020), 'Voice communication in noisy environments in a smart house using hybrid lms+ ica algorithm', *Sensors* **20**(21), 6022.

Masci, J., Meier, U., Cireşan, D. & Schmidhuber, J. (2011), Stacked convolutional autoencoders for hierarchical feature extraction, *in* 'Artificial Neural Networks and Machine Learning–ICANN 2011: 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14-17, 2011, Proceedings, Part I 21', Springer, pp. 52–59.

Mase, K. (1991), 'Recognition of facial expression from optical flow', *IEICE TRANSACTIONS on Information and Systems* **74**(10), 3474–3483.

Masurovsky, A., Chojecki, P., Runde, D., Lafci, M., Przewozny, D. & Gaebler, M. (2020), 'Controller-free hand tracking for grab-and-place tasks in immersive virtual reality: Design elements and their empirical study', *Multimodal Technologies and Interaction* **4**(4), 91.

McInnes, L., Healy, J. & Melville, J. (2018), 'UMAP: Uniform manifold approximation and projection for dimension reduction', *arXiv preprint arXiv:1802.03426* .

Meng, M., Fallavollita, P., Blum, T., Eck, U., Sandor, C., Weidert, S., Waschke, J. & Navab, N. (2013), Kinect for interactive ar anatomy learning, *in* '2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)', IEEE, pp. 277–278.

Mitsopoulos-Rubens, E., Trotter, M. J. & Lenné, M. G. (2011), 'Effects on driving performance of interacting with an in-vehicle music player: A comparison of three interface layout concepts for information presentation', *Applied ergonomics* **42**(4), 583–591.

Mittal, Y., Toshniwal, P., Sharma, S., Singhal, D., Gupta, R. & Mittal, V. K. (2015), A voice-controlled multi-functional smart home automation system, *in* '2015 Annual IEEE India Conference (INDICON)', IEEE, pp. 1–6.

Mlakar, S., Alida Haberfellner, M., Jetter, H.-C. & Haller, M. (2021), Exploring affordances of surface gestures on textile user interfaces, *in* 'Designing Interactive Systems Conference 2021', pp. 1159–1170.

Modaberi, M. (2024), 'The role of gesture-based interaction in improving user satisfaction for touchless interfaces', *International Journal of Advanced Human Computer Interaction* **2**(2), 20–32.

Mousavi Hondori, H. & Khademi, M. (2014), 'A review on technical and clinical impact of microsoft kinect on physical therapy and rehabilitation', *Journal of medical engineering* **2014**(1), 846514.

Muceli, S. & Farina, D. (2011), 'Simultaneous and proportional estimation of hand kinematics from emg during mirrored movements at multiple degrees-of-freedom', *IEEE transactions on neural systems and rehabilitation engineering* **20**(3), 371–378.

Murray-Smith, R. (2017), Stratified, computational interaction via machine learning, *in* 'Eighteenth Yale Workshop on Adaptive and Learning Systems (New Haven, CT, USA. 95–101'.

Myers, B., Hudson, S. E. & Pausch, R. (2000), 'Past, present, and future of user interface software tools', *ACM Transactions on Computer-Human Interaction (TOCHI)* **7**(1), 3–28.

Nacenta, M. A., Kamber, Y., Qiang, Y. & Kristensson, P. O. (2013), Memorability of predesigned and user-defined gesture sets, *in* 'Proceedings of the SIGCHI conference on human factors in computing systems', pp. 1099–1108.

Nasution, M. I. P., Nurbaiti, N., Nurlaila, N., Rahma, T. I. F. & Kamilah, K. (2020), Face recognition login authentication for digital payment solution at covid-19 pandemic, *in* '2020 3rd International Conference on Computer and Informatics Engineering (IC2IE)', IEEE, pp. 48–51.

Neimark, D., Bar, O., Zohar, M. & Asselmann, D. (2021), Video transformer network, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 3163–3172.

Nguyen, H.-Q., Le, T.-H., Tran, T.-K., Tran, H.-N., Tran, T.-H., Le, T.-L., Vu, H., Pham, C., Nguyen, T. P. & Nguyen, H. T. (2023), 'Hand gesture recognition from wrist-worn camera for human–machine interaction', *IEEE Access* **11**, 53262–53274.

Novack, M. & Goldin-Meadow, S. (2015), ‘Learning from gesture: How our hands change our minds’, *Educational psychology review* **27**, 405–412.

O’Brien, K., Liggett, A., Ramirez-Zohfeld, V., Sunkara, P. & Lindquist, L. A. (2020), ‘Voice-controlled intelligent personal assistants to support aging in place’, *Journal of the American Geriatrics Society* **68**(1), 176–179.

O’Hara, K., Gonzalez, G., Sellen, A., Penney, G., Varnavas, A., Mentis, H., Criminisi, A., Corish, R., Rouncefield, M., Dastur, N. et al. (2014), ‘Touchless interaction in surgery’, *Communications of the ACM* **57**(1), 70–77.

Paik, J., Kim, J. W., Ritter, F. E. & Reitter, D. (2015), ‘Predicting user performance and learning in human–computer interaction with the herbal compiler’, *ACM Transactions on Computer-Human Interaction (TOCHI)* **22**(5), 1–26.

Panger, G. (2012), Kinect in the kitchen: testing depth camera interactions in practical home environments, *in* ‘CHI’12 Extended Abstracts on Human Factors in Computing Systems’, pp. 1985–1990.

Panwar, M. & Mehra, P. S. (2011), Hand gesture recognition for human computer interaction, *in* ‘2011 International Conference on Image Information Processing’, IEEE, pp. 1–7.

Parilusyan, B., Teyssier, M., Martinez-Missir, V., Duhart, C. & Serrano, M. (2022), ‘Sensurfaces: A novel approach for embedded touch sensing on everyday surfaces’, *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **6**(2), 1–19.

Partridge, M. & Calvo, R. A. (1998), ‘Fast dimensionality reduction and simple pca’, *Intelligent data analysis* **2**(3), 203–214.

Patil, A. K., Kim, S. H., Balasubramanyam, A., Ryu, J. Y. & Chai, Y. H. (2019), Pilot experiment of a 2d trajectory representation of quaternion-based 3d gesture tracking, *in* ‘Proceedings of the ACM SIGCHI Symposium on Engineering Interactive Computing Systems’, pp. 1–7.

Pavlovic, V. I., Sharma, R. & Huang, T. S. (1997), ‘Visual interpretation of hand gestures for human-computer interaction: A review’, *IEEE Transactions on pattern analysis and machine intelligence* **19**(7), 677–695.

Pearson, J., Bailey, G., Robinson, S., Jones, M., Owen, T., Zhang, C., Reitmaier, T., Steer, C., Carter, A., Sahoo, D. R. et al. (2022), Can't touch this: Rethinking public technology in a covid-19 era, *in* 'Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems', pp. 1–14.

Pierdicca, R., Paolanti, M., Naspetti, S., Mandolesi, S., Zanoli, R. & Frontoni, E. (2018), 'User-centered predictive model for improving cultural heritage augmented reality applications: An hmm-based approach for eye-tracking data', *Journal of imaging* **4**(8), 101.

Potts, D., Dabravalskis, M. & Houben, S. (2022), Tangibletouch: A toolkit for designing surface-based gestures for tangible interfaces, *in* 'Sixteenth International Conference on Tangible, Embedded, and Embodied Interaction', pp. 1–14.

Prekop, P. & Burnett, M. (2003), 'Activities, context and ubiquitous computing', *Computer communications* **26**(11), 1168–1176.

Pterneas, V. (2023), 'Measuring distances using kinect – the right way'. Accessed: 18 March 2025.
URL: *https://pterneas.com/2016/08/11/measuring-distances-kinect/*

Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., Stevens, A. & Carin, L. (2016), 'Variational autoencoder for deep learning of images, labels and captions', *Advances in neural information processing systems* **29**.

Pukari, A., Nivalainen, J., Alavesa, P. & Korkiakoski, M. (2023), Learn flags: Assessment of the learning curve of inexperienced users with gesture interactions with hands-free augmented reality, *in* 'Proceedings of the 26th International Academic Mindtrek Conference', pp. 340–343.

Qi, C. R., Su, H., Mo, K. & Guibas, L. J. (2017), Pointnet: Deep learning on point sets for 3D classification and segmentation, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 652–660.

Qi, J., Ma, L., Cui, Z. & Yu, Y. (2024), 'Computer vision-based hand gesture recognition for human-robot interaction: a review', *Complex & Intelligent Systems* **10**(1), 1581–1606.

Qian, X., Ju, W. & Sirkin, D. M. (2020), 'Aladdin's magic carpet: Navigation by in-air static hand gesture in autonomous vehicles', *International Journal of Human–Computer Interaction* **36**(20), 1912–1927.

Qin, K., Chen, C., Pu, X., Tang, Q., He, W., Liu, Y., Zeng, Q., Liu, G., Guo, H. & Hu, C. (2021), 'Magnetic array assisted triboelectric nanogenerator sensor for real-time gesture interaction', *Nano-micro letters* **13**, 1–9.

Rajanna, V. & Hammond, T. (2022), 'Can gaze beat touch? a fitts' law evaluation of gaze, touch, and mouse inputs', *arXiv preprint arXiv:2208.01248* .

Razavi, A., Van den Oord, A. & Vinyals, O. (2019), 'Generating diverse high-fidelity images with vq-vae-2', *Advances in neural information processing systems* **32**.

Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Kaluri, R., Rajput, D. S., Srivastava, G. & Baker, T. (2020), 'Analysis of dimensionality reduction techniques on big data', *IEEE Access* **8**, 54776–54788.

Ren, Z., Meng, J. & Yuan, J. (2011), Depth camera based hand gesture recognition and its applications in human-computer-interaction, *in* '2011 8th International Conference on Information, Communications & Signal Processing', IEEE, pp. 1–5.

Renaud, K. & Cooper, R. (2000), 'Feedback in human-computer interaction-characteristics and recommendations', *South African Computer Journal* **2000**(26), 105–114.

Rezende, D. J., Mohamed, S. & Wierstra, D. (2014), Stochastic backpropagation and approximate inference in deep generative models, *in* 'International conference on machine learning', PMLR, pp. 1278–1286.

Rieger, B. & Van Vliet, L. J. (2004), 'A systematic approach to nD orientation representation', *Image and Vision Computing* **22**(6), 453–459.

Rusu, M. M., Schött, S. Y., Williamson, J. H., Schmidt, A. & Murray-Smith, R. (2021), 'Low-dimensional embeddings for interaction design', *Advanced Intelligent Systems* p. 2100045.

Rusu, M. M., Schött, S. Y., Williamson, J. H., Schmidt, A. & Murray-Smith, R. (2022), 'Low-dimensional embeddings for interaction design', *Advanced Intelligent Systems* **4**(2), 2100045.

Ryoo, M. S. & Aggarwal, J. K. (2007), Robust human-computer interaction system guiding a user by providing feedback., *in* 'IJCAI', pp. 2850–2855.

Sambrooks, L. & Wilkinson, B. (2013), Comparison of gestural, touch, and mouse interaction with fitts' law, *in* 'Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration', pp. 119–122.

Sara, U., Akter, M. & Uddin, M. S. (2019), 'Image quality assessment through FSIM, SSIM, MSE and PSNR — a comparative study', *Journal of Computer and Communications* **7**(3), 8–18.

Saxena, A., Driemeyer, J. & Ng, A. Y. (2009), Learning 3-D object orientation from images, *in* '2009 IEEE International Conference on Robotics and Automation', pp. 794–800.

Seaborn, K., Miyake, N. P., Pennefather, P. & Otake-Matsuura, M. (2021), 'Voice in human–agent interaction: A survey', *ACM Computing Surveys (CSUR)* **54**(4), 1–43.

Shakeri, G., Williamson, J. H. & Brewster, S. (2017), Novel multimodal feedback techniques for in-car mid-air gesture interaction, *in* 'Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications', pp. 84–93.

Shapiro, B. R., Hall, R. P. & Owens, D. A. (2017), 'Developing & using interaction geography in a museum', *International Journal of Computer-Supported Collaborative Learning* **12**, 377–399.

Sharif, K. & Tenbergen, B. (2020), 'Smart home voice assistants: a literature survey of user privacy and security vulnerabilities', *Complex Systems Informatics and Modeling Quarterly* (24), 15–30.

Sharma, S., Henderson, J. & Ghosh, J. (2019), 'Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models', *arXiv preprint arXiv:1905.07857* .

Shi, R., Wei, Y., Li, Y., Yu, L. & Liang, H.-N. (2023), Expanding targets in virtual reality environments: A fitts' law study, *in* '2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)', IEEE, pp. 615–618.

Shoemake, K. (1985), Animating rotation with quaternion curves, *in* 'Proceedings of the 12th annual conference on Computer graphics and interactive techniques', pp. 245–254.

Silva, V. S., Freitas, A. & Handschuh, S. (2019), 'On the semantic interpretability of artificial intelligence models', *arXiv preprint arXiv:1907.04105* .

Sinha, S. & Dieng, A. B. (2021), 'Consistency regularization for variational auto-encoders', *Advances in Neural Information Processing Systems* **34**, 12943–12954.

Sluÿters, A., Lambot, S., Vanderdonckt, J. & Vatavu, R.-D. (2023), ‘Radarsense: Accurate recognition of mid-air hand gestures with radar sensing and few training examples’, *ACM Transactions on Interactive Intelligent Systems* **13**(3), 1–45.

Song, S. & Xiao, J. (2014), Sliding shapes for 3d object detection in depth images, *in* ‘European conference on computer vision’, Springer, pp. 634–651.

Soukoreff, R. W. & MacKenzie, I. S. (2004), ‘Towards a standard for pointing device evaluation, perspectives on 27 years of fitts’ law research in hci’, *International journal of human-computer studies* **61**(6), 751–789.

Stančić, I., Musić, J. & Grujić, T. (2017), ‘Gesture recognition system for real-time mobile robot control based on inertial sensors and motion strings’, *Engineering Applications of Artificial Intelligence* **66**, 33–48.

Starner, T., Weaver, J. & Pentland, A. (1998), ‘Real-time american sign language recognition using desk and wearable computer based video’, *IEEE Transactions on pattern analysis and machine intelligence* **20**(12), 1371–1375.

Strachan, S., Murray-Smith, R. & O’Modhrain, S. (2007), Bodyspace: inferring body pose for natural control of a music player, *in* ‘CHI’07 extended abstracts on Human factors in computing systems’, pp. 2001–2006.

Su, H., Maji, S., Kalogerakis, E. & Learned-Miller, E. (2015), Multi-view convolutional neural networks for 3d shape recognition, *in* ‘Proceedings of the IEEE international conference on computer vision’, pp. 945–953.

Suk, H.-I., Sin, B.-K. & Lee, S.-W. (2010), ‘Hand gesture recognition based on dynamic bayesian network framework’, *Pattern recognition* **43**(9), 3059–3072.

Sun, J.-H., Ji, T.-T., Zhang, S.-B., Yang, J.-K. & Ji, G.-R. (2018), Research on the hand gesture recognition based on deep learning, *in* ‘2018 12th International Symposium on Antennas, Propagation and EM Theory (ISAPE)’, IEEE, pp. 1–4.

Sun, W., Li, F. M., Huang, C., Lei, Z., Steeper, B., Tao, S., Tian, F. & Zhang, C. (2021), ‘Thumbtrak: Recognizing micro-finger poses using a ring with proximity sensing’, *arXiv preprint arXiv:2105.14680* .

Taka, E., Stein, S. & Williamson, J. H. (2022), ‘Does interacting help users better understand the structure of probabilistic models?’, *arXiv preprint arXiv:2201.03605* .

Tidwell, J. (2010), *Designing interfaces: Patterns for effective interaction design*, " O'Reilly Media, Inc.".

Tran, L., Khasahmadi, A. H., Sanghi, A. & Asgari, S. (2021), 'Group-disentangled representation learning with weakly-supervised regularization', *arXiv preprint arXiv:2110.12185* .

Tseng, W.-J., Huron, S., Lecolinet, E. & Gugenheimer, J. (2023), Fingermapper: Mapping finger motions onto virtual arms to enable safe virtual reality interaction in confined spaces, *in* 'Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems', pp. 1–14.

Ur Rehman, M., Ahmed, F., Attique Khan, M., Tariq, U., Abdulaziz Alfouzan, F., M Alzahrani, N. & Ahmad, J. (2021), 'Dynamic hand gesture recognition using 3D-CNN and LSTM networks', *Computers, Materials & Continua* **70**(3).

Vad, B., Boland, D., Williamson, J., Murray-Smith, R. & Steffensen, P. B. (2015), 'Design and evaluation of a probabilistic music projection interface'.

Valtakari, N. V., Hooge, I. T., Viktorsson, C., Nyström, P., Falck-Ytter, T. & Hessels, R. S. (2021), 'Eye tracking in human interaction: Possibilities and limitations', *Behavior Research Methods* pp. 1–17.

Vatavu, R.-D. & Bilius, L.-B. (2021), Gesturing: A web-based tool for designing gesture input with rings, ring-like, and ring-ready devices, *in* 'The 34th Annual ACM Symposium on User Interface Software and Technology', pp. 710–723.

Vogiatzidakis, P. & Koutsabasis, P. (2021), Mid-air gestures for manipulation of multiple targets in the physical space: comparing the usability of two interaction models, *in* 'CHI Greece 2021: 1st International Conference of the ACM Greek SIGCHI Chapter', pp. 1–9.

Vogiatzidakis, P. & Koutsabasis, P. (2022), ''address and command': Two-handed mid-air interactions with multiple home devices', *International Journal of Human-Computer Studies* **159**, 102755.

Vogler, C. & Metaxas, D. (1999), Parallel hidden markov models for american sign language recognition, *in* 'Proceedings of the seventh IEEE international conference on computer vision', Vol. 1, IEEE, pp. 116–122.

Wang, J., Mueller, F., Bernard, F., Sorli, S., Sotnychenko, O., Qian, N., Otaduy, M. A., Casas, D. & Theobalt, C. (2020), 'Rgb2hands: real-time tracking of 3d hand interactions from monocular rgb video', *ACM Transactions on Graphics (ToG)* **39**(6), 1–16.

Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M. & Solomon, J. M. (2019), 'Dynamic graph cnn for learning on point clouds', *ACM Transactions on Graphics (tog)* **38**(5), 1–12.

Wästlund, E., Sponseller, K., Pettersson, O. & Bared, A. (2015), 'Evaluating gaze-driven power wheelchair with navigation support for persons with disabilities.', *Journal of Rehabilitation Research & Development* **52**(7).

Weise, M., Zender, R. & Lucke, U. (2020), 'How can i grab that? solving issues of interaction in vr by choosing suitable selection and manipulation techniques', *i-com* **19**(2), 67–85.

Wen, H., Ramos Rojas, J. & Dey, A. K. (2016), Serendipity: Finger gesture recognition using an off-the-shelf smartwatch, *in* 'Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems', pp. 3847–3851.

Wilhelm, M., Krakowczyk, D. & Albayrak, S. (2020), 'Perisense: Ring-based multi-finger gesture interaction utilizing capacitive proximity sensing', *Sensors* **20**(14), 3990.

Wilson, A. D. & Benko, H. (2010), Combining multiple depth cameras and projectors for interactions on, above and between surfaces, *in* 'Proceedings of the 23nd annual ACM symposium on User interface software and technology', pp. 273–282.

Wobbrock, J. O., Morris, M. R. & Wilson, A. D. (2009), User-defined gestures for surface computing, *in* 'Proceedings of the SIGCHI conference on human factors in computing systems', pp. 1083–1092.

Woodfill, J. & Von Herzen, B. (1997), Real-time stereo vision on the parts reconfigurable computer, *in* 'Proceedings. The 5th Annual IEEE Symposium on Field-Programmable Custom Computing Machines Cat. No. 97TB100186)', IEEE, pp. 201–210.

Wu, B., Jiang, T., Yu, Z., Zhou, Q., Jiao, J. & Jin, M. L. (2024), 'Proximity sensing electronic skin: Principles, characteristics, and applications', *Advanced Science* **11**(13), 2308560.

Wu, C.-L. & Fu, L.-C. (2011), 'Design and realization of a framework for human–system interaction in smart homes', *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* **42**(1), 15–31.

Wu, H., Luo, W., Pan, N., Nan, S., Deng, Y., Fu, S. & Yang, L. (2019), 'Understanding freehand gestures: a study of freehand gestural interaction for immersive vr shopping applications', *Human-centric Computing and Information Sciences* **9**, 1–26.

Xu, B., Wang, N., Chen, T. & Li, M. (2015), 'Empirical evaluation of rectified activations in convolutional network', *arXiv preprint arXiv:1505.00853* .

Xu, C., Pathak, P. H. & Mohapatra, P. (2015), Finger-writing with smartwatch: A case for finger and hand gesture recognition using smartwatch, *in* 'Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications', pp. 9–14.

Xu, S., Ge, X., Kaul, C. & Murray-Smith, R. (2024), Hpeis: Learning hand pose embeddings for multimedia interactive systems, *in* '2024 IEEE International Conference on Multimedia and Expo (ICME)', IEEE.

Xu, S., Kaul, C., Ge, X. & Murray-Smith, R. (2023), Continuous interaction with a smart speaker via low-dimensional embeddings of dynamic hand pose, *in* 'ICASSP', IEEE, pp. 1–5.

Yan, B., Wang, P., Du, L., Chen, X., Fang, Z. & Wu, Y. (2023), 'mmgesture: Semi-supervised gesture recognition system using mmwave radar', *Expert Systems with Applications* **213**, 119042.

Yang, L., Huang, J., Feng, T., Hong-An, W. & Guo-Zhong, D. (2019), 'Gesture interaction in virtual reality', *Virtual Reality & Intelligent Hardware* **1**(1), 84–112.

Yang, X., Sun, X., Zhou, D., Li, Y. & Liu, H. (2018), 'Towards wearable a-mode ultrasound sensing for real-time finger motion recognition', *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **26**(6), 1199–1208.

Yeo, H.-S., Lee, B.-G. & Lim, H. (2015), 'Hand tracking and gesture recognition system for human-computer interaction using low-cost hardware', *Multimedia Tools and Applications* **74**(8), 2687–2715.

Yin, R., Wang, D., Zhao, S., Lou, Z. & Shen, G. (2021), 'Wearable sensors-enabled human–machine interaction systems: From design to application', *Advanced Functional Materials* **31**(11), 2008936.

Yousefi, S., Kidane, M., Delgado, Y., Chana, J. & Reski, N. (2016), 3d gesture-based interaction for immersive experience in mobile vr, *in* '2016 23rd International Conference on Pattern Recognition (ICPR)', IEEE, pp. 2121–2126.

Yu, C.-C., Cheng, C.-H. & Fan, K.-C. (2014), 'A gait classification system using optical flow features.', *J. Inf. Sci. Eng.* **30**(1), 179–193.

Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R. & Toderici, G. (2015), Beyond short snippets: Deep networks for video classification, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 4694–4702.

Zahra, R., Shehzadi, A., Sharif, M. I., Karim, A., Azam, S., De Boer, F., Jonkman, M. & Mehmood, M. (2023), 'Camera-based interactive wall display using hand gesture recognition', *Intelligent Systems with Applications* **19**, 200262.

Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D. & Saeed, J. (2020), 'A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction', *Journal of Applied Science and Technology Trends* **1**(2), 56–70.

Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C.-L. & Grundmann, M. (2020), 'Mediapipe hands: On-device real-time hand tracking', *arXiv preprint arXiv:2006.10214* .

Zhang, F., Chu, S., Pan, R., Ji, N. & Xi, L. (2017), Double hand-gesture interaction for walk-through in vr environment, *in* '2017 IEEE/ACIS 16th international conference on computer and information science (icis)', IEEE, pp. 539–544.

Zhang, Z. & Sabuncu, M. (2018), 'Generalized cross entropy loss for training deep neural networks with noisy labels', *Advances in Neural Information Processing Systems* **31**.

Zhao, H., Cheng, M., Huang, J., Li, M., Cheng, H., Tian, K. & Yu, H. (2023), 'A virtual surgical prototype system based on gesture recognition for virtual surgical training in maxillofacial surgery', *International Journal of Computer Assisted Radiology and Surgery* **18**(5), 909–919.

Zhongcheng, W., Le, K., Fei, S. & Bin, F. (2005), The closed-loop human-computer interface: active information acquisition for vision-brain-hand to computer (vbh-c) interaction based on force tablet, *in* 'Proceedings. 2005 First International Conference on Neural Interface and Control, 2005.', IEEE, pp. 1–5.

Zhou, H., Wang, D., Yu, Y. & Zhang, Z. (2023), ‘Research progress of human–computer interaction technology based on gesture recognition’, *Electronics* **12**(13), 2805.