GesPlayer: Using Augmented Gestures to Empower Video Players

Xiang Li
University of Cambridge
Cambridge, United Kingdom
xl529@cam.ac.uk

Yuzheng Chen
Xi'an Jiaotong-Liverpool University
Suzhou, China
yuzheng.chen18@student.xjtlu.edu.cn

Xiaohang Tang
University of Liverpool
Liverpool, United Kingdom
sgxtang4@liverpool.ac.uk

Figure 1: Example: Using GesPlayer to control the video player.

ABSTRACT

In this paper, we introduce GesPlayer, a gesture-based empowered video player that explores how users can experience their hands as an interface through gestures. We provide three semantic gestures based on the camera of a computer or other smart device to detect and adjust the progress of video playback, volume, and screen brightness, respectively. Our goal is to enable users to control video playback simply by their gestures in the air, without the need to use a mouse or keyboard, especially when it is not convenient to do so. Ultimately, we hope to expand our understanding of gesture-based interaction by understanding the inclusiveness of designing the hand as an interactive interface, and further broaden the state of semantic gestures in an interactive environment through computational interaction methods.

ISS Companion '22, November 20–23, 2022, Wellington, New Zealand

ACM ISBN 978-1-4503-9356-0/22/11.
https://doi.org/10.1145/3532104.3571456

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**; **Gestural input**.

KEYWORDS

semantic gestures, augmented reality, gesture input, video player

ACM Reference Format:
Xiang Li, Yuzheng Chen, and Xiaohang Tang. 2022. GesPlayer: Using Augmented Gestures to Empower Video Players. In *Companion Proceedings of the 2022 Conference on Interactive Surfaces and Spaces. (ISS Companion '22), November 20–23, 2022, Wellington, New Zealand*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3532104.3571456

1 INTRODUCTION

Hand-based interaction is one of the most commonly used interaction methods in human-computer interaction (HCI) and intelligent interactive systems [12]. Currently, most of the AR and VR head-mounted displays also implement the function of mid-air gesture interaction for more natural interaction. In addition, semantic gestures are being explored to understand the underlying gestural behavior during user interactions in order to design and propose more intuitive ways of interaction. To fully investigate the importance of semantic gestures, Adam Kendon has summarized how the gestured component contributes to the meaning or expression of the utterance [7].

However, gestures are more than just an input to the process of interaction, but also instructions based on changes and activities of the body (especially the fingers). In the process of moving our fingers, we often overlook the expressive interface that is part of the body and the vehicle of gesture: the hand itself. Thus most gestures are designed to please the movement itself, which does improve the experience of interaction to some extent, but ignores the independence of the hand as an interface.

In this paper, we present GesPlayer, an augmented gesture-based video player which explores how users can also experience their hands as an interface with their gestures. We articulate the technical setup of GesPlayer, its design, and present preliminary findings, and give our discussions in the form of two themes: (a) how the hand as an interface expresses the interaction of gestures, and (b) the state of hand interaction. Ultimately, we hope to expand our understanding of gesture-based interaction by understanding the inclusion of designing hands as an interactive interface and further broaden the state of semantic gestures in interactive environments via the computational interaction method.

2 RELATED WORK

2.1 Mid-air Gestures

Koutsabasis and Vogiatzidakis point out that mid-air interactions are characterized by (a) touchless interactions, (b) real-time sensor tracking of (parts of) the user's body, and (c) body movements, postures, and gestures need to be recognized and matched to specific user intentions, goals, and commands [8].

Most previous research on gesture-based interaction has been based on the use of one or more RGB cameras [1]. For example, Dani et al. proposed a low-cost approach using only a single monocular RGB camera [3]. Similarly, Jain et al [6] proposed a low-cost framework to manipulate objects in mid-air using only one RGB camera. In summary, with the recent advances in low-cost depth cameras and RGB cameras, many algorithms and techniques (see [2]) have been developed to enable gesture recognition for mid-air interaction.

2.2 Semantic Gestures

Gestures are a vehicle for conveying semantic information and an input for the external environment to understand the gestures. Thus, the research of semantic gestures focuses on semantics itself rather than on gesture design.

In semantic gestures, there is a term, Utterance, which refers to any combination of actions that counts to others as an attempt by the actor to "provide" some kind of information [7]. Drawing on Goffman's formulation [5], he notes that although whenever people are present together with each other they cannot avoid providing each other with information about their intentions and involvement, their status as social beings, and their own personal character, and thus can be said to be "giving out" information, people also engage in actions that are considered to be explicitly aimed at providing information.

2.3 Experiencing Body in HCI

Game research in HCI has always been interesting in the human body. Mueller et al. suggest that the field of game research is evolving from using keyboards to play digital content to using the body to play digital content, moving toward a future where we experience the body as a digital game [13]. To guide designers interested in supporting players to experience their bodies as games, Mueller et al. present two phenomenological perspectives on the human body (Körper and Leib) [13]. Mueller et al. argue that before new sensors entered the realm of game design, users primarily used mice and keyboards, joysticks and gamepads to play computer games. With the advancement of sensors such as Kinect, users are beginning to use their bodies to experience digital content (Körper). Ultimately, the vision presented by Mueller et al. is that we are able to experience our bodies as digital games (Körper & Leib). In addition, another paper by Mueller et al. investigates the possibility of "limited control of the body" as an interesting design resource for body play systems and proposes four strategies for limited control of the body: Exploration, Reflection, Learning, and Embracement [4].

3 GESPLAYER

3.1 Technical Setup

We used Unity 2021.2.13f1 for developing our system. The detection algorithm and the parameters collection were empowered by the MediaPipe [1] APIs and its Unity Plugin [2].

3.2 Design

In GesPlayer, we introduced three different states of gestures: "trigger gestures", "baseline gestures", and "interaction gestures": (a) a trigger gesture is defined in GesPlayer as a prerequisite for starting and triggering a judgment. This safeguards the balance between gestural interactions and normal bodily movements, and reduces misdetection for our system; (b) the baseline gesture is defined as a benchmark to calibrate some behaviors that require relatively precise operations: for example, adjusting the progress bar to control the progress of the currently playing video. When traditionally using the mouse or keyboard to control the progress of the video playback, we would see a reference at the bottom of the video (i.e., the progress bar), but the lack of an intuitive control for mid-air interactions would also make it difficult to move the video or perform other relatively precise operations. Therefore, the purpose of baseline gestures is to support truly interaction gestures to achieve more precise results (i.e., the act of dragging and dropping); and (c) the definition of interaction gestures is the same as in previous studies. That is, the action in effect performs the user's real command. For example, we manipulate objects in real life [15], use gestures for text input [10], translate the objects in dynamic tasks [18], experience the world in VR and navigate using the senses of touch, motion restriction, and proprioception [17], fight the virus

[1] MediaPipe is a framework for building pipelines to perform inference over arbitrary sensory data. With MediaPipe, a perception pipeline can be built as a graph of modular components, including model inference, media processing algorithms and data transformations, etc. [11]
[2] https://github.com/homuler/MediapipeUnityPlugin

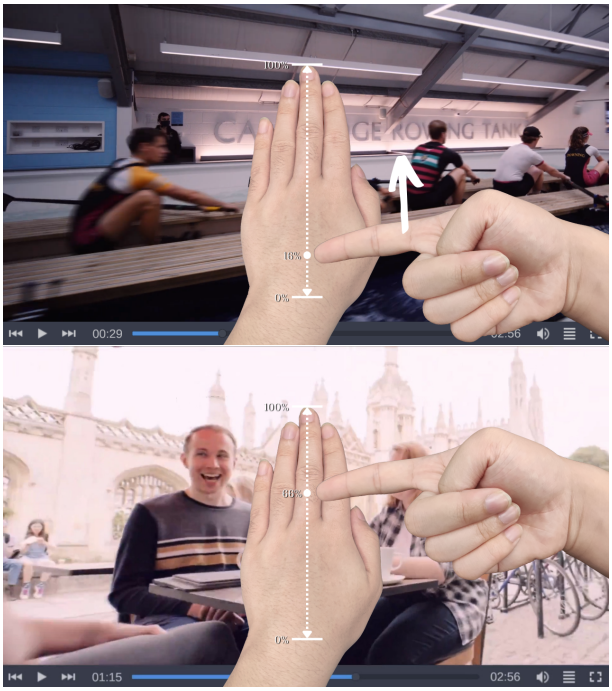

Figure 2: In the video playback control, we used the left hand as the progress bar and the right hand as the pointer. Once the left hand stays in front of the camera, the directed line segments between the wrist and the middle finger tip would stand for the progress bar of the video.

Figure 3: In the video volume control, we keep the left hand as the volume bar, however, use the "thumb", "index" and "middle" finger gestures on the right hand and use the middle fingertip as the pointer to adjust the volume.

in VR boxing games [19], or use our bodies for gesture detection in virtual environments [9].

3.2.1 Progress of Video Playback. In the video playback control, we used the left hand as the progress bar and the right hand as the pointer. Once the left hand stays in front of the camera, the directed line segments between the wrist and the middle finger tip would stand for the progress bar of the video. The wrist and the middle finger tip stand for 0% and 100% of the progress of video playback. Once the progress bar is set up, use the right-hand index fingertip as the pointer to adjust the video progress by hiding the right-hand thumb and touching the directed line segments on the left hand physically. Touching for starting the adjustment and untouching to finish the adjustment. Real-time progress bar feedback will be shown on the screen while the adjustment is activated (see Figure 2).

3.2.2 Video Volume. In the video volume control, we keep the left hand as the volume bar, however, we use the "thumb", "index" and "middle" finger gestures on the right hand and use the middle fingertip as the pointer to adjust the volume. The left hand should also keep distance from the right hand in this case, and the volume value also depends on the closest point from the right-hand index fingertip to the directed line segments between the left hand's wrist and the middle fingertip. The volume would be also changed in real-time while adjusting to the feedback (see Figure 3).

3.2.3 Screen Brightness. Similar to the video progress control, in the screen brightness control, we used the left hand as the brightness bar and the right hand as the pointer. Similarly, once the left hand appears in front of the screen, the directed line segments between the wrist and the middle finger tip would be recognized as the brightness bar and the wrist and the middle finger tip stands for 0% and 100% of the brightness of the screen. Once the brightness bar is set up, make the right hand in the gesture of only revealing the thumb and index finger. Keep distance between left hand and right hand and use the right-hand index fingertip as the pointer to adjust the brightness. The closest point from the right-hand index fingertip to the directed line segments between the left hand's wrist and the middle finger tip would be the value of the screen brightness. The brightness would be changed in real-time while adjusting to the feedback (see Figure 4).

4 PRELIMINARY FINDINGS AND DISCUSSIONS

4.1 Hands As an Interface to Express Gestures

HCI research has recognized human skin as a promising interface for interacting with intelligent computing devices [16]. Its use as an interface helps to overcome the limited surface space of today's wearable devices and allows for input to multiple smart devices. Most existing work treats the skin as a hypothetical flat surface, with the principles and models for designing interactions shifting from existing touch-based devices to the skin. In addition, current

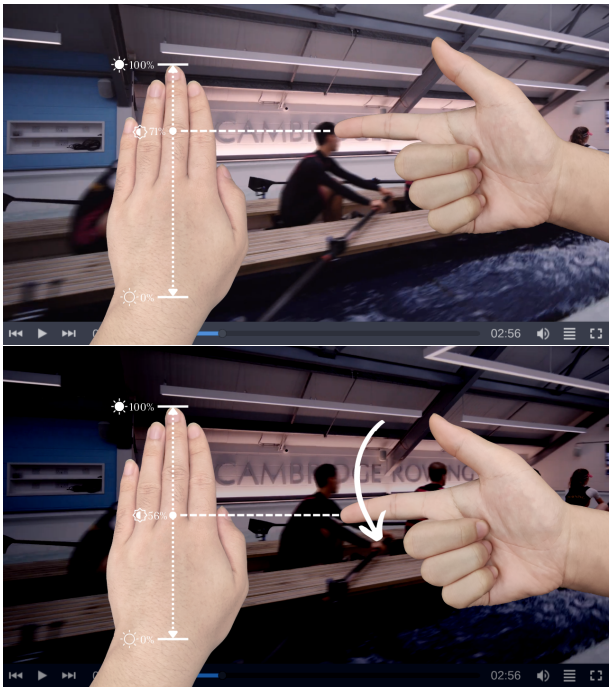

Figure 4: In the screen brightness control, we used the left hand as the brightness bar and the right hand as the pointer. Similarly, once the left hand appears in front of the screen, the directed line segments between the wrist and the middle finger tip would be recognized as the brightness bar and the wrist and the middle finger tip stands for 0% and 100% of the brightness of the screen.

skin interactions typically allow only touch gestures or taps in several different distinct locations, thus greatly limiting the possible interactions to expressive interactions with a wide range of user interfaces and applications. However, in GesPlayer, instead of directly bringing the two hands into physical contact, our design allows one hand to assist the hand that needs to provide the gesture to better express the semantic information. We believe that a visual benchmark will significantly improve the accuracy of dynamic tasks that require gestural movement.

4.2 The State of Gestural Interaction

According to the concept of "Computational Interaction" [14], the essence of interaction is the direct interplay of information expressed by behaviors and actions in different states. In previous studies, the state of gestural input/output (I/O) has been defined as only one step in the overall interaction process. In fact, the description of the state should be more clearly defined: in GesPlayer, we introduced "trigger gestures", "baseline gestures", and "interaction gestures", which represented the "start" state, the "benchmark" state, and the "input/output" state. For the user, this series of state changes is the real behavioral analysis of interaction, which could be helpful for researchers to understand the underlying behavioral semantic information.

5 CONCLUSION AND FUTURE WORK

In this paper, we introduce a gesture-based empowered video player, GesPlayer, which explores how users can experience their hands as an interface through gestures. We provide three semantic gestures based on the camera of a computer or other smart device to detect and adjust the progress of video playback, volume, and screen brightness, respectively. We articulate the technical setup of GesPlayer, its design, and present preliminary findings, and give our discussions in the form of two themes: (a) how the hand as an interface expresses the interaction of gestures, and (b) the state of hand interaction. Our goal is to enable users to control video playback simply by their gestures in the air, without the need to use a mouse or keyboard, especially when it is not convenient to do so. Ultimately, we hope to expand our understanding of gesture-based interaction by understanding the inclusiveness of designing the hand as an interactive interface, and further broaden the state of semantic gestures in an interactive environment through computational interaction methods.

In the future, we plan to conduct a user study to evaluate our GesPlayer in three areas: (a) usability and efficiency, (b) advantages and potential disadvantages of GesPlayer compared to gesture interaction without a baseline state, and (c) how semantic gestures can support a better mastery of our control over the video player.

ACKNOWLEDGMENTS

Xiang Li is supported by the China Scholarship Council (CSC) International Cambridge Scholarship (No. 202208320092).

REFERENCES

[1] Kanad K Biswas and Saurav Kumar Basu. 2011. Gesture recognition using microsoft kinect®. In *The 5th international conference on automation, robotics and applications*. IEEE, 100–103.
[2] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. 2017. A survey of depth and inertial sensor fusion for human action recognition. *Multimedia Tools and Applications* 76, 3 (2017), 4405–4425.
[3] Meghal Dani, Gaurav Garg, Ramakrishna Perla, and Ramya Hebbalaguppe. 2018. Mid-air fingertip-based user interaction in mixed reality. In *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 174–178.
[4] Florian 'Floyd' Mueller, Rakesh Patibanda, Richard Byrne, Zhuying Li, Yan Wang, Josh Andres, Xiang Li, Jonathan Marquez, Stefan Greuter, Jonathan Duckworth, et al. 2021. Limited control over the body as intriguing play design resource. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.
[5] Erving Goffman. 1981. *Forms of talk*. University of Pennsylvania Press.
[6] Varun Jain, Gaurav Garg, Ramakrishna Perla, and Ramya Hebbalaguppe. 2019. Gestarlite: An on-device pointing finger based gestural interface for smartphones and video see-through head-mounts. *arXiv preprint arXiv:1904.09843* (2019).
[7] Adam Kendon. 2004. *Gesture: Visible action as utterance*. Cambridge University Press.
[8] Panayiotis Koutsabasis and Panagiotis Vogiatzidakis. 2019. Empirical research in mid-air interaction: A systematic review. *International Journal of Human–Computer Interaction* 35, 18 (2019), 1747–1768.
[9] Xiang Li, Yuzheng Chen, Rakesh Patibanda, and Florian'Floyd' Mueller. 2021. vr-CAPTCHA: exploring CAPTCHA designs in virtual reality. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–4.
[10] Xueshi Lu, Difeng Yu, Hai-Ning Liang, Wenge Xu, Yuzheng Chen, Xiang Li, and Khalad Hasan. 2020. Exploration of hands-free text entry techniques for virtual reality. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 344–349.
[11] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172* (2019).
[12] Shahzad Malik, Chris McDonald, and Gerhard Roth. 2002. Hand tracking for interactive pattern-based augmented reality. In *Proceedings. International Symposium*

on Mixed and Augmented Reality. IEEE, 117–126.
[13] Florian'Floyd' Mueller, Richard Byrne, Josh Andres, and Rakesh Patibanda. 2018. Experiencing the body as play. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.
[14] Antti Oulasvirta, Per Ola Kristensson, Xiaojun Bi, and Andrew Howes. 2018. *Computational interaction*. Oxford University Press.
[15] Rakesh Patibanda, Xiang Li, Yuzheng Chen, Aryan Saini, Christian N Hill, Elise van den Hoven, and Florian Floyd Mueller. 2021. Actuating Myself: Designing Hand-Games Incorporating Electrical Muscle Stimulation. In *Extended Abstracts of the 2021 Annual Symposium on Computer-Human Interaction in Play*. 228–235.
[16] Jürgen Steimle, Joanna Bergstrom-Lehtovirta, Martin Weigel, Aditya Shekhar Nittala, Sebastian Boring, Alex Olwal, and Kasper Hornbæk. 2017. On-skin interaction using body landmarks. *Computer* 50, 10 (2017), 19–27.
[17] Lokesh Kumar VM. 2021. Touch and Explore: A VR Game Exploration, Based on Haptic Driven Game-play. In *Interactive Surfaces and Spaces*. 12–15.
[18] Wenge Xu, Hai-Ning Liang, Yuzheng Chen, Xiang Li, and Kangyou Yu. 2020. Exploring visual techniques for boundary awareness during interaction in augmented reality head-mounted displays. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 204–211.
[19] Wenge Xu, Hai-Ning Liang, Xiaoyue Ma, and Xiang Li. 2020. VirusBoxing: A HIIT-based VR boxing game. In *Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play*. 98–102.

Received 2022-09-30; accepted 2022-10-07