CLOUD HANDS: THREE EMANATIONS OF A TOUCHLESS MUSICAL ASSEMBLAGE

Tim-Tarek GRUND (grund@mdw.ac.at) [1] and **Alex HOFMANN** (hofmann-alex@mdw.ac.at) [1]

[1] *Department of Music Acoustics*, **University of Music and Performing Arts Vienna**, Anton-von-Webern-Platz 1/II, Vienna, 1030 Austria

ABSTRACT

Touchless control of human-computer interfaces in science-fiction has inspired several musical applications. This paper presents *Cloud Hands*, a touchless musical assemblage that controls sample-based granular synthesis via hand motion. The paper discusses how the assemblage manifests in three emanations: as a sound interface, a digital music instrument and an interactive multimedia composition. A dedicated Python script extracts the hand landmark positions and distances from a camera frame, which are mapped to control parameters of granular synthesis of a sample in Pure Data. A Wekinator patch may extract symbolic hand poses to change samples and playback speed. The emanations use varying modes of visual feedback, from no feedback at all to wire-frame models of the hand to sound material-related segmentation of video material as a reference map. Performance recollections from the first author and other users of Cloud Hands are analyzed in order to identify conveniences and constraints of the sensing technology, mapping, and sound production. Finally, the paper discusses the influence of conveniences and constraints on aesthetics, form and performance behavior.

1. INTRODUCTION

Controlling human-computer interfaces (HCI) by hand motion has been an ongoing topic of interest in science fiction film and literature, which has inspired numerous interfaces. The touchless HCI design literature frequently mentions films such as Johnny Mnemonic (1995), the X-Men series (since 2000), Minority Report (2002) and the Iron Man Series (since 2008) as influential work for hand gesture-based interfaces [1–3]. In addition, other circumstances further enhanced the general interest in touchless HCI. Leap Motion and Xbox Kinect brought hand gesture interfaces and touchless motion control into the realm of consumer electronics. Incidentally, the Covid-19 pandemic demanded an increased effort in developing touchless HCI.

Camera sensor technology is among the more commonly investigated touchless sensor technologies within the field of digital music instruments. Kinect devices [4, 5], multi-camera motion capture systems (MoCap) [6], and smartphone cameras [7,8] are some examples of this. The choice of a certain camera system (and computer environment) can greatly influence the mapping, aesthetic and performance. While dedicated motion-sensing devices, especially with high-resolution multi-camera MoCap or additional sensors, allow for a high-fidelity capture of body or hand gestures [4–6], smartphones may have to turn to alternative means of analyzing movement due to the lack of computational capabilities, for example tracking colorful reference points [8]. However, thanks to recent technical advances in this area, more computationally-heavy algorithms such as optical flow are able to be implemented on mobile devices [7].

Jensenius [9] distinguishes between the concepts of motion, action and gesture in the context of music. Here, motion is described as the "physical displacement of an object in time and space", an objectively measurable signal without beginning or end. Action then is the subjective segmentation of motion by an observer, human or machine. Gesture and pose however are concepts that ascribe some sense of meaning to actions, poses being stationary, gestures dynamic. In addition to being measured, gestures and poses are symbols that need to be understood by an observer. It should be noted however, that some of the works cited in this paper use the term pose as a synonym for position, without the symbolic component. The term "pose estimation" is one example of this. Therefore, the term "symbolic pose" is used here to differentiate positional from meaningful data. Camera-based landmark detection is a popular method for pose or gesture recognition, both positional and symbolic. Here, a computer model extracts positional data of recognizable features of the hand, face or body, usually from a camera frame or a depth sensor. This data may be further processed to extract symbolic poses and gestures. Human pose estimation or pose landmark detection (non-symbolic) has been employed to beat detection from musicians visual cues [10]. Instrument pose estimation has been applied to studying violin-bow gestures [11]. Hand landmark detection has found several use cases in sound and music programming, both in academic and commercial projects. Pet and Maezawa [12] combined MediaPipe's Hand Landmarker [13] with audio feature extractors to create a multi-modal piano score follower. Kritsis et al. [14] developed a gesture recognition system for playing virtual instruments via the Leap Motion controller.

ISBN 978-3-200-10642-0

In the communities surrounding digital music practices, a number of ways have been proposed, in which a digital music instrument might defy traditional notions of a musical instrument as an object. One of these is the conception, that a musical instrument exists as a combination of different mediations (e.g. technological, social, but also temporal). Born [15] defines a combination of this type as a musical assemblage. Drawing on Deleuze, Delanda and Latour, Born [16] discusses themes like contribution of non-human actors, relations between heterogenous components (e.g. software, hardware), and modularity and autonomy of components. Snape and Born [17] apply the concept later to the graphical programming environment Max. In a reasoning close to temporal mediation, Jack et al. [18] call into questions of finality and improvement as necessary notions in digital music instrument design. McPherson and Tahiroğlu question, whether we could "say that any instrument, including traditional acoustic instruments, might be 'composed' in the sense that its identity is inherently bound up with the pieces at the time of its creation?" [19]. These lines of thinking resonated with the way the first author experienced the design of *Cloud Hands*. The mediation, especially through technology, intentions and time, compelled the first author to regard it as a musical assemblage, from which three emanations radiate into realization.

In this paper we introduce *Cloud Hands*, a touchless musical assemblage using hand motion extracted from camera frames to control parameters of sample-based granular synthesis. We present its development through three *emanations* as a sound interface, a digital music instrument and an interactive multimedia composition. Finally, we discuss the impact of conveniences and constraints on aesthetics and performance.

2. CLOUD HANDS

Cloud Hands extracts hand landmark positions from a live camera feed and maps these to control parameters of granular synthesis in Pure Data. It maps the horizontal position of the user's hands to the playback position of a sample and the distances between index finger, little finger, and thumb to synthesis parameters. This allows users to choose a location within a sample, play it back in a continuous, granular fashion, and alter the size of the synthesis grains to unearth the local sonic content. Control via symbolic poses was implemented to facilitate changing samples and playback speed. *Cloud Hands* has been subject to changes during the course of its existence. It was originally conceptualized and presented as an interactive sound interface that allows users to investigate different sections of an audio signal, but it has since been used as a digital music instrument in performances and recently rewritten as an interactive multimedia composition.

2.1 *Cloud Hands* as a Sound Interface

Cloud Hands was conceived in 2022 during a university seminar on investigating alternative approaches to sound-aided data exploration beyond sonification by the Inter-

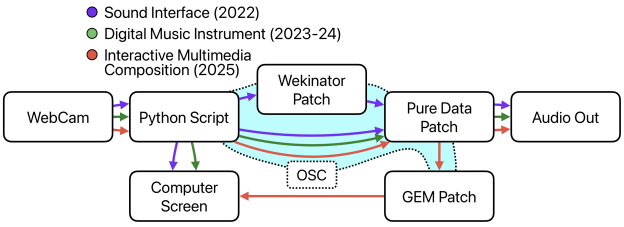

Figure 1. Data flow of the three different emanations of *Cloud Hands* between 2022 and 2025. Arrows represent data flow with colors according to the corresponding emanation. Arrows within the blue area represent transmission via OSC.

specifics collective [1] at TU Berlin, that the first author, who wrote this manuscript, attended. Inspired by various works of science-fiction films and literature, the first author wanted to create an interface for the exploration of sound material using hand motion. Similar to the iconic user interface in Minority Report (2002), where hand gestures allow the user to scroll through a video feed and zoom into images, the aim of *Cloud Hands* was to investigate the sound by selecting a playback position using the horizontal displacement of the hand and alter the grain size by opening or closing the hand. Sound material from different parts of a sound file ought to be layered by using both hands, creating a collage of sound and enabling investigation by ear of different sections at once.

The basic operating principle of *Cloud Hands* is shown in Figure 1. A Python script receives input from a camera feed, frame by frame. Using MediaPipe's hand landmark model, it extracts hand landmark positions, calculates distances between thumb, index finger, and pinkie of each hand (intra-hand distances) and between fingers of the same categories across hands (inter-hand distances). Positional and distance data is sent via Open Sound Control (OSC) to a Wekinator patch [20] and a Pure Data patch [21]. Wekinator then uses a k-Nearest Neighbor model to classify, whether the hand landmark positions correspond to a trained symbolic pose and whether it is performed in one of two areas relative to the frame. A symbolic pose similar to the ILY hand sign from the American Sign Language was chosen due to performance considerations. It is possible to move from an open hand to the ILY hand sign simply by closing the middle and ring finger, while keeping thumb, index and pinkie in the same location. Wekinator sends the classification result using OSC to the Pure Data patch. Here, a positive classification result from one area is mapped to changing the sample and from another to changing the playback speed. Hand landmark positions and distances extracted from the Python script are then mapped to synthesis parameters of a sample-based granular synthesis algorithm.

The aesthetics of sound production of *Cloud Hands* are dominated by the idiosyncratic sonic characteristic of a particular method of granular synthesis, where "grains", small, windowed sections of a source sound material, are looped. The range of possible grain lengths spans from

[1] https://interspecifics.cc/work/

ISBN 978-3-200-10642-0

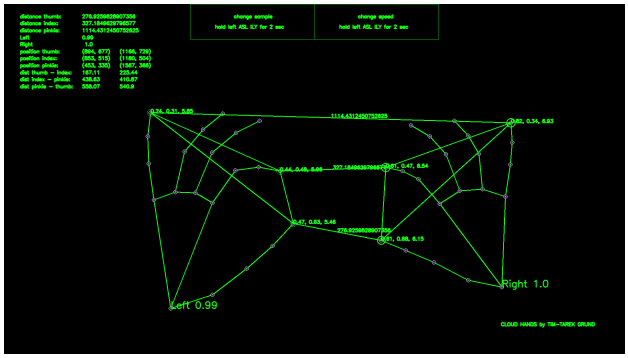

Figure 2. The visual scene of early *Cloud Hands* iterations. A wire-frame model of hand landmarks with additional wires connecting thumbs, index fingers and pinkies provides visual feedback for the performer. The top left corner features positional and distance data of hand landmarks alongside detection confidence values of left- and right-handedness. The top middle area contains two positional indicators for hand gesture placement.

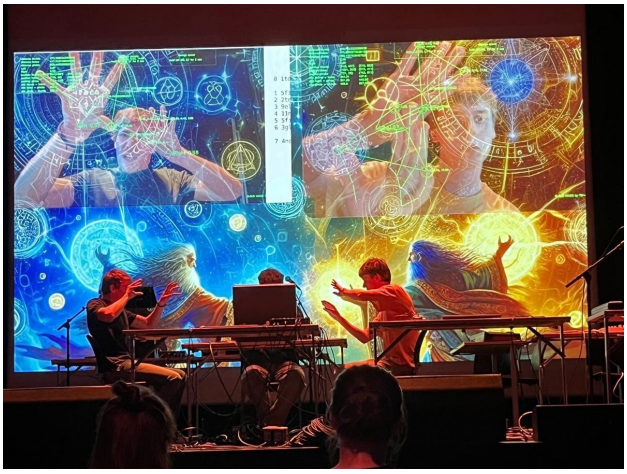

Figure 3. Schuster, Meling and Hegdahl performing with two instances of *Cloud Hands* and an Octatrack MKII in a concert performance in Trondheim, Norway, 2024. The image overlay of the fighting wizards serves as a graphical notation. © Malte Strewick.

around 40 ms to around 1000 ms from the smallest detectable finger distance between thumb and index finger positions to the largest, as tested on the webcam of an Apple M2 MacBook Pro. This is only in part in the realm of what is traditionally accepted as granular synthesis, a grain length of one to 100 ms [22]. As a result, *Cloud Hands* can move seamlessly between a perceptually granular sound on the smaller end of scale, and regular, looped sample playback on the larger end.

The Pure Data patches are designed in such a way that each hand controls three abstractions called grain.pd, that each contain three linear ramp generators from start to end position of a grain, modulated by low-frequency oscillators with their phases shifted by thirds of a period. Then, a table lookup using the generated values is performed and subsequently windowed using a cosine window. Each grain.pd has an individual stereo pan and a positional offset, both of which are modulated by the distance between thumb and index finger. One is always panned to the center, one to the left, one to the right. Moving the hand from left to right across a transient event, the sound will first play from the grain.pd patch placed in the middle, then from the right, then from the left channel. This produces a sonic ambiguity confined to a small interval, which serves to blur the transitions from one region of the sound material to another.

In the interface version of *Cloud Hands*, the depiction of hand landmark positions and the distances between them as monochrome circles and lines on top of the live video feed were the defining element of visual feedback. The visual elements and simple color choices served as a stylistic references to the era of monochrome computer monitors and The Matrix (1999). This visual scene, as created from the Python script, can be seen in Figure 2. A wire-frame display is the most prominent element of the scene. It features circular markers at the joint positions and uses lines to visually highlights intra- and inter-hand connections. From the top left corner, the scene displays several parameters of the pose recognition: Intra-hand distances, the detection confidence of left- and right-handedness, positional data on the three fingers and inter-hand distances. All positional data is given in pixels. Some of these values are displayed next to selected markers of the wire-frame model, either in values relative to the frame dimensions or in pixels. The top of the scene features two demarcated areas. These are the positional indicators for the areas, in which one can use the ILY hand gesture to change the sound material (left) or the playback speed (right). It is important to note, that the hand landmark positions used are in reference to the image coordinate system of the video frame instead of a world coordinate system. Thus, the distance of hand landmarks to the center of the frame increases with shrinking distance to the camera, which has implications for interaction. The ambiguity between smaller motion close to the camera and bigger motion further away requires some attunement, especially when changing camera systems. As a consequence, the first author developed the habit of either controlling the openness of his hand or the distance to the camera, while keeping the hand parallel to the image plane. The audio feedback also plays an important role in informing about the measured openness.

2.2 *Cloud Hands* as a digital music instrument

Cloud Hands as a digital music instruments [2] uses the same Pure Data patches as the sound interface, both mapping and sound production were kept mainly as they were. However, new sound material was added and occasionally reverberation was added to the signal path. *Cloud Hands* requires camera peripherals and the ability to execute Python scripts and pure data patches, and has been performed with on several systems. In the first author's personal artistic practice, it runs on an Apple M2 MacBook Pro.

[2] https://github.com/grundton/cloud_hands

ISBN 978-3-200-10642-0

During the first and second installation several users commented on its potential as a musical instrument. The first author also noticed during the development process that he would find himself improvising without end, when he simply wanted to test incremental changes to the synthesis code. In consequence, *Cloud Hands* inspired a series of short video and live performances [3]. The Python script was expanded to include command line arguments that provided options to change the color, disable the live video feed, or disable visual feedback altogether to promote customization and to invite performers to test their sense of proprioception instead of depending on visual feedback. Through presentation in workshops, it has also sparked the interest of other artists, such as Paul Schuster, Aslak Meling and Hugo Hegdahl, who used it in *Dueling Wizards*. It has been played individually, in a duo setting, and alongside other (live-electronically enhanced) instruments.

2.2.1 *Dueling Wizards*

An interesting performance concept was displayed by Schuster, Meling and Hegdahl in their concert performance *Dueling Wizards*, Fig. 3. Here, two musicians enter a musical battle as opposing magicians that duel in front of a third musician, the "high court". After an introductory section, the musicians cast different "spells" by performing with *Cloud Hands* with different samples. The sonic output of these spells is processed by the "high court", who in the end gets to decide the winner of the duel. In this composition, an image of two wizards surrounded by magical circles is used as a graphical score to guide performers and the audience through the duel. Each magic circle denotes a different spell, which requires the performer to work with a different sound material. The size indicates the duration that a wizard has to perform with this sample. Once a winner is decided, sounds and gestural expressions of the performers should reflect the outcome.

2.2.2 `GL00M` *and Leiyla and the Poet*

`GL00M` [4] is a duo performance with Cloud Hands, saxophone and live-electronics with live video projection, which was premiered in a Nightclub during the *Kilele Summit 2024*, a music technology summit in Nairobi, Kenya. The second author plays the tenor saxophone, which places motifs, electro-acoustically enhanced through Csound, over a dense pad of granular piano sounds which the first author produces with *Cloud Hands*. The visual feedback of *Cloud Hands* played an important role, as the scene (Fig. 2) was projected on a wall in the Nightclub in order to immerse the audience in the performance. *Cloud Hands* was further introduced to the participants of the Kilele Summit during a workshop on "Building live-electronic instruments with Pure Data derived from tape-music compositions". [5] Parts of it, most importantly the Python script, were used to control a synth emulation of a reed flute from Halim El-Dabh's "Leiyla and the Poet" (1959) using hand landmark detection. This touchless reed flute emulation instrument was later played by composer and sound artist Nyokabi Kariũki in a live performance of "Leiyla and the Poet".

2.2.3 *Personal instrumental practice*

During the first author's personal instrumental practice with *Cloud Hands*, an influential factor is the delay between analyzed frames. While the MacBook Pro camera is certainly capable of recording at 28 frames per second, an inter-frame duration of around 36 ms, the hand landmark estimation process and the Python script almost double the inter-frame duration to around 70 ms. Performing with such an amount of delay enforces some behavioral practices. As a result, the first author tends to perform *Cloud Hands* with slow hand movements, which actively promotes conscious listening. The sound generation process also enforces slow motion of the hands and fingers. The ramp generator that reads from the sound material has a freely controllable playback speed, but the grain length is updated only after every cycle in order to avoid discontinuities. If the grain length is altered by closing thumb and index finger, this leads to immediate jumps in the playback speed. These are mitigated through a timed ramp generator, which instead creates pitch speed artifacts that produce a tape acceleration or deceleration effects. With slow hand motion, this effect is barely noticeable. The first author therefore adjusted his performance speed to suppress this effect. Furthermore, the Wekinator script was abandoned. Changing the playback speed and sound material is more easily accomplished by pressing a button, changing it by using a symbolic pose feels incongruous, since it is the only one. The first author adopted the practice of using a single sound material and a fixed playback speed throughout each piece. To this end he started experimenting with arranging sound material specifically for performance. While he had previously selected sound material on the basis of exploring its sonic content, the first author now recorded and arranged sound material in such a way that it allowed for moving the hands to the relevant positions ergonomically, mentally projecting content of sound material to sections in space. Some fruitful albeit at times mutually exclusive strategies for arranging content of the sound material were employed, placing content to be played in succession next to each other, placing the content most suitable to layering in the middle, doubling the content so each hand can access each part, and concatenating pitch shifted content from multiple sources. Also, the playback direction is set in such a way, that a sample will be played forward by moving from left to right. When already playing in the middle of a waveform section, one knows that transient content is to the left, more decaying content to the right. However, with the hands out of the frame a performer has no way of knowing the spatial relationship between their body and the sonic content except for proprioception. As this alone did not fully support the need for intentionality and repeatability, due to varying performance conditions, it compelled the first author to develop *Cloud Hands* further from an instrument to an

[3] `https://www.youtube.com/watch?v=pOW0q13Kjdw&list=PLNiGqDliYwa8nJAKfLl-LcDxoJCke5jZV`

[4] `https://www.youtube.com/watch?v=ktEx26GQkPw`

[5] `https://github.com/grundton/Building-live-electronic-instruments-kilele`

ISBN 978-3-200-10642-0

Figure 4. A frame from the visual scene of *Cloud Hands* as an interactive multimedia composition. The discontinuities in the visual display divide the image into seven segments that correspond to sections with different sonic content of the sound material.

interactive multimedia composition, which provides visual information on the position of sonic content within in the sound material.

2.3 *Cloud Hands* as an interactive multimedia composition

The main addition to the sonic aesthetic of *Cloud Hands* as a digital music instrument was a revision of the volume or gain mapping. In the previous version, the vertical position of the thumb was mapped to a (linear) gain with a quadratic function with a negative coefficient of the quadratic term of -4 and a positive coefficient of the constant term of 1, so that the maximum volume was placed in the middle of the vertical range of the camera frame. Values above one and below zero were clipped. This however resulted in some problems. If the thumb appears to be out of screen, the hand landmark model has to estimate an approximate position, which can result in positional noise. The thumb position may alternate between inside and outside areas of the frame camera frame, creating a stuttering playback. Positional noise may also produce pitch variation artifacts. Due to the quadratic functions' gradient increasing in magnitude further away from its vertex point, the places with the highest magnitude of the gradient are at the borders of the input range. Incremental changes in horizontal position produce here the largest (linear) gain difference, resulting in undesired boosts of sound and occasionally failure to shut off. In performances, this was mitigated by a separate, external volume control using MIDI devices.

For the new volume mapping, a function was designed that is symmetrical over the input range, had a gradient of zero at the borders of the input range and additionally a gradient of zero at the peak in the middle of the input range as well. To this end, the volume mapping was devised by a concatenation of two sigmoid functions. Each covers one half of the input range, while one is mirrored. The sigmoid functions are stretched before the output is clipped between zero and one to suppress their asymptotic behavior. From a performance perspective, this meant that the positional uncertainty is less likely to produce any noticeable sound artifacts when moving the hands in and out of the frame. Lastly, a reverb effect was added to the final output to add depth to the clouds of grains.

Now, the visual feedback formerly handled by the Python script is instead implemented in a Pure Data patch using the Graphics Environment for Multimedia (GEM). This has several advantages. It provides access to GEM's library of visual effects that facilitate video playback and enable OpenGL Shading language (GLSL), but also simplifies the integration of visual feedback and signal information. All these points facilitate the inclusion of a dynamic, visual score to be displayed on the screen, which consolidates the continuous reconfiguration of *Cloud Hands* from an auditory display to an instrument to an interactive multimedia composition.

In order to act with intention and repeatably in a musical manner, as opposed to the previous mainly exploratory interaction, it is necessary to have a known frame of reference of how bodily motion relates to musical output. Since *Cloud Hands* is performed in the air, there is no physical representation of it from which a performer can receive haptic feedback. Proprioception thus becomes the main sensory feedback channel. However, in order to aid realizing the physical distances involved in playing, especially under conditions where the distance to the camera or the camera equipment itself might vary between performances, a visual feedback channel was incorporated. In essence, representations of hand landmark positions and information on sections of the sound material were tied together in the visual scene.

The main scene consists of a single video feed, wireframe models of the players hand and the connection elements. The left and right borders of the video serve as visual delimiters of the playable area. Areas of interest in the sound material form the basis for a segmentation of the video, so that the boundaries between video segments coincide with border positions of areas of interest within the sound material. These visual boundaries are created only by modulating vertical position and horizontal cropping offset of these segments in random directions, affected in scale by the vertical position of the thumbs. This creates perceivable discontinuities in the video material, which can be interpreted as a reference map of sound material. Yet, it also relies on interaction, as it is only activated, if a hand is detected. To this end, one has to move their hands into the detection range, which further promotes slow and conscious transitions from inactive to active. Without a detected hand, the segments form a consistent video that plays and loops. However, once a hand is detected, the video frame index is modulated by the horizontal index finger position, as if time stopped for both audio and video material.

The background that frames the main scene is constituted of the same video. This video is unsegmented, but its effects are modulated by three parameters. First, a zoom effect is applied on the video based on horizontal index finger displacement. Then, a motion blur effect is applied, whose strength is not modulated by hand landmark posi-

ISBN 978-3-200-10642-0

tions but by the volume envelope instead. The angle of the motion blur is controlled by the angle between the vector difference of the index finger positions and the horizontal axis. The relative hand landmark positions are displayed at the top left, volume at the bottom left. Additional indicators display the relative horizontal index position inside the main scene.

There is a defining coupling of themes between visuals and sound. The present application of granular synthesis technique subjects existing sound material to a process of segmentation and looping. The visual domain follows this approach by looping a video and segmenting it into zones. These zones however are not segmented by pronounced borders, but instead by the discontinuity of elements, by movement in opposite directions (Fig. 4). Morphing between grain loops in the audio domain, between a single video and segmented zones in the visual one, but also performing with a hand motion just below the threshold of musical gestures: The exploration of the perceptual thresholds between neighboring states is an immanent theme throughout *Cloud Hands*.

3. CONCLUSIONS

This paper presented *Cloud Hands*, an assemblage that morphs between an interface, an instrument and an interactive composition. Conveniences and constraints of the sensing technology, mapping and sound production were identified that encourage and enforce musical hand motion which assists an attentive aural interaction with the sound material. To this end, performance recollections from the first author and others were included, which seem to have, despite their different approaches, some concurring themes.

The paper described the relationship between the heterogenous elements that make up *Cloud Hands*, their materialities, and their detachment and reattachment which realizes in the different emanations. The reed flute emulation instruments used for the reperformance of "Leiyla and the poet" arguably contained the smallest common denominator of *Cloud Hands*, the hand landmark detection Python script. For the first author, *Cloud Hands* presents itself as an assemblage. The "Cloud" in *Cloud Hands* is therefore not only a reflection of its fluidity of sound but also of its fluidity of form. During *Cloud Hands* performances, performers tend to keep their hands on the respective side of the screen, with the hand plane almost parallel to the image plane. In his reflections on using *Cloud Hands* in his performance on 'Dueling Wizards' Schuster notes that "it is very difficult to make precise instructions and inputs, as perfect tracking of the hands is never achieved despite the very high camera quality" (personal communication). He further describes that when holding the hand at an angle so that it is almost perpendicular to the image plane the pose estimation becomes unreliable and unpredictable. This coincides with he first author's own performance habits of controlling the openness of the hand. The unreliability of hand landmark detection apparently enforces certain performance practices with camera interfaces. In an effort to maximize detection accuracy, performers may tend to avoid ambivalent angles and occlusion of the other hand. This reduces the space of possible hand positions, which may require strategic composition of the sound material.

In order to structure a performance with *Cloud Hands* through time, Schuster, Meling, and Hegdahl use a graphic score, organizing different samples and performance phases. Schuster notes on the graphical score that it aids both improvisation and composition by providing a guideline for musical variety in improvisation but also precise, "strict sequences" (personal communication). The visual presentation of the graphic score, of *Cloud Hands* visual landmark presentation but also of the performers' facial expressions guide the performers but also help the audience comprehend the performance. On the other hand, the first author used a visual video score in the interactive multimedia composition to organize the layout of *Cloud Hands* through space. The discontinuities in the video playback of *Cloud Hands* inform the performer but also the audience of the location of different sound material segment. However, the performers' symbolic gestures are of little importance to this approach, the only indicator of the performers presence in this composition is the representation through the hand landmarks.

It is striking, how these two very different approaches of composing for *Cloud Hands* work with its conveniences and constraints. The former uses the visual capabilities to organize time, The latter organizes space. The former puts the performer in an almost theatrical position, the latter removes everything but their digital traces from the performance.

4. AUTHORS' CONTRIBUTIONS

TTG wrote this manuscript, developed the software for *Cloud Hands*, and led the "Building live-electronic instruments with Pure Data derived from tape-music compositions" workshop. AH supervised the research, co-organized the Kilele 2024 summit and the aforementioned workshops. The authors read and approved the final manuscript.

Acknowledgments

The first author wants to thank the Interspecifics collective for the lecture course that inspired *Cloud Hands*, Paul Schuster for his detailed explanation of *Dueling Wizards* and the Kilele workshop attendees for their feedback. We would like to thank the reviewers for their helpful feedback. This research was funded in whole by the Austrian Science Fund (FWF) [10.55776/AR743].

5. REFERENCES

[1] M. Schmitz, C. Endres, and A. Butz, "A Survey of Human-Computer Interaction Design in Science Fiction Movies," May 2010.

[2] J. Iio, S. Iizuka, and H. Matsubara, "The Database on Near-Future Technologies for User Interface Design from SciFi Movies," in *Design, User Experience, and Usability. Theories, Methods, and Tools for Designing*

ISBN 978-3-200-10642-0

the User Experience, A. Marcus, Ed. Cham: Springer International Publishing, 2014, pp. 572–579.

[3] C.-K. Yang and Y.-C. Chen, "A HCI interface based on hand gestures," *Signal, Image and Video Processing*, vol. 9, no. 2, pp. 451–462, Feb. 2015.

[4] J. M. Fernandez, T. Köppel, N. Verstraete, G. Lorieux, A. Vert, and P. Spiesser, "Gekipe, a gesture-based interface for audiovisual performance," in *Proceedings of the International Conference on New Interfaces for Musical Expression*. Copenhagen, Denmark: Aalborg University Copenhagen, 2017, pp. 450–455.

[5] R. Pritchard and I. Lavery, "Inexpensive colour tracking to overcome performer id loss," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, R. Michon and F. Schroeder, Eds. Birmingham, UK: Birmingham City University, July 2020, pp. 89–92.

[6] D. Sardana, W. Joo, I. I. Bukvic, and G. Earle, "Introducing locus: a NIME for immersive exocentric aural environments," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, M. Queiroz and A. X. Sedó, Eds. Porto Alegre, Brazil: UFRGS, June 2019, pp. 250–255.

[7] C. Arslan, F. Berthaut, J. Martinet, I. M. Bilasco, and L. Grisoni, "The phone with the flow: Combining touch + optical flow in mobile instruments," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, T. M. Luke Dahl, Douglas Bowman, Ed. Blacksburg, Virginia, USA: Virginia Tech, June 2018, pp. 148–151.

[8] H. de Souza Nunes, F. Visi, L. H. W. Coelho, and R. Schramm, "Sibilim: A low-cost customizable wireless musical interface," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, M. Queiroz and A. X. Sedó, Eds. Porto Alegre, Brazil: UFRGS, June 2019, pp. 15–20.

[9] A. R. Jensenius, *Sound Actions: Conceptualizing Musical Instruments*, ser. The MIT Press. Cambridge: The MIT Press, 2022.

[10] S. Chakraborty, S. Aktaş, W. Clifford, and J. Timoney, "Beat estimation from musician visual cues," Jun. 2021.

[11] Z. Vamvakousis, A. P. Carrillo, and R. Ramirez, "Acquisition of violin instrumental gestures using an infrared depth camera," Jul. 2018.

[12] K. Pet and A. Maezawa, "Combining Hand Pose Estimation with Audio for Noise-robust Piano Score Following," Jul. 2024, publisher: Zenodo.

[13] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "MediaPipe: A Framework for Building Perception Pipelines," Jun. 2019, arXiv:1906.08172 [cs].

[14] K. Kritsis, A. Gkiokas, M. Kaliakatsos-Papakostas, V. Katsouros, and A. Pikrakis, "Deployment of lstms for real-time hand gesture interaction of 3d virtual music instruments with a leap motion sensor," Jul. 2018.

[15] G. Born, "On Musical Mediation: Ontology, Technology and Creativity," *Twentieth-Century Music*, vol. 2, no. 1, pp. 7–36, Mar. 2005.

[16] ——, *Music and the Social*, 2nd ed. Routledge, 2012.

[17] J. Snape and G. Born, *Max, music software and the mutual mediation of aesthetics and digital technologies*, ser. A planetary anthropology. UCL Press, 2022, p. 220–266.

[18] R. Jack, J. Harrison, and A. McPherson, "Digital musical instruments as research products," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, R. Michon and F. Schroeder, Eds. Birmingham, UK: Birmingham City University, July 2020, pp. 446–451.

[19] A. McPherson and K. Tahıroğlu, "Idiomatic patterns and aesthetic influence in computer music languages," *Organised Sound*, vol. 25, no. 1, p. 53–63, Apr. 2020.

[20] R. Fiebrink, D. Trueman, and P. R. Cook, "A metainstrument for interactive, on-the-fly machine learning," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, Pittsburgh, PA, United States, 2009, pp. 280–285.

[21] M. Puckette *et al.*, "Pure data: another integrated computer music environment," *Proceedings of the second intercollege computer music concerts*, pp. 37–41, 1996.

[22] C. Roads, *Microsound*. Cambridge, Mass: MIT Press, 2001.

ISBN 978-3-200-10642-0