Hindawi
Human Behavior and Emerging Technologies
Volume 2023, Article ID 8417012, 15 pages
https://doi.org/10.1155/2023/8417012

Research Article

On the Multimodal Resolution of a Search Sequence in Virtual Reality

Nils Klowait

Faculty of Mechanical Engineering, Department of Technology and Diversity, Paderborn University, Germany

Correspondence should be addressed to Nils Klowait; nils.klowait@gmail.com

Received 15 August 2022; Revised 30 January 2023; Accepted 17 February 2023; Published 27 March 2023

Academic Editor: Zheng Yan

In virtual reality (VR), participants may not always have hands, bodies, eyes, or even voices—using VR helmets and two controllers, participants control an avatar through virtual worlds that do not necessarily obey familiar laws of physics; moreover, the avatar's bodily characteristics may not neatly match our bodies in the physical world. Despite these limitations and specificities, humans get things done through collaboration and the creative use of the environment. While multiuser interactive VR is attracting greater numbers of participants, there are currently few attempts to analyze the in situ interaction systematically. This paper proposes a video-analytic detail-oriented methodological framework for studying virtual reality interaction. Using multimodal conversation analysis, the paper investigates a nonverbal, embodied, two-person interaction: two players in a survival game strive to gesturally resolve a misunderstanding regarding an in-game mechanic—however, both of their microphones are turned off for the duration of play. The players' inability to resort to complex language to resolve this issue results in a dense sequence of back-and-forth activity involving gestures, object manipulation, gaze, and body work. Most crucially, timing and modified repetitions of previously produced actions turn out to be the key to overcome both technical and communicative challenges. The paper analyzes these action sequences, demonstrates how they generate intended outcomes, and proposes a vocabulary to speak about these types of interaction more generally. The findings demonstrate the viability of multimodal analysis of VR interaction, shed light on unique challenges of analyzing interaction in virtual reality, and generate broader methodological insights about the study of nonverbal action.

1. Introduction

Two themes are challenging the status quo of video-based interaction analysis: firstly, we see the rise of technologically mediated interaction [1–3]. We can no longer rely on participants being physically copresent nor can we expect interactants to deploy the full scope of embodied resources: in Zoom, microphones may be muted, parts of the body obscured, and mutual pointing may be practically impossible [4, 5]; in short, interactants are facing fractured ecologies [6, 7]. Moreover, new resources may become available: emojis, chats, whiteboards, etc. In short, not all analytical conventions can be imported from the physical world [8, 9], and considerable adaptation is required to render new forms of action visible [10, 11].

Secondly, the rise of multimodality—with its focus on *embodiment* and its move away from a standardized analytical treatment of all social situations—has increasingly penetrated into realms that were traditionally unavailable to the somewhat talk-centric analytical toolset of conversation analysis [12, 13]. We are now trying to transcribe silent actions [14], study multisensorial and haptic modalities [13, 15, 16], and are even attempting to include nonhumans within our analytic scope [11, 17–20].

This paper will attempt to investigate a case where technological mediation, as well as unorthodox interactional resources, is particularly pronounced: fully nonverbal, embodied, multiparty interaction in immersive virtual reality (VR), where both parties are present, with virtual body analogues, in a three-dimensional virtual space, and where their bodily motions are accurately translated into the bodily motions of their avatars (Figure 1). In short, VR makes it possible for two physically disparate persons to use avatars for embodied interaction within the same virtual space. Using

Figure 1: VR system with controllers and a headset.

videodata collected from within VR, this paper attempts to make initial forays into multimodal virtual interaction analysis.

We will move from a brief introduction of the technology to the analysis of a multimodal sequence in a specific virtual space. Due to the unusual nature of the data, we have opted to introduce the general technological context—and its fit with our methodology—before introducing our concrete analytic case.

1.1. The Rise of Affordable Virtual Reality. Although virtual reality systems are not entirely new, they were too impractical to be afforded a place in mundane everyday interactive environments. Bulky, technologically limited, and expensive, they were largely situated in academic research labs and therapeutic environments [21].

With the advent of affordable high-tech consumer electronics, particularly the miniaturization of processors, we are now at the forefront of VR becoming a more ubiquitous part of everyday life and research [22–27].

Until very recently, only enthusiasts could afford immersive virtual reality systems: these required powerful computers for their processing and needed special sensors to be installed in a fixed environment. In mid-2019, Facebook's recent acquisition—Oculus—launched the first Oculus Quest; much like its more expensive enthusiast counterparts, it was capable of both head and hand presence: users could put on the helmet and appear in a virtual environment, with the system keeping track of head and hand movement accurately and other VR users being able to observe their embodied action. After a brief setup phase, any physical location could become an entry point into a multitude of virtual spaces, without the need of external sensors or any additional hardware [28, 29].

In short, virtual reality has become affordable enough to represent a visible, albeit marginal, component of the modern multimedia landscape (for a critical examination, see [30]). Though it is currently largely marketed for entertainment or training purposes, with ample "experiences" that let a single user walk through an imaginary landscape, escape a burning building, or wield lightsabers, a focus on VR's potential for furnishing multiuser interactions is increasing (for a review of the diversification of VR, see [31]). This is evidenced by the emergence of technologically advanced multiplayer titles, moves to create virtual social networks, and a greater availability of multi-user VR experiences overall [32, 33]. Indeed, Meta's heavy involvement in the VR market hints at a plan to facilitate tele-copresent social interaction between physically distant users [34]. Moreover, with the recent relevance of telemediated education [35, 36], mixed reality social situations are likely to become more ubiquitous [37, 38], with substantive consequences for learning [39].

With this emergence of virtual interactional encounters, challenges arise, many of which pertain to social aspects of the organization of a shared VR space: how to organize locomotion and deixis, which virtual objects are mutually visible, which of the user-tracking information (such as gaze direction, voice, bodily positions, and gestures) is rendered as visible to all participants, and many others, down hitherto unexplored avenues. These issues cannot be tackled in isolation, as they are not purely subservient to the whims of designers or to the constraints of technologies: they pertain chiefly to the realm of local social interaction, the copresence of interactants in an unfolding encounter, and the locally generated—and sustained—norms and meanings. In other words, they mark the territory of interaction analysis.

Before proceeding to an analysis of our specific data, the following section will briefly outline the chosen methodological approach.

2. Methodology

2.1. Introduction to Multimodal Analysis. Around the turn of the millennium, fields occupied with the microanalysis of situated social interaction experienced a turn towards multimodality.

> *Within perspectives inspired by gesture studies and the study of social interaction, such as conversation analysis, the term is used to refer to the various resources mobilized by participants for organizing their action – such as gesture, gaze, facial expressions, body postures, body movements, and also prosody, lexis and grammar. The plurality of "modalities" referred to in this term treats multimodality as constitutive and primary. This encourages a view of modalities as constitutively intertwined, and language as integrated within this plurality as one among other resources, without any a priori hierarchy.* ([40], p. 338)

In other words, fields like conversation analysis (CA) moved towards a terminology, data-gathering and transcription method that could analyze unfolding interaction without the a priori prioritization of language [41–45]. Practically, this had the consequence that videography became a

more substantial component of CA (see [46]) and that the standard Jeffersonian method of CA transcription [47]—with its focus on the transcription of audible interactional devices—was extended to account for inaudible conversational elements such as the body, gaze direction, and the material environment. This turn is by no means complete, with several transcription conventions being developed in parallel—and with different points of analytic emphasis [12, 40, 46].

The multimodal turn did not change the fundamental focus of analysis: the study of unfolding interactional order and the resources deployed by the copresent participants to transform, uphold, or challenge it [48, 49]. In other words, even in the absence of traditional talk-based turn-taking arrangements, moments of interaction have a temporal-normative structure capable of being investigated by attending to the way the *interactants themselves* attend to this process of structuring, no matter whether the interaction is a game of hopscotch [41], a physician's visit [42] or, indeed, a play session in virtual reality. Following the progenitor of the field, Garfinkel [50], the focus is squarely on a detailed analysis on the ethnomethods, i.e., the locally co-constructed sensemaking and action strategies rather than the generation of external categories for understanding what is going on.

2.2. Multimodal Analysis of VR Interaction. Interaction in virtual reality has a number of characteristics that make it a peculiar object of study. As already established, the interactants do not have a physical body. For all practical purposes, they are two-handed, fingerless heads floating in space. The position of their virtual body—which does not *have* to be humanoid in form—is approximated using inverse kinematics (IK), where the known positions of the tracked elements (head and hands) are used to orient a virtual skeleton and body.

Additionally, the things that are mutually visible to the copresent interactants might be subject to varying degrees of mismatch. Since the purpose of, for example, a gestural sequence may be contingent upon the mutual availability of the precise trajectories of hand and head movements, any deviations due to latency, lag, different rendering modes, etc. might radically transform how this gestural sequence may appear to an observer. A "HELLO" traced in virtual space might, for example, become visible as a "H-I\\." to an interactant logged in from a great physical distance, since the transmission rate of the writer's movement can modify how their actions become visible to their interlocutor. If, for example, it takes me two seconds to draw an "L" shape in the air, and the given virtual space tracks motion trajectory every two seconds, the motion might be extrapolated based on the initial and final positions of my hand, thus resulting in an entirely different shape being displayed to an observer (Figure 2).

This circumstance can have quite pragmatic consequences for mundane embodied actions. Due to the liberties taken by inverse kinematic computations, a physical *bow* might be rendered as a squat in VR.

In the below illustration (Figure 3), the IK computation does not have any information about the position, rotation,

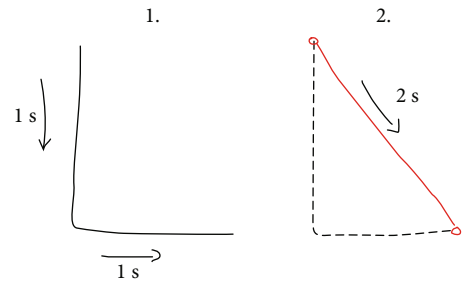

Figure 2: Distorted drawing due to latency.

and translation of anything but the controllers and headset (highlighted in blue), which are used to pinpoint the location of the hands and head, respectively. As such, it needs to make an educated guess about the logical position of the legs and torso. In the physical world, the person wearing the headset is performing a bow by keeping the knees locked and having the upper body follow the descending arcing of the head. In VR, the avatar is made to take a step forward, into a squat that leaves the arms trailing behind. From the perspective of the system, the position of the arms and head can be used to infer a divergent set of body poses. For users in VR, this can produce ambiguities in certain interactional situations. For instance, it may create the impression that a person decided to perform a squat as a formal greeting, rather than a bow.

On the turn-organizational level, similar issues may be present. A second-pair part [51] might, due to latency, be audible much later than when it was originally produced, potentially leading to additional ambiguities [9]. More fundamentally, it is not a given that each interactant sees the same number of coparticipants in any particular space, since some VR-based spaces allow the individual user to selectively delete, mute, or otherwise hide other participants.

With all these caveats, however, there are a few good arguments for using specifically *multimodal* CA for VR. Firstly, VR is arguably the first truly embodied, mediated experience: interactants may, through mutually visible attention, orient towards same or different objects, and even the rudimentary nature of controllers tracked in 3D space makes it possible to indexically incorporate (see [12]) ongoing action, the "material" environment, and the different possible interactional ecologies available in the heterogenous spaces of VR. Secondly, precisely because the spaces are so heterogenous, multimodal analysis—with its move away from the assumption of a talk-like sequential structure as the primary interactional playspace—is well suited to the exploration of the idiosyncratic modal configurations that may obtain in any specific VR environment. Lastly, and most importantly, the above-described difficulties and strangenesses of VR are not purely analytical categories—they are ethnomethodological considerations available as objects to the interactants themselves. As such, the exploration of interaction in unorthodox realms is also the exploration of the interactants' means of facing these realms. Indeed, the subfield of atypical interaction analysis, which has historically been concerned with the analysis of

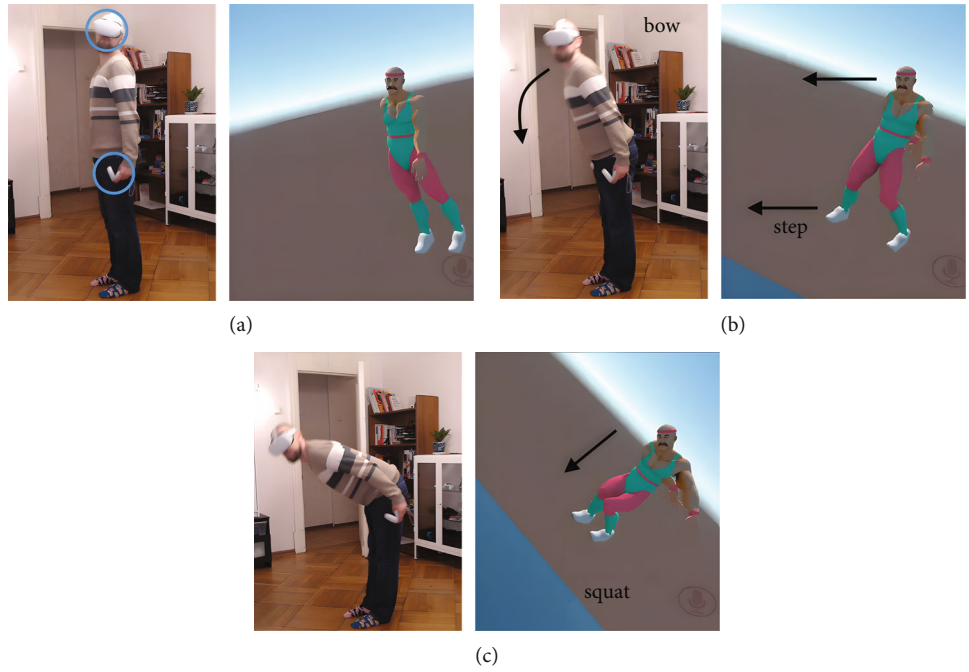

Figure 3: Three-point IK. Physical recording (left) and the way it translates into VR (right).

interaction involving people with disabilities [52], is both an inspiration and a partial addressee of this work, since it is similarly concerned with how people overcome interactional challenges in unorthodox environments.

With these preliminary caveats out of the way, we may now move on to the introduction of our specific data.

2.3. Data Collection. This paper will present a case study of VR interaction in the cooperative multiplayer game *Arizona Sunshine*. At the time of writing, approximately 8 hours of in-game video recordings, across 16 sessions and different participants, have been collected within that space. Volunteers were asked to play the game in our VR laboratory, with the in-game footage being recorded throughout. At the time of the recording, our lab was equipped with an enthusiast-level HTC Vive Pro Wireless VR system, which is what was used to generate this particular set of recordings.

Arizona Sunshine was chosen because it was, at the time of the data collection, one of the more ubiquitous VR experiences available to regular consumers. While there are systems that allow the user to use hand-tracking, eye-tracking, or even realistic haptics, they would also require a substantial investment of time and money. Arizona Sunshine, conversely, remains a favorite experience in mall-based or arcade VR settings [53], with a correspondingly low barrier to entry. As such, it is an example of the current challenges, and peculiarities, of "consumer-level VR."

Due to the technical specificities of VR, the video footage was recorded from the first-person view of the person wearing the HMD (head-mounted display). This style of recording has the limitation that only the things visible to the HMD wearer are available for analysis. Moreover, since there is an inherent asymmetry of perspectives, a "nod" would be visible through the down-up movement of the video footage from the first-person view while being available as a more straightforward head movement for another interactant.

The following section will investigate the interactional resources of Arizona Sunshine.

3. Analysis

Before proceeding with an analysis of any given interaction, it is not very common to describe the physical properties of the space the interactants inhabit. This may change when multimodal conversation analysts enter the age of interstellar travel; right now, we tend to operate on a number of commonsense assumptions about gravity, Euclidean space-time relations, and the kind of bodies humans typically inhabit, along with the material properties of these bodies and the environment. Interaction in VR is less dogmatic about physical constraints, making it necessary to introduce some basic facts about the kind of properties that obtain for any given virtual space.

3.1. Interaction in Arizona Sunshine. In "Arizona Sunshine," players navigate a postapocalyptic zombie-infested landscape in search of supplies and survivors. Much of the game involves searching cupboards, car trunks, and abandoned buildings for various in-game items while fending off enemy zombies.

As can be seen in the below screenshot (Figure 4), both characters (usually) have guns equipped in both hands. The large objects on both sides of the screenshot are the guns held by the person whose view is being recorded. Things can still be grabbed with the guns in hand, but the guns themselves tend to be held at all times. Guns can be pointed in any direction; as they can be moved through space, they

Figure 4: First-person perspective. Arizona Sunshine. Desaturated to enhance detail.

make it principally possible to produce mutually visible movement trajectories through the air. Even though the eyes are not tracked in this game, the position of the head is tracked by the VR headset. If a player turns their head physically, the avatar's head moves accordingly. Since, in current-gen HMDs, only the central spot on each of the screens is appreciably in focus, and the direction of the head can be used as a way to approximate gaze direction.

Lastly, although the game allows players to communicate through speech, both players may choose to mute their microphone. In our data, players frequently chose to keep the voice communication channel turned off. In the absence of audible speech, much of the interaction between players occurred through a combination of head and hand movement, with the specific restriction that both hands were usually occupied by guns; these guns did not, however, prevent players from interacting with in-game objects. Compared to the physical world, Arizona Sunshine does not track fingers—the controllers are effectively monodirectional sticks and are tracked as such. This limitation has the consequence of reducing the gestural repertoire to a linear pointing: the guns can be oriented in any direction in 3D space and be moved through it without restrictions. Thus, it is impossible to do a "victory sign," a "thumbs up," and a "middle finger gesture" through the usual combination of individual finger arrangements. This does not, however, mean that gestures could not be developed. After all, the temporal-historical movement of sticks through space, coordinated and witnessable in situ, allows for a rather complex sequence of constellations.

In some of our data, for instance, the "arm pump" movement—a mutually coordinated twice-repeated lowering of the controllers—was locally produced as a form of affirmation or agreement (see Figure 5). Moreover, players sometimes used guns as means of attracting the attention of another player; this was particularly important in cases of muted microphones. When a player shoots another player, the shot player's screen visibly indicates this fact, in addition to the sound of a gunshot. A virtual gun can thus be used as a means of long-distance communication or as a means of prompting attention from a visually disengaged player. In other words, the seemingly restricted and predetermined VR space can become the scaffolding for complex local at-hand conventions that frequently disregard the original intent of the game mechanics.

As a side activity, the game allows players to put on masks and hats scattered throughout the game world. These masks have no in-game significance beyond the cosmetic change they impart on otherwise identical avatars. In order to put on the mask, a player has to grab it and drag it onto their own face (mimicking the way a real-life mask would be put on). If the mask is close enough to the face when it is released, it automatically equips. In order to remove a mask, the same procedure is repeated in reverse. Notably, the mask becomes invisible for the person wearing it—it is only visible for the coparticipant; furthermore, only one mask can be worn at a time—and the normal procedure for putting on a mask will *fail* if the person in question is already wearing a mask; the mask will fall to the ground without equipping. The game has a tutorial section where the process of mask equipping is explained.

3.2. Removing a Mask. While I will do my best to render the following sequence through description and transcription, I highly encourage the reader to watch the original recording that this analysis is built upon (the clip is also attached in the supplementary materials (available here)): https://youtu.be/WTMaJ2AAj-s

In the analyzed fragment, one player, Fred (for *first-person perspective*), is putting on a mask while another player, Terry (for *third-person perspective*), observes from the side; both player's microphones are muted, so any interaction is nonverbal (Figure 6). After spotting and moving towards a mask on the floor (Figure 6(a)) and grabbing it (Figure 6(b)), Fred fails to equip the mask (Figures 6(c) and 6(d)).

Fred repeats the sequence (Figures 7(a)–7(c)). As the equip fails again, Fred is shot in the head by Terry (Figure 7(d)). This produces a white flash and noise from the direction where Terry stands (Figure 7).

Fred, after being shot (Figure 8(a)), proceeds to drop the mask (Figure 8(b)) and turns (Figure 8(c)) to face Terry (Figure 8(d)). As it turns out later in the sequence, Fred was already wearing a mask, which was the reason why the standard mask equip sequence failed. Terry, observing from the side, would be the only participant who could visually ascertain that Fred was *already wearing a mask*, and that *in order to put on a mask*, *Fred needs to remove the mask she is already wearing* (Figure 8).

In the concluding sequence (Figure 5), Fred and Terry successfully achieve the mutual understanding of what was going on. Fred removes the old mask (Figure 5(a)), successfully equips the new mask (Figures 5(b) and 5(c)) and exchanges a reciprocal arm pump gesture sequence with Terry (Figure 5(d)).

This paper is dedicated to the analysis of the cooperative work that occurs between the visible technical trouble and the resolution of the interactional ambiguities that follow.

3.3. The Trouble with Symmetry. After gaining the attention of Fred, Terry launches into a sequence of operations with the mask that she is wearing. As a gloss, we may preliminarily say that this sequence relates to the trouble of the mask

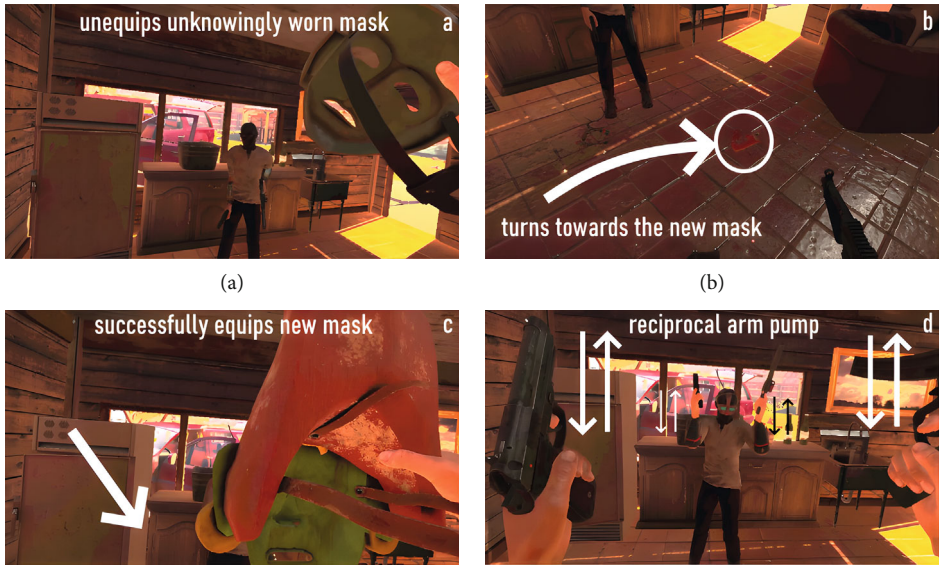

Figure 5: Successful resolution of the search sequence.

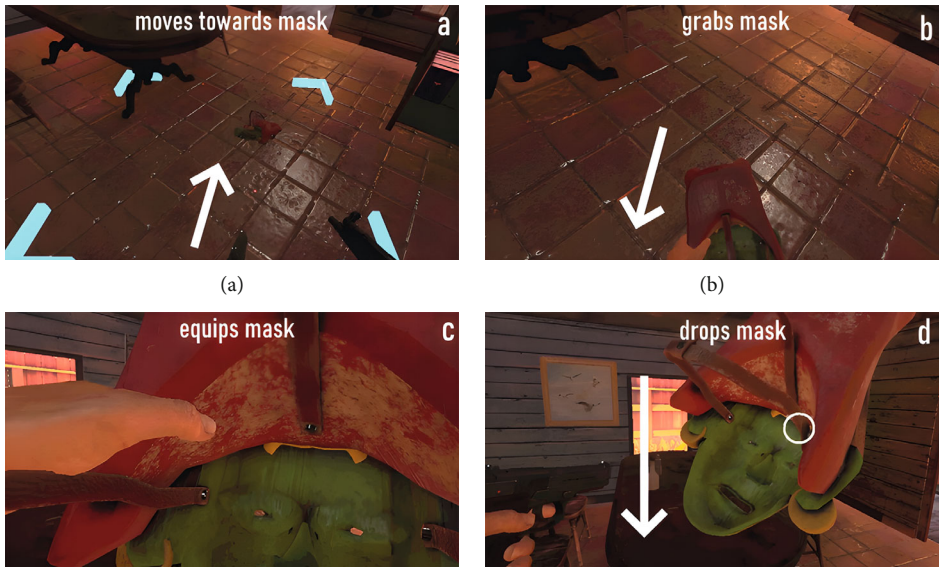

Figure 6: Unsuccessful attempt to equip a mask.

equipping demonstrated by Fred and Terry responds to this in an instructional manner.

We can formalize the above sequence in a purely descriptive way (Figure 9), where in 1a, Terry's right hand moves towards the mask, grabs it in 1b, moves it away in 1c, and moves it to the head in 1d, equipping it.

The sequence illustrates a fundamental interactional trouble that emerges around that specific gestural sequence: the resolution of the *technical* problem, namely, that a mask cannot be equipped on top of an unknowingly worn mask, turns into an *interactional* problem through the resources that are employed to convey the quite complex matter of "in order to put on a mask, you must first remove the mask that you are already wearing."

As it is performed, the gestural sequence is inherently symmetrical. That is, it can be taken as "this is how you put on a mask" and "this is how you take off a mask." At first, Fred proceeds with an attempt to put on a mask, demonstrating a preference for the first interpretive framework by treating Terry's preceding gestural sequence as being an instruction of how to put on a mask. This interpretation is further favored by the context in which the sequence is embedded: the originally visible technical problems relate to putting on a mask, which favors the treatment of the

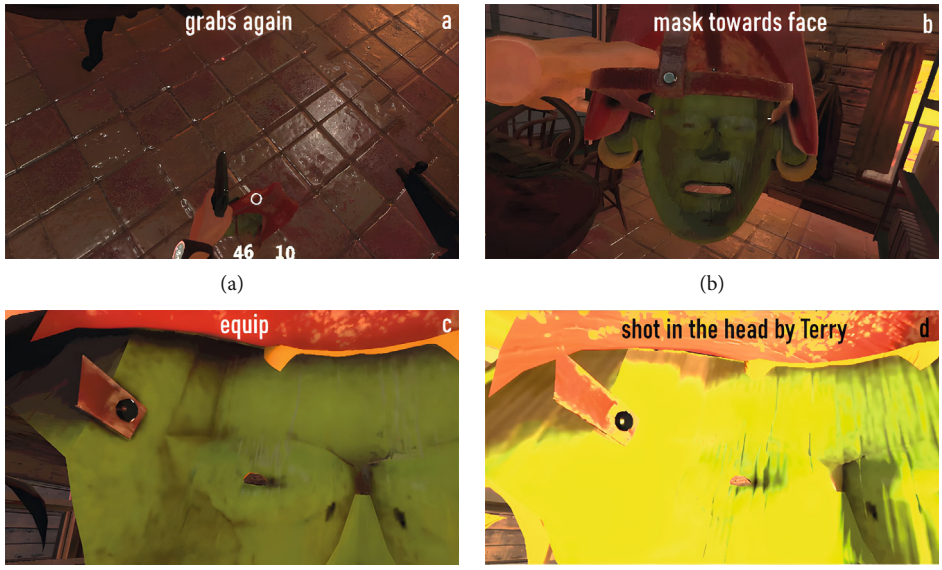

Figure 7: Sequence repeat. Attention-seeking firearm use.

Figure 8: Establishment of participation framework.

self-selected instructional activity on the part of Terry as an attempt to demonstrate the correct way to equip a mask. That is, Terry is faced with the problem of having to fight an uphill battle against a favored—yet problematic—interpretive framework.

If we, following Mondada [40], section an action's trajectory into a preparatory phase, an apex, and retraction, we get this treatment of the action sequence from Fred's perspective (Figure 10). 1a to 1c is treated as the preparation of the action of putting on a mask. This makes sense if we consider prop use: the fact that Terry is already wearing a mask does not have to index the mask removal as relevant — it may easily be treated as Terry using her own mask to demonstrate the proper technique of mask equipping, the "worn" mask being as relevant as a mask lying on the floor.

In contrast, an alternative treatment of the action sequence would take 1a and 1b as the preparation, 1c as the apex—"this is how you remove a mask"—and 1d as the retraction that would allow a repetition of the sequence (Figure 11). The preceding sequence repeats a number of times, without evidence that Fred treated the sequence as anything but "this is how you put on a mask." The last action sequence prior to the successful resolution displays several peculiar features which help us understand the kinds of resources invoked to disambiguate the preceding sequence.

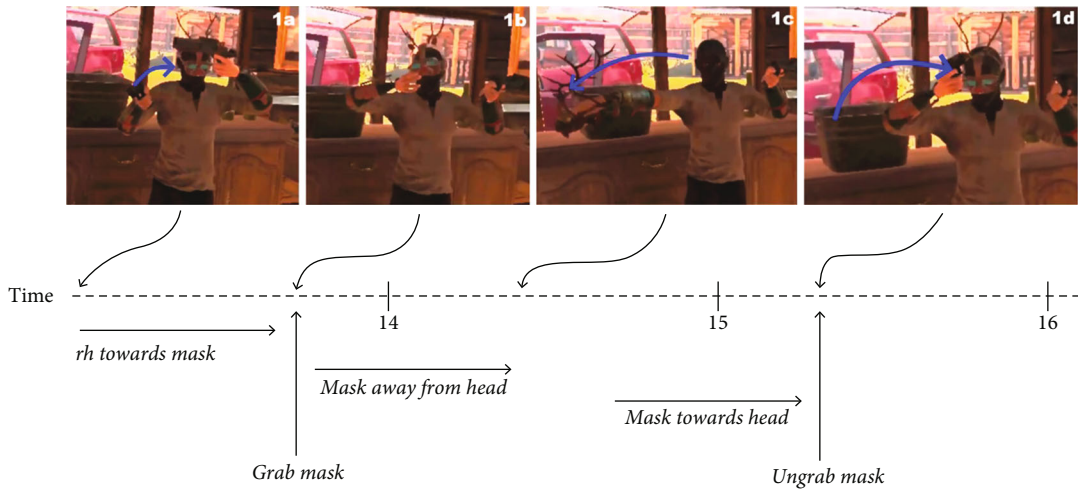

Figure 9: Removing and reequipping the mask. Neutral rendering.

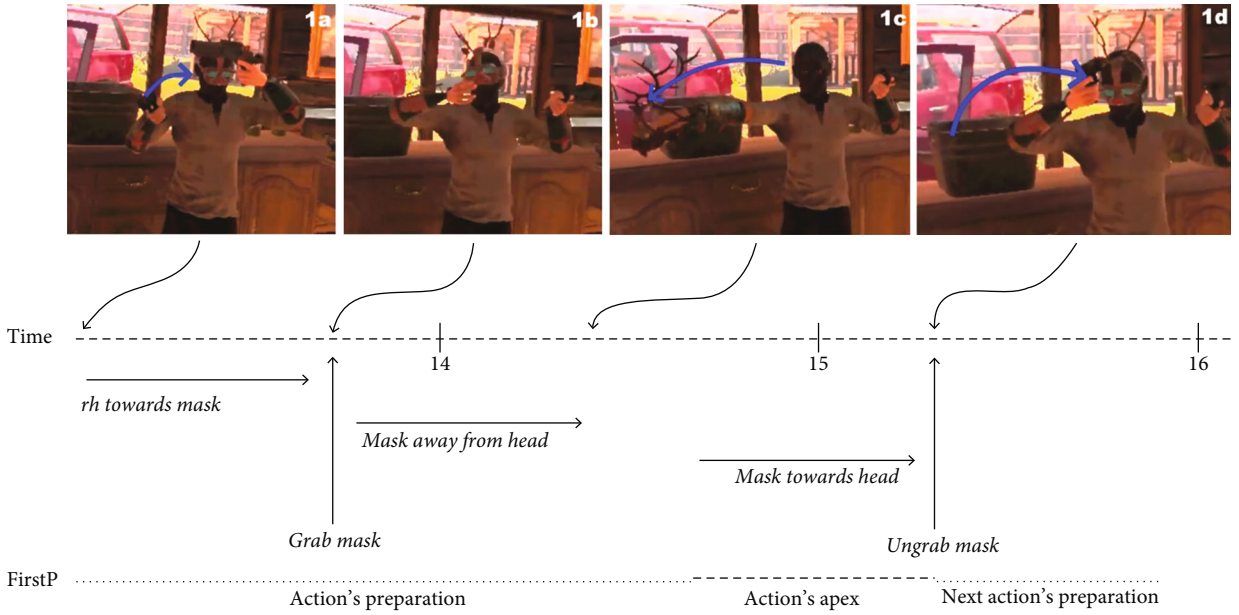

Figure 10: Removing and reequipping the mask. "This is how you put on the mask". Time in seconds.

Before, however, proceeding with an analysis of said interaction, we need to introduce a specialized framework from the "real world." For reasons that may become clearer later, the presently analyzed interaction bears a number of important similarities with a case analyzed in Goodwin [54] investigation of interaction involving a person with aphasia. The following section will recover a range of necessary distinctions that will enable us to relate Goodwin's case to ours.

3.4. The Multimodal Analysis of Search Sequences. Charles Goodwin, one of the fathers of contemporary multimodal analysis [41], did a substantial amount of research on the video-recorded interaction between Chil—his father and successful lawyer—and his family. After Chil "suffered a stroke in the left hemisphere of his brain"([55], p. 60), he was diagnosed with aphasia; in his case, it meant the near-total inability to produce speech beyond three words—*yes*, *no*, and *and*—as well as partial bodily paralysis. Much of Goodwin's work was dedicated to demonstrate that—in the right interactional environment—Chil could recover much of his interactional agency through a combination of rich prosody, gesture, tactically deployed speech, and the incorporation of other-produced speech in his actions for his own purposes. Goodwin would go on to formulate this type of incorporation as a key component of co-operative action as the reuse with modification of existing at-hand material, be it speech, abstract semiotic resources, or the physical material environment [12].

One particularly relevant piece of Goodwin's research is his work on the structural characteristics of search sequences [54, 56]. Since Chil could not produce complex utterances,

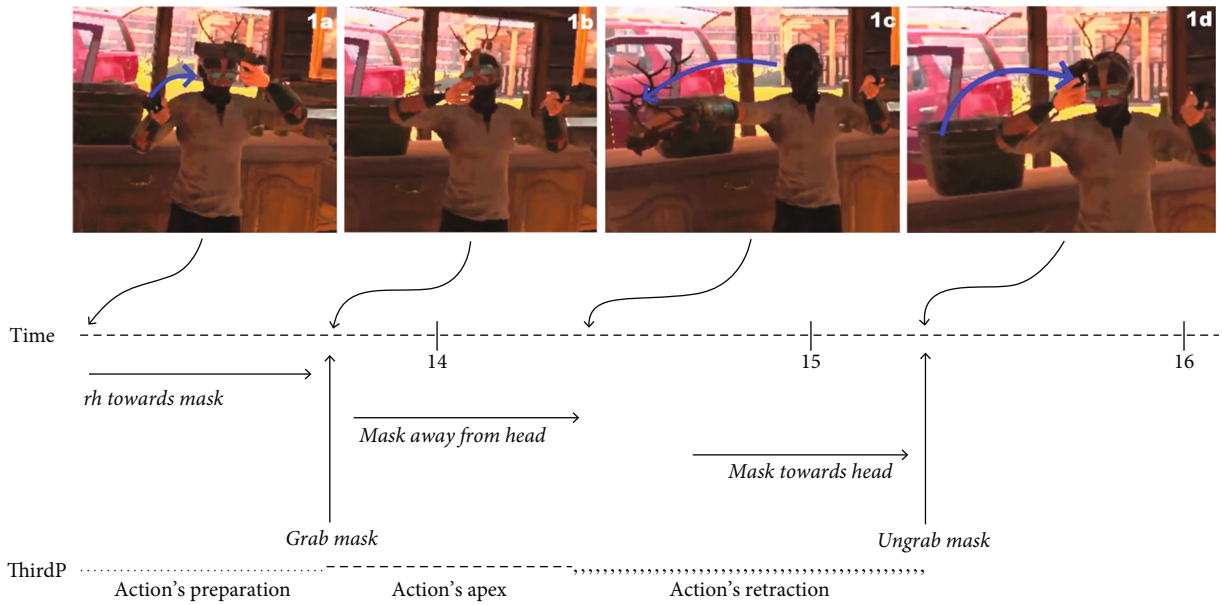

Figure 11: Removing and reequipping the mask. "This is how you remove the mask". Time in seconds.

he was forced to work with the immediate context of his environment. For instance, he could point out objects, actively engage in the speech-production process by being an "active listener," and display a visible orientation to ongoing action, thereby modifying its course. However, Goodwin reports that "[frequently] the process [of interacting with Chil] has a game-like quality […] as a consequence of this, and the restrictions on what [Chil] is able to say, there is a strong division of labor; the activity generates a set of structurally different kinds of participants who perform different kinds of action: [Chil] accepts or rejects proposals about what he might be trying to say, while his interlocutors provide relevant guesses." ([54], p. 7).

In a seminal paper, Goodwin [54] analyzes the substantial amount of interactional work necessary for Chil to formulate a request for *English muffins* after being asked whether he wanted *some toast* for breakfast: without speech, the process of asking Chil what he would like for breakfast involves a guessing "game"; another party proposes an object "toast," and Chil may either confirm or deny that this is what he wants. If he says "yes" to an item, the guessing game ends and he receives the item that preceded the "yes." If he says "no" to an item, the guessing game continues, with the other party producing guesses of another order. Thus, if Chil says "yes" to toast, he will receive toast. If Chil says "no" to toast, he will be offered other breakfast-related objects, most likely something notably different from the "toast" category (e.g., condiments and cereal). With enough "no"s produced, the guessing game may end up moving beyond the category of "Guessing what Chil wants for breakfast"; it is possible that it may move on to "Guessing whether Chil wants breakfast at all" and "Guessing the alternative activity Chil may want to engage in." So, a simple desire for English muffins—a toast-*like* thing—may conceivably result in no breakfast at all.

Systematically, then, Chil has a number of options. Firstly, he may engage the listing activity on the level of mutual knowledge about its rules. He may—as he indeed does—produce a "yes…no" or give the yes or no specific prosodic contours to indicate "not quite, but close." More fundamentally, Chil may choose to disengage from the activity and its specific local ordering; he may visibly turn from the person listing, instead facing another copresent interactant; in Goodwin [54], Chil *visibly reorients* towards his wife while disengaging from the two-person back-and-forth guessing activity. This conversational move allows Chil to pursue alternative courses of actions on his way to his currently preferred breakfast food.

Ultimately, Chil is successful in visibly producing a preference for an English muffin. In the absence of complex language, the process is organized as a prolonged sequence of candidate solutions on the part of the lister and confirmatory/nonconfirmatory takings by Chil. More generally, the sequential distribution of proposer/confirmer is organized by the mutually visible coupling between a proposal and a response to it, either by an orientation towards the proposed object or the activity in general. This response does not need to be verbally produced but needs to be accomplished as being relevant—*as a response*—within the participation framework.

A number of papers have extended Goodwin's original work on search sequences in non- or partially verbal interaction [57–66]. For our purposes, Laakso and Klippi [64] paper is particularly instructive, since it attempts to formalize what they call, following Lubinski et al. [67], "hint-and-guess" sequences into distinct stages with distinct interactional properties—"a problem establishment phase, a phase for establishing the collaborative co-participation framework, a 'hint and guess' phase, and a confirmation phase" ([64], p. 350). We will abbreviate the phases as PEP

(for problem establishment phase), CFP (for coparticipation framework phase), HGP (for "hint-and-guess" phase), and CP (for confirmation phase), respectively.

3.5. Reformalizing Search Sequences. In the PEP, a trouble is visibly produced. In the context of a word search, this could come in the form of a repeated, outwardly visible production of the same. This may be achieved, for example, through explicit remarks ("what is this thing called?") or through interruptions in the flow of speech and attempts of self-repair over several turns. As Laakso and Klippi [64] note, the PEP is not principally collaborative: it is possible to resolve the trouble without assistance from copresent listeners. In the CFP, the personal trouble is transformed into a matter where coparticipation is invited or self-selected. This can be achieved, for example, by the word searcher visibly reorienting themselves to face a copresent listener as a coparticipant in the activity. The HGP matches what was previously discussed in Goodwin's [54] analysis of the collaborative activity of Chil and the candidate-producing coparticipant. Guesses can be rejected with varying vehemence; hints can provide feedback on the candidate solution's relation to the category within which the sought object is situated. To take an example from charades, if I have "dog" in mind, then "telephone" could be actively distanced from "cat" as a candidate solution. This may be achieved by imparting various amount of prosodic emphasis to the rejection of either guesses.

Lastly, Laakso and Klippi separate the HGP from the CP, where the work of collaboratively establishing the successful completion of the HGP is performed. This may be done by producing agreement tokens (e.g., "exactly!" and "that's it!"), producing a visible recognition that the guesser's guess was correct, or through an overall shift in the cadence of the interaction, a successful turn completion and a move away from the search sequence activity.

3.6. Mapping the Phases to the Case of the Masks. The preceding separation of collaborative search sequence activities makes it possible to consider the case of the masks with greater granularity.

3.6.1. Establishing the Problem in VR (PEP). What, exactly, is the problem that is being collaboratively resolved in the mask case? If we look primarily at the visible display of the trouble, then the problem would be Fred's evident failure to equip the mask. However, it is *Terry* who produces multiple largely identical gestural sequences, and it is *Fred* who, through her attempts to equip the mask post-instruction, produces candidate solutions that are subsequently rejected or confirmed.

Thus, there is a somewhat modified distribution of "troubles" in this case. The visible production of a practical problem (not being able to equip a mask) proceeds to a self-selected instructional sequence, which turns into a nonverbal search sequence where Fred guesses and Terry hints.

An additional layer of complexity lies within the categorical misunderstanding related to Fred's treatment of Terry's hints, as they are taken as instructions to equip—not *remove*—a mask. This would be akin to a sequence where upon rejecting toast, the guesser would go on to produce guesses along the line of "fine, what kind of salad do you want, then?". Later sections will discuss the difficulty of escaping these kinds of categorical misconstruals.

3.6.2. Establishing the Collaborative Coparticipation Framework in VR (CFP). The coparticipation framework creation seems to be overdetermined, as it is produced *prior* to the PEP. Terry launches an instructional sequence by shooting Fred in the head, whereupon Fred shifts her gaze towards Terry and enters into a participation framework delineated by reciprocal body and gaze coorientation.

This case points towards a possible nestedness of participation frameworks, where a collaborative orientation may be inherited from a structurally different ordering arrangement (instructional sequence to search sequence), which, in turn, may lead to a tension in the distribution of roles and expectations.

3.6.3. The "Hint-and-Guess" Sequence in VR (HGP). The HGP has two peculiar characteristics that distinguish it from the previously considered cases.

Firstly, both parties are nonverbal. This means that, in addition to the regular difficulties of managing the unusual segmentation of interaction that is the characteristic of interacting with nonverbal persons, the mirrored lack of resources may contribute to difficulties on the process of indexical incorporation, which seems to depend on at least one speaker being present for the kind of "active listening" that Chil is capable of doing in order to be able to cooperatively act.

Secondly, though the discussed cases [54, 64] evidenced troubles with categories—most notably the tension between the proximity of "toast" and "toast-like things" juxtaposed to the categorically inproximate steps typically found in a guessing game—the case of the masks represents an instance where the very purpose of the hint-and-guess game turns out to be the exact opposite. This issue can be related to the contextual configuration of the place of the searching activity within the overall sequence of unfolding interaction: since the search sequence follows a trouble with the equipping of a mask, the way Fred takes this sequence to index prior turns—and thereby as being characterizeable as "demonstrating the correct procedure for equipping a mask"—is perfectly in line with the turn-by-turn sensemaking. A closer equivalent case would be a situation where Chil is presented with breakfast items, all the while trying to communicate that he does not want any breakfast at all.

3.6.4. Confirming the Resolution in VR (CP). The mutually visible resolution of the HGP is produced both through gestural means—in the form of a mutual confirmatory "arm pump" movement—and through the technical specificities of the search itself. In the case of a word search, the successful resolution of the search is entirely contingent upon the judgment of the person hinting. In the case of the masks, the successful resolution is contingent upon the reinterpretation of the nature of the search sequence—"*you are wearing*

Figure 12: Disambiguational work through sequential modifications.

a mask, remove it first, then equip the new one," which has two distinct moments where technological feedback provides indications of a resolution. Firstly, when Fred removes an unknowingly worn mask, the prior sequence of instruction may be seen in a new light as having had the purpose of producing this effect. Secondly, the successful equipping of the second mask may furnish the link between the original possible take of the instruction as "how to equip a mask," since the mask removal becomes its demonstrable condition.

In sum, the case of the masks bears structural similarities to hint and guess sequences: there is a distribution of roles, where one participant has the authority to successfully conclude the search (i.e., the searched element is within their epistemic domain, see [68]), and another may produce guesses which may be variably taken up by the person searching. Furthermore, the mask-searching activity suspends all other activities and becomes the primary ordering principle of the activity—another feature shared with many HGPs.

That said, the mask removal case is peculiar in a number of fundamental and incidental ways. The mutual nonverbality, for instance, *does* introduce additional difficulties, but it does not fundamentally transform the activity—the search goes on, and candidate solutions are produced and rejected/taken up. After all, the activity is not a *word*-search sequence, it is a search for a *gestural procedure*. Similarly, though the PEP is established as an outflux from a different visible trouble (failing to equip a mask), the interactants visibly reorient themselves towards the guessing of what Terry ultimately wants Fred to do.

More fundamentally, however, the HGP appears to be made considerably more difficult by the categorical nature of the reinterpretation necessary to take the instructional activity as what it is. Terry's inability to establish the fact that—in addition to the two masks visible to Fred—a third mask is present in the equation forces her to produce more than just hints—she is forced to modify the interactional ecology to disfavor a treatment of the building blocks of the hint production in a way that precludes a successful resolution. Terry must do this while also sustaining the HGP; repeating the same gestural hint could very well lead to the abandonment of the entire activity, since it could be taken as a hint that is not understood—rather than a hint that is interpreted based on incomplete information about the situation. The following section will analyze how Terry maintains the HGP as an activity while modifying the hints in a way that disfavors Fred's preferred interpretation.

3.7. Shifting the Favorability of an Interpretation. We will focus on the last sequence of hints and guesses prior to the successful resolution of the HGP (Figure 12).

After removing the mask in 2a, Terry proceeds to drop it in 2b. This drop occurs precisely at the moment when, according to the action sequence evidently favored by Fred (i.e., "this is how to put on the mask," Figure 10), the action would move from the preparatory phase to the apex. The drop also happens to occur precisely at the end of the apex of the competing action sequence ("this is how to remove a mask"). This move may serve to reconfigure the

interpretive framework, since the mask drop is arguably more compatible as a marker of a sequence's conclusion. In simpler terms: "Why would Terry drop the mask at the moment where the action sequence *begins*?"

Furthermore, Terry visibly shifts her gaze towards the floor and mask in 2c, reengages in 2d, and launches into a continuous pointing sequence that in 2e is followed by reciprocal nodding. This move may serve to underline the continuation of the overall hint and guess sequence and may have been designed to create a link between the dropping of the mask and a relevant course of action for Fred. As has been pointed out previously, the ever-present possibility of Fred rejecting the hint-and-guess activity is exacerbated by the possibility of treating a repeated hint sequence as an unsuccessful attempt to produce the right guess—rather than an attempt to achieve a fundamental categorical shift regarding the *kind* of search that is going on.

Lastly, the final repeat of the mask removal sequence (2f to 2 h) is notably different from all previous sequences: first, the sequence proceeds without the mask, possibly as a means to insinuate the presence of an already-worn mask. Second, the trajectory of the hand moving towards the head is shortened and proceeds with a diminished visible arc (2f), almost vertically upwards from the home position (see [12], p. 80). This may serve to reduce the symmetry of trajectories. The move towards the mask is deemphasized, while the movement away from the head is thereby highlighted.

Third, the sequence end is followed by a visual and gestural disengagement (2 h), precluding an immediate sequence repeat that may otherwise introduce further ambiguity.

This sequence of actions turns out to be successful. Fred throws away the old mask and successfully equips the new one. Through the use of purely nonverbal multimodal resources, Terry's work proceeded on a level of granularity that goes beyond the treatment of actions, instead achieving a disambiguation by indexically incorporating specific segments of the preceding flow of gestures.

4. Conclusion

This paper has attempted to make some first forays into the multimodal investigation of social interaction in fully immersive virtual reality. It is aimed at demonstrating that (a) VR is a radically strange interactional space characterized by unexpected difficulties in unexpected places, (b) nevertheless, a multimodal analysis of this interaction is possible, and (c) that the interactional analysis bears relevance on a range of methodological questions in contemporary multimodality studies in general.

More generally, the paper is aimed at contributing to the availability of emic perspectives on interaction within virtual reality. The remarkable complexity of the interaction within a comparatively sparse interactional ecology demonstrates how even the most clearly demarcated at-hand resources may find themselves being used in unexpected constellations for the accomplishment of unanticipated actions.

5. Discussion

A number of relevant methodological issues have cropped up during the process of analyzing the mask case. This section will highlight several prominent issues in turn.

5.1. Epistemics. Due to the greater intersubjective opacity of telemediated interactional spaces—simply put, we are less sure what the others see and what they see us seeing—we might speculate that VR is conducive to a greater number of epistemic imbalances between interactants. This is partially due to the novelty of the space but partially a fundamental property of the mode of embodied interaction typical of contemporary VR systems. On the first level, VR users find themselves with two bodies and an unclear relationship between them. On the second level, there is a general lack of certainty about the degree to which either interactant is present in any given space.

This circumstance makes the recent turn to the investigation of epistemic gradients, domains, and stances as organizational elements of interaction [68] potentially relevant to virtual reality: if we need to track the distribution of knowledge across interactants for the accomplishment of certain activities, then a space where mutual knowledge is less certain will require more epistemic-focused interactional work.

5.2. Sequentiality and Temporality. Multimodal CA, as the avant-garde of this kind of analysis, is still firmly rooted in the Sacksian tradition of the turn-by-turn analysis of temporally unfolding interaction. This focus is supported by the physical characteristics of produced speech in time. With purely nonverbal interaction, Goodwin's shift from sequentiality to synchronicity [12] and Mondada's analysis of multiple interactional temporalities [69] are reasserted in its importance.

Without words, there is no straightforward timeline onto which to map multimodal actions.

More practically, there does not seem to be a strong multimodal tradition of analyzing *completely* nonverbal interaction [14]. The ordering principle of most multimodal transcription methods (Goodwin's being a possible exception) is mapped to talk. This makes the process of transcribing and analyzing nonverbal interaction trickier and calls for the development of a greater range of transcription-analytic tools.

Moreover, the shift towards a greater granularity of action analysis brings with it the potential of shifting the distinction between formal for-research noticing and the *things* that are available to interactants as interactional resources (see especially [70]). Thus, while the granularity of preparation, apex, and retraction (to take the level of granularity used in the present paper) might be useful for analytical purposes, it might also be the location of interpretive work and action: the "apex" thus moves from a descriptive-analytical category to a contentious fulcrum in the local meaning-making process.

5.3. Contentious Privacy in Virtual Spaces. Lastly, I would like to highlight a lingering issue with privacy and anonymity. As a field that deals with videography, multimodal

conversation analysis always had to balance analytical precision with protecting the personal identity of the people's activities being investigated. One common way of anonymizing the published images was to use a sketch filter or to redraw the interaction entirely. In this light, virtual reality interaction could be seen as anonymous by default, especially when we consider spaces where avatars are identical, and no other identifying marks are present. However, I would argue for caution. Our fixation on the peculiarities of the body and face for identification is largely tied to the institutions that classify identity in the physical world: we have *bio*metric passports, have profile *pictures*, and even use DNA to identify suspects. In short, we use the physical body to trace identity.

However, body-centric identity is not the only way to identify persons. For example, one could identify and classify a person's activity by correlating the traces they leave on various websites. One could look at the ways a person writes or identify them by their gait. In short, identity, and therefore anonymization, is a matter of understanding the ways these are related to a person's activity and embodiment. In virtual reality, we may well be entering grounds that make the mere depiction of motion a source of identity. We should therefore not be presumptuous about the way identification may be produced in these spaces.

Data Availability

The videographic data used to support the findings of this study is included within the article.

Conflicts of Interest

The author declares that he has no conflicts of interest.

Acknowledgments

Research for this article was funded by the Collaborative Research Centre "Constructing Explainability" (DFG TRR 318/1 2021–438445824) at Paderborn University and Bielefeld University.

Supplementary Materials

"Arizona Sunshine Mask Removal" video fragment. This is a slightly brightened video fragment upon which this paper's analysis is based. The recording was created through the screen capture of an HTC Vive virtual reality headset. It approximates the first-person perspective of one of the participants of the studied interaction. *(Supplementary Materials)*

References

[1] I. Arminen, C. Licoppe, and A. Spagnolli, "Respecifying mediated interaction," *Research on Language & Social Interaction*, vol. 49, no. 4, pp. 290–309, 2016.

[2] J. Mlynář, E. González-Martínez, and D. Lalanne, "Situated organization of video-mediated interaction: a review of ethnomethodological and conversation analytic studies," *Interacting with Computers*, vol. 30, no. 2, pp. 73–84, 2018.

[3] S. Zhao, "Toward a taxonomy of copresence," *Presence: Teleoperators and Virtual Environments*, vol. 12, no. 5, pp. 445–455, 2003.

[4] Y. Gan, C. Greiffenhagen, and C. Licoppe, "Orchestrated openings in video calls: getting young left-behind children to greet their migrant parents," *Journal of Pragmatics*, vol. 170, pp. 364–380, 2020.

[5] C. Licoppe and J. Morel, "Video-in-interaction: "talking heads" and the multimodal organization of mobile and Skype video calls," *Research on Language & Social Interaction*, vol. 45, no. 4, pp. 399–429, 2012.

[6] M. Erofeeva and N. O. Klowait, "The impact of virtual reality, augmented reality, and interactive whiteboards on the attention management in secondary school STEM teaching," in *7th international conference of the immersive learning research network (ILRN)*, Eureka, CA, USA, 2021.

[7] P. Luff, C. Heath, H. Kuzuoka, J. Hindmarsh, K. Yamazaki, and S. Oyama, "Fractured ecologies: creating environments for collaboration," *Human–Computer Interaction*, vol. 18, no. 1-2, pp. 51–84, 2003.

[8] I. Hutchby, *Conversation and Technology: From the Telephone to the Internet/Ian Hutchby*, Polity, 2001.

[9] L. M. Seuren, J. Wherton, T. Greenhalgh, and S. E. Shaw, "Whose turn is it anyway? Latency and the organization of turn-taking in video- mediated interaction," *Journal of Pragmatics*, vol. 172, pp. 63–78, 2021.

[10] N. O. Klowait, "Interactionism in the age of ubiquitous telecommunication," *Information, Communication and Society*, vol. 22, no. 5, pp. 605–621, 2019.

[11] A. I. Egorova and N. Klowait, "How to say good-bye to a robot? The matter of conversational closing," *Monitoring of Public Opinion: Economic and Social Changes*, vol. 161, no. 1, pp. 241–270, 2021.

[12] C. Goodwin, "Why multimodality? Why co-operative action? (transcribed by J. Philipsen)," *Social Interaction Video-Based Studies of Human Sociality*, vol. 1, no. 2, 2018.

[13] L. Mondada, "Contemporary issues in conversation analysis: embodiment and materiality, multimodality and multisensoriality in social interaction," *Journal of Pragmatics*, vol. 145, pp. 47–62, 2019.

[14] L. Mondada, "Transcribing silent actions: a multimodal approach of sequence organization," *Video-Based Studies of Human Sociality*, vol. 2, no. 1, 2018.

[15] S. Goico, Y. Gan, J. Katila, and M. H. Goodwin, "Capturing multisensoriality," *Social Interaction. Video-Based Studies of Human Sociality*, vol. 4, no. 3, 2021.

[16] C. Jewitt, D. Chubinidze, S. Price, N. Yiannoutsou, and N. Barker, "Making sense of digitally remediated touch in virtual reality experiences," *Discourse, Context & Media*, vol. 41, article 100483, 2021.

[17] L. Caronia and F. Cooren, "Decentering our analytical position: the dialogicity of things," *Discourse & Communication*, vol. 8, no. 1, pp. 41–61, 2014.

[18] B. L. Due, "RoboDoc: semiotic resources for achievingface-to-screenface formationwith a telepresence robot," *Semiotica*, vol. 2021, no. 238, pp. 253–278, 2021.

[19] M. Erofeeva, "On multiple agencies: when do things matter?," *Information, Communication & Society*, vol. 22, no. 5, pp. 590–604, 2019.

[20] N. Klowait and M. A. Erofeeva, "The rise of interactional multimodality in human-computer interaction," *Monitoring of Public Opinion: Economic and Social Changes*, vol. 161, no. 1, pp. 46–70, 2021.

[21] J. L. Maples-Keller, B. E. Bunnell, S.-J. Kim, and B. O. Rothbaum, "The use of virtual reality technology in the treatment of anxiety and other psychiatric disorders," *Harvard Review of Psychiatry*, vol. 25, no. 3, pp. 103–113, 2017.

[22] P. Araiza-Alba, T. Keane, and J. Kaufman, "Are we ready for virtual reality in K–12 classrooms?," *Technology, Pedagogy and Education*, vol. 31, no. 4, pp. 471–491, 2022.

[23] G. Colombini, M. Duradoni, F. Carpi, L. Vagnoli, and A. Guazzini, "Leap motion technology and psychology: a mini-review on hand movements sensing for neurodevelopmental and neurocognitive disorders," *International Journal of Environmental Research and Public Health*, vol. 18, no. 8, p. 4006, 2021.

[24] M. El Beheiry, S. Doutreligne, C. Caporal, C. Ostertag, M. Dahan, and J.-B. Masson, "Virtual reality: beyond visualization," *Journal of Molecular Biology*, vol. 431, no. 7, pp. 1315–1321, 2019.

[25] P. Haddington and T. Oittinen, "11. Interactional spaces in stationary, mobile, video-mediated and virtual encounters," in *Pragmatics of Space*, A. H. Jucker and H. Hausendorf, Eds., pp. 317–362, De Gruyter, 2022.

[26] P. McIlvenny, "The future of 'video' in video-based qualitative research is not 'dumb' flat pixels! Exploring volumetric performance capture and immersive performative replay," *Qualitative Research*, vol. 20, no. 6, pp. 800–818, 2020.

[27] M. Zhang, H. Ding, M. Naumceska, and Y. Zhang, "Virtual reality technology as an educational and intervention tool for children with autism spectrum disorder: current perspectives and future directions," *Behavioral Sciences*, vol. 12, no. 5, p. 138, 2022.

[28] A. Carnevale, I. Mannocchi, M. S. Sassi et al., "Virtual Reality for Shoulder Rehabilitation: Accuracy Evaluation of Oculus Quest 2," *Sensors*, vol. 22, no. 15, 2022.

[29] V. Holzwarth, J. Gisler, C. Hirt, and A. Kunz, "Comparing the accuracy and precision of SteamVR tracking 2.0 and oculus quest 2 in a room scale setup," in *ICVARS 2021: 2021 the 5th International Conference on Virtual and Augmented Reality Simulations*, pp. 42–46, Melbourne, VIC, Australia, 2021.

[30] D. Pimentel, M. Foxman, D. Z. Davis, and D. M. Markowitz, "Virtually real, but not quite there: social and economic barriers to meeting virtual reality's true potential for mental health," *Virtual Reality*, vol. 2, 2021.

[31] C. Chen, K. Liu, H. Yang, and M. Wu, "The development characteristics of virtual reality after "the year of VR"," in *2020 International Conference on Innovation Design and Digital Technology (ICIDDT)*, pp. 152–155, Zhenjing, China, 2020.

[32] J. Lin and M. E. Latoschik, "Digital body, identity and privacy in social virtual reality: a systematic review," *Virtual Reality*, vol. 3, 2022.

[33] M. Wang, "Social VR: a new form of social communication in the future or a beautiful illusion?," *Journal of Physics: Conference Series*, vol. 1518, no. 1, p. 012032, 2020.

[34] B. Egliston and M. Carter, "Oculus imaginaries: the promises and perils of Facebook's virtual reality," *New Media & Society*, vol. 24, no. 1, pp. 70–89, 2022.

[35] M. A. Erofeeva and N. Klowait, "Dei ex machina: the interaction order of gamified distance learning," *Sociology of Power*, vol. 32, no. 3, pp. 189–220, 2020.

[36] U. M. Kimstach, N. Klowait, and M. A. Erofeeva, "Talking without a voice: virtual co-speakership in an educational webinar," *Sociology of Power*, vol. 33, no. 4, pp. 198–216, 2021.

[37] F. Arici, R. M. Yilmaz, and M. Yilmaz, "Affordances of augmented reality technology for science education: views of secondary school students and science teachers," *Human Behavior and Emerging Technologies*, vol. 3, no. 5, pp. 1153–1171, 2021.

[38] R. T. Azuma, "The road to ubiquitous consumer augmented reality systems," *Human Behavior and Emerging Technologies*, vol. 1, no. 1, pp. 26–32, 2019.

[39] A. Skulmowski and G. D. Rey, "Realism as a retrieval cue: evidence for concreteness-specific effects of realistic, schematic, and verbal components of visualizations on learning and testing," *Human Behavior and Emerging Technologies*, vol. 3, no. 2, pp. 283–295, 2021.

[40] L. Mondada, "Challenges of multimodality: language and the body in social interaction," *Journal of SocioLinguistics*, vol. 20, no. 3, pp. 336–366, 2016.

[41] C. Goodwin, "Action and embodiment within situated human interaction," *Journal of Pragmatics*, vol. 32, no. 10, pp. 1489–1522, 2000.

[42] C. Heath and P. Luff, "Embodied action and organisational interaction: establishing contract on the strike of a hammer," *Journal of Pragmatics*, vol. 46, no. 1, pp. 24–38, 2013.

[43] L. Mondada, "Multiple temporalities of language and body in interaction: challenges for transcribing multimodality," *Research on Language & Social Interaction*, vol. 51, no. 1, pp. 85–106, 2018.

[44] J. Streeck, "Interaction and the living body," *Journal of Pragmatics*, vol. 46, no. 1, pp. 69–90, 2013.

[45] J. Streeck, C. Goodwin, and C. D. LeBaron, "Embodied Interaction: Language and Body in the Material World," in *Learning in Doing: Social, Cognitive and Computational Perspectives*, J. Streeck, C. Goodwin, and C. LeBaron, Eds., Cambridge University Press, 2011.

[46] C. Heath, J. Hindmarsh, and P. Luff, *Video in Qualitative Research: Analysing Social Interaction in Everyday Life/Christian Heath, Jon Hindmarsh, Paul Luff*, SAGE, 2010.

[47] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974.

[48] J. M. Atkinson and J. Heritage, *Structures of Social Action: Studies in Conversation Analysis. Studies in emotion and social interaction. Editions de la Maison des sciences de l'homme*, Cambridge University Press, 1984.

[49] H. Sacks and G. Jefferson, *Lectures on Conversation*, Blackwell, 1992.

[50] H. Garfinkel, *Studies in Ethnomethodology*, Prentice-Hall, 1967.

[51] E. A. Schegloff and H. Sacks, "Opening up closings," *Semiotica*, vol. 8, no. 4, 1973.

[52] R. Wilkinson, J. P. Rae, and G. Rasmussen, *Atypical interaction*, Springer International Publishing, 2020.

[53] V. R. Springboard, "Arizona Sunshine Launches New VR Arcade Edition," 2018, https://medium.com/@springboardVR/arizona-sunshine-launches-new-vr-arcade-edition-c252e4edaf22.

[54] C. Goodwin, "Co-constructing meaning in conversations with an aphasic man," *Research on Language & Social Interaction*, vol. 28, no. 3, pp. 233–260, 1995.

[55] C. Goodwin, *Co-Operative Action. Learning in Doing*, Cambridge University Press, 2017.

[56] M. Harness Goodwin and C. Goodwin, "Gesture and coparticipation in the activity of searching for a word," *Semiotica*, vol. 62, no. 1-2, pp. 1-2, 1986.

[57] E. Adami and R. Swanwick, "Signs of understanding and turns-as-actions: a multimodal analysis of deaf–hearing interaction," *Visual Communication*, 2019.

[58] A. Blackledge and A. Creese, "Translanguaging and the body," *International Journal of Multilingualism*, vol. 14, no. 3, pp. 250–268, 2017.

[59] E. Davitti, "Methodological explorations of interpreter-mediated interaction: novel insights from multimodal analysis," *Qualitative Research*, vol. 19, no. 1, pp. 7–29, 2019.

[60] L. Doak, "'But I'd rather have raisins!': Exploring a hybridized approach to multimodal interaction in the case of a minimally verbal child with autism," *Qualitative Research*, vol. 19, no. 1, pp. 30–54, 2019.

[61] I. Hörmeyer and G. Renner, "Confirming and denying in co-construction processes: a case study of an adult with cerebral palsy and two familiar partners," *Augmentative and Alternative Communication*, vol. 29, no. 3, pp. 259–271, 2013.

[62] I. Koshik and M.-S. Seo, "Word (and other) search sequences initiated by language learners," *Text & Talk*, vol. 32, no. 2, 2012.

[63] A. Kusters, "Gesture-based customer interactions: deaf and hearing Mumbaikars' multimodal and metrolingual practices," *International Journal of Multilingualism*, vol. 14, no. 3, pp. 283–302, 2017.

[64] M. Laakso and A. N. Klippi, "A closer look at the 'hint and guess' sequences in aphasic conversation," *Aphasiology*, vol. 13, no. 4-5, pp. 345–363, 1999.

[65] G. Renner, I. Hörmeyer, and L. Hoffer, "Ko-Konstruktion erkennen und verstehen – eine Analyse verschiedener Ko-Konstruktionstechniken in der Unterstützten Kommunikation," *Sprache · Stimme · Gehör*, vol. 43, no. 2, pp. e1–e7, 2019.

[66] S. Tetzchnervon and C. Basil, "Terminology and notation in written representations of conversations with augmentative and alternative communication," *Augmentative and Alternative Communication*, vol. 27, no. 3, pp. 141–149, 2011.

[67] R. Lubinski, J. Duchan, and B. Weitzner-Lin, "Analysis of breakdowns and repairs in aphasic adult communication," in *Clinical Aphasiology Conference Proceeding*, R. Brookshire, Ed., pp. 111–116, BRK, 1980.

[68] J. Heritage, "Epistemics in action: action formation and territories of knowledge," *Research on Language & Social Interaction*, vol. 45, no. 1, pp. 1–29, 2012.

[69] L. Mondada, "The temporal orders of multiactivity," in *Multiactivity in Social Interaction*, P. Haddington, T. Keisanen, L. Mondada, and M. Nevile, Eds., pp. 33–76, John Benjamins Publishing Company, 2014.

[70] R. Watson, "Comparative sociology, laic and analytic: some critical remarks on comparison in conversation analysis," *Cahiers de Praxématique*, vol. 50, no. 50, pp. 203–244, 2008.