Collaborating with Computer Vision Systems: An Exploration of Audio Feedback

Cecily Morrison, Neil Smyth, Robert Corish, Kenton O'Hara, Abigail Sellen
Microsoft Research Cambridge
21 Station Road, Cambridge, UK, CB1
(cecilym, t-nsmyth, rcorish, kohara, asellen) @microsoft.com

ABSTRACT

Computer visions (CV) systems are increasingly finding new roles in domains such as healthcare. These collaborative settings are a new challenge for CV systems, requiring the design of appropriate interaction paradigms. The provision of feedback, particularly of what the CV system can "see," is a key aspect, and may not always be possible to present visually. We explore the design space for audio feedback for a scenario of interest, the clinical assessment of Multiple Sclerosis using a CV system. We then present a mixed-methods experimental study aimed at providing some first insights into the challenges and opportunities of designing audio feedback of this kind. Specifically, we compare audio feedback that differentiates which body parts the CV system can see to audio feedback that is undifferentiated. The findings reveal that it is not enough to simply convey that something might be out of view of the camera as what the camera can "see" depends on the specific configuration of participants and the peculiarities of the skeleton inference algorithms. The results highlight the importance of providing feedback which more naturally conveys spatial information in developing CV systems for collaborative use.

Author Keywords
Computer vision technologies; Kinect; audio feedback; co-present interaction.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI):

INTRODUCTION

Computer vision (CV) systems, such as the Microsoft Kinect depth camera and associated algorithms, are increasingly enabling new ways for us to use bodily movement and gestures to engage with the digital world. While many of the first applications were games [9,28], CV systems are being used in a range of other settings, including surgery [11], rehabilitation [6], the kitchen [29] and magic shows [23]. In many of these settings, the person using the CV system is simultaneously interacting with other, co-present people. For example, a surgeon may gesture at a CV system to navigate through medical images while instructing a colleague in how to manipulate an instrument inside a patient [11].

Such collaborative settings are a new challenge for CV systems and call for appropriate interaction paradigms to be designed with these uses in mind. An important aspect of this is the provision of feedback in the interface so that multiple users can better understand how and why the system is responding in the way it does. In other words, users need to understand various aspects of what the CV system "sees". These may include: the boundaries of the zone of interaction, possible actions the system can understand, and the system's interpretation of actions. Feedback may also be used to communicate inferred positional information, indicating for example if people are too close to each other for the system to work properly. Indeed research is starting to show that providing effective feedback is vital to support collaborative contexts [11].

Perhaps the most obvious approach is to provide feedback through visual information on a display. However, screens may not be available, or may be obstructed in collaborative situations. For this reason, audio feedback is an interesting alternative. Indeed, the use of non-speech audio or "sonification" (the use of sound to communicate information) has its own affordances. For example, it has been highlighted as particularly appropriate for presenting information with minimal attention requirements [14,15]. It is therefore potentially suitable as an addition to on-going human-human interaction. Studies have shown that the audio channel is effective in supporting navigation in variety of settings, such as location-based games [18] and audio tourist guides [24]. Yet to our knowledge there is no HCI research that has explored the use of audio feedback to facilitate collaborative interaction with CV systems.

To that end, we present a mixed-methods experimental study aimed at providing some first insights into the challenges and opportunities for designing audio feedback of this kind. Specifically, we ask the question:

How can we design audio feedback to support co-present collaborators who are simultaneously interacting with CV systems?

- Release statement

This paper contributes to the HCI literature by extending research on interaction paradigms for sensing systems [1] to a situation of co-present interaction with a focus on audio feedback.

We begin by detailing the scenario that inspired this study: a project that involved using a CV system to carry out the clinical assessment of patients with Multiple Sclerosis. We then draw together a range of related work, including research that addresses: the theory of co-present interaction, co-presence and CV systems, and sonification. Before presenting a study that compares two types of audio feedback for the Kinect, we detail the design and evaluation of the audio feedback used. We conclude with a discussion of the challenges of designing audio feedback to communicate spatial concepts in CV systems.

USE OF CV TO DIAGNOSE MULTIPLE SCLEROSIS

This study draws inspiration from a project that is exploring the potential of CV systems to help clinicians more accurately monitor the progression of Multiple Sclerosis (MS), a neurological disease that affects 2 million people. MS can affect many systems in the body, but clinicians often focus on cerebellar dysfunction as manifested in changes in movement patterns. This project aims to use CV systems to detect these changes more rapidly, objectively and accurately than human assessors.

To do this, the project is using depth images captured by a Kinect. The depth images represent 3-dimensional point clouds of people and objects within the range of the camera's infrared sensor. The camera is often used in conjunction with the Kinect for Windows software development kit (SDK) which provides, among other things, real-time skeleton tracking for two people. In this project, the SDK is used only to support the provision of positional feedback and not for image analysis.

Currently, examinations are carried out by a neurologist who guides a patient through a range of simple exercises, including stretching out one arm to the side and then touching the nose or walking on a pretend tight rope. These exercises are then marked on the Expanded Disability Status Scale (EDSS), an ordinal rating-scale from 0 – 4 [19]. The clinician would usually stand opposite the patient and instruct him in these movements. If the patient's balance has been affected by the disease, such as paralysis of one side of the body, the clinician must stand next to him while he performs them to ensure the patient does not fall. The examination with the Kinect aims to capture the same movements, with the camera placed about 1.6 meters in front of where the patient is sitting or standing.

This scenario presents some interactional challenges. The Kinect must capture a complete image of the patient to maximise the likelihood of algorithmic success. However, changes in a patient's position can lead to body parts that are out of view during parts of the recording. Moreover, the clinician's normal position – standing in front of, or next to, the patient -- can block the camera's view of the patient. This means the clinician must negotiate many aspects of the examination, such as keeping the patient relaxed and safe while prioritising the Kinect's view of the patient.

Why Audio Feedback?

Our decision to focus on audio feedback resulted from observational work of an early prototype in the clinical setting. The prototype provided basic visual feedback through color-coded depth images displayed on an 11" laptop screen plugged into the Kinect. This feedback helped the initial positioning of the patient, ensuring the patient was the correct distance from Kinect and all limbs were visible. However, it did not address the visibility of the patient throughout the examination. We saw that the clinician moves around the space a great deal, finding the optimal position for supporting or guiding the patient in the confined space of the examination room. Very attentive to their interaction with the patient and their safety needs, the clinician is rarely well positioned to see the screen. Even when facing the screen, its small size and the impracticality of a larger screen in this clinical context makes visual feedback during the examination problematic.

Approach

Due to the sensitive nature of working with patients, it was inappropriate to do these preliminary explorations in the clinical setting. Consequently, we decided to explore the key issues in a laboratory setting. While we draw inspiration from the clinical assessment of MS with the Kinect, instead of simulating the clinical experience we chose instead to focus on the key features of interaction we were trying to support. Namely, we wanted to explore whether audio feedback could be effectively used to help convey information about the camera view of a CV system while people were otherwise engaged in doing a primary, physically mobile task with each other. In this case, we wanted a situation in which one person was clearly the instructor, engaged in trying to instruct someone else to carry out a set of standardized movements. We also wanted the physical task to be prioritized, with the secondary goal being to make sure that the person performing the movements remained as much as possible in view of the camera at all times.

RELATED WORK

Co-Presence and Spatial Configuration

The clinical assessment of MS is just one of a growing number of examples in which technology is used to support the main interaction between two or more people in the same physical space. We refer to this configuration of people and technology as co-presence. It differs from explorations of co-located interaction in which the technology is the central part of the human-human interaction [30] or discussions of proxemics when designing interactions between a human and a sensing system [22].

Figure 1: Collaborative settings in healthcare in which CV systems are used.

The literature on collaborative interaction and work practices is vast (see [20] for a seminal example). However, most useful in our analysis is that which considers how the embodied nature of human-human interaction affects the dynamic rearranging of spatial configuration in a setting. Some early theoretical work in this area is Goffman's discussion of frame analysis [10]. He argues that when two or more people are co-present, that is, within each other's perceptual fields, they participate in spatial and postural arrangements of their bodies that reflect their negotiated relationship. This theoretical lens can be useful to understand the impact technological systems can have on clinical practice [13].

More specifically, recent work in clinical settings shows how spatial configuration is shaped both by the demands of technology and other shared artefacts, and by the needs of human-human interaction. For example, Morrison et al [27] showed how a lack of design effort to support the desired spatial configurations in co-present interaction around an electronic patient record led to a decrease in multi-disciplinary interaction in an intensive care setting, a change correlated with increased patient fatality. In a study of imaging technology use in neurosurgery, Mentis et al [25] describe a conflict between the positions two surgeons take either side of a patient's body and their need to make reference to images outside this interactional space. As such this can disrupt work, or leave them attempting to create new interaction spaces through assuming awkward postures.

CV Systems and Co-Presence

Only recently have researchers begun to explore situations of co-presence and its implications for interaction design of sensing systems. Downs et al [7] describe the notion of "paraplay", the playful activities that take place between players and audience members while interacting with console games. Other literature has analyzed crowd interaction with public large screen displays that use edge detection computer vision algorithms to create simple games [28]. Both of these papers highlight the social context of CV system usage.

CV systems are now being used in a variety of social settings other than gaming. Figure 1 shows three collaborative interactions in healthcare settings: surgery, rehabilitation, and clinical assessment. In all three cases, there are multiple people in the line of sight of the Kinect.

O'Hara et al [11] present the first detailed, academic account of the issues that arise when using a CV system in a collaborative setting. They describe the deployment of a Kinect-based system to provide touchless manipulation of medical images in vascular surgery. A substantial part of the analysis in this account addresses spatial concerns that arise in a collaborative setting using Kinect. One vignette highlights how the positioning of the surgical team for the purposes of the operation influences the placement of the imaging screens and consequently the placement of the Kinect. As such, either the person standing in front of the camera is delegated the task of managing the images, or a more senior colleague needs to leave the operating table to correctly position himself in front of the sensor. In both cases, the positioning needed to carry out the collaborative task of surgery must be adjusted to incorporate the Kinect.

A second scenario details the efforts of a clinician to get the system to track only his body. As he was in close proximity to another surgeon while interacting with the Kinect, the system inferred the wrong skeleton, extending it across to the arm of the colleague next to him. The clinician used the visual feedback of the abstracted skeleton overlaid on the depth images of the people in the room to move into a position in which the system "saw him" correctly. This took multiple attempts, illustrating the importance of feedback in working collaboratively with the Kinect.

Sonification

In such situations it is interesting to consider how audio feedback might be usefully applied. Auditory displays use sound as the primary channel for communicating and transmitting information, capitalising on the unique strengths of the auditory system [14]. Examples of auditory displays range from the verbal instructions provided by car navigation systems to the support of new interaction experiences when mobile [31]. There is also much work in HCI on augmenting visual displays with audio feedback, using audio icons and "earcons" [3,4].

Sonification is one type of audio display which "seeks to translate relationships in data or information into sounds that exploit the auditory perceptual abilities of human beings such that the data relationships are comprehensible [32]." Sonification is thought to be particularly useful: (1) to recognise temporal patterns and changes; (2) when a visual interface cannot be seen; and (3) when background processing of information is necessary [32]. As such,

sonification is commonly used for monitoring tasks, such as financial data [16] or patient status in intensive care units [26]. Sonification has been most explicitly explored in HCI for navigation tasks [15,17], but not for monitoring tasks.

Similar to visual displays, it is important to iteratively test auditory displays [2]. While it is not possible to test experimentally all aspects of the auditory design space, it is important to choose sounds carefully, and to test the parameter mapping and information density conveyed by those sounds. The parameter mapping describes the relationship between the attributes of the sound and the information presented. For example, one can map higher temperature data to a higher pitch. The density of information conveyed corresponds to the number of parameters mapped into the audio display (e.g. temperature or temperature and level of rain fall).

STUDIES

In order to explore the design space for audio feedback in a way relevant to the original use scenario (clinical assessment of MS), we began by considering the key aspects of information to be conveyed. We wanted to indicate with sound any time that the Kinect did not have a full view of a person performing a movement protocol. For the purposes of this research, we defined a "full view" as the Kinect being able to track the head, torso, and both hands of the performer. This choice accounts for the unreliability of leg tracking, and intrinsic inclusion of the arms as part of the skeleton geometry of the hands.

The simplest feedback that we could provide would be a single sound that is played when a Kinect does not have a full view of the performer. Making the feedback more fine-grained, we could signal, or differentiate, which body parts cannot be seen. We refer to these two options as *undifferentiated* and *differentiated* feedback respectively. Differentiated audio feedback will, by necessity, contain more information and thus potentially require more mental effort to process. This is especially true in a context in which the audio feedback is secondary to the task. Differentiated feedback is only of value then, if it increases people's understanding of what the camera can see whilst not overloading them. Accordingly, we address two specific research questions in this study:

Can audio feedback effectively help people understand the field of view of a CV system as they work together collaboratively on a movement task?

In this context, does differentiated audio feedback enable people to more effectively determine what the camera can see than undifferentiated audio feedback?

There are many aspects of the audio feedback design that could be explored and tested in this research situation. We judged that the above two questions would be a good starting point to identify important issues in applying audio feedback to CV systems.

In what follows, we describe three stages of the research. We begin by exploring different ways that we can construct differentiated audio feedback. We then assess these for their intelligibility in a pilot study. Using these results, we describe the main study that directly addresses the above research questions.

Designing the Differentiated Audio Feedback

We began by considering what kinds of sounds would be best for the differentiated audio feedback -- the feedback that would convey which body parts were in the field of view of the Kinect. A sound designer (one of the authors) developed six differentiated audio displays drawing from research in this area [5,21]. Each display consisted of four sounds, one for each body part that needs to be seen to meet our definition of "full view" (see Table 1). Recordings and longer descriptions are included in the appendix.

The sound designer explored two ways of mixing sounds: ensembles and sequences. Ensembles are sounds intended to be played together, while sequences are played in a series. Ensemble sounds must be distinct enough for recognition, but mix pleasantly together. In this case, the designer used orchestral, natural, and metaphoric sounds. The sequences utilised pitch, rhythm as well as pitch and rhythm together to indicate presence of a body part. All sounds were created in Ableton and Max for Live. These six combinations were chosen because they represented a diverse cross-section of options.

Pilot Study

Following on from this design exercise, we carried out a pilot study following guidance specific for audio testing [2]. We tested whether the four sound elements in the differentiated audio feedback displays were identifiable, discriminable and pleasing. We did not explicitly connect the sound elements to body parts at this stage as the Kinect was not in use. Specifically, we wanted to know:

RQ1: Which audio display has the most discriminable components?

RQ2: Which audio display components are easiest to interpret?

RQ3: Which audio display is most aesthetically pleasing?

Participants

Twenty-one participants were recruited from a large research laboratory with a range of job descriptions, ages, and backgrounds. The sample was 35% female. Ages ranged from 25 – 60: 50% were 25-31, 25% were 32 – 38, and 25% were over 39. No other personal data was collected. We did not collect data on musical experience due to the difficulty of determining musicianship [2]. One participant's data was removed due to hearing loss and consequent outlying data.

Ensemble	
Orchestral	Continuous sounds that provide immediate feedback violin, tuba, clarinet, cello
Natural	Sounds that utilise the ear's ability to distinguish environmental sounds (Auditory Scene Analysis) birdsong, seashore, dog barking, cicada
Metaphoric	Sounds with metaphoric relationship to the body choir, heartbeat, single-clap, double-clap
Sequence	
Pitched	A series of four pitches of which non-detected ones are left out leaving a gap. FM patch
Rhythmic	A series of four sounds with one representing presence of body part and the other absence FM synthesis (woodblock, sawtooth)
Arpeggio	A series of pitches that represent which body-parts are seen, the rhythm changing for different configurations Physical modelling (glockenspiel, vibraphone)

Table 1: Description of differentiated audio displays tested

Procedure

Participants were played 8 audio displays. This included two repetitions, an approach recommended to ensure consistency of response by participants, and thus validity, of results [2]. Participants were first introduced to the audio display and its four component sound elements. They then heard five random combinations of sound elements and were asked to identify them. For the sequential sounds, participants used numbers (e.g. first and third) to identify what they heard. For ensemble sounds, participants were given a card that listed the sounds (e.g. violin, cello etc). These positions/names were emphasised in the introduction. When participants were able to identify what they heard, they verbally gave their answers to the experimenter who was managing the audio displays. Before switching to the next audio display, participants were prompted to fill in a 3-item questionnaire. Audio display order was randomized.

Measures

To answer Research Question 1 (RQ1), we collected the number of correct identifications for each audio display. For RQ2 and RQ3, we provided a 3-item questionnaire based on the NASA Task Load Index [12]. Participants were given a 7-point Likert-type scale for each of three questions: "How successful do you think you were in identifying the sounds?" (RQ2); "How much mental effort did it take to identify the sounds?" (RQ2); and, "How un/pleasant were the sounds?" (RQ3).

	#Correct	Q1	Q2	Q3
Metaphor	**4.90±.31**	**6.7±.47**	**2.65±1.18**	5.00±1.12
Natural	4.6±.75	5.95±.69	3.3±1.45	4.80±1.32
Rhythmic	4.6±.94	5.55±1.69	3.95±1.90	*3.80±1.24**
Arpeggio	*3.9±.97**	*4.5±2.12**	*4.3±1.72**	4.60±1.19
Pitch	*3.55±1.19**	*4.25±1.21**	*4.75±1.29**	**5.15±.93**
Orchestral	*3.15±1.50**	*4.5±1.67**	*4.35±1.18**	4.65±1.31
F-score	*9.31/2.29***	*8.70/2.29***	*5.48/2.29***	3.14/2.29*

Table 2: Pilot study results. The highest score is in bold; significant differences are marked as p<.001 and p<.01*.**

Results

One factor ANOVAs were performed for each measure. The results of this test were significant with large effect size across measures as shown in Table 2. The *Metaphor* audio display ranked first for conveying information (4.9 out of 5 possible correct answers) and interpretability (6.7 out of 7 on perceived success and 2.65 out of 7 on perceived mental effort). It was ranked second for being aesthetically pleasing (5 out of 7). Interestingly, the *Pitched* audio display was ranked first for being aesthetically pleasing, but last or near last on all other measures.

Further post-hoc analysis was carried using Tukey tests. As we were interested in significance between the top ranked audio display and the others, we used only 5 degrees of freedom (comparing first ranked to all other tests). There was a significant difference between the first ranked marked in bold and those starred. Those audio displays that used pitch (*Arpeggio*, *Pitched*, *Orchestral*) conveyed less information and were more difficult to interpret. The post-hoc analysis also showed there was only a significant difference between the *Pitched* and *Rhythmic* audio displays for RQ3: aesthetic preference.

Main Study

Building on the design explorations and pilot, we developed an experimental study to compare differentiated and undifferentiated audio feedback with a more contextual task based on our clinical scenario. We chose the *Metaphor Ensemble* as the differentiated feedback. It was both distinguishable and pleasant in the pilot study, enabling us to focus on whether the additional information of which body parts could not be seen helped the negotiation of position. For the undifferentiated feedback we used a C-major chord played by orchestral instruments. It was chosen as a continuous sound that would not be confused with differentiated feedback and is generally considered aesthetically pleasing.

The collaborative task we designed was specifically created to force an "instructor" to negotiate the guiding and supporting of a "performer" without blocking the camera view. This is the main interactional issue that we identified as problematic in the assessment of MS with the Kinect. Here we were more concerned with developing an

experimental task to induce this negotiation process than trying to simulate the clinical situation.

Collaborative task

The participants' basic task was to instruct a performer to complete two movement sequences in front of the Kinect while hearing audio feedback indicating whether the performer was in full view of the camera. We developed the details of the task: adding constraints and a motivating scenario to create the negotiation process of interest.

Temporal, physical, and verbal constraints were added to the task to keep the emphasis on this negotiation process. The instructor was given 5 minutes to instruct the performer in the movement sequences. He or she could talk as well as demonstrate, but their vocabulary was restricted. Body part words (e.g. hand) and directional words (e.g. left) could not be used. This forced the physical correction seen in the MS scenario that would otherwise be avoided through the use of verbal instructions. The instructor was also required to stay in a triangular area marked on the floor (see Figure 2 for depiction of experimental setup), a constraint that reflects the small size of the examination rooms and physical support needs of some patients.

We motivated the participants by asking them to carry out a meaningful task and by providing a reward. Participants were told that they would be involved in animating cartoon characters using the Kinect, instructing a performer in the movements to do for the camera. We then explained that for the animation to work the performer must be in "full view" of the camera, their head, torso, and hands clearly seen. An important part of their job as instructors was to guide the performer to be in full view without blocking the camera view themselves. In order to help them, we explained that they would hear an audio display providing information about what the Kinect could see.

The scenario described to participants was intended to provide a realistic context, and thus motivation, for instructing a movement protocol in front of a non-interactive camera. Presented as a game, participants with the highest number of points based on a motivational scoring system won two £50 amazon vouchers.

The task focuses on the negotiating of interaction that the instructor must do. To avoid confounding factors, the performer was a confederate and their response carefully controlled. The performer followed the same set of eight rules for each instructor. These were developed to mimic interactions observed in a clinical setting. For example, the performer turned to face the instructor and mirrored all movements. Several rules were generated to force the instructor to correct the performer. The performer always did the first attempt with the wrong arm, for example. The performer was professionally trained in physical theatre and therefore able to copy movements exactly and to adhere to the rules consistently.

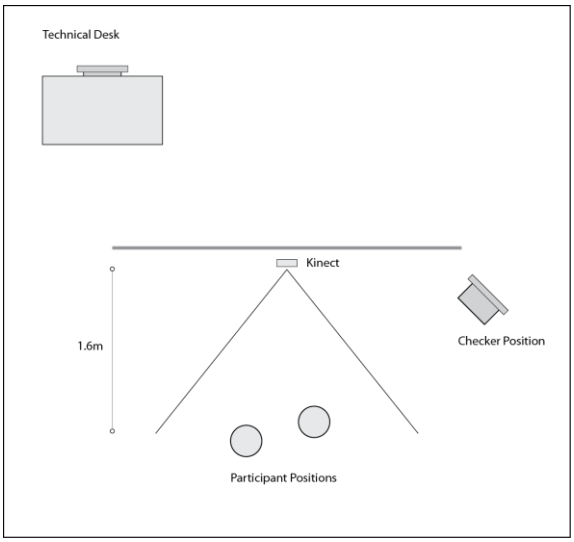

Figure 2: Diagram of experimental set-up

Movements

The movement protocol drew on movements from a neurological examination. Small changes were made to the movements and titles to be more suggestive of cartoon animation, e.g. walking zombie. Movements in space, steps side wards or forwards were added to increase the complexity of demonstration. Two sets of movements were developed, matched for type of movement and spatial complexity. For example, a matched movement sequence changes between arms the same number of times and has the same number of steps. See multimedia appendix for full movement sequences.

Participants

Pairs of friends were recruited to the study. We felt that this contributed to the image of the study as something fun to do and reduced any embarrassment that comes with adults doing a physical movement task. This approach had the further advantage of increasing recruitment because each interested individual had to ask a friend.

Twenty-two participants in all were recruited from a local business. They had a range of occupations; 27% were male; age ranged from 25 – 59 with 45% being 25 – 32. Only three participants had used a Kinect before. The pairs were required to know each other for at least 6 months.

Procedure

Once introduced to the task, participants were provided training in the differentiated feedback. They listened to all components of the audio feedback and were then asked to identify several combination of components. All participants were able to do this. Next participants were instructed in the specifics of the CV system. They were shown the visual representation of the skeleton tracking in the Kinect SDK and demonstrated the three most likely triggers for the loss of full view: (1) the performer being out of range; (2) the performer's hands being too close to the

body; and (3) the instructor being was too close to, or blocking, the performer. The training was completed with a brief task. Participants were given as much time as they needed to adjust to the system and complete this first task.

The first instructor was given written instructions and accompanying videos of two movements sequences that they needed to instruct the performer to do [8]. The instructor could take as much time as needed to feel confident that they knew the movement sequences. They could refer back to the written instructions throughout. When completed, the second participant became the instructor. The order of the movement sequences remained unaltered and the audio feedback was counter-balanced.

To involve the second participant, they were asked to check concrete movement criteria that indicated whether a movement was done correctly (e.g. done on the correct leg) and note whether the performer moved out of the box. They could only say correct, incorrect, and out. Their participation was minimal, but enough to maintain a relaxed atmosphere.

System design

The Kinect for Windows SDK was used to build an interface to show RGB images overlaid by the four "bones" of interest from the skeleton tracking, the head, torso, and two hands. Smoothing algorithms were used to reduce sensor volatility. The disappearance of a body part in the interface was communicated to the Sound Program, Max for Live, through a Midi device to support a Wizard of Oz method. The Midi device had four buttons which the operator pressed when he saw a body part disappear on the screen, making this real-time. This approach was necessary because the SDK does not maintain consistent identifications for the skeletons, which made it impossible to ensure we would have always tracked the performer.

Data collection and analysis

Our research questions address overall response to the audio feedback and more specifically, the ability to respond versus perceived mental effort. We also wanted to check that the audio feedback did not disturb elements of the collaborative setting (e.g. conversation). We captured log-data of when and what audio feedback was triggered as well as video recordings of the sessions. This data gives us two perspectives on people's responses to the audio feedback. We also collected a five-item, 7-point Likert-scale questionnaire for measures of perceived experience. A between-subjects analysis was carried out using Student's t-tests when appropriate.

Quantitative Results

Performance Measures

We calculated several measures of feedback to explore our research question of whether differentiated feedback more effectively enables people to understand what the camera can see than undifferentiated feedback. We postulated that if the feedback was triggered fewer times or was of a shorter duration this would suggest a positive answer to our question. Table 3 shows these measures, including frequency of evocation, average length, and adjusted average length. Frequency and average length are created by aggregating per participant and then taking the mean of the group. Adjusted length is calculated by subtracting the amount of time needed to convey all intended information for each response and then taking the mean of the group. This is 0 ms for the undifferentiated feedback and 2000 ms for the differentiated feedback.

	Freq.	Length	Adj Length
Differentiated Feedback	19.2 (13-28)	3398 ms** (2065 – 5608)	1398 ms (65 – 3608)
Undifferentiated Feedback	24 (6 – 47)	1267 ms (495 – 2100)	1267 ms (495 – 2100)

Table 3: Log-data showing frequency and average length of elicited feedback as well as average estimated response time for each type of feedback.

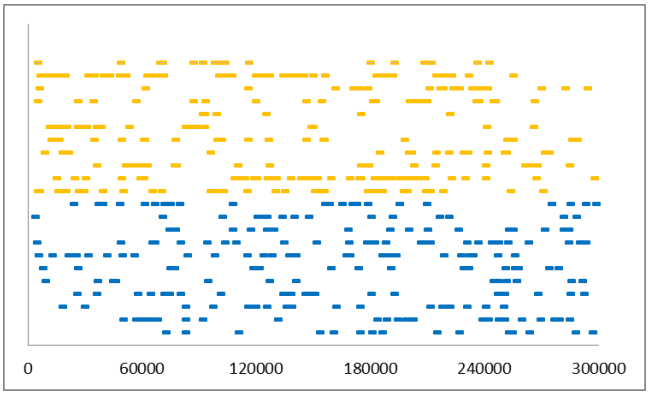

Figure 3: Temporal visualization of audio feedback in ms

The results indicate that the only significant difference ($p<.001$) is length of feedback and this disappears once adjustment is made for the mechanics of the feedback. This suggests that differentiated feedback may not provide a more effective way of helping people understand what the camera can see. If anything, the undifferentiated feedback was of shorter average duration, suggesting a quicker response time.

Prompted by the substantial variation in length of feedback per individual, we next explored whether there was any learning effect. That is, did the length of feedback decreased over time as people began to understand what the camera could see. The video recordings for example, showed efforts to diagnose the cause of the audio feedback. Figure 3 illustrates triggers for each participant across the 5-minute interval of the task. Blue participants heard differentiated feedback and yellow ones, undifferentiated.

The figure suggests no obvious temporal pattern in the data and thus no trend demonstrating that learning had occurred.

Questionnaire Data

The questionnaire data was collected to determine the helpfulness of the audio feedback. Participants were asked to respond to the questions, listed in Table 4, on a 7-point Likert-type scale (1 being "strongly disagree" and 7 "strongly agree"). T-tests were used to compare conditions.

Mean scores between 3 and 4 suggest that neither condition was perceived to be particularly irritating, disruptive to conversation or distracting. However, the differentiated feedback required more mental effort, showing a statistical difference ($p<.001$). While the differentiated feedback also showed a trend towards providing greater understanding of what the Kinect could see, this was not significant. On this basis, we might conclude that the differentiated audio required more effort, but did not add benefit.

Qualitative Results

We examined the video data to address our more general research question of whether audio feedback of any kind supported how people understood the camera view.

Inspection of the videos indicated that participants found it challenging to respond to the audio feedback at all, regardless of type. Indeed, many participants wholly, or partially, ignored it. If we examine participants' responses in the first two minutes of the task, more than two-thirds of the participants ignored the audio entirely or after one failed attempt to understand it (n=6 differentiated and n=9 undifferentiated). Even for the few people who verbally identified the body-part not in view, e.g. the left hand, they found it difficult to diagnose the cause of the problem. The next three examples look at some of the difficulties of understanding the field of view of the camera.

Who is the cause?

Participants were most concerned, often becoming quite agitated, with long stretches of the audio feedback. They would move quickly around the space to make it stop.

Participant 10 started in the far right corner of the triangle (from the camera's perspective). When she heard the audio feedback, she inched along towards the middle of the triangle, leaped across to the other side, and then inched herself towards to the far corner again. The feedback did not disappear so she chose to ignore it.

This participant, like many others, had difficulty determining who should adjust to improve the camera view of the performer. The most common assumption was that the instructor must be blocking the performer and should move as the performer was situated in front of the camera. However, in this case, the log-data showed that the feedback was triggered by the performer's hands, most likely because they were too close to the body. This example illustrates the first challenge of responding to the feedback – who should move?

Question	Undif	Dif
The audio feedback was irritating.	3.27	3
I needed a lot of mental effort to listen to the audio feedback.	**2.27****	**4.45****
The audio feedback disrupted my conversation.	3.10	3.72
I knew which body part the Kinect camera could see.	2.91	3.45
The audio feedback was distracting.	2.8	4

Table 4: Questionnaire data from Likert scale (1 being "strongly disagree" and 7 "strongly agree")

Field of View

Determining who should respond is dependent to a large extent on understanding the field of view. Many participants used the triangular marker on the floor as a guide for the camera's field of view; however, this did not always prove helpful.

Participant 3 stood with her back to the camera on the right side of the triangle, half way between the camera and the performer. She wanted to move closer to the centre of the space so that she could demonstrate stepping to the right without stepping outside the box. She peeked over her left shoulder and inched a bit inwards and stopped, waiting for the feedback. Although her expression and hunched posture indicated that she felt she was blocking a substantial part of the camera view, the audio feedback did not come.

Attempts to be sensitive to the audio feedback, like the one above, illustrate the uncertainty of participants about the field of view. It is not a simple physical mapping between line of sight and the ability to see the performer. It is one affected by the relative movements of both instructor and performer. As such, field of view is dynamically configured in collaborative movement tasks such as this one.

Skeleton Inference

The dynamic configuration of a CV system's field of view is also affected by the inferences that it makes, which are not always bound by the laws of the physical world.

Participant 11 decided to stand next to and slightly angled toward the performer while demonstrating the movements. The performer, as per choreographed rules, turned slightly towards the instructor. When the movements were carried out, the full differentiated feedback triggered. Confused, Participant 11 tried various different angles, which changed the feedback, but did not make it stop. She did not realise that she was too close to the performer for the skeleton tracking to work.

Participant 21 was standing three-quarters of the way down the left triangle side, facing into the triangle. She wanted to demonstrate the zombie but was afraid of blocking the camera view. She stretched one arm out in front of her and nothing happened. She stretched the other arm out in front of her and still nothing happened. Her arms were placed such that the system could still infer the performer's arms even though a horizontal section was blocked.

It was difficult for participants to grasp how their spatial configuration with their collaborator did, or did not, affect what the camera could see. Using knowledge of the physical mapping of the field of view, Participant 11 would have expected to be seen and Participant 21, not seen.

DISCUSSION

This paper describes an initial exploration of the design space for audio feedback to support collaborative interaction with computer vision systems. We developed a differentiated audio feedback display and compared it to undifferentiated audio feedback in a collaborative task inspired by the scenario of the clinical assessment of MS. As such, we were able to ask whether providing information about the body parts the Kinect could see afforded any benefit over the simpler feedback. Our quantitative results showed no benefit of differentiated over undifferentiated audio. Indeed indications were that the former required more mental effort for people to interpret. While this result is dependent upon the particular properties of the audio feedback used and should be treated cautiously, the study highlighted more general challenges in designing audio feedback for these types of systems.

Specifically, the qualitative data help articulate why participants found it challenging to understanding the camera view of a CV system while doing a collaborative task. The analysis indicates that it is not enough to simply convey that something might be out of view of the camera. As illustrated through the video vignettes, there are a variety of complex reasons why a CV system may not be able to "see" someone during a dynamic and physical collaborative interaction. The physical mappings that participants might use to see whether they are blocking the camera view are not reliable. We demonstrated that what the camera can see depends on the specific configuration of participants and the peculiarities of the skeleton inference algorithms. As such, "instructing" participants were unable to determine whether to move themselves or their "performing" partners and thus incapable of instituting even a basic trial-and-error strategy to learn the unexpected behavior of the CV system.

These findings suggest that, independent of medium, feedback for CV systems that support collaborative interaction must communicate the ramifications of the algorithms these systems use to parse the 3D scene (in this case, skeleton tracking). Any interaction paradigm needs to provide users with a quick way of gleaning what a system can see and how it has been interpreted. Reflecting upon our exploration of audio feedback has prompted us to contemplate whether audio, as an independent medium, may be inappropriate for providing this kind of feedback. It is easier to convey the spatial positioning and dynamics within a camera's view, and the system's interpretation of those relationships, using a visual medium. So while it is dangerous to generalize from these initial explorations, we conjecture that while sound can be successfully used to convey some aspects of spatial information for navigational tasks of a single person [15], it is not clear how this might be adapted to address the issues that arise from the collaborative context.

Returning to our scenario of inspiration, the results have led us to take a more pragmatic approach when considering the design of the CV system. This study was motivated by finding a way for a clinician to seamlessly negotiate the support of the patient while maintaining the camera view in order to achieve high quality data capture for the clinical assessment of MS using a CV system. Our findings suggest that this is unlikely to be feasible with sound alone, particularly as we cannot assume that every clinician will have the opportunity to use these systems enough to become very expert users. Instead, the results have prompted us to consider how we might combine visual and auditory feedback. For example, we are considering using an audio cue to prompt the clinician to check the position on her personal screen. Visual cues will indicate that there is a problem and what kind of strategy might solve it (e.g. move farther apart). When tested, we hope this will address the information needs of the clinician without disrupting the interaction.

CONCLUSIONS

This paper articulates the importance of feedback to convey not just the zone of interaction of a CV system, but how the system interprets data within its field of view, and the dynamic spatial relationships that impact that view. This is especially important for tasks, based in the real world, where people collaborate with each other in rich, dynamic and often complicated ways. In this sense, this work provides added confirmation of some of the design principles first explicated by Bellotti et al [1]. This work raises important questions about how what a computer vision system can "see" can be portrayed to users through different kinds of feedback channels in real-time. As CV systems begin to establish themselves as a new interaction modality, there is a rich space to explore appropriate interaction paradigms.

BIBLIOGRAPHY

1. Bellotti, V., Back, M., Edwards, W.K., Grinter, R.E., Henderson, A., and Lopes, C. Making sense of sensing systems. *CHI '02*, 415.

2. Bonebright, T.. and Flowers, J.H. Evaluation of auditory display. In *The sonification handbook*. 2011.

3. Brewster, S. Nonspeech auditory output. In *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications, Third Edition*. CRC Press, 2002, 221–237.

4. Buxton, W. Introduction to this special issue on nonspeech audio. *Human-Computer Interaction 4*, (1989), 1–9.

5. Campo, A. de. Toward a data sonification design space map. *ICAD2007*, 342 – 347.

6. Chang, Y.-J., Chen, S.-F., and Huang, J.-D. A Kinect-based system for physical rehabilitation: a pilot study for young adults with motor disabilities. *Research in developmental disabilities 32*, 6 (2011), 2566–70.

7. Downs, J., Vetere, F., and Howard, S. Paraplay: Exploring Playfulness Around Physical Console Gaming. *INTERACT'13*, 682–699.

8. Fothergill, S., Mentis, H., Kohli, P., and Nowozin, S. Instructing people for training gestural interactive systems. *CHI'12*, 1737.

9. Gerling, K., Livingston, I., Nacke, L., and Mandryk, R. Full-body motion-based game interaction for older adults. *CHI'12*, 1873–1882.

10. Goffman, E. *Frame analysis: An essay on the organization of experience*. Harper and Row, 1974.

11. Hara, K.O., Gonzalez, G., Carrell, T.O.M., et al. Interactional Order and Constructed Ways of Seeing with Touchless Imaging Systems in Surgery. *Transactions on Computer Human Interaction*, 1–34.

12. Hart, S. and Staveland, L. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Human mental workload*, (1988).

13. Heath, C. and Nicholls, K. *Body Movement and Speech in Medical Interaction (Studies in Emotion and Social Interaction)*. Cambridge University Press, 2006.

14. Hermann, T., Hunt, A., and Neuhoff, J. *The sonification handbook*. 2011.

15. Holland, S. and Morse, D. Audio GPS: spatial audio in a minimal attention interface. *Pers Ubiquit Comput 6*, 4 (2002), 253–259.

16. Janata, P. and Childs, E. Marketbuzz: Sonification of real-time financial data. *ICAD04*.

17. Jones, M., Jones, S., Bradley, G., Warren, N., Bainbridge, D., and Holmes, G. ONTRACK: Dynamically adapting music playback to support navigation. *Pers Ubiquit Comput 12*, 7, (2008) 513–525.

18. Kurczak, J. and Graham, T. Hearing is believing: Evaluating ambient audio for location-based games. *ACE'11*.

19. Kurtzke, J.F. Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS). *Neurology 33*, 11 (1983), 1444–1444.

20. Luff, P., Hindmarsh, J., and Heath, C. *Workplace Studies: Recovering Work Practice and Informing System Design*. Cambridge University Press, 2000.

21. Ma, X.M., Fellbau, C., and Cook, P.R. Environmental Sounds as Concept Carriers for Communication. *ICAD2010*.

22. Marquardt, N. and Greenberg, S. Informing the Design of Proxemic Interactions. *IEEE Pervasive Computing 11*, 2 (2012), 14–23.

23. Marshall, J., Benford, S., and Pridmore, T. Deception and magic in collaborative interaction. *CHI'10*, 567.

24. McGookin, D., Brewster, S., and Priego, P. Audio bubbles: employing non-speech audio to support tourist wayfinding. In M.E. Altinsoy, U. Jekosch and S. Brewster, eds., *Haptic and Audio Interaction Design*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, 41–50.

25. Mentis, H.M., O'Hara, K., Sellen, A., and Trivedi, R. Interaction proxemics and image use in neurosurgery. *CHI'12*, 927.

26. Momtahan, K., Hetu, R., and Tansley, B. Audibility and identification of auditory alarms in operating room and intensive care unit. *Ergonomics 36*, (1993), 1159–1176.

27. Morrison, C., Fitzpatrick, G., and Blackwell, A. Multi-disciplinary collaboration during ward rounds: Embodied aspects of electronic medical record usage. *International Journal of Medical Informatics 80*, 8 (2011), e96–111.

28. O'Hara, K., Glancy, M., and Robertshaw, S. Understanding collective play in an urban screen game. *CSCW'08*, 67–76.

29. Panger, G. Kinect in the kitchen: testing depth camera interactions in practical home environments. *CHI'12 Extended Abstracts*, 1985–1990.

30. Reitmaier, T., Benz, P., and Marsden, G. Designing and theorizing co-located interactions. *CHI'13*, 381.

31. Vazquez-Alvarez, Y., Oakley, I., and Brewster, S. a. Auditory display design for exploration in mobile audio-augmented reality. *Pers Ubiquit Comput 16*, 8 (2011), 987–999.

32. Walker, B. and Nees, M. Theory of sonification. In *The sonification handbook*. 2011, 9–40.