Multi-Viewer Gesture-Based Interaction for Omni-Directional Video

Gustavo Rovelo[1,2], **Davy Vanacken**[1], **Kris Luyten**[1], **Francisco Abad**[3], **Emilio Camahort**[3]
[1]Hasselt University - tUL - iMinds, Expertise Centre for Digital Media
Wetenschapspark 2, 3590 Diepenbeek, Belgium
[2]Dpto. de Sistemas Informáticos y Computación.
[3]Inst. Universitario de Automática e Informática Industrial
Universitat Politècnica de València - Camino de Vera S/N, Valencia, Spain
{gustavo.roveloruiz, davy.vanacken, kris.luyten}@uhasselt.be, {fjabad, camahort}@dsic.upv.es

ABSTRACT

Omni-directional video (ODV) is a novel medium that offers viewers a 360° panoramic recording. This type of content will become more common within our living rooms in the near future, seeing that immersive displaying technologies such as 3D television are on the rise. However, little attention has been given to how to interact with ODV content. We present a gesture elicitation study in which we asked users to perform mid-air gestures that they consider to be appropriate for ODV interaction, both for individual as well as collocated settings. We are interested in the gesture variations and adaptations that come forth from individual and collocated usage. To this end, we gathered quantitative and qualitative data by means of observations, motion capture, questionnaires and interviews. This data resulted in a user-defined gesture set for ODV, alongside an in-depth analysis of the variation in gestures we observed during the study.

Author Keywords

Gesture Elicitation; User-Defined Gestures;
Omni-Directional Video; Multi-User Interaction

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

General Terms

Human Factors; Design; Measurement.

INTRODUCTION

ODV is an emerging media format that offers viewers a 360° panoramic video (Figure 1). To create an immersive experience, ODV is typically shown in a CAVE-like setup, or a personal display (e.g. a head-mounted display) in combination with a tracking system to calculate the viewer's correct viewpoint. Recent efforts such as Microsoft's Illumiroom [15] provide interesting possibilities for ODV, as they show how a living room environment could be turned into a small CAVE-like theatre. Benko and Wilson [4] show different scenarios in which ODV can be used, as they describe a portable dome setup in which users can interact with applications such as a 360° video conferencing system, a multi-user game or an astronomical data visualization system.

Although capturing and rendering ODV have been widely investigated and optimized over time, little attention is given to interaction with ODV content. Interaction with ODV includes triggering typical control operations we know from regular video (e.g. play, pause, fast forward and go backward), but also includes changing viewpoint by means of typical spatial interactions such as zooming and panning. These spatial interactions are, however, somewhat constrained, since spatial manipulations are always relative to the original camera position that was used while recording the ODV.

Bleumers et al. [6] recently presented a number of interesting findings regarding users' expectations of ODV. Their research highlights the uncertainty among users about how to interact with ODV and puts forward mid-air gestural interfaces as a possible solution, although they did not explore such interfaces in their work. Mid-air gesturing has been used since the early nineties for controlling television sets [2, 8], and nowadays television sets with a built-in camera and simple gestural interface are commercially available.

We envision ODV content becoming more and more common in the future and accessible within the context of our living rooms. As a result, traditional television watching experiences will change, since multiple viewers no longer have the same region of focus (i.e. the television screen in front of them), but are able to watch video content in any direction. This change also implies that traditional interaction methods, such as a remote control or the current gesture-based TV interfaces, need to be re-evaluated.

Our aim is to understand which mid-air gestures are the most appropriate for interacting with ODV, not only when users are

CHI 2014, April 26 - May 01 2014, Toronto, ON, Canada

ACM 978-1-4503-2473-1/14/04...$15.00. http://dx.doi.org/10.1145/2556288.2557113

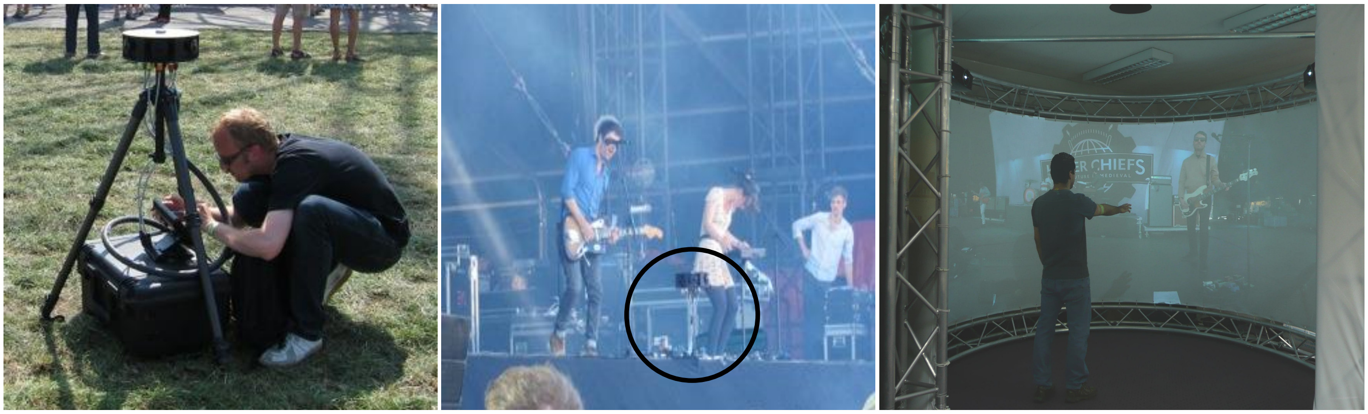

Figure 1. ODV workflow. From left to right, an omni-directional camera, the recording process during a concert, and projection of the resulting omni-directional video in a CAVE-like setup.

on their own, but also in a collocated scenario, in the presence of other viewers who might want to interact with the ODV. More specifically, we investigate which factors users take into account when eliciting gestures for the most frequent ODV control operations. For this purpose, we gather both qualitative and quantitative data through observations, motion capture, questionnaires and interviews.

The main contributions of this paper are: (1) a quantitative and qualitative study to capture user-defined gestures for ODV, (2) an analysis and classification of these gestures and (3) an analysis of the changes in gestures when used in two different configurations: single and collocated settings. Our results inform the design of gestures for future ODV systems and highlight the expected flexibility in tracking and recognizing these gestures when being used in collocated settings. Although our contributions are focused on designing appropriate gesture-based interactions for ODV, our findings can also be useful in other domains that require spatial or time-related operations (e.g. interacting with home cinema systems or controlling the viewpoint during predefined navigation sequences inside a virtual environment).

RELATED WORK

ODV and mid-air gestures are both very active and broad research areas. In this section, we describe the related work that is most relevant to the work presented in this paper.

User-Defined Gestures

In the areas of user-defined gestures and exploration of users' preferences, researchers often focus on finding the best set of gestures for specific tasks, such as grabbing and rearranging a set of objects [12], or pan-and-zoom operations [19]. Others analyse the gestures for a very specific action such as rotation [14], evaluate users' behaviour when interacting with zoomable video [1], study how users rate the appropriateness of the gestures they observe [7], or compare the acceptance level of different gesture sets (i.e. one set created by HCI experts and another by "inexperienced" users) [18].

To generate a set of user-defined gestures, Nielsen et al. [21] and Wobbrock et al. [27] propose similar elicitation approaches: define what operations have to be executed through gestures, ask participants to perform gestures for those operations, and finally extract the gesture set from the collected data. Wobbrock et al. also use Likert scales to gather qualitative feedback, while Nielsen et al. benchmark the set in a second round of trials. Nielsen's et al. methodology has for instance been applied to find gestures that can be used to interact with music players [11]. Another approach is proposed by Grandhi et al. [9], who ask a group of users to describe and mimic different daily tasks that can be extrapolated to human-computer interactions.

Interaction with Omni-Directional Video

Benko et al. [3, 5] describe the challenges of interactive curved and spherical displays, which we believe to be representative for CAVE-like ODV setups. These challenges include developing walk-up-and-use interaction techniques, creating a transparent environment where users can interact with the appropriate device for each task, and devising compelling applications for this type of device. Our focus lies on the challenge of designing appropriate interaction in the context of ODV.

Researchers already investigated several aspects of ODV interaction. Macq et al. [16] implement ODV navigation using the camera of a tablet PC as the orientation tracker and the screen as the display device (i.e. a peephole display). Neng and Chambel [20], on the other hand, describe the use of 360° "hypervideos", which provide extra information through embedded navigational links. These videos are watched over the Internet, on a regular computer screen. We are, however, specifically interested in mid-air gestural interfaces.

Bleumers et al. [6] describe user expectations on ODV, and how users think gestural interfaces would be appropriate for this new media format. Zoric et al. [28] present a user study in which they observed pairs of participants interacting with high definition panoramic TV through gestures. Their observations suggest that the design of multi-user gesture systems should allow for socially adapted gestures for controlling and navigating video content. However, Zoric et al. consider this study to be merely a first step in exploring how users interact with such content using a gesture-based system. We investigate this topic more in-depth.

STUDY ON ODV GESTURES

The aim of our study is to determine how users conceive gesture-based ODV interaction, when alone and when interacting with other participants in a collocated scenario. We not only look into interactions that map on control operations typically performed with video content, either on television or digital video players, but also on some control operations that are typical for spatial exploration. As a result, the control operations considered in this study are commands that manipulate time (i.e. play, pause, skip scene, fast forward and go backward) or space (i.e. panning and zooming).

Methodology

Since user-generated gesture sets tend to have a higher acceptance level among users [18], we adapted the gesture elicitation methodology of Nielsen et al. [21] to gather the gestures that participants consider most appropriate for the aforementioned control operations. The study consisted of two sessions: first, a participant was asked to perform the gestures alone, and in the second session, two participants had to perform the gestures in a collocated setting.

Participants were seated on a couch, inside a CAVE-like ODV setup, as seen in Figure 2. This kind of setup helps participants to explore the interaction possibilities, since it clearly reveals the spatial properties of the ODV content. It also prevents participants from being influenced by the form factor of the output device (e.g. when using a rectangular screen, participants are more likely to unnecessarily frame gestures within a rectangle in front of them). To evaluate the impact of the collocated setting, participants were seated reasonably close to each other during the second session (as it would happen in a living room, sitting next to each other on a couch). Participants were not forced to sit uncomfortably close to each other, however, and had sufficient space to sit without invading each other's personal space.

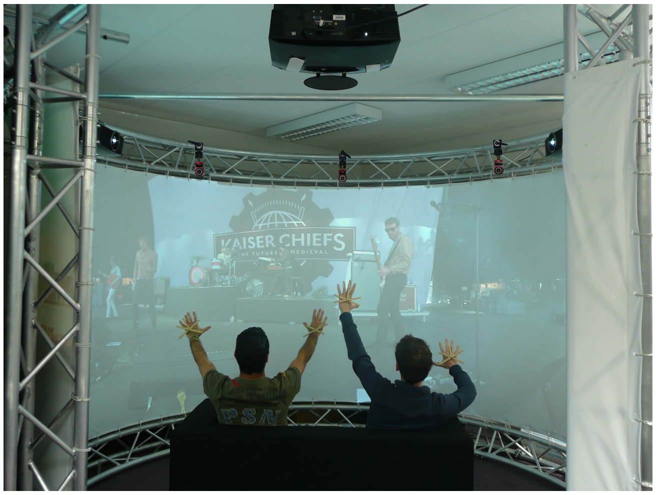

Figure 2. Participants performed the experiment while sitting on a couch inside a CAVE-like ODV setup.

The ODV did not respond to the gestures of the participants. Similar to Wobbrock et al. [27], we decided against a Wizard of Oz to avoid that participants constrain or adapt their gestures according to the feedback they receive (e.g. to compensate for a delay or mismatch). A Wizard of Oz approach would also be impractical in the collocated scenario, because providing feedback for each participant simultaneously would inevitably result in inconsistencies. Before the sessions, an observer explained the list of control operations by showing an actual ODV to the participants. During the sessions, however, only still images of an ODV were shown to avoid unnecessary distractions.

Participants were asked to perform one easy to repeat and easy to understand gesture for each operation. They were informed that they had complete freedom of action to devise a gesture or posture using hand(s) and/or finger(s), and that the same gesture could be repeated for more than one operation, if considered appropriate. Before the collocated session, the observer explained that they would be interacting with the ODV independently, but at the same time.

After the short explanation, the observer asked the participants to devise an appropriate gesture for each operation, one by one. The observer did not impose any time constraints. Participants simply had to signal when they were ready to perform a gesture, and next, the observer gave the go-ahead to execute the gesture. During the collocated session, both participants had to perform their gesture at the same time, so the observer waited to give the go-ahead until both participants were ready.

When both participants finished performing the gestures for all control operations during the collocated session, they were asked to swap positions on the couch and repeat the trial. In other words, participants performed gestures for each control operation three times in total: once alone, and twice when sitting next to another participant.

To control order effects, we divided the participants in two groups: each group received the control operations in a different order during the sessions. However, we did not simply randomize the order of the control operations, but decided to maintain a logical structure (e.g. by grouping related operations such as fast forward and go backward), to make it easier for the participants to devise gestures.

Personal information such as age, gender and experience was recorded before the first session. After completing both sessions, participants filled out a questionnaire regarding their experience and discussed their opinions with the observer.

Participants

Sixteen participants took part in our study: twelve male and four female, with ages ranging from 23 to 52 years old (average age 31.5). All of them are colleagues at our research centre. Two participants are left handed, two ambidextrous, and the others are right handed. Most participants are experienced touch screen users (twelve participants use them daily), but merely two participants play video games on consoles like the Nintendo Wii or Microsoft Xbox with Kinect more than once a month. Only two participants make regular use of gestures to interact with their PC, either by performing mouse gestures to control the web browser, or by using a multi-touch mouse. None of the participants have interacted

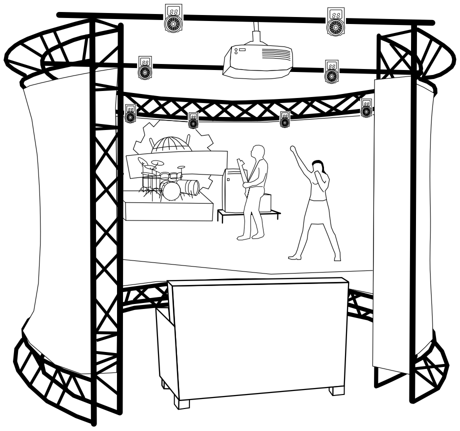

Figure 3. Motion capture setup for our study. An 8-camera OptiTrack system and a set of rigid body markers were used to track hand movements. The tracking system was set up in a 360° CAVE-like environment, with a couch in the centre of the tracking volume.

with gesture-based TVs. Finally, twelve participants knew beforehand what ODV is, but only two of them had previously interacted with an ODV system.

For the collocated session of the study, we formed eight pairs according to the following criteria: four pairs with participants who are used to interact with each other and four pairs with participants who rarely interact with each other. We based our grouping criteria on the *"friendship ties"* described by Haythornthwaite and Wellman [10]:

Close friendship: people who work in the same office, usually have lunch together, and would go to the movie theatre together.

Working together: people who know each other, but rarely interact with each other in the work environment.

Apparatus

To gather data on the gestures that participants performed, we used motion capture, allowing us to measure the spacial dimensions of the gestures. For this purpose, we used eight OptiTrack V100:R2 cameras and the Natural Point Tracking Tools software (Figure 3). The OptiTrack cameras have a 640x480 pixels image resolution and a maximum capture frame rate of 100 fps. They are capable of tracking markers with sub-millimetre accuracy. We also used a normal video camera to record the sessions, to make classifying the different gesture easier during analysis.

To track participants, a rigid body marker composed of small IR reflective balls had to be attached to each hand. Before the actual study, we ran a pilot study for two purposes: (1) to verify whether the instructions and study design were clear and (2) to uncover limitations and issues with our apparatus. It allowed us to optimise the rigid body markers in order to avoid occlusion problems when participants turned their hands. We therefore built the markers with wooden sticks that exceed the size of the participant's hand (Figure 4). In this manner, only the hands' centres are tracked and not the small finger movements, but this suffices for our purposes, since we have complementary video recordings.

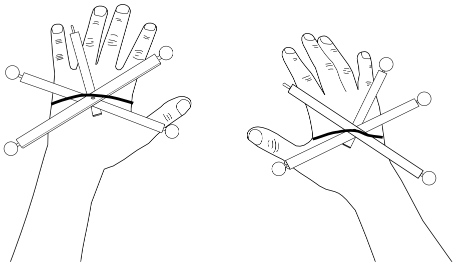

Figure 4. Rigid body markers composed of small IR reflective balls, used for motion capture. The markers are sufficiently large to ensure that they always exceed the size of the participant's hands.

RESULTS

We analysed the video recordings of all the sessions, following the strategy proposed by Peltonen et al. [23]: first extracting the video segments with useful data and then extracting all the required information. For this purpose, we defined parameters to annotate the videos of each gesture according to the suggestions of Nielsen et al. [21]. The parameters we used are *hand usage (one or two hands)*, *trajectory of the movement* (*linear or circular*), *type of gesture* (*movement or steady posture*) and *granularity* (*fine-grained finger movements or coarser hand movements*). We also described all the gestures in natural language using these parameters, in such way that others would be able to understand and reproduce them.

The *Hand usage* column of Table 1 shows that participants had **no clear preference for using one or both hands for most gestures**. They did prefer to use one hand for performing a pause gesture and both hands for zooming (these differences are statistically significant, based on a non-parametric binomial test between the two possibilities). Furthermore, participants **preferred to use linear movements rather than circular movements**. As indicated in the *Trajectory* column of Table 1, the difference is statistically significant for all control operations. This confirms the findings previously reported for pan-and-zoom interaction with wall-sized displays [19]. People prefer linear movements when they are asked to devise easy to perform, easy to remember and easy to repeat gestures.

We classified all the gestures as either static (e.g. a steady hand posture that uses both index fingers to represent the typical pause symbol), or dynamic (e.g. performing a "push" gesture by moving a hand away from the body and back, with the palm outwards). Both examples are depicted in Figure 5. The *Gesture type* column of Table 1 shows a **clear preference for using dynamic movements rather than static hand postures to represent most control operations**. Pause and stop are control operations for which the participants' preference is less clear, but for the other operations, the differences are statistically significant. We believe that the number of participants using steady postures is higher for pause and stop, because they both implicitly denote turning the video into a standstill state. A number of participants used the same gesture with different speed/timing to represent different control operations. P15, for instance, explicitly mentioned that he did the same gesture for play and fast forward, moving his right

	Hand usage		Trajectory		Gesture type		Granularity	
Control operation	One	Two	Linear	Circular	Static	Dynamic	Fine	Coarse
Play	31	17	48	0	14	34	14	34
Pause	36	12	48	0	27	21	9	39
Stop	19	29	48	0	26	22	0	48
Skip scene	24	24	33	15	3	45	2	46
Fast forward	31	17	41	7	3	45	7	41
Go backward	29	19	39	9	6	42	9	39
Pan	18	30	46	2	1	47	4	44
Zoom	3	45	48	0	0	48	10	38

Table 1. The gesture elements that were used for analysis of the video recordings. The value in each cell represents the number of participants who used the element during the study (16 participants performed a gesture for each control operation three times, resulting in 48 samples per control operation in total). Coloured cells represent a statistically significant difference between levels (non-parametric binomial test, $\alpha = 0.05$)

hand to the right, but varying the time he kept pointing in that direction (he used more time for fast forward).

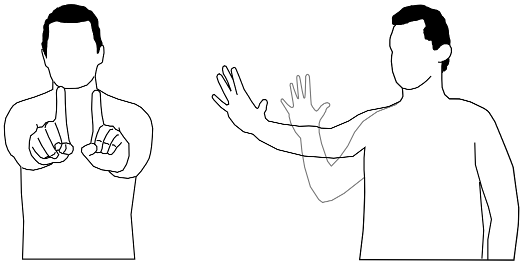

Figure 5. A steady hand posture and a dynamic hand movement to represent the pause operation. Grey lines represent early states of the gesture, black lines represent the final state.

We also found statistically significant differences comparing the usage of fingers (fine granularity) and whole hands (coarse granularity) to perform the gestures (*Granularity* column of Table 1). Participants **preferred to use coarser hand movements instead of fine-grained finger movements**, even though they were informed that they could use finger movements to represent the control operations.

As expected, participants extrapolated their knowledge from real-life devices and software applications (in this case, mostly video or DVD players), an observation that was also made by Henze et al. [11] in the context of gestures for music playback. This was especially true for play, pause, stop and zoom. Participants for instance tried to transform a symbolic representation into a gesture or posture, such as a triangle for play or a square for stop (e.g. P14 explicitly asked *"... do I have to do the square for stop?"*). Another form in which participants extrapolated their real-life knowledge is when they considered that play, pause and sometimes stop should be represented by the same gesture, as these control operations are often mapped to the same button on devices or in software applications (e.g. a lot of media players use the same button for play/pause and do not have a stop button). For zooming, twelve of the sixteen participants employed the typical spread-and-pinch gesture, even those participants who mentioned they are not frequent multi-touch users.

Collocated Interaction

We asked participants to choose one or several factors that influenced their decision on gestures when sitting next to another person. "Avoid invading the other participant's private space" and "Avoid colliding with the other participant's gestures" were chosen by seven out of the sixteen participants. "Avoid blocking the other participant's view of the video", however, only received two votes, one of which belongs to P14, who felt his field of view was blocked and reported collisions with his fellow participant while performing the gestures. We believe the fact that "Avoid blocking the other participant's view of the video" received few votes is due to the absence of a particular task to perform with the ODV. Participants did not need to be engaged with the content and thus did not consider blocking the other participant's view an important factor.

Analysis of our study notes and video recordings shows that participants of the four pairs of the *"close friendship"* category had no problems performing gestures side by side. One pair of participants even made jokes about synchronized dancing, because they performed nearly identical movements for some control operations. Participants who were part of a *"working together"* pair, on the other hand, were more uncomfortable and some of them expressed that feeling during an informal interview after the study. P5 reported, for instance, that *"It was not comfortable doing the gestures with the other participant."* and P16 reported that *"I felt limited by the presence of the other participant. She invaded my private space."*

By analysing the video recordings and motion capture data, we identified three interesting situations that resulted from the collocated interaction: participants adapted the size of their gesture, changed hands to perform the same gesture, and chose a completely different gesture for the same control operation. We discuss each of these gesture adaptations in the next sections.

Size Adjustment

A number of scripts were implemented to automatically analyse the motion capture data gathered by the OptiTrack system. We first measured the space participants can cover when they completely stretch their arms to the side, to the front and to the top. The areas created on each plane (*XY - frontal*, *XZ - top* and *YZ - lateral*) represent the maximum distances that a participant is able to reach. These areas were used to create baseline bounding boxes. Next, we decomposed the captured hand movements and determined the 2D bounding boxes that represent the areas covered by each gesture on the three planes. Finally, we calculated the ratios between the sizes of the bounding boxes for each gesture and the participant's baseline bounding boxes.

	Lateral adjustment (X)			Vertical adjustment (Y)			Depth adjustment (Z)		
Control operation	S vs A	S vs B	A vs B	S vs A	S vs B	A vs B	S vs A	S vs B	A vs B
Play	3	3	2	2	0	1	4	7	2
Pause	1	1	0	4	3	0	2	3	0
Stop	2	3	1	1	1	0	4	3	0
Skip scene	6	3	1	2	3	0	7	2	1
Fast forward	5	5	0	2	3	0	5	5	2
Go backward	4	5	1	4	3	1	5	6	3
Pan	8	4	1	1	2	0	7	5	3
Zoom	8	7	0	1	1	0	6	5	0

Table 2. Size adjustments. Values in each cell represent the number of participants adapting the size of the gesture by more than 10% between the specified sessions. Column names stand for S-Single participant session, A-First collocated trial, and B-Second collocated trial.

We used these ratios to detect changes in size of a gesture for each control operation across the sessions, to investigate if participants used this as a strategy to adapt gestures in the collocated setting. Table 2 presents the number of participants who reduced the size of their gestures by more than 10%, for each of the three axes. The size adjustment is especially noticeable for control operations that typically involved lateral movements (e.g. fast forward, go backward, pan, zoom), due to the presence of the other participant.

Analysis of the friendship ties revealed an expected trend: participants of *"working together"* pairs used the size adjustment strategy more often. They adjusted 42.2% of all the gestures performed during the sessions (for all the control operations and in any of the three movement directions), while participants of the *"close friendship"* pairs adjusted 17.2% of their gestures. We did not find a statistically significant correlation between friendship ties and size adjustments, which can probably be attributed to the limited number of pairs per friendship tie.

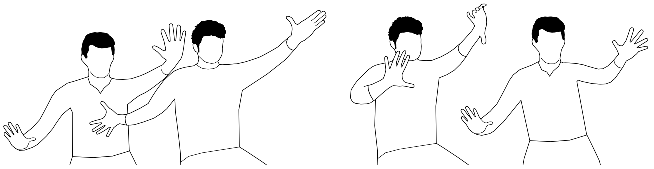

Figure 6. An example of how participants adapted the size of their gestures. P8 and P13 displaced their movements when doing the zoom gesture, after their hands collided during the first collocated trial.

As an example, we briefly discuss the gestures for the zoom operation of a *"working together"* pair. When P8 and P13 performed the zoom gestures for the first time, their hands collided. The second time, after switching positions, both participants adjusted their gesture by displacing the movements to the free space (Figure 6). While looking into another *"working together"* pair, P4 and P16, we clearly noticed both participants reducing the movement of their hands between the single and collocated session to represent the skip scene operation.

Gesture Mirroring

Analysis of the video recordings shows that participants also adapted their gestures by using a different hand to perform the same gesture. Table 3 depicts how many participants used gesture mirroring across the different sessions. In total, five participants adapted their gestures in this manner. Only one of those participants was ambidextrous, and four were part of a *"working together"* pair. In total, the *"working together"* pairs used gesture mirroring for 17.2% of the gestures performed during the sessions and the *"close friendship"* pairs for 4.17% of their gestures, but again, no statistically significant correlation was found between friendship ties and the adaptations.

Control operation	S vs A	S vs B	A vs B
Play	1	2	3
Pause	1	2	1
Stop	1	1	0
Skip scene	1	1	2
Fast forward	0	0	1
Go backward	1	1	0
Pan	2	1	3
Zoom	0	0	0

Table 3. Gesture mirroring. Values in each cell represent the number of participants who mirrored gestures between the specified sessions. Column names stand for S-Single session, A-First collocated trial, and B-Second collocated trial.

To illustrate the gesture mirroring strategy, we briefly discuss three examples. P8 used his left hand for the fast forward gesture when performing the gesture during the single session, but he used his right hand when sitting to the right of P13 (Figure 7). Similarly, P14 used his right hand when he was sitting to the right of P7 when doing the skip scene gesture, and then his left hand when he was sitting to the left of P7. Finally, P1 did the play gesture using her left hand when P10 was sitting at her right, and her right hand after exchanging positions on the couch.

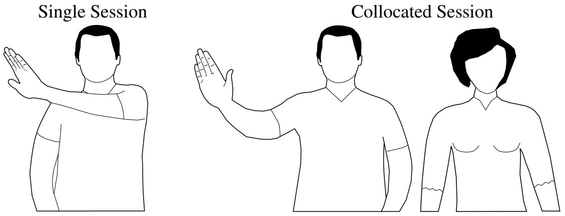

Figure 7. An example of how participants mirrored gestures. P8 used a different hand to perform the same gesture when another participant was present.

Although only a limited number of participants adopted this mirroring strategy, it is still interesting to note that users will expect a gesture to be recognized by the system in both cases, whether they are using their left or right hand. The Microsoft Kinect development guidelines already suggest this strategy to create flexible gestural interfaces [17].

Choosing New Gestures

Table 4 depicts the number of participants who changed gestures across the sessions. Eleven participants changed at least one of their gestures. In total, 11.5% of the gestures performed by participants of a *"working together"* pair were changed across the sessions, and 9.38% in case of *"close friendship"* pairs. No statistically significant correlation was found between friendship ties and choosing new gestures.

The main reason for this behaviour is the extra time participants spent thinking about the gestures. P15, for example, mentioned that the second time he had to perform the gestures, he *"tried to put some logic"* in them, and P2 mentioned that he tried *"to do more energy efficient"* gestures.

Control operation	S vs A	S vs B	A vs B
Play	2	3	1
Pause	1	1	0
Stop	2	2	1
Skip scene	1	3	2
Fast forward	3	3	1
Go backward	4	3	1
Pan	3	2	1
Zoom	0	0	0

Table 4. Choosing new gestures. Values in each cell represent the number of participants who chose a new gesture between the specified sessions. Column names stand for S-Single session, A-First collocated trial, and B-Second collocated trial.

Most pairs (all the *"close friendship"* pairs and two of the *"working together"* pairs) discussed the reasons and implications of their gestures, with comments like *"your gesture is not energy efficient"* or *"your gesture is error prone"*. This interaction between participants sometimes resulted in them changing the gesture. For instance, P4 used both hands for panning when she represented the operation the first time. She completely stretched both arms to the front, making a clockwise circle with her right arm to pan right and a counter clockwise circle with her left arm to pan left. The second time, she used only one hand, copying the gesture of her fellow participant: moving her right hand, pointing with the index finger to the left and then to the right.

Agreement Level

We classified all the gestures for every control operation into groups of similar gestures, based on the parameters *hand usage (one or two hands)*, *trajectory of the movement* (*linear or circular*), and *type of the gesture* (*movement or steady posture*). In addition to these parameters, we considered the overall movement pattern of the gesture (e.g. the directions of the movements). We did not include *granularity* (*fine-grained finger movements or coarser hand movements*), so we classified a spread-and-pinch performed with two fingers or with both hands as similar gestures.

Next, we calculated the percentage of participants who used a particular type of gesture and the agreement level for each operation. For this purpose, we used the formula of Wobbrock et al. [26]:

$$A_i = \sum_{j=1}^{n} \left(\frac{G_{ij}}{G_i} \right)^2 \qquad (1)$$

A_i is the agreement level of the ith operation, G_i the total number of gestures performed for the ith operation and G_{ij} the number of elements in the jth group of gestures for the ith operation. Park and Hand [22] used this formula in a similar manner.

We illustrate the formula's usage by applying it to the gestures used for panning (Equation 2). We found three groups of similar gestures, with sizes 30, 2 and 16. As a result, the agreement level was 0.5036 for the panning operation.

$$A_{Panning} = \left(\frac{30}{48}\right)^2 + \left(\frac{2}{48}\right)^2 + \left(\frac{16}{48}\right)^2 = 0.5036 \qquad (2)$$

We also calculated the percentage of participants who chose a particular type of gesture, taking into account the 48 gestures performed for each operation. Table 5 depicts both these percentages and the agreement levels for the top-rated gesture for each operation.

DISCUSSION

We propose the user-defined gesture set described in Table 5. We believe these gesture will lead to a high acceptance level among users. Figure 8 gives a graphical representation of this gesture set. The different states of gestures that require movements are represented with different line colours: grey colours represent early states of a gesture and the black line the final state.

We considered the most repeated gestures across all the sessions to assemble our gesture set. In 62.5% of the cases, the difference between the most repeated gesture and the second most repeated gesture was very large. For the fast forward and go backward operations, however, the differences were small. There were six groups of similar gestures for the fast forward operation (representing 35.42%, 29.16%, 18.75%, 10.42%, 4.17% and 2.08% of the participants) and eight groups for the go backward operation (representing 41.67%, 33.33%, 8.33%, 6.26%, 4.17%, 2.08%, 2.08% and 2.08% of the participants). In case of the stop operation, performing the halt gesture with one hand was the most repeated one (performed by 33.33% of the participants), but as we already chose this gesture to represent the pause operation, we decided to use the second most repeated gesture: the halt gesture using both hands (performed by 22.92% of the participants).

We can observe in Figure 9 that the large variety of gestures to represent certain control operations sometimes causes low agreement levels, for instance for the play and skip scene operations. The zoom operation, on the other hand, was the one with the most consensus. This can be explained by the fact that participants regularly relied on existing mental models to devise gestures to represent control operations. The following three examples illustrate this behaviour:

Zoom and pan. Not surprisingly, the most repeated gesture for zoom was the widely used spread-and-pinch gesture, as indicated by the agreement level. For panning, "grabbing" the video and moving the hand was the most repeated gesture. Similar gestures have been proposed by Fikkert et al. [7] for zooming and by Stellmach et al. [24] for panning, in the context of large display control.

Play and pause. Some participants mentioned that video players use the same button to represent play and pause,

Control operation	Gesture	Rate (%)	Agreement level
Play	Push gesture, moving the hand with the palm outwards toward the front and back in a fluent movement.	35.41	0.18
Pause	Halt gesture, holding the arm completely stretched with the palm outwards for a few seconds.	41.67	0.22
Stop	Halt gesture, holding both arms completely stretched with the palm outwards for a few seconds.	22.92	0.21
Skip scene	Moving the hand from right to left or from left to right one time and returning to the starting position.	33.33	0.16
Fast forward	Left to right movement, holding the hand pointing to the right for a few seconds.	35.42	0.26
Go backward	Right to left movement, holding the hand pointing to the left for a few seconds.	41.67	0.30
Pan	Using one hand to "grab" the video and then move it from left to right or from right to left.	62.50	0.50
Zoom	Using the spread-and-pinch gesture, moving two hands apart (spread) and bringing them back together (pinch).	75.00	0.63

Table 5. Top-rated gestures for the eight control operations. The rate represents the % of participants who performed this type of gesture during the study. The agreement level gives an indication about the variety of gestures that were performed for an operation.

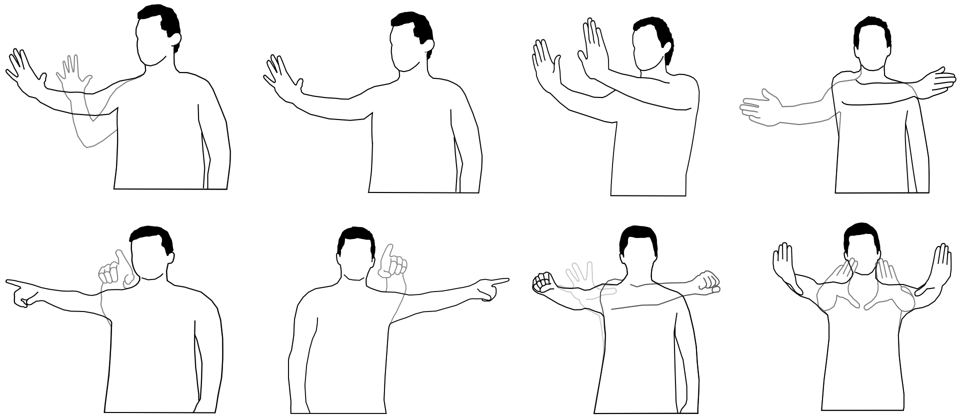

Figure 8. User-defined gesture set for ODV. From left to right and top to bottom: play, pause, stop, skip scene, fast forward, go backward, pan and zoom. Grey lines represent early states of the gesture, black lines represent the final state.

and thus they also used the same gesture for both control operations. This behaviour was also reported by Henze et al. [11] in the context of gestures for music playback. Overall, there was a great diversity in gestures to represent play and pause, leading to lower agreement levels.

Fast forward and go backward. Most media players and timelines associate "the future" with the right side (e.g. arrows pointing to the right in video players for fast forward), and "the past" with the left side. Participants in our study represented both control operations following this established mental model, which is consistent with the observations of Henze at al. [11].

Although we considered the most repeated gestures to assemble our gesture set, not all gestures are necessarily the most optimal solution. Participants sometimes changed their gesture to copy their fellow participant (imitative behaviour was also reported by Walter et al. [25], in the context of a public display game), so they considered their first gesture to be suboptimal. Eleven participants also chose a new gesture for at least one of the control operations, due to reasons such as wanting a more "energy efficient" gesture. This implies that using a gesture set over a prolonged period of time might lead to a different prioritisation of gestures. Our gesture set thus needs further validation and refinement before its actual implementation, and the next step is to benchmark the chosen gestures, as suggested in the methodology of Nielsen et al. [21]. Another component to consider before implementing the set, is how to discriminate between gestures and other movements, for instance by indicating the start of an interaction with a specific body pose [25].

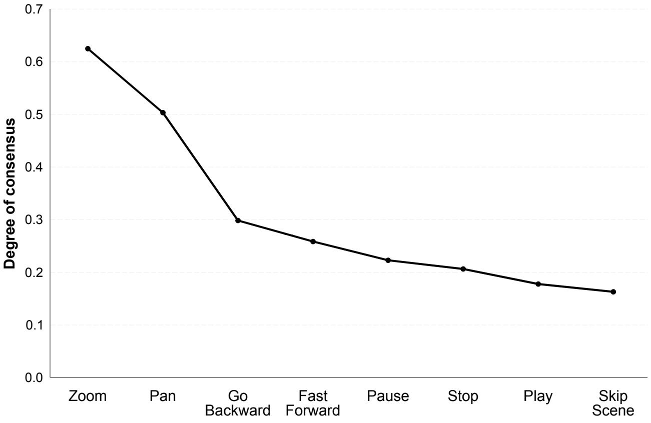

Figure 9. Gesture set agreement level in descending order.

We noticed that minimal friendship ties between participants had a negative effect on their experience. Participants of the *"working together"* category often felt uncomfortable performing gestures close to each other. This must be taken into consideration when designing a gestural interface: when users are likely to be unfamiliar with each other, less invasive gestures might need to be considered, while such gestures might be a source of fun for close friends.

We identified a number of gesture adaptations caused by the collocated setting: participants used a different hand to perform the same gesture, changed the size of the gesture, or performed the gesture more to the left or to the right to avoid colliding with their fellow participant. Participants expect their gestures to be recognized in all cases, regardless of the hand they use or the scale of their gesture. Therefore, the system has to be designed to recognize all (or at least the most common) forms of a gesture (e.g. recognize a spread-and-pinch gesture performed with two hands, but also one performed with two fingers). The need to support gesture variations was already observed in other contexts, such as multi-touch surfaces [13] and interactive public displays [25].

We only looked into which gestures participants found to be the most appropriate for every operation. We neither took into account the parameterization of control operations (e.g. how users express how many degrees to pan with their gesture), nor the focus point of those control operations (e.g. how users express on which area they want to zoom in). These factors also need to be investigated, because they can influence how gestures are scaled. When a single user performs a small panning gesture, for instance, it probably means that she wants to move only a little. In a collocated setting, on the other hand, the same small panning gesture might be the result of the presence of others. A gesture recognizer should take this into account by scaling the panning operation according to the situation. Care has to be taken, however, that this kind of adaptation does not confuse users. The system needs to provide sufficient feedback about the scale of control operations.

During our study, participants did not need to be engaged with the actual ODV content, nor did they have different points of focus, which is likely to happen in a CAVE-like setup. As a result, they rarely considered aspects such as blocking the other participant's view. Such aspects will become an important factor when participants do engage with the content, which might lead to additional adaptation strategies when performing gestures. However, participants not having different points of focus during the study does allow us to generalise our results beyond CAVE-like setups.

CONCLUSIONS

We presented the results of a gesture elicitation study that we carried out with the goal of understanding which mid-air gestures users consider to be the most appropriate for interacting with ODV. We not only considered interacting with ODV when users are on their own, but also when they are together with other viewers in a collocated setting. We gathered both quantitative and qualitative data by means of observations, motion capture, questionnaires and interviews. Based on an in-depth analysis of this data, we described a user-defined gesture set for ODV, and gesture variations and adaptations that came forth from individual and collocated usage.

The user-defined gesture set contains the most repeated gestures in our study. We observed a clear preference for using linear movements to represent easy to perform and easy to remember gestures. Participants also preferred to use dynamic movements rather than static hand postures to represent most control operations, and coarser hand movements instead of fine-grained finger movements. We also found that participants tried to extrapolate their knowledge from interaction with real-life devices or software applications.

Analysis of the collocated interactions revealed interesting behaviours that participants exhibited while devising and performing gestures. They adapted their gestures in several ways because of the presence of another participant. The most prominent adaptations were changing the size of the gesture and shifting the hand movements to the opposite side of where the other participant was sitting. Other gesture adaptations were using a different hand for the same gesture or devising a new gesture. These adaptation strategies highlight the importance of a good system design. The ODV system must be able to interpret the user's actions (e.g. adapting the scale of a gesture because of the proximity of another person versus adapting the scale to make smaller adjustments), give sufficient feedback about the scale, and provide sufficient flexibility to cope with different variations of gestures (e.g. a spread-and-pinch with both hands or two fingers).

By forming pairs of participants with different friendship ties, we observed that the level of comfort between participants was an important factor. Participants who are familiar with each other enjoyed the study and even started making jokes about a choreographed dance when they made similar gestures. Participants who are not used to close interaction were less comfortable and made comments such as *"I felt limited by the presence of the other participant. She invaded my private space."* The analysis revealed a trend showing that those participants also used the adaptation strategies more often. The effect of gestures on the level of comfort between users should be taken into account when deciding on a gesture set.

Although our findings are based on a gesture elicitation study regarding control operations for ODV in a CAVE-like setup, we believe the user-defined gesture set and user expectations can also be useful in other setups and domains that require spatial or time-related operations. Furthermore, this study is only a first step in the exploration of ODV gestures, with many interesting avenues for future research, such as studying collaborative tasks with users who each have different points of focus, or an in-depth analysis of the reasons for choosing a specific gesture, which might reveal certain cultural, educational or generational influences.

ACKNOWLEDGMENTS

We thank the participants of our study, Miguel Nacenta for feedback on early drafts, Raf Ramakers for his help with the OptiTrack system and Philippe Bekaert for the use of the ODV infrastructure. The iMinds ICON AIVIE project, with project support from IWT, is co-funded by iMinds, a research institute founded by the Flemish Government. This research was also partially supported by the National Council of Science and Technology of México as part of the Special Program of Science and Technology.

REFERENCES

1. Axel, C., Ravindra, G., and Tsang, O. W. Towards characterizing users' interaction with zoomable video. In *Proc. of the ACM Workshop on Social, Adaptive and Personalized Multimedia Interaction and Access*, SAPMIA (2010), 21–24.

2. Baudel, T., and Beaudouin-Lafon, M. Charade: remote control of objects using free-hand gestures. *Communications of the ACM 36*, 7 (1993), 28–35.

3. Benko, H. Beyond flat surface computing: challenges of depth-aware and curved interfaces. In *Proc. of the ACM Int. Conference on Multimedia*, MM (2009), 935–944.

4. Benko, H., and Wilson, A. D. Multi-point interactions with immersive omnidirectional visualizations in a dome. In *Proc. of the ACM Int. Conference on Interactive Tabletops and Surfaces*, ITS (2010), 19–28.

5. Benko, H., and Wilson, A. D. Pinch-the-sky dome: freehand multi-point interactions with immersive omni-directional data. In *CHI Extended Abstracts on Human Factors in Computing Systems* (2010), 3045–3050.

6. Bleumers, L., den Broeck, W. V., Lievens, B., and Pierson, J. Seeing the bigger picture: a user perspective on 360° TV. In *Proc. of the European Conference on Interactive TV and Video*, EuroiTV (2012), 115 – 124.

7. Fikkert, W., van der Vet, P., van der Veer, G., and Nijholt, A. Gestures for large display control. In *Proc. of the Int. Gesture Workshop*, GW (2010), 245–256.

8. Freeman, W. T., and Weissman, C. D. Television control by hand gestures. In *Proc. of the Int. Workshop on Automatic Face and Gesture Recognition*, FG (1995), 179–183.

9. Grandhi, S. A., Joue, G., and Mittelberg, I. Understanding naturalness and intuitiveness in gesture production: insights for touchless gestural interfaces. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI (2011), 821–824.

10. Haythornthwaite, C., and Wellman, B. Work, friendship, and media use for information exchange in a networked organization. *Journal of the American Society for Information Science 49*, 12 (1998), 1101–1114.

11. Henze, N., Löcken, A., Boll, S., Hesselmann, T., and Pielot, M. Free-hand gestures for music playback: deriving gestures with a user-centred process. In *Proc. of the Int. Conference on Mobile and Ubiquitous Multimedia*, MUM (2010), 1–10.

12. Hespanhol, L., Tomitsch, M., Grace, K., Collins, A., and Kay, J. Investigating intuitiveness and effectiveness of gestures for free spatial interaction with large displays. In *Proc. of the Int. Symposium on Pervasive Displays*, PerDis (2012), 1–6.

13. Hinrichs, U., and Carpendale, S. Gestures in the wild: Studying multi-touch gesture sequences on interactive tabletop exhibits. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI (2011), 3023–3032.

14. Hoggan, E., Williamson, J., Oulasvirta, A., Nacenta, M., Kristensson, P. O., and Lehtiö, A. Multi-touch rotation gestures: performance and ergonomics. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI (2013), 3047–3050.

15. Jones, B., Benko, H., Ofek, E., and Wilson, A. IllumiRoom: peripheral projected illusions for interactive experiences. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI (2013), 869–878.

16. Macq, J.-F., Verzijp, N., Aerts, M., Vandeputte, F., and Six, E. Demo: omnidirectional video navigation on a tablet pc using a camera-based orientation tracker. In *Proc. of the ACM/IEEE Int. Conference on Distributed Smart Cameras*, ICDSC (2011), 1–2.

17. Microsoft. Kinect for Windows Human Interface Guidelines v1.8.0 (2013). Retrieved Jan. 2014, from: msdn.microsoft.com/en-us/library/jj663791.aspx.

18. Morris, M. R., Wobbrock, J. O., and Wilson, A. D. Understanding users' preferences for surface gestures. In *Proc. of Graphics Interface*, GI (2010), 261–268.

19. Nancel, M., Wagner, J., Pietriga, E., Chapuis, O., and Mackay, W. Mid-air pan-and-zoom on wall-sized displays. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI (2011), 177–186.

20. Neng, L. A. R., and Chambel, T. Get around 360° hypervideo. In *Proc. of the Int. Academic MindTrek Conference*, MindTrek (2010), 119–122.

21. Nielsen, M., Störring, M., Moeslund, T. B., and Granum, E. A procedure for developing intuitive and ergonomic gesture interfaces for HCI. *Gesture Based Communication in Human-Computer Interaction 2915* (2004), 409–420.

22. Park, W., and Han, S. H. Intuitive multi-touch gestures for mobile web browsers. *Interacting with Computers 25*, 5 (2013), 335–350.

23. Peltonen, P., Kurvinen, E., Salovaara, A., Jacucci, G., Ilmonen, T., Evans, J., Oulasvirta, A., and Saarikko, P. It's mine, don't touch!: interactions at a large multi-touch display in a city centre. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI (2008), 1285–1294.

24. Stellmach, S., Jttner, M., Nywelt, C., Schneider, J., and Dachselt, R. Investigating freehand pan and zoom. In *Mensch & Computer* (2012), 303–312.

25. Walter, R., Bailly, G., and Müller, J. StrikeAPose: revealing mid-air gestures on public displays. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI (2013), 841–850.

26. Wobbrock, J. O., Aung, H. H., Rothrock, B., and Myers, B. A. Maximizing the guessability of symbolic input. In *CHI Extended Abstracts on Human Factors in Computing Systems* (2005), 1869–1872.

27. Wobbrock, J. O., Morris, M. R., and Wilson, A. D. User-defined gestures for surface computing. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI (2009), 1083–1092.

28. Zoric, G., Engström, A., Barkhuus, L., Hidalgo, J. R., and Kochale, A. Gesture interaction with rich TV content in the social setting. In *CHI workshop on Exploring and Enhancing the User Experience for TV*, TVUX (2013).