Spatiality in Videoconferencing: Trade-offs between Efficiency and Social Presence.

ABSTRACT

In this paper, we explore ways to combine the video of a remote person with a shared tabletop display to best emulate face-to-face collaboration. Using a simple photo application we compare a variety of social and performance measures of collaboration when using two approaches for adding spatial cues to videoconferencing: one based on immersive 3D, the other based on traditional 2D video-planes. A Face-to-Face condition is included as a 'gold-standard' control. As expected, social presence and task measures were superior in the Face-to-Face condition, but there were important differences between the 2D and 3D interfaces. In particular, the 3D interface positively influenced social- and co-presence measures in comparison to 2D, but the task measures favored the two-dimensional interfaces.

Categories and Subject Descriptors

H.5.3 [**Information Systems**]: Information Interfaces and presentation(e.g. HCI) – Group and Organization Interfaces

General Terms

Design, Experimentation, Human Factors.

Keywords

Videoconferencing, Social Presence, Collaborative Virtual Environment, photo-ware, distributed collaboration

1. INTRODUCTION

There is a growing demand for real time telecommunication systems that support effective collaboration between physically dispersed teams. To meet this need, many CSCW researchers are developing video mediated communication (VMC) systems that allow distant colleagues to accomplish tasks with the same, or better, efficiency and satisfaction than when collocated [1].

VMCs provide a rich medium where distant people can see and hear each other in real time while sharing both verbal and non-verbal cues such as speech and facial expressions. Unfortunately, however, VMC teleconferencing has proven to be more similar to audio only conferencing than unmediated face-to-face collaboration [2], [3], leading to a research push to improve VMC's support. One particular approach is to provide a shared spatial frame of reference, where users combine individual locations and individual views into a common space [4], [5], [6]. To date, the research in this area has primarily focused on the development of systems that support and demonstrate immersive 3D VMC environments that offer shared spatially rich perspectives. However, there has been a lack of empirical analysis of their effectiveness.

This paper presents the results of an experiment that investigates the impact of spatial contexts on social presence and on parameters of task performance. Our work is significant because it is one of the first papers that empirically studies the use of multiple display surfaces for supporting remote collaboration, and compares interaction in such a system with Face-to-Face and 2D interface conditions. We also discuss the implications of our findings for the iterative refinement of VMC systems in general.

CSCW '06. November 2006. Banff. Alberta. Canada.

2. RELATED WORK

In our work we are interested in comparing 2D and 3D video mediated collaboration to unmediated face-to-face communication. In doing this we want to consider the effect of spatial cues, and also how an interactive tabletop display can be added to increase the naturalness of the collaboration. Thus we draw on several areas of related work.

2.1 Video-mediated versus non-mediated communication

Traditional 2D video-conferencing systems provide a compressed 2D representation of a 3D space, constraining many of the rich cues available in face-to-face collaboration, including depth cues, resolution, and field of view. More importantly, the natural and fluid human controls for directing attention (rotating the eyes, turning the head, etc.) are replaced with crude mechanical surrogates that require explicit control. All these factors reduce the quality of visual input and inhibit perceptual exploration [7].

Vertegaal [6] argued that the disparity between audio-visual and face-to-face communication is caused by the absence of several nonverbal channels that are normally used in face-to-face meetings, particularly gaze awareness. Some of this is due to the

lack of a common spatial reference frame in conventional video conferencing. For example, without the ability to establish a relative position between him/her and the remote person, speakers can not negotiate a mutual distance between them [8]. Without a spatial reference frame gaze awareness is difficult, i.e. the remote person cannot infer from the video image of the other, where s/he is looking at. This is indeed a problem, as gaze has been identified as an essential part in verbal communication [9]. Speakers and auditors use gaze during face-to-face conversations to exchange and maintain roles, to regulate turn taking behavior, to signal attention or boredom, or to give and seek feedback in form of short glances [10].

If communication channels are missing, speakers automatically by-pass them through supported channels (mostly verbal) in order to adhere to basic coordination mechanisms that can be used in non-mediated communication (described as grounding in [11]. However, this comes at the cost of a higher collaborative effort. For example, if turn taking behavior cannot be regulated through gaze, the name of the attended person may be spoken before turning over the floor, or a dedicated moderator could control the floor. These necessary workarounds contribute to what we might perceive as the unwanted artificial, distanced, or mediated character that is frequently associated with conventional videoconferencing systems today.

2.2 Spatial Approaches to VC

In order to overcome the problem of missing spatial cues, various spatial approaches to videoconferencing have been developed.

One way of creating a shared reference frame is to make videoconferencing consistent within a fixed room or hardware configuration. This approach is applied in the "Office of the Future" work at UNC [12], or the TelePort [13]. In both cases projectors are used to create a spatially immersive AR display that supports remote collaboration in a office environment.

Collaborative Virtual Environments (CVEs) offer another way of creating a shared spatial reference frame. Artificial representations (avatars) of participants "meet" each other in a computer generated, shared 3D environment. There are a number of different types of CVEs. For example, in Vertegaal's GAZE groupware system [6], "personas" (2D image) of every participant are arranged around a virtual table in a shared room. Every user is equipped with an eye tracker that detects the fixation points of the person on the screen and is used to rotate the virtual persona. As a consequence, participants can easily infer from the orientation of each others persona where that person is looking at.

The general principle of personas was adopted and extended by other three dimensional CVEs like "FreeWalk"[4],"AliceStreet" [14] or cAR\PE! [5]. In the latter, users can freely navigate their persona through a virtual conferencing room and interact with others and shared documents in a number of different ways. Spatial visual and audio cues can combine in natural ways to aid communication [15]. Users can freely move through the space setting their own viewpoints and spatial relationships; enabling crowds of people to inhabit the virtual environment and interact in a way impossible in traditional video or audio conferencing [16]. Even a simple virtual avatar representation and spatial audio model enables users to discriminate between multiple speakers.

2.3 Spatiality in Table-Top Scenarios

Table-top interfaces are becoming more and more popular and used not only because of the inexpensive digital projector technology available nowadays, but also because of the advantages of a horizontal interface. People are used to work on tables, so it therefore is an obvious option to use this surface as an interaction space, especially for co-located collaboration. The increasing amount of digital information (in particular digital photography) becomes increasingly the object for communication and collaboration. Table-top interfaces allow for embodied, media-rich, fast and fluid interaction in co-located collaboration. Scott et al. [17] give an overview on the history of table-top interfaces including guidelines for the design.

Collaborative table-top systems provide a spatial reference frame for the interactions which do not need to be learned by the users. In addition, the placement of physical, tangible objects on the table follows the same ease of use. When bringing virtual objects into the scene, either a tangible user interface metaphor [18] should be used or some other metaphors have to be developed or adapted. As Krueger et al. [19] point out; the orientation of the objects on the table is a significant HCI factor for comprehension, coordination, and communication. While using a single vertical display groupware orientation is clearly defined, due to the limited options for arrangement and position of the co-located users, table-top interfaces have to provide interfaces to move and orient the virtual objects.

3. USER STUDY

Although there have been many examples of 2D and 3D collaborative systems, there have been few empirical studies comparing collaboration between such systems and with unmediated cace-to-face collaboration. We are interested in the impact of (added) spatiality in three dimensional systems on social presence and on task performance. Furthermore we want to investigate the effect of table-top interfaces as an additional shared spatial frame compared to collaborative virtual environments displayed on a vertical screen only. By doing this, we hope to contribute to a better understanding of the issues related to the design of effective VMC systems that eventually will be able to provide real alternatives for physical travel.

We narrow our interest to two specific dimensions: (1) The extent to which a person feels being together with the other persons and (2) the usability of actual state-of-the-art systems in terms of efficiency.

The first dimension is best described with the term Social Presence. Common definitions of Social Presence include the sense of "being together" [20], the sense of "Being There with others" [21], or the "perceptual illusion of non-mediation"[22]. Measuring Social Presence can be done in a very reliable and elegant way using the semantic differential measure by Short et al. [23]. The reliability of this instrument for comparisons of videoconferencing interfaces to has been proven in earlier studies [24], [25]. As a sub-factor within the social presence construct, Co-Presence is of further particular interest here too, because it refers to the interpersonal sub-dimensions of isolation/inclusion and mutual awareness [26].

To explore the second dimension, the usability and efficiency of the interface can be measured using several metrics such as the time needed to complete a certain task, the confusion an interface introduces, the errors and misunderstandings it produces, and the speed of conversation including referencing to objects to be discussed. We are interested in all these factors and have adopted a mixed measurement approach using subjective ratings and video observation.

3.1 Experimental Design

The experiment used a one-factor, repeated measures design, comparing different variables of the communication and collaboration across four conditions. The order of conditions was randomized in each experiment following a Latin square scheme.

3.2 User Interfaces

To be able to explore our dimensions of interest in different conditions we developed four collaborative interfaces, suitable for a photoware task, where participants have to talk-about, point at, move, and rotate digital pictures on a table:

A) unmediated face-to-face collaboration around a real table (figure 1), labeled as "FTF". Here, the digital pictures are printed onto paper and allow for natural tangible interaction.

Figure 1. Condition Face-To-Face ("FTF")

B) mediated remote collaboration around a shared interactive table (figure 2), labeled as "3D-local", because spatial cues are supported within the local, real world reference frame. The digital photos are displayed and pre-arranged on the table surface, while a touch sensitive surface allows for interaction with the pictures.

Figure 2. Videoconferencing with touch table ("3D-local")

C) mediated collaboration through a standard 2D-video conferencing interface (figure 3), labeled as "2D", as no three-dimensional reference frame is given. This setup uses a state-of-the-art videoconferencing system involving relatively large-sized video streams of the participants displayed on the screen as well as a screen area for application sharing operated with a standard computer mouse.

Figure 3. Standard Videoconferencing ("2D")

D) mediated collaboration around a virtual table in an immersive desktop collaborative virtual environment (figure 4), labeled as "3D-remote" as the given spatial reference frame within spatial cues are supported is a remote space different from the real world. While the interaction with digital photos is done with the standard computer mouse, the representation of the table to share the pictures as well as the representations of the participants' video streams are shown in the three-dimensional space. A special tracking device was used to allow for virtual head-movement within the environment.

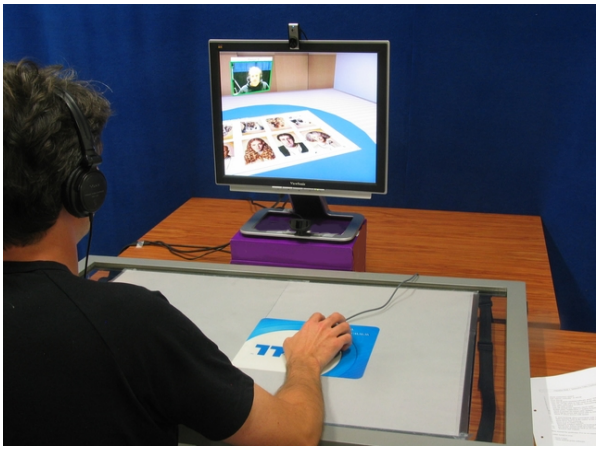

Figure 4. Immersive Videoconferencing ("3D-remote")

As can be seen the main difference between these conditions is in how the user's partner is represented, either in unmediated face-to-face collaboration, or using a variety of 2D and 3D cues. Table 1 outlines the main differences of the conditions, including whether it was possible for the users to have their individual spatial perspective onto the pictures, the spatial reference frame provided, whether digital or printed media were used, and what form of interaction was applied.

Table 1. Main differences of the conditions

	Face-to Face	***3D-local***	***2D***	***3D-remote***
Gaze supported	*Yes*	*Yes*	*No*	*Yes*
Table Interaction	*Yes*	*Yes*	*No*	*Yes*
Input	*Natural Gesture*	*Natural Gesture*	*Mouse*	*Mouse*
User View	*Individual*	*Individual*	*Shared*	*Individual*

3.3 Participants

Thirty subjects (22 male and 8 female) participated in the experiment. In 15 sessions, teams of two subjects took part in four trials for a total of 120 trials. The age of the participants ranged from 22 to 45 years (median age 26 years).

Participants had no prior knowledge of the experiment except for the fact that the objective was to compare videoconferencing systems. The participants were recruited among post-grad students and staff members from different departments at the local university. To exclude mixed gender effects and to make sure that all team members already knew each other before the experiment, we asked every participant we invited by email to bring along a same-gender friend as his or her team partner. All participants had normal, or corrected to normal vision.

3.4 Task

In order to obtain realistic results on collaborative behaviour the design of an appropriate task is crucial. To provoke a rich communication between participants that would reveal the limits of different videoconferencing systems, a task was designed with a highly ambiguous content. This follows from Media Richness Theory [27], in which more communication cues are required to resolve tasks with a high level of uncertainty.

In this case the task was for participants to work together matching photographs of dogs to pictures of their owners. Participants were told during the introduction that one side result of this experiment should reveal if a study that showed that dogs and their owners resemble each other [28] could be replicated successfully for local dogs and owners. In each of four rounds, a set of four photos of owners and four photos of their dogs were presented in random arrangements. The challenge for the participants was to find the correct matches by discussing which dog might resemble which owner the most Each team was allowed to take as much time as they needed to come up with an answer that both team members agreed upon, but they were also encouraged to take as little time as possible.

The photographs were taken especially for this by the first author, with consent of all dog owners. The pictures of the owners showed the face of the person, the pictures of the dog showed either a portrait or a full body perspective of the dog, depending on its size. Out of a total of 30 pairs, five sets of four dog and owner pairs each were formed with an equal balance of female and male owners, as well as a mixture of different dog breeds.

3.5 Experimental Conditions and Apparatus

As mentioned in section 3.1 we explored social presence across four conditions:

1. Condition "Face-to-Face". In this condition, both participants collaborated about a set of paper photographs in the same room, sitting on two opposite sides of a table (figure 1). The photos were of a standard format (9x7 inch, resolution 1024 x 1280 pixels).

2. Condition "3D-local". Each participant was seated in front of a horizontally aligned, touch sensitive panel which in turn was placed in front of a LCD monitor (figure 2). A projector under the table projects the photo application onto the touch panel. Using a single finger photos could be easily moved across the panel or rotated by dragging the rotation handles of a selected photo. The LCD monitor behind the touch panel shows live video of the remote person. That person was seated in front of an identical setup, but with an upside down version of the photo application running on the touch screen. Both participants had a clear idea of their own side of the panel and had their own individual view of the table. Half the photos were initially placed in a way facing towards participant 1, and the other part facing towards participant 2, upside down for participant 1.

3. Condition "2D". In this condition, a conventional videoconferencing system (Conference XP [29]) was used. Two video windows were placed at the top segment of the LCD screen, one showing one's own video and one showing the other person's video. A shared photo application window was positioned underneath (see figure 3). Both participants could interact with the photos at the same time using a simple mouse click and drag interface. At all times, both users should see exactly the same things on the screen, just as in most conventional video conferencing tools. Photos that were uploaded at the beginning of the trial were all facing the same way (upright).

4. *Condition "3D-remote*". In this condition, participants met in a 3D virtual room, represented as virtual video-personas around

a virtual table, on top of which was running a shared photo-application (figure 4). The interface was implemented using the "cAR\PE!" virtual tele-collaboration space [5]. The head orientation of the participants was tracked with a 2DOF infrared tracker [30] that was positioned close to the web camera. Head tracking information was used to control the view into the virtual room. That way, participants could easily change their viewpoint from the table towards the other person's persona, and from the orientation of the virtual persona it could be inferred what the other person was currently paying attention to. The position of the virtual characters was fixed and could not be changed by the participant. Half the photos were flipped in the initial layout, so that half the photos could be seen in the right orientation by each participant. To manipulate the photos, both participants used a standard mouse that controlled the shared mouse pointer displayed on the virtual table.

Audio and video recordings were made of the subjects using two DV-cameras with external microphones that were placed close to the speakers. For all mediated conditions, two visually and acoustically separated rooms were prepared with identical standard desktop PCs (P4, 2.80 GHz), monitors (LCD, 17'', 1280x 1024), headsets (stereo with mono microphone) and webcams (USB, CIF resolution). All computers involved in the setup were connected through a one megabit network switch.

The shared photo viewing application was based on the open source graphics editor Inkscape [31].Shared access to the application was implemented using the desktop sharing software UltraVNC [32]. Both participants shared the same mouse pointer with equal manipulation privileges. The photo application as well as the UltraVNC Server and UltraVNC client ran on extra two laptop computers, which were also connected through the network switch. In order to capture the activity on the shared Inkscape window, one further PC was connected to the network switch which ran another UltraVNC client window that was captured in real time by the screen capturing software.

3.6 Procedure

For every one-hour session a group of two subjects were present. Upon arrival the participants were given a sheet with the *Participant Information*, explaining (1) the goal of the experiment, (2) the general procedure, (3) the anonymity of the experiment, and (4) a participant consent text, which was to be signed by them. Additionally, the document contained the *General Demographics Questionnaire*.

A second sheet was handed out, describing the task according to 3.3. After potential questions with regards to the task description were answered, each participant took part in four rounds, one for each condition (FTF, 3D-local, 2D, 3D-remote). The order of conditions was randomized beforehand (Latin Square). The task in each condition was the same, however new sets of photos with different dogs and owners were used in each round.

In the videoconferencing conditions, participants were given training in the use of the interface using a special set of photos of dogs and owners that was shown on the photo application window during every "warm-up" phase. In the "2D" condition, participants were explicitly made aware that the other person sees exactly the same view as them at all times.

In the two 3D conditions, the individual view aspect of the interface was emphasised and the ability to infer the other person's gaze direction was pointed out. No instructions on the general strategy how to find the matching pairs were given.

In all mediated conditions, the subjects wore audio head-sets which were explained and adjusted for best comfort. The head tracking in the 3D-remote condition was adjusted individually for every participant, so that all parts of the virtual table and the other participant's persona could be viewed within a comfortable head posture range.

Once both participants signalled that they had understood the interface and how to use it, a set of the actual experiment photos was opened on the shared photo-application. That was the official start of that round. It was now up to the participants to discuss and manipulate all the pictures that were on display and come up with a solution as to what the possible correct pairs might be. Suggested pairs could be indicated simply by moving a photo of a dog close to the photo of an owner. Once the team found four pairs that both team members agreed on, the round was finished.

Subjects were then brought back to the same room and were asked to fill out a questionnaire addressing different communication and usability parameters. After the questionnaires were filled out, the actual number correct dog-owner pairs found in the last round were told to the team. After the fourth and final round was over and the fourth questionnaire was filled out by the participants, they were briefly interviewed about how they liked the task and were then asked to give their personal preference ranking of all four conditions they had just collaborated with. At the end of the experiment, the participants were thanked, and two blocks of chocolate were given to them as a reward.

3.7 Expected Results

Social presence and co-presence scores are predicted to be higher in the spatial interfaces than in the two dimensional one because of the additional spatial cues. Furthermore, we expect that the spatial cues in the 3D interfaces would have a positive impact on the participants' ability to create a common ground that would also show in their communication patterns. We therefore predict a higher use of deictic references in the 3D interfaces. On the other side, completion times in the 2D interface are expected to be shorter, as the photos did not need to be rotated as often as in the 3D interfaces. Finally, we assume that face-to-face communication will be the most effective and efficient in all dimensions of interest.

4. RESULTS

In the following, questionnaire results as well as data from video analysis are presented. The questionnaire results have been analyzed using the statistical package SPSS version 11. Main effects were first tested with a repeated measures analysis of variance (ANOVA). If a significant effect was found, post-hoc pair wise comparisons were calculated using the Bonferroni adjustment for multiple comparisons. The significance level was set to 0.05 during the entire analysis.

4.1 Questionnaire Results

According to the procedure described in 3.5, 15 sessions with 2 participants each were run, where session 1 and 2 were initial

pilot trials whose results have not been considered in this statistical analysis. Therefore, 13 sessions form the basis for our results. All questionnaires of the 26 subjects have been valid. No values were missing. The questionnaires included a total of 24 seven point Likert scale items addressing usability parameters as well social presence and copresence.

4.1.1 Copresence:

In total four items addressed perceived copresence:

"I was always aware that my partner and I were at different locations."

"I was always aware of my partner's presence"

"It was just like being face to face with my partner"

"It felt as if my partner and I were in the same room."

Subject marked how much they agreed or disagreed with each of these statements on a Likert scale of 1 (disagree) to 7 (agree). A reliability analysis for the factor "Copresence" was calculated which showed that all four items measure a uni-dimensional construct sufficiently well (Cronbach's Alpha = 0.84). Therefore, the individual scores of those four items could be combined to one single Copresence score. The results of that score in the different conditions is shown in Figure 5.

A significant main effect was found, $F(3,75)=64.3$, $p<0.01$.While Face-to-Face was rated the highest in copresence, both 3D-conditions received higher average scores than the 2D-condition. Post-hoc analysis furthermore showed a significant difference between Conditions "3D-local" and "2D" ($p=0.04$). Subjects felt significantly more co-present in the Face-to-Face condition that the other conditions, and also in the 3D-local condition than the 2D condition.

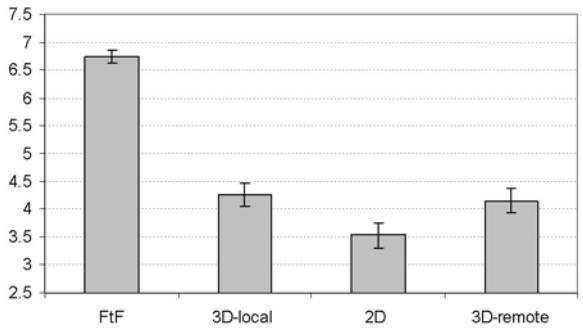

Figure 5. Average score and std. error for Copresence

4.1.2 Social Presence

Social presence was measured with a semantic differential technique like that suggested in Short et al. [33] using a total of eight bi-polar pairs. Participants were asked to rate the communication media on a seven point scale between each of the following pairs:

cold	-	*warm*
insensitive	-	*sensitive*
small	-	*large*
spontaneous	-	*formal*
impersonal	-	*personal*
passive	-	*active*
unsociable	-	*sociable*
open	-	*closed*

Reliability analysis on these eight items revealed a high Cronbach's alpha result of 0.89. Again, one single combined social presence score could therefore be formed from the average of the individual item scores. The results of the social presence factor score in the different conditions is shown in Figure 6.

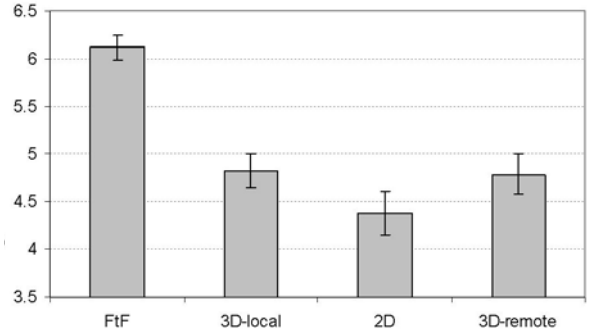

Figure 6. Average score and std. error for Social Presence

There was a significant main effect, $F(3,75)=20.8$, $p<0.01$, showing that social presence in the Face-to-Face condition was significantly higher than other conditions. Both 3D conditions were rated higher in social presence than the 2D condition, however, none of the mediated conditions showed differences in post-hoc comparisons.

4.1.3 Usability Parameters

Eleven items addressed different aspects of the usability of the system. As these items were not expected to measure a single construct, the results were calculated for every item individually. The questions and their scores are shown in table 2.

Except for questions 5, 6, and 7, all results showed a significant main effect. Post-hoc comparisons revealed, that many of these effects reside in the big difference of the scores between the Face-to-Face and the mediated conditions. However, two significant differences between the spatial and the 2D videoconferencing interface could be found. The score for question 4, "I could easily tell where my partner was looking" was significantly higher in the 3D-local condition than in the 2D condition ($p=0.02$), and also significantly higher in the 3D-remote condition than in the 2D condition ($p=0.03$). Furthermore, the results of question 6, "I was often confused", uncovered, that participants felt more often confused in the 3D-local condition than in the 2D condition ($p=0.045$) and also more often confused in the 3D-remote condition than in the 2D condition ($p=0.05$). The results in all other usability and communication related items show the trend for condition 2D to be closer to Face-to-Face than the three dimensional conditions.

Table 2: Average scores and standard deviations for the eleven usability questions in the questionnaires on a 7-point, Likert-type scale ranging from 1 (strongly disagree) to 7 (strongly agree)

	Question	*FtF*	*3D-local*	*2D*	*3D-remote*	*Results post-hoc comparisons*
1.	It was very easy to make myself understood.	6.4 (1.2)	5.4 (1.1)	5.9 (1.2)	4.9 (1.6)	FtF > 3D-loc; FtF > 3D-rem
2.	I could easily tell where my partner was pointing at.	6.7 (0.7)	4.9 (2.0)	4.3 (2.2)	4.7 (1.8)	FtF > 3D_local; FtF > 2D; FtF > 3D_remote
3.	I could not contribute anything to the solution we came up with.	1.6 (0.7)	2.1 (1.2)	1.8 (0.8)	2.2 (1.1)	FtF < 3D_remote
4.	I could easily tell where my partner was looking at.	5.8 (1.5)	4.6 (1.7)	2.9 (1.7)	4.2 (2.0)	FtF > 3D_remote; FtF > 2D 3D_local>2D, 3D_remote>2D
5.	There was a lot of time when no-one spoke at all.	2.4 (1.6)	3.3 (1.7)	2.6 (1.5)	3.0 (1.7)	*
6.	I was often confused.	1.7 (1.0)	3.1 (1.8)	2.1 (1.2)	3.5 (1.9)	FtF < 3D_local; FtF < 3D_remote; 2D < 3D_local; 2D<3D_remote
7.	We were never talking over one another.	5.0 (2.0)	4.4 (1.6)	4.2 (1.6)	4.4 (1.6)	*
8.	I hardly looked at my partner's face.	4.0 (2.2)	3.3 (1.7)	4.5 (2.1)	3.9 (1.9)	*
9.	I knew exactly when it was my turn to speak.	5.8 (1.0)	4.5 (1.5)	4.9 (1.5)	4.7 (1.4)	FtF > 3D_local; FtF > 2D; FtF > 3D_remote;
10.	I could always clearly hear my partner's voice.	6.7 (1.0)	5.2 (1.6)	5.9 (1.1)	5.6 (1.5)	FtF > 3D_local; FtF > 2D; FtF > 3D_remote;
11.	When I looked at my partner, I could always clearly see his or her face.	6.8 (0.4)	5.5 (1.4)	5.3 (1.9)	5.9 (1.0)	FtF > 3D_local; FtF > 2D; FtF > 3D_remote;

Note: Standard deviation in parentheses, Asterisk = no significant differences

4.1.4 Preference:

After every condition had been used by the participants, they were asked to rank them from one to four according to their personal preference. From there ranks, a preference score was calculated from 0 to 1, where the rank 4 results in a preference score of 0, and a ranking of 1 results preference score of 1. The results are shown in figure 7. All participant preferred the Face-to-Face over any of the mediated ones. This condition significantly won this category. Within the mediated conditions, the 2D condition was slightly preferred over both spatial approaches, although the differences did not reach significant levels in the post-hoc analysis.

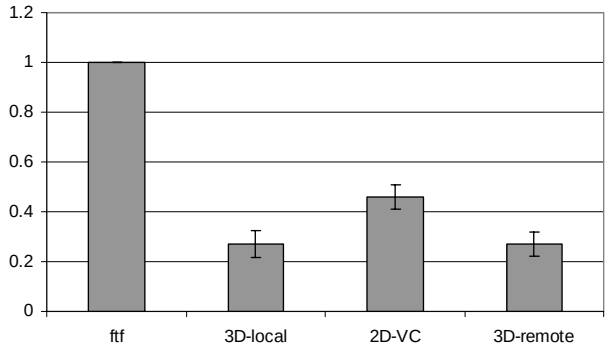

Figure 7. Mean Preference Score and Standard Error.

4.2 Video analysis results

The video observation analysis was done by the first author. The captured video streams were combined into a single, synchronized video with separate audio tracks. The original audio streams of the two participants have been assigned to the left and right audio channel in the final video.

Due to technical difficulties only 12 out of 13 videos were completely captured and available for analysis. The outside views of the experiment at each station as well as the shared photo application window were rendered into a single video as shown in figure 8. Video editing was done with the video editing package Adobe Premiere Professional 1.5.

Figure 8. Still sample from combined observation video.

Video analysis was done with these videos that integrated all the information in one file. The following occurrences were of interest: (1) task completion time, (2) turns per minute, (3) technology and process versus task related turns, and (4) deictic versus descriptive references.

4.2.1 Task Completion Time

The task completion time was defined from the moment when the participants first saw the dogs and owners until the moment when they signaled that they found the solution both agreed with. Results varied significantly across the four conditions, $F(3,33)=9.1$, $p<0.01$, where condition "2D" was the fastest, followed by condition Face-to-Face, condition "3D-local" and at the end, taking more than twice as long on average, condition "3D-remote". Post-hoc analysis found significant differences

Table 3: Mean values and standard deviation of video analysis parameters

	Variable	FtF	3D_local	2D	3D_remote	Results post-hoc comparisons
1.	Task Completion Time (seconds)	192 (131)	306 (165)	163 (59)	414 (414)	3D_remote > FtF, 3D_remote > 2D
2.	Total turns per minute	5.4 (2.3)	4.2 (0.5)	5.0 (1.7)	4.1 (1.7)	*
3.	Technology and process related turns out of total turns	0.12 (0.1)	0.26 (0.08)	0.12 (0.08)	0.40 (0.18)	3D_remote > FtF, 3D_remote > 2D 3D_local > FtF, 3D_local > 2D
4.	Ratio deictic references to total references	0.98 (0.04)	0.78 (0.17)	0.70 (0.25)	0.65 (0.20)	FtF > 3D_local, FtF > 2D, FtF > 3D_remote

Note: Standard deviation in parentheses, Asterisk = no significant differences

between condition "3D-remote" and "Face-to-Face" ($p<0.01$), and between condition "3D-remote" and "2D" ($p<0.01$)

4.2.2 Turns per Minute

The spoken turns of both participants were counted during the video analysis. The same definition of a turn as in [3]was used following which "a turn consists of a sequence of talk spurts and pauses by a speaker that holds the floor." During video analysis, turns were counted for one person at a time and the number of turns of both participants were then added up to determine the number of total turn. As the absolute number of turns would not be comparable to other conditions due to the different durations of the rounds, the number of total turns was divided by the task completion time. The so gained value of total turns per minute can be considered as a variable that indicates the quality of the communication flow. "Face-to-Face" and "2D" had slightly more turns per minute on average, suggesting a higher communication flow. However, these differences did not reach significance in the test for the main effect.

4.2.3 Turn Content

Besides the frequency of the turns we were also interested if the content of each turn was related either to the collaborative task, or if it was instead related to the use of the technology involved or the collaborative process. For example the content of the statement: "I think this dog doesn't look at all like this guy" is clearly task related, whereas statements like "Did you just move your mouse" or, "I think you should first rotate the dogs so you can see them, and then I will do the same afterwards" fit more in the technology or process related category. By constructing the ratio of all the non-task related turns by the total number of turns an indicator as to what extend the technology got in the way during the collaboration could be obtained. The calculated numbers showed a significant main effect across the four conditions, $F(3,33)=17.7$, $p<0.01$. Post-hoc comparisons revealed that the occurrence of non-task related turns was significantly higher in the condition "3D-local" than in the conditions Face-to-Face ($p=0.01$) and "2D" ($p=0.03$). The occurrence of non-task-related turns was furthermore found to be higher in the condition "3D-remote" than in conditions Face-to-Face ($p<0.01$) and "2D" ($p=0.03$).

4.2.4 Deictic References vs. Descriptive References

Deictic references are expressions that only make sense within a certain context within they are used. For example in the sentence

"I want you to put that over there", the words I, you, that, and over there are all deictic references as their meaning depends entirely on who said them to whom, where they are, what they are talking about etc. The meaning of descriptive references like "the red book" instead of "that book" are independent of the context occur. They are considered a higher communication effort but are used if a context is getting ambiguous and deictic references might easily be misunderstood. In mediated communication deictic references are less frequently used, as it is harder to maintain a shared context with the absence of certain communication cues. Therefore, their occurrence in mediated collaborations can be an indicator for a more or less established common ground.

In all 12 Videos, all references to either dogs or owners were registered during the video analysis and were counted either as deictic like in "that dog", "him"," her", "that guy" or as descriptive like in "the girl with the glasses", "the Labrador", "the third dog from the left". Out of the total number of references, the ratio of deictic references was calculated and compared between all conditions. A significant main effect was found, $F(3,33)=18.2$, $p<0.01$. Further post-hoc analysis showed that the relative occurrence of deictic references out of all registered references was significantly higher in Face-to-Face than in conditions "3D-local" ($p<0.01$), "2D" ($p=0.01$), and "3D-remote" ($p<0.01$).

5. Discussion

The results of our experiment showed some benefits of our 3D-interfaces. They were able to support more spatial cues like gaze awareness than the 2D-interface and that they could produce higher social presence and co-presence scores. However, these benefits came at the cost of a significantly higher mental load that lead to more confusion, more distraction from the task and overall reduced task performance scores in the 3D conditions. Although we predicted a longer task completion time, we did not expect the overall tendency of our measured task performance to be closer to Face-to-Face in the 2D and not in the 3D conditions.

Our initial assumption, that we can improve a collaborative system by adding a new spatial dimension while keeping the other dimensions proved to be oversimplified. Adding spatiality is capable of creating a collaborative context that is closer to Face-to-Face, but at the same time looses the efficiency of a task focused two-dimensional interface. In our experiment, that trade did not pay off as could be seen in particular at the low preference scores of the 3D interfaces.

In general, the lessons that can be learned from this are:

Supporting the right cues: Our spatial interfaces proved to support better gaze awareness than the 2D condition. However, the ability to infer where the other person was looking seemed to be of no significant benefit to solve the task. Instead, it emerged from observing the participants that, once participants were immersed in the shared spatial reference frame, they started to use their hands to gesture and to point in space. Supporting these cues could have probably resulted in a better performance and could have better exploited a spatial context's ability to support the task process. It is therefore important to know beforehand what sort of cue is required by a certain type of task.

Process before context: The higher preference scores for the 2D interface suggests, that people's satisfaction with an interface starts with its usability. If an interface does not allow the user to solve their task fast and easily, then it seems that the way it supports a sense of sitting around the same table seems to be of minor importance. This has to be kept in mind when it comes to compromising task support for context support. For example, if, like in our 3D-remote setup, the collaboration around documents is strictly emulated in a collaborative virtual environment, then there will always be the problem of a distorted view of the documents by the participants, whose avatars are sitting around the table. Relaxing the strict emulation of the real room metaphor in favor of task efficiency for example by displaying an undistorted version of each document in the foreground of each participants application window would therefore improve overall satisfaction, even if it might compromise the experience of immersion of being present in a real room might suffer.

In this sense, new interaction mechanisms have to be thought of for spatial interfaces that are different from what is strictly done in a real Face-to-Face meeting, as long as they can support the task process. Participants for example repeatedly asked if the whole table could be rotated by 180 degrees in order to avoid the need of rotating every single photograph when a whole set of pictures was facing one person who wanted to show them altogether to the other person. Another way of solving this problem in the same non-real world manner could be to implement a button that as long as being held down would allow a person to see through the eyes of the other person, and thus temporarily leave the concept of individual views.

3D interaction for 3D Videoconferencing. The Face-to-Face condition clearly won all categories we investigated in our experiment. That was not only because of the high communication bandwidth of Face-to-Face communication, but also because of the simultaneous, two-handed interaction participants were able to use when sitting around the real table discussing the real photographs. Future systems that want to better exploit the benefits of spatial interfaces should therefore avoid a primitive mouse based interaction concept and should instead try to support tangible, simultaneous, and lightweight manipulation mechanisms that can reduce the mental load and keep up with the highly interactive path of face-to-face-like communication. The fact that more relative deictic references were found in the 3D-local interface with the touch screen input than in the mouse based conditions "2D" and "3D-remote" can be seen as an indicator that a light weight mechanism for example for pointing can have impact on the communication patterns and moves them closer towards Face-to-Face.

Handling navigation. Adding spatiality adds the need for users to navigate in the shared space. This necessarily creates additional mental load compared with the 2D interface. In our experiment, we tried to keep that mental load as small as possible in the "3D-remote" condition by restricting the degrees of freedom to rotation only, and by using a head tracker to control the individual view into the space. However, the high score in confusion, the results of task completion time and the high ratio of non-task related turns show that the mental overhead of the system still was relatively high. In order to further reduce that mental load, a restriction of the rotation into only one degree of freedom, for example only looking down at the table and up to the other persona might have reduced the overhead, while at the same time limiting the feeling of immersion. Again, the right decision on the granted degrees of freedom should depend on a given task.

Quantity of information. In our task, two people had a discussion about one given set of pictures. In this scenario, managing the collaborative process might not be too challenging. However, if it were 6 people that had to discuss 10 different sets of photos at the same time, the confusion score of a user of a 2D interface is likely to be much higher. Although at this point only hypothetical, it seems likely that spatial approaches can resolve confusion if the amount of information does not fit onto one monitor window any more. This case, however, needs to be investigated in a future study.

6. Conclusions & Future Work

We presented the results of a study comparing two videoconferencing interfaces that support spatial cues with a conventional 2D system as well as with a same room Face-to-Face condition. We found various differences between the conditions which suggest that the spatial character of an interface can support a higher gaze awareness sense of social presence, while at the same time compromising a two-dimensional interface's task focus and efficiency. From our results it becomes clear that, in order to fully exploit the benefits of a spatial approach remote collaboration it is necessary to guarantee its usability first.

Therefore our next steps concentrate on the research in improving the interface with respect to its task focus while maintaining the three-dimensionality of the context. From the lessons we learned in this experiment we will draw our particular interest into (a) fast and robust view changes (head movement), (b) support of pointing with the hands, (c) natural object handling (moving, rotating, flipping, etc.), and (d) new interaction metaphors suitable and tailored for virtual environments.

Our general approach, based on the lessons learned in this and related studies will be the provision of appropriate interfaces

regarding process and context. We conclude, that a 3D context deserves appropriate 3D interfaces and can hardly be understood and operated with 2D interfaces. Bringing together approaches for efficient and natural support for the processes as well as the context is a promising way to follow in research.

7. ACKNOWLEDGMENTS

[redacted]

8. REFERENCES

[1] J. Hollan and S. Stornetta, "Beyond being there " in *Proceedings of the SIGCHI conference on Human factors in computing systems* Monterey, California, United States ACM Press, 1992 pp. 119-125

[2] E. Williams, "Experimental Comparisons of Face-to-Face and Mediated Communication: A Review," *Psychological Bulletin*, vol. 84, pp. 963-976, 1977.

[3] A. J. Sellen, "Remote Conversations: The Effects of Mediating Talk With Technology," *Human-Computer Interaction*, vol. 10, pp. 401-444, 1995.

[4] H. Nakanishi, C. Yoshida, T. Nishimura, and T. Ishida, "FreeWalk: A Three-Dimensional Meeting-Place for Communities," in *Community Computing: Collaboration over Global Information Networks*, T. Ishida, Ed.: John Wiley and Sons, 1998, pp. 55-89.

[5] H. Regenbrecht, T. Lum, P. Kohler, C. Ott, M. T. Wagner, W. Wilke, and E. Mueller, "Using Augmented Virtuality for Remote Collaboration," *Presence: Teleoperators and Virtual Environments*, vol. 13, pp. 338-354, 2004.

[6] R. Vertegaal, "The GAZE groupware system: mediating joint attention in multiparty communication and collaboration " in *Proceedings of the SIGCHI conference on Human factors in computing systems: the CHI is the limit* Pittsburgh, Pennsylvania, United States ACM Press, 1999 pp. 294-301

[7] W. W. Gaver, "The affordances of media spaces for collaboration " in *Proceedings of the 1992 ACM conference on Computer-supported cooperative work* Toronto, Ontario, Canada ACM Press, 1992 pp. 17-24

[8] A. J. Sellen, "Speech patterns in video-mediated conversations," presented at Computer Human Interaction, Monterey, CA, 1992.

[9] M. Argyle, *Gaze and mutual gaze*. London: Cambridge University Press, 1976.

[10] A. Kendon, "Some functions of gaze-direction in social interaction," *Acta Psychologica*, vol. 26, pp. 22-63, 1967.

[11] H. H. Clark and S. E. Brennan, "Grounding in Communication.," in *Perspectives on socially shared cognition*, J. M. L. L. Resnick, e S.D. Teasley, Ed. Washington, DC: APA, 1991, pp. 127-149.

[12] R. Raskar, G. Welch, M. Cutts, A. Lake, L. Stesin, and H. Fuchs, "The Office of the Future: A Unified Approach to Image-Based Modeling and Spatially Immersive Displays," presented at SIGGRAPH 98, Orlando, Florida, 1998.

[13] S. J. Gibbs, C. Arapis, and C. J. Breiteneder, "TELEPORT -- Towards immersive copresence.," *ACM Multimedia Systems*, vol. 7, pp. 214-221, 1999.

[14] AliceStreet, Online product description. http://www.alicestreet.com, last accessed 17/03/06

[15] S. Benford and L. Fahlen, "A Spatial Model of Interaction in Large Virtual Environments," presented at European Conference on Computer Supported Cooperative Work (ECSCW'93), Milano, Italy, 1993.

[16] S. Benford, C. Greenhalgh, and D. Lloyd, "Crowded Collaborative Virtual Environments.," presented at CHI'97, Atlanta, Georgia, USA, 1997.

[17] S. D. Scott, K. D. Gant, and R. L. Mandryk, "System Guidelines for Co-located Collaborative Work on a Tabletop Display," presented at ECSCW'03, European Conference Computer-Supported Cooperative Work 2003, Helsinki, 2003.

[18] H. Ishii and B. Ullmer, "Tangible Bits: Towards Seamless Interfaces between People, Bits andAtoms," presented at CHI'97, Atlanta, Georgia, USA, 1997.

[19] R. Kruger, S. Carpendale, S. D. Scott, and S. Greenberg, "Roles of Orientation in Tabletop Collaboration: Comprehension, Coordination and Communication," *Journal of Computer Collaborative Work*, vol. 13, pp. 501-537, 2004.

[20] P. de Greef and W. IJsselsteijn, "Social Presence in the PhotoShare Tele-Application," presented at Presence 2000 - 3rd International Workshop on Presence, Delft, The Netherlands, 2000.

[21] R. Schroeder, "Social interaction in virtual environments: Key issues, common themes, and a framework for research.," in *The social life of avatars: Presence and interaction in shared virtual environments.*, R. Schroeder, Ed. London: Springer, 2002.

[22] M. Lombard and T. Ditton, "At the heart of it all: The concept of presence," *Journal of Computer Communication*, vol. 3, 1997.

[23] J. Short, E. Williams, and B. Christie, *The Social Psychology of Telecommunications*. New York: Wiley, 1976.

[24] J. Hauber, H. Regenbrecht, A. Hills, A. Cockburn, and M. Billinghurst, "Social Presence in Two- and Three-dimensional Videoconferencing," presented at 8th Annual International Workshop on Presence, London, 2005.

[25] A. Hills, J. Hauber, and H. Regenbrecht, "Videos in Space: A study on Presence in Video Mediating Communication Systems.," presented at 15th International Conference on Artificial Reality and Telexistence (ICAT 2005), Christchurch, 2005.

[26] F. Biocca, C. Harms, and J. Gregg, "The Networked Minds Measure of Social Presence: Pilot Test of the Factor Structure and Concurrent Validity," presented at Presence 2001, 4th international workshop, Philadelphia, 2001.

[27] R. L. Daft and R. H. Lengel, "Information richness: a new approach to managerial behavior and organizational design," in *Research in organizational behavior 6*, L. L. Cummings and B. M. Staw, Eds. Homewood, IL: JAI Press, 1984, pp. 191-233.

[28] M. M. Roy and N. J. S. Christenfeld, "Do dogs resemble their owners?," *Psychological Science*, vol. 15, pp. 361-363, 2004.

[29] ConferenceXP, Online product description, *http://www.conferencexp.com/community/default.aspx* last accessed 17/03/2006

[30] TrackIR, Online product description. http://www.naturalpoint.com/trackir/, last accessed 17/03/2006

[31] Inkscape, Online product description. http://www.inkscape.org/, last accessed 17/03/2006

[32] UltraVNC, Online product description. http://www.ultravnc.com/, last accessed 17/03/2006

[33] J. Short, E. Williams, and B. Christie, *The social psychology of telecommunications*. London: John Wiley & Sons, 1976.

Columns on Last Page Should Be Made As Close As Possible to Equal Length