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

Jörg Hauber * Holger Regenbrecht ** Mark Billinghurst * Andy Cockburn *

* University of Canterbury
Christchurch, New Zealand
64-3-364-2349

** University of Otago
Dunedin, New Zealand
64-3-479-8322

{joerg.hauber, mark.billinghurst} @hitlabnz.org,
holger@infoscience.otago.ac.nz, andy@cosc.canterbury.ac.nz

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 of a standard non-spatial 2D interface with two approaches for adding spatial cues to videoconferencing: one based on simulated immersive 3D, the other based on video streams in a physically fixed arrangement around an interactive table. 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 also important differences between the 2D and spatial interfaces. In particular, the spatial interfaces positively influenced social presence and copresence measures in comparison to 2D, but the task measures favored the two-dimensional interface.

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, Remote Collaboration, Social Presence, Photo-ware, Collaborative Virtual Environment

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 trying to develop video-mediated communication (VMC) systems that allow distant colleagues to accomplish tasks with the same if not better efficiency and satisfaction than when collocated [14]. 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 [27, 33], 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 [21, 23, 32]. 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.

2. RELATED WORK

In the context given we first consider related work in the field of VMC in general. Then, we narrow down our focus to spatial approaches to VMC, and finally we give a brief overview of the role of spatiality in collaborative table-top settings.

2.1 Video-Mediated versus Non-Mediated Communication

Traditional 2D video-conferencing systems provide a compressed 2D representation of a 3D space. This constrains 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 [10].

Vertegaal [32] argued that the disparity between audio-visual and face-to-face communication is caused by the absence of several nonverbal channels. This impedes the use of certain non-verbal utterance types that we normally use to coordinate our communication processes (described as "grounding" in [6]). Fussell at al. [9] identify the participants' heads and faces,

CSCW'06, November 4–8, 2006, Banff, Alberta, Canada.
Copyright 2006 ACM 1-59593-249-6/06/0011...$5.00.

participants' bodies and actions, shared task objects, and a shared work context as non-verbal resources for grounding in co-present settings.

Some of the non-verbal channels are missing in conventional video conferencing due to the lack of a common spatial reference frame. 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 [28]. 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 [2]. 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 [17].

If communication channels are missing, speakers automatically compensate for their absence through a more extensive use of supported channels (mostly verbal) in order to adhere to the grounding mechanisms that can be used in non-mediated communication. 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. Same holds for maintaining a common ground when talking about shared objects such as documents or pictures. In face-to-face collaboration the context of the collaboration is normally clear. Within this context it is possible to easily refer to an item with a deictic reference that typically combines an utterance (e.g. "this one" or "that one") with a clarifying gesture. However, if certain communication channels are not supported, the context of the communication can be ambiguous. Then, deictic references need to be replaced by longer and potentially awkward descriptive references (e.g. "the tall guy in the second row from the right").

Necessary workarounds like these 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 VMC

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 [22], or the TelePort [11]. In both cases projectors are used to create a spatially immersive AR display that supports remote collaboration in an 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 [32], "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. This information is then used to orient the virtual persona towards the source of attention. 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"[21],"AliceStreet" [1] or cAR\PE! [23]. In the latter, users can freely navigate their personas through a virtual conferencing room and interact with others and shared documents in a number of different ways. Spatial visual and audio cues are combined in natural ways to aid communication [3]. For instance, 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 [4]. 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 widespread 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 around tables, so it therefore is an obvious option to use a table-top surface as an interaction space, especially for collocated collaboration. Table-top interfaces allow for embodied, media-rich, fast and fluid interaction in collocated collaboration. Scott et al. [26] 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 [16] should be used or some other metaphors have to be developed or adapted. As Krueger et al. [18] 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 collocated 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 spatial collaborative systems, there have been few empirical studies comparing collaboration between such systems and with unmediated face-to-face collaboration. We are interested in the impact of (added) spatial aspects such as individual views and gaze awareness 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 in the future.

We narrow our interest to two specific dimensions: (1) The extent to which persons feel "being together" with a remote person 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 There with others" [25], the "salience of the other in mediated communication" [29], or the "perceptual illusion of non-mediation"[19]. Although being consistent at a general level, these definitions present different concepts and operationalizations of the phenomenon. In our study, we are especially interested in the perceived capability of the communication medium to transmit non-verbal utterance types. Therefore, we subscribe to the definition of social presence given by Short et al. [29], who see social presence as a fixed property of the medium, mainly depending on the richness of non-verbal cues that are supported. Media with a high level of social presence are typically perceived as warm, sensitive, personal, and sociable. Following this conception, social presence can be measured using the semantic differential technique. The reliability of this instrument for comparisons of different videoconferencing interfaces has been found in earlier studies [12, 13].

In addition, we are also interested in copresence as a sense of spatial co-location, or "the feeling that the people with whom one is collaborating are in the same room" [20] which has been identified as a sub-factor of the experience of social presence by [5].

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 flow of conversation including references 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 Experiment 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 or real pictures on a table:

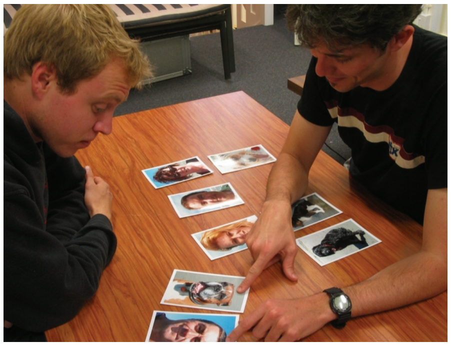

Figure 1. Condition "Face-to-Face".

I) Unmediated face-to-face collaboration around a real table (Figure 1), labeled as "Face-to-Face". Here, the digital pictures are printed onto paper and allow for natural tangible interaction.

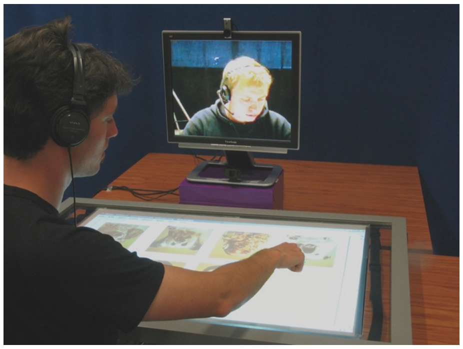

Figure 2. "Spatial-Local" Videoconferencing.

II) Mediated remote collaboration around a shared interactive table (Figure 2), labeled as "Spatial-Local", because spatial cues are supported within a local, real-world reference frame. The digital photos are displayed and pre-arranged on a touch sensitive table surface that allows for interaction with the pictures.

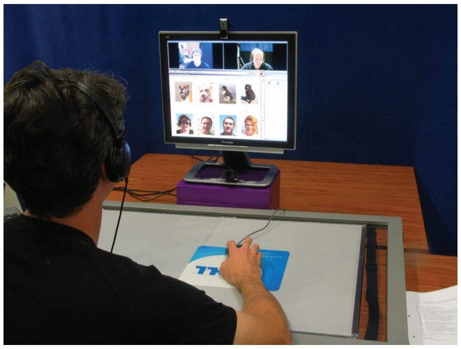

Figure 3. "2D" Videoconferencing.

III) Mediated collaboration through a standard 2D-video conferencing interface (Figure 3), labeled as "2D", as no aspects of a shared three-dimensional reference frame is given. This setup uses a state-of-the-art videoconferencing system involving video streams of both participants displayed on the screen as well as a shared application window which is operated with a standard computer mouse.

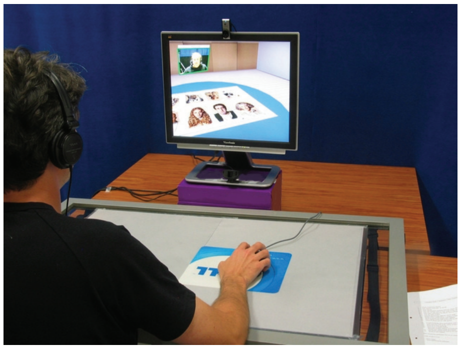

Figure 4. "Spatial-Remote" Videoconferencing.

IV) Mediated collaboration around a virtual table in an immersive desktop collaborative virtual environment (Figure 4), labeled as "Spatial-Remote" as the given spatial reference frame within which spatial cues are supported is a remote space different from the real world. While the interaction with digital photos is done with a standard computer mouse, the representations of the table and of the participants' video streams are shown in the simulated three-dimensional space. A special head tracking device is used to allow for consistent virtual head-movement within the virtual environment.

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	*Spatial-Local*	*2D*	*Spatial-Remote*
Gaze supported	Yes	Yes	No	Yes
Table Interaction	Yes	Yes	No	Yes
Input	Gesture	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 judgemental task was designed with a highly ambiguous content. This follows from Media Richness Theory [8], 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 [24] 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 Experiment Conditions and Apparatus

As mentioned in Section 3.1 the experiment involved four conditions:

1. Condition "Face-to-Face". In this condition, both participants collaboratively examined 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 (5x7 inch, resolution 1024x1280 pixels).

2. Condition "Spatial-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 projected the photo application onto the touch panel. Using a single finger photos could be moved across the panel or rotated by dragging rotation handles of a selected photo. The LCD monitor behind the touch panel showed 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 [7]) 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 saw exactly the same content on the screen, just like 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 "Spatial-Remote*". In this condition, participants met in a virtual 3D room, represented as 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 [23]. The head orientation of the participants was tracked with a 2DOF infrared tracker [30]. Head tracking information was used to control the individual view into the virtual room. That way, person A could e.g. change his/her viewpoint between the table and the persona of person B by moving his/her head up and down. At the same time, the orientation of person A's persona consistently followed the head movements, allowing person B in turn to infer what was in the view field and thus the point of attention of person A. The positions of the virtual characters could not be changed by the participants. Half the photos were flipped in the initial layout, so that half the photos could be seen in the correct 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 participants. For all mediated conditions, two visually and acoustically separated rooms were prepared with identical standard desktop PCs (P4, 2.80 GHz), monitors (LCD, 17'', 1280x1024), headsets (stereo with mono microphone) and webcams (USB, CIF resolution). All computers involved in the setup were connected through a 100 megabit network switch.

The shared photo viewing application was based on the open source graphics editor Inkscape [15]. Shared access to the application was implemented using the desktop sharing software UltraVNC [31]. 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 screen capturing software.

3.6 Procedure

For every one-hour session a group of two subjects was present. Upon arrival the participants were given a sheet with the *Participant Information*, which outlined (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 a *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 round for each condition (FtF, Spatial-Local, 2D, Spatial-Remote). The order of conditions was randomized beforehand following a Latin Square scheme. 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 instructions on 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 spatial videoconferencing 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 three mediated conditions, the subjects wore audio head-sets which were explained and adjusted for best comfort. The head tracking in the Spatial-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 explicitly expressed they would agree with, 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 of correct dog-owner pairs found in the last round was 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 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 chocolate was given to them as a reward.

3.7 Expected Results

We assumed that face-to-face communication will be the richest, most familiar, and most effective mode of collaboration and would thus result in the best scores in our dimensions of interest. However, as the spatial conditions supported some of the cues that are present in face-to-face talk which were not available in the 2D interface, we generally expected that the spatial interfaces would afford a collaborative behaviour that is closer to that of face-to-face. As such, we predicted social presence and copresence to be higher in the spatial interfaces than in the 2D-condition. Furthermore, we expected that the additional cues in the spatial interfaces would have a positive impact on the participants' ability to create common ground that would also show in their communication patterns. We therefore predicted a higher use of deictic references in the spatial interfaces.

In terms of task completion times, we anticipated that the 2D interface would allow fastest task completion, as the photos would not need to be rotated as often as when using the individualised views of the spatial conditions.

4. RESULTS

All participants (except one "cat person") liked the task and quickly became engaged in finding the matching pairs. The most common judgment criteria were whether a dog would be a woman's or a man's dog, if a dog would match the more active or passive lifestyle inferred from the photos of the owners, and matching hair color and facial features between owners and dogs. For the total of 16 dogs presented in each experiment, on average 5.15 correct owners were found. This is slightly more than would be expected in a total random scenario and could indicate a small correlation between dogs and there owners[1].

The teams' strategy of handling the photograph orientation was consistent over all three conditions that involved individual viewpoints. The two main strategies were to either rotate all pictures to be correctly oriented for person A first, and then rotate them all back so person B could have a look; or, to place the photos in the middle of the table and rotate them about 90 degrees

[1] Mentioned here for completeness only. This result was not further investigated as it is not within the focus of this paper.

into a more neutral sideways position where both could examine them sideways at the same time.

Occasionally, in the condition “Spatial-Local”, the participants had difficulty rotating photos using the touch-sensitive table due to problems acquiring the rotation handle.

4.1 Questionnaire Results

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.

According to the procedure described in Section 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.” (negative loading)

“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.”

Subjects 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 were averaged to one single copresence score, summarized for the four conditions in Figure 5.

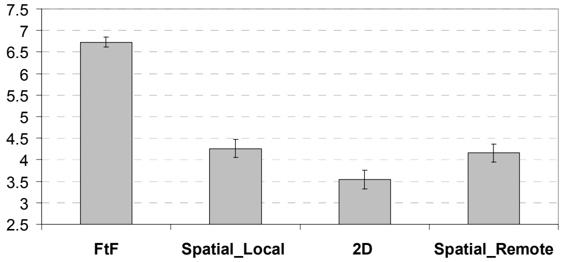

Figure 5. Average score and std. error for copresence.

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 spatial videoconferencing conditions received higher average scores than the 2D condition. Post-hoc analysis showed a significant difference between conditions “Spatial-Local” and “2D” (p=0.04). Furthermore, not surprisingly, subjects felt significantly more copresent in the Face-to-Face condition than in the other conditions.

4.1.2 Social Presence

Social presence was measured with the semantic differential technique like suggested in Short et al. [29]. In our case, in total eight bi-polar pairs were used. 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*, *formal – spontaneous*, *impersonal – personal*, *passive – active*, *unsociable – sociable*, and *closed – open*.

Reliability analysis on these eight items revealed a high Cronbach’s alpha 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 scores in the different conditions are shown in Figure 6. There was a significant main effect, F(3,75)=20.8, p<0.01. Post-hoc comparisons showed that social presence was significantly higher in the Face-to-Face condition than in all the other conditions. However, none of the mediated conditions showed differences in pair-wise comparisons. Although not significant, the average of social presence was rated higher in both spatial conditions than in the 2D condition.

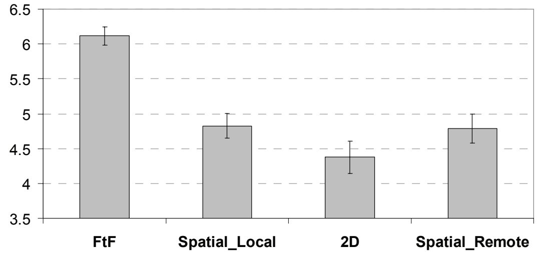

Figure 6. Average score and std. error for social presence.

4.1.3 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 their ranks, a normalized preference score was calculated from 0 to 1, where the rank 4 is normalized to score 0 and a ranking of 1 is normalized to 1. The results are shown in Figure 7. All participants significantly preferred the Face-to-Face over any of the mediated conditions. 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.

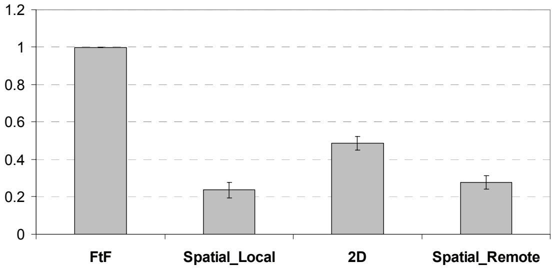

Figure 7. Average and std. error for Preference score.

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	*Face to Face (FtF)*	*Spatial-Local (Sp-loc)*	*2D (2D)*	*Spatial-Remote (Sp-rem)*	*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 > Sp-loc; FtF > Sp-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 > Sp-loc; FtF > 2D; FtF > Sp-rem
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 < Sp-rem
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 > Sp-rem; FtF > 2D; Sp-loc>2D; Sp-rem > 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 < Sp-loc; FtF < Sp-rem; 2D < Sp-loc; 2D < Sp-rem
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 > Sp-loc; FtF > 2D; FtF > Sp-rem
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 > Sp-loc; FtF > 2D; FtF > Sp-rem
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 > Sp-loc; FtF > 2D; FtF > Sp-rem

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

4.1.4 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, 7, and 8, 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 Spatial-Local condition than in the 2D condition (p=0.02), and also significantly higher in the Spatial-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 Spatial-Local condition than in the 2D condition (p=0.05). They also felt more often confused in the Spatial-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 both spatial videoconferencing conditions.

4.2 Video Analysis Results

The video observation analysis was done by the first author. 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. The original audio streams of the two participants were assigned to the left and right audio channel in the final video. Video editing was done with the video editing package Adobe Premiere Professional 1.5.

In these combined videos, 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. The results of the Video analysis are summarized in Table 3.

4.2.1 Task Completion Time

The task completion time was defined from the moment when the participants first saw the photos of the dogs and owners until the moment when they explicitly signaled that they found a 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 "Spatial-Local" and at the end, taking more than twice as long on average than condition 2D, condition "Spatial-Remote". Post-hoc analysis found significant differences between conditions Spatial-Remote and Face-to-Face, and between conditions Spatial-Remote and 2D.

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 [27] 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 was then summed to determine the total turns. 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.

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

Variable	*Face to Face (FtF)*	*Spatial-Local (Sp-loc)*	*2D (2D)*	*Spatial-Remote (Sp-rem)*	*Results post-hoc comparisons*
1. Task completion time (seconds)	192 (131)	306 (165)	163 (59)	414 (206)	Sp-rem > FtF; Sp-rem > 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)	Sp-rem > 2D; Sp-rem > FtF; Sp-loc >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 > Sp-loc; FtF > 2D; FtF > Sp-rem

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

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 "Spatial-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 "Spatial-Remote" than in conditions "Face-to-Face" (p<0.01) and "2D" (p=0.03).

4.2.4 Deictic References vs. Descriptive References

As pointed out in chapter 2, deictic references are less frequently used in mediated communication, as it is harder to maintain a shared context with the absence of certain communication cues. Therefore, their occurrence in mediated collaboration can be seen as 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 "Spatial-Local" (p<0.01), "2D" (p=0.01), and "Spatial-Remote" (p<0.01).

5. DISCUSSION

The results of our experiment showed some benefits of our spatial videoconferencing interfaces. They were able to support more spatial cues like gaze awareness than the 2D interface and they could produce higher social presence and copresence 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 spatial conditions. Although we predicted a longer task completion time for the spatial interfaces, we did not expect the overall tendency of our measured task performance parameters to be closer to Face-to-Face in the 2D and not in the spatial 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 loses 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 spatial interfaces.

At this point it is legitimate to ask whether our results are able to inform about the general value and the affordances of added spatial cues in videoconferencing, or, if they simply reflect the usability of the systems that we used here. While this is a non-trivial question that every advanced interface evaluation has to face, we tried to compensate this bias with the provision of two substantially different spatial conditions that addressed dissimilar spatial reference frames. Based on the general consistency of the patterns that we observed and obtained for both spatial conditions in our study, we think we learned the following general lessons while keeping the implementation details in the back of our minds.

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 which cues are primarily required to solve a certain type of collaborative task.

Process before context: The higher preference score 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 is of minor importance. This has to be kept in mind when it comes to compromising task support for context support. 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.

New interaction for spatial 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 Spatial-Local interface with the touch screen input than in the mouse based conditions “2D” and “Spatial-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 “Spatial-Remote” condition by allowing 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 up to the other persona and down at the table might have reduced the overhead, while on the other hand limiting the feeling of immersion.

Quantity of Information: In our task, two people had a discussion about one given set of pictures. Managing the collaborative process in such a scenario 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 degree of gaze awareness, social presence, and copresence, while at the same time compromising a two-dimensional interface’s task focus and efficiency. Surprisingly, despite the better results for social presence and copresence of the spatial interfaces, participants slightly preferred the two-dimensional interface.

As a consequence of these findings we see it as a promising, yet challenging approach to combine the social context that can be created with a spatial setup with the efficiency and usability of two dimensional interfaces. Therefore, our next steps will concentrate on improving an interface with respect to its task focus while maintaining the spatial aspects 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.

Furthermore, we will also replace the standard monitors we used in this experiment with immersive stereoscopic projection displays to investigate how different display parameters affect the behavior of remote collaborators who meet around a (virtual) table.

7. ACKNOWLEDGMENTS

We would like to thank Daniel Wickeroth, Manuela Waldner, and Sandra Bejarano for their help before and during the experiment. Furthermore, thanks to all the subjects for participating as well as the dog owners who agreed to the photos being taken. In addition, we would like to thank DaimlerChrysler Research and Technology for supporting our work. The German Academic Exchange Service, DAAD, financially supports the first author.

8. REFERENCES

[1] AliceStreet - Online product description, last accessed in June 2006. *http://www.alicestreet.com*

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

[3] Benford, S. and Fahlen, L. A Spatial Model of Interaction in Large Virtual Environments. In *Proceedings of the Third European Conference on Computer Supported Cooperative Work (ECSCW 93)*, (Milano, Italy, Sep. 13-17, 1993), Kluwer Acad., 1993, 109-124.

[4] Benford, S., Greenhalgh, C. and Lloyd, D. Crowded collaborative virtual environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '97)*, (Atlanta, Georgia, USA, Mar. 22-27, 1997), ACM Press, New York, 1997, 59-66.

[5] Biocca, F., Harms, C. and Gregg, J. The Networked Minds Measure of Social Presence: Pilot Test of the Factor Structure and Concurrent Validity. *Presented at PRESENCE 2001, 4th Annual International Workshop on Presence*, (Philadelphia, USA, May 21-23, 2001).

[6] Clark, H.H. and Brennan, S.E. Grounding in Communication. In Resnick, L.B., Levine, J.M. and Teasley, S.D. (Eds.). *Perspectives on socially shared cognition*, APA, Washington, 1991, 127-149.

[7] ConferenceXP - Online product description, last accessed in March 2006. *http://www.conferencexp.com/community/default.aspx*

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

[9] Fussell, S.R., Kraut, R.E. and Siegel, J. Coordination of communication: effects of shared visual context on collaborative work. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW '00)*, (Philadelphia, Pennsylvania, USA, Dec. 2 - 6, 2000), ACM Press, New York, 2000, 21-30.

[10] Gaver, W.W. The affordances of media spaces for collaboration. In *Proceedings of the ACM conference on Computer-Supported Cooperative Work (CSCW '92)*, (Toronto, Ontario, Canada, Nov. 01-04, 1992), ACM Press, New York, 1992, 17-24.

[11] Gibbs, S.J., Arapis, C. and Breiteneder, C.J. TELEPORT -- Towards immersive copresence. *ACM Multimedia Systems*, *7*, 3 (May 1999), 214-221.

[12] Hauber, J., Regenbrecht, H., Hills, A., Cockburn, A. and Billinghurst, M. Social Presence in Two- and Three-dimensional Videoconferencing. In *Proceedings of the 8th Annual International Workshop on Presence*, (London, September 21-23, 2005), 2005.

[13] Hills, A., Hauber, J. and Regenbrecht, H. Videos in Space: A study on Presence in Video Mediating Communication Systems. In *Proceedings of the 15th International Conference on Artificial Reality and Teleexistence (ICAT 2005)*, (Christchurch, Dec. 5 - 8, 2005), 2005.

[14] Hollan, J. and Stornetta, S. Beyond being there. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '92)*, (Monterey, California, USA, May 03-07, 1992), ACM Press, New York, 1992, 119-125.

[15] Inkscape - Online product description, last accessed in March 2006 *http://www.inkscape.org/*

[16] Ishii, H. and Ullmer, B. Tangible bits: towards seamless interfaces between people, bits and atoms. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '97)*, (Atlanta, Georgia, USA, Mar. 22-27, 1997), ACM Press, New York, 1997, 234-241.

[17] Kendon, A. Some functions of gaze-direction in social interaction. *Acta Psychologica*, *32* (1967), 1-25.

[18] Kruger, R., Carpendale, S., Scott, S.D. and Greenberg, S. Roles of Orientation in Tabletop Collaboration: Comprehension, Coordination and Communication. *Comput. Supported Coop. Work*, *13*, 5-6 (Dec. 2004), 501-537.

[19] Lombard, M. and Ditton, T. At the heart of it all: The concept of presence. *J. of Computer-Mediated Communication*, *3*, 2 (Sep. 1997).

[20] Mason, R. *Using Communications Media in Open and Flexible Learning.* Stylus Publishing, LLC, 1994.

[21] Nakanishi, H., Yoshida, C., Nishimura, T. and Ishida, T. FreeWalk: A Three-Dimensional Meeting-Place for Communities. In Ishida, T. (Ed.). *Community Computing - collaboration over global information networks*, Wiley, 1998, 55-89.

[22] Raskar, R., Welch, G., Cutts, M., Lake, A., Stesin, L. and Fuchs, H. The office of the future: a unified approach to image-based modeling and spatially immersive displays. In *Proceedings of the 25th Annual Conference on Computer Graphics and interactive Techniques (SIGGRAPH '98)*, (Orlando, Florida, USA, July 19-24, 1998), ACM Press, New York, 1998, 179-188.

[23] Regenbrecht, H., Lum, T., Kohler, P., Ott, C., Wagner, M.T., Wilke, W. and Mueller, E. Using Augmented Virtuality for Remote Collaboration. *Presence (Camb)*, *13*, 3 (Jun. 2004), 338-354.

[24] Roy, M.M. and Christenfeld, N.J.S. Do dogs resemble their owners? *Psychological Science*, *15*, 5 (May 2004), 361-363.

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

[26] Scott, S.D., Gant, K.D. and Mandryk, R.L. System Guidelines for Co-located Collaborative Work on a Tabletop Display. In *Proceedings of the Eighth European Conference on Computer Supported Cooperative Work (ECSCW2003)*, (Helsinki, Finland, Sep. 14-18, 2003), Kluwer Acad., 2003, 159-178.

[27] Sellen, A.J. Remote Conversations: The Effects of Mediating Talk With Technology. *Human-Computer Interaction*, *10*, 4 (1995), 401-444.

[28] Sellen, A.J. Speech patterns in video-mediated conversations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '92)*, (Monterey, California, USA, Jun. 03-07, 1992), ACM Press, New York, 1992, 49-59.

[29] Short, J., Williams, E. and Christie, B. *The social psychology of telecommunications*. Wiley, London, 1976.

[30] TrackIR - Online product description, last accessed in March 2006. *http://www.naturalpoint.com/trackir/*

[31] UltraVNC - Online product description, last accessed in March 2006. *http://www.ultravnc.com/*

[32] Vertegaal, R. The GAZE groupware system: mediating joint attention in multiparty communication and collaboration In *Proceedings of the Proceedings of the SIGCHI conference on Human factors in computing systems (CHI '99)*, (Pittsburgh, Pennsylvania, USA, May 15-20, 1999), ACM Press, New York, 1999, 294-301

[33] Williams, E. Experimental comparisons of face-to-face and mediated communication: A review. *Psychol Bull*, *84*, 5 (1977), 963-976.