Interacting with Embodied Agents that can see: How Vision-Enabled Agents can assist in Spatial Tasks

Arjan Geven, Johann Schrammel and Manfred Tscheligi
CURE - Center for Usability Research and Engineering
Hauffgasse 3-5, 1110 Vienna, Austria
{geven;schrammel;tscheligi}@cure.at

ABSTRACT
In this paper, we describe user experiences with a system equipped with cognitive vision that interacts with the user in the context of personal assistance in the office. A cognitive vision computer can see the user and user responses and react to situations that happen in the environment, crossing the boundary between the virtual and the physical world. How should such a seeing computer interact with its users? Three different interface styles – a traditional GUI, a cartoon-like embodied agent and a realistic embodied agent – are tested in two tasks where users are actively observed by a (simulated) cognitive vision system. The system assists them in problem solving. Both the non-embodied and the embodied interaction styles offer the user certain advantages and the pros and cons based on the experiment results are discussed in terms of performance, intelligence, trust, comfort, and social presence.

Author Keywords
Cognitive vision, embodied agent, personal assistant, intelligent systems, user experience, personality type.

ACM Classification Keywords
H.5.2. Information Interfaces and Presentation: User Interfaces – Interaction Styles.

INTRODUCTION
Cognitive vision equips computer systems with cameras and allows them to make sense out of what they see, enabling computer systems to acquire knowledge about objects and activities in the environment and use this knowledge to improve the interaction and better serve users needs. This technology is still under heavy development and can be seen as the next step in computer development [6]. It connects the real physical world with the virtual computational world and allows for systems that can detect, locate, recognize and understand objects and situations in the real world [27]. Additionally, a cognitive vision system can show purposive goal-directed behaviour, can adapt to unforeseen changes, and can anticipate the occurrence of objects and events [9]. The introduction of systems that can understand their environment requires a paradigm shift in the way we interact with a system. As computers acquire more human capabilities, human-machine-interactions can more and more approach human to human interaction instead of the more traditional way of interaction.

A reasonable approach to realize a more human form of interaction is by means of an agent; a program that acts as the intelligent 'personality' of the computer system [7]. An intelligent search agent e.g. can search and recommend interesting articles based on users' previous searches. An intelligent agent that is not embodied has the form of an algorithm and interacts using a traditional GUI. Embodied agents do have some kind of graphical representation and have a more or less human look and feel. The combination of spoken language with an embodied agent introduces human-like interaction capabilities to the system. Using a human-like representation sets expectations for human-like behaviour and responses. Users have a mental model of how to interact with other people and another model of how to interact with computers. These models converge as the interaction becomes more similar.

What happens when an agent looks human? People are known to attribute emotions and feelings to computers and interact socially with computers already when it does not look human. Seminal results were found by Nass and colleagues [20, 23], who found that users apply social norms to their interaction with computers and that users attribute gender stereotypes to computers based on the voice that is used. It was also found that an agent can really build and maintain a relationship with a user, as indicated by those users [10]. The advantage of an embodied agent is that it gives a sense of identity and personalization to an otherwise abstract system and enhances the user experience to be more meaningful and effective. As humans attribute emotions to the agents, agents are intentionally designed to show emotions or react to emotions, which gave rise to the

NordiCHI 2006: Changing Roles, 14-18 October 2006, Oslo, Norway

entire field of affective computing and affective agents [21]. Such agents can e.g. enhance students' perception of a learning experience and increase learning engagement [17]. Users talk more to an embodied interface [5] and animated presentation agents are rated as entertaining and as more helpful than a system without such an agent [1]. On the other hand, users are also more aroused (less confident, less relaxed) when the interface they are working with has a face and users like to present themselves in more positive light to the face than when the face is not seen [26].

The amount of realism that is displayed by agents can alter significantly the perception and reactions of users on such an agent. Mori theorized about a principle he called the "Uncanny Valley" of realism [18, 19], stating that on a scale from very unlike human to very human, there is a certain area very close to realism that is perceived as creepy and scary by users: the situation is "too realistic" but still has certain unrealistic features. Findings from [22] seem to confirm this theory in the area of agents, who saw that when the interface is too realistic, users describe such 'real' agents as being scary, whereas more abstract characters are seen as more friendly and pleasant (but also less interesting or even boring) [22].

The combination of cognitive vision with an embodied conversational agent allows us to create a perceptive animated agent, an agent that sees and can be seen, and can proactively and reactively assist the user with his tasks and adds a social dimension to the computer. Conversational systems that always wait for the user to initiate interaction present users an inconsistent personality, especially when after the user initiated the conversation, the agent controls all further interaction [10]. The question remains how such a system should communicate with its users. On the one hand, there is the more human capability of the system to observe and react, which suggests a more human interface. On the other hand, there are also negative sides to such a more human interface. Embodied agents can give users the impression that the system will act rationally, similar to human beings and will be able to take responsibility for its actions, which can move the feeling of responsibility away from the user and towards the computer [15]. Although the topic of embodied conversational agents is an established area, the application to a cognitive vision system is new and spatial interaction (in the context of personal assistance) remains to be explored.

In our present research, we study the merits of this form of personal and spatial interaction. Our basic question is to find out which presentation and interaction style can be used for interaction with intelligent vision systems, in the context of personal assistance. The function of the cognitive vision system in this context is to assist the user in performing spatial tasks in an office environment, where interaction can take place off-screen. The system can help the user remember certain things or give advice on how to solve a problem. The system observes what the user is doing in the office, and helps by remembering the location of documents or utensils, by interpreting user actions and support problem solving or spatial tasks. We therefore chose two tasks that match these potential application scenarios of a cognitive vision system to see how users react on different kinds of agents that represent such a system.

RESEARCH QUESTIONS

The goal of the study was to get more information on the optimal interaction style for cognitive-vision-enabled computers, and to gain a deeper understanding of how users interact with such a cognitive vision system. We especially wanted to find an answer to the following questions:

1) Does the efficiency (task completion time) depend on the style of the interface? Previous research generally suggests that efficiency is not or only very marginally affected by the chosen style [e.g. 28]. Although we don't assume large differences between conditions, we are interested in whether or not such an effect occurs. Related to this objective measure of efficiency, there's also the user's idea of which system supports the user the most.

2) Does the interface style influence the interpreted intelligence and trust of the system? If the embodied agent is perceived to have a higher degree of intelligence than the non-embodied agent, this can cause problems when the system might occasionally make mistakes. King and Ohya [13] found that more humanoid forms can lead to higher interpreted intelligence, which also resulted in more trust in the system, also when this trust was not justified. Related to the question of intelligence, is the question of trust: does the interface style affect the user's trust and reliance on the system? Trust is ideally based on a careful assessment of the successes and failures that are made by a system followed by a decision on how much to trust a system. On the other hand, humans already make decisions about trust based on the first impression of a system. What happens after they got used to the system? Does an anthropomorphic representation of a system seem more trustworthy than a non-embodied representation of the same system? In a relatively small period of time, users build up an idea of a system. Additionally, it is interesting to see what happens when the system does make a mistake and whether this influences the trust in the system. Do users lower their trust in the system? Or do stop relying on the results on altogether?

3) The third question we want to investigate is whether users appreciate subtle hints from the system, like movements of the eyes and head that guide them. The embodied agents are able to look at the user who is sitting in front of the screen, but also turn away their heads to look at something else that is happening in the office. The question is: do users recognize that the system follows them with its 'eyes' and can they meaningfully interpret gaze changes and do they appreciate it?

Figure 1: The three interaction-styles; in condition A, a map was shown, in condition B a cartoon-like agent, and in condition C a realistic agent.

4) Does the interface style influence the experienced social presence of the system? Social presence can be described as the degree of salience of an interaction partner in a mediated communication and the consequent salience of their interpersonal interactions [cf. 25, p. 65], or the sense of awareness of the presence of an interaction partner. With higher social presence, users are better able to perceive the interaction partner and know more about their qualities and inner states. Therefore, social presence is one of the factors that make interaction 'real'.

We are also interested to find out whether the preferences for an interaction style and feelings of presence are influenced by the user's personality. Literature shows that social presence is influenced by whether a user's personality is more extraverted or more introverted [16]. These character traits might very well influence the general assessment of the three different interface styles.

EXPERIMENT

The experiment we performed was designed to find out how users interact with a personal (embodied) assistant that can see what the user is doing. The cognitive vision system that observes and interprets user actions was simulated by a Wizard of Oz [8]. This allowed us to see how users interact with this new technology before the technology is developed far enough to be used in real applications. None of the participants recognized that they were not really interacting with a computer system.

Three different interface styles were introduced to the user: an agent without embodiment, a cartoon-like embodied agent and a realistic embodied agent. These agents helped the users to complete two tasks in an office environment. The three interaction styles are represented in Figure 1.

The first of the two tasks that the participants had to perform was to find back items that had been hidden in the office: the hide-and-seek scenario. In this task, the system knew where the items were and gave this information to participants; the participants then had to find each object as quickly as possible. Each user worked with each of the three interface styles and answered questions after task completion about the help they received.

The second task consisted of a complex cognitive task in which participants had to solve a 3D puzzle. The vision system could be asked for advice and it also actively gave hints during the task, e.g. "piece 2 is not in the right position". Each user had to solve two puzzles: one with the cartoon-like embodied agent and one with the realistic embodied agent.

In the first task, the three interaction styles can convey equal amounts information regarding the location of the hidden object. The GUI shows the location in 2D (without altitude information), the embodied agents show the direction of the object (without distance information). In the second task, the situation is different. Using gaze, is it impossible to provide information about which piece of the cube is meant, which would have led to an unfair advantage for the GUI-style interaction., which is why the non-embodied version was not used in this task. Both tasks were performed largely "off-screen", in the whole office in task 1 and on the desk in task 2, during which the cognitive vision system interprets the users' actions.

Method

Participants

The participants of the experiment were 12 people from Austria, with average age 26.4, the youngest participant being 21, the oldest participant 37. Participants received a monetary compensation of € 36 for their time and cooperation. A test took between 45 and 75 minutes, depending on the speed with which the participants solved the puzzles.

Materials

All user tests took place at the local labs, where video and

Figure 2a and b: Test shot made from a third camera behind the participant, and a screen shot of the computer screen

audio equipment was available for the necessary recordings. The lab room where the test took place was equipped with four video-cameras and a microphone, which allowed the operator (the "Wizard of Oz") to observe everything from a separate control room, and create a realistic atmosphere where the user had the idea that the system was fully operational and that he or she was being observed by a computer system (Figure 2). The cartoon-like embodied agent was generated with the aid of a Logitech Quickcam with video effect enabled. The realistic agent was generated with software from Haptek, to display an animated "talking head" (see also Figure 1). To see whether these preferences are related to users' personalities they were asked to complete a translated self-report personality type test after the session [12].

Procedure

Participants first received a briefing giving a short explanation of how the system works and that it might make some mistakes, but that they should not be bothered too much by it, as well as ethical information. After this introduction, the actual test begun.

The goal of the first task was to test user responses in a setting where the system had knowledge about the location of certain objects throughout the office (e.g. sticky tape). The participants were told that objects had been hidden by a previous participant under the careful eye of the cognitive vision system, which remembered the position of the objects. The system then told the user where to look for the specific item (e.g. "in the drawer on your right"). Figure 2 shows a still of how the screen looked in one of the conditions. In each condition, the computer-voice said where the target object could be found. In the first condition (A), the voice was either male or female; in the second condition (B), the voice was always male; in the third condition (C) the voice was always female. In condition A, a graphical representation of the office was shown - a 2D-map showing the office from the top. The position of the target object the corresponding place on the map was highlighted. Condition B showed a cartoon-like representation of a male, which looked in the direction where the item could be found. Condition C showed a realistic representation of a female, which also looked in the direction of the item. In each of the conditions, a photo of the object was shown in the bottom half of the screen.

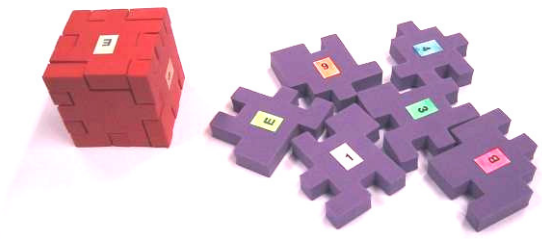

Figure 3: The two happy cubes (red and purple versions) that were used in the second task.

For the voice engine, we had two electronic voices at our disposal that converted written text to speech, one female and one male voice. The female (realistic) agent was logically combined with the female voice and the male (cartoon-like) agent with the male voice. To make sure no ordering effects occurred, the voice in the map-condition was female for six of the participants and male for the other six.

In addition, the sequence of the conditions was varied between participants: some participants started with the map-style interface, some with the cartoon-like agent and some with the realistic agent, to avoid ordering effects.

In each condition, each participant was asked to locate six objects as fast as possible. After a participant found all objects in a condition, he or she was asked to draw a line which length represented the *support, intelligence, trust* and *support* of the system and their feeling of *comfort* during the interaction with the system. These items were formulated as follows: "please draw a line that represents:

- Your level of trust in the system giving you correct information
  0 = absolutely no trust, 100 = complete trust
- Your estimate of the intelligence of the system
  0 = no intelligence at all, 100 = human-like intelligence
- The level of support the system gives you in the task
  0 = absolutely no support, 100 = very good support
- How pleasant you felt during the interaction with this system?
  0 = very unpleasant, 100 = very pleasant

A reference line was shown with the values 0 and 100.

Participants were also asked to rate the social presence of the interaction style by filling out the eleven semantic differential items of the IPO Social Presence Questionnaire, [11], resulting in a rating of 1 to 7 on *social presence* (translated to German for this experiment).

After all trials in the three conditions were completed, the participant was asked to complete a short questionnaire comparing the three conditions on subjective preference. Then, the participant received a briefing on the second task.

In the second task, participants were asked to assemble a 3d-structure puzzle, a so-called happy cube, see Figure 3. The puzzle consisted of six separate pieces, which, if fitted together in the right way, formed a cube. There was only one way to solve the puzzle. The participants were asked to solve the puzzle as fast as possible. Participants always used the purple cube first and the red cube after solving the purple cube (half the participants started with the cartoon-like agent, the other half with the realistic agent). According to the vendor's difficulty ratings, the difficulty levels of the two cubes were close to each other. Informal pre-testing with six users confirmed this. In the test,

participants were seated in front of the computer and everything was observed by the system (controlled by the Wizard of Oz); the system gave hints on solving the puzzle. Each piece of the puzzle was numbered and colour-coded for ease of recognition.

The hints that were given were either reactive or proactive in nature. Reactive hints were given when the participant held their finger over a piece of paper on the table that said "Hint". This was then observed by the Wizard, who gave a hint through the embodied agent. Proactive hints were given without the need for user action, but comprised the same kind of hints. A proactive hint would be given after the user had not initiated a reactive hint for more than 45 seconds. After this time, the system first asked whether it could be of any help to the user, and if nothing happened, it would give a hint from itself after a while. Hints that did not need a vision system were of the type "piece A and B border on each other" or "side A belongs on the inside of the cube". Other hints that did require a vision system were of the type "piece A is not on the right position" or "piece A is in the right position, but is not turned in the right direction". There was no difference in content between proactive and reactive hints.

After solving the each cube, the participant was asked to complete a similar questionnaire as during the first task to measure *support, intelligence, trust, feeling,* and *social presence*. After solving both puzzles, participants also were asked to complete a short questionnaire comparing the two interface styles.

After these questions an open interview was held to find out more precisely about how the participant felt during the interactions, which system he preferred and why. Finally, the participant was asked to complete the personality questionnaire, which gave a rating on four different personality scores.

RESULTS

Efficiency and Support

The time required to complete the assignment was measured in both tasks. Finding the objects in the office generally went very fast with 12.8 seconds on average. As expected this time did not vary significantly between conditions.

Aside from this objective measure, users were also asked to rate how much they thought that the system supported them in finishing the task. It turns out that Condition A actually scores higher than the other two conditions ($F(2,20) = 3.9$, $p<.05$): users prefer the map-like interaction style over the two styles with embodied agents in terms of support.

In the second task, finding the solution to the puzzle took the participants on average 8:01 minutes in condition B and 6:12 minutes in condition C; condition A was not used in this task. These times also did not vary significantly (due to the very large variability within each condition, ranging from 2:04 minutes to 12:40 minutes). The ratings for support are close to each other, and the differences are not significant. What can be seen though is that the support ratings in task 1 are much higher than the ratings in task 2.

Intelligence, Trust, and Comfort

Participants were asked questions with respect to the intelligence of the system, their trust in the system, and how comfortable they felt during interaction with the system on a scale from 0 to 100. The results for the first task did not directly match our expectations. We found that neither intelligence ($F(2,20) = 0,84$, $p = .41$), nor trust ($F(2,20) = 2.5$, $p = .11$) or comfort ($F(2,20) = 1.0$, $p = .38$) ratings varied significantly between the three conditions, although we expected different ratings for the different systems. This means that the three interface styles were rated rather similarly on the factors intelligence and trust, and that users' feelings about the interaction might be more affected by the general test setup and being observed by cameras than by the difference between the three conditions. What can be noted though is that scores on system trust and system support are generally very high in the first task (see Table 1). Users rated their trust in the three styles very high, as high as 91 out of 100 on average in the case of interface-style C. The trust ratings are even more interesting when they are compared to the reliability of the system. As said before, the system was not completely error-free and made mistakes with each user. In the second task, users rated their trust in the correctness of the system a bit lower (Table 1).

Social Presence and Personality

Social presence, however, varies significantly between the three conditions in the first task, with $F(2,20) = 6,12$, $p < .01$. This indicates that the participants felt more connected with the realistic embodied agent than with the interaction style that was not embodied. The ratings for social presence for the three different interaction styles are graphically depicted in Figure 4a and 4b. In the second task, similar results were found, with $F(1,10) = 4,98$, $p = 0.05$, which implies that interaction with the cartoon-like embodied

	Support		Trust		Intelligence		Comfort	
	Task 1	**Task 2**	**Task 1**	**Task 2**	**Task 1**	**Task 2**	**Task 1**	**Task 2**
A	89.9	-	76.8	-	52.0	-	63.9	-
B	71.5	51.0	82.1	64.9	53.5	55.8	58.8	43.2
C	83.8	54.3	91.2	72.7	56.3	58.1	67.2	49.2

Table 1: Average ratings (0-100) for Support, Trust, Intelligence and Comfort

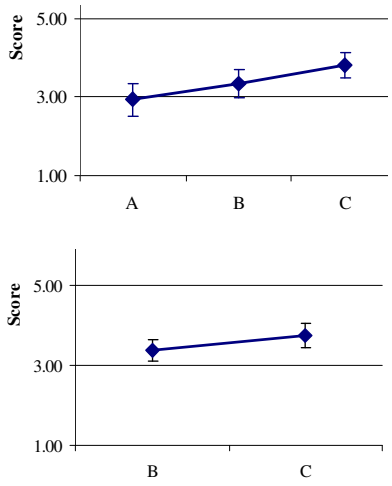

Figure 4a and 4b: Average social presence scores for all participants for tasks 1 task 2 (on a scale from 1 to 7).

agent is not perceived to be as real as interaction with the realistic embodied agent.

We saw that people with a more introverted personality generally gave lower scores in general in all three conditions than the more extraverted people, which was not unexpected [10]. We were more interested in possible interaction effects. We did not find any significant effect with the four questions about support, trust, intelligence and comfort, but what we did find was *an interaction effect between social presence and personality type* in the first task, with $F(2,20) = 4{,}36$, $p < .05$, indicating that extraverted individuals react stronger on an interaction style with a face than more introverted individuals (see Figure 5). In the second task, we found similar results, with $F(1,10) = 6{,}92$, $p < .05$, also indicating that extraverts react stronger on the realistic interface than on the cartoon-like interface (Figure 6).

Preferences

In addition to the questions that were asked after every trial, users were also asked to complete a questionnaire that contained six comparison questions. These questions were about which system was the most intelligent, trusted and supportive, as well as the friendliest, and which system made the fewest mistakes. These questions were 'forced choice' questions. As can be seen in the figures 7 (task 1) and 8 (task 2), condition A scores is chosen by more than 50% of the participants in the first five questions (prefer to work with this system, is the most supportive, is the most trusted, makes fewest mistakes, is the most intelligent system), but scores lower than 50% with the last question: "which system is the friendliest system?". Here can be seen that most users prefer condition C, the realistic embodied agent.

In the second task, condition A was not tested (as it would have lead to very unequal comparisons). Instead, only conditions B and C were compared, the two embodied agents. As can be seen in the figure, condition C scores higher on all six questions compared to condition B: users decidedly prefer the more realistic female agent over the cartoon-like male agent.

Help Requests

During the second task, the system gave the users hints about how they could solve the puzzle. This task took place off-screen, which meant that the users did not have to look continuously at the agent. Instead, they could direct their attention to the cube, and receive the hints mainly over audio.

The differences in problem-solving strategies were quite large between the participants. Some participants continuously asked for hints and received as many as 28 hints before solving the cube, where another participant solved the cube with as few as 3 hints from the system. Here has to be noted though that the user who got so many hints also needed much more time to solve the puzzle than the user with the few hints (who might have made a few lucky guesses).

Both proactive and reactive hints were given to the user. Additionally, when hints were not asked for, the system gave reminders of its own ability to help. After the system

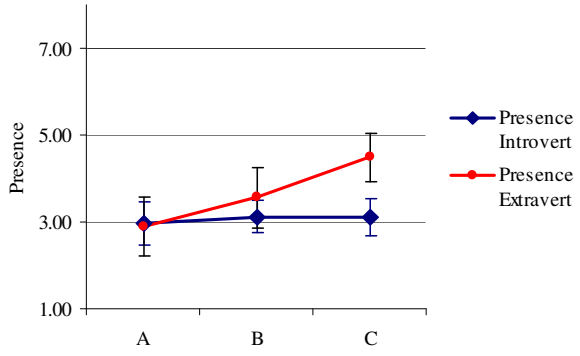

Figure 5: Social Presence Scores in task 1, for introverted & extraverted participants (on a scale from 1 to 7).

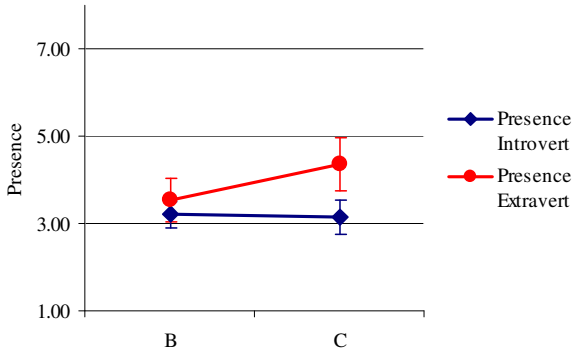

Figure 6: Social Presence Scores in task 2, for introverted and extraverted participants (on a scale from 1 to 7).

had asked "May I help you?", the participant usually pressed the "Hint-button" very fast and a user-initiated hint would be given, indicating either that the participants 'forgot' about the system and were reminded of its existence and found the reminder useful, or that the reminder was interrupting their thought process and they then asked for a hint because the interruption broke their concentration. Both explanations were mentioned by participants.

Participants used both proactive and reactive hints in their problem-solving strategy. They did use them in different ways. Hints were ignored when the participant could not make sense out of it in proactive situations, whereas in reactive situations a new hint or a repetition was asked. Noticeably, not all users wanted to receive help from the system because they liked to solve the puzzle by themselves. Most participants accepted the offer to help though and asked for a hint after the suggestion.

Open Interviews

The open interview that followed after the tasks gave us more insight in the answers that were given in the questionnaire. The open interview was held after both tasks were completed, and comprised all three tested interface-styles.

Users generally saw the most utility in the system that presented them with a clear overview of the office when they were trying to find something: although they appreciated that face-representation, eleven of the twelve participants found that the map gave them more support in retrieving the objects than the embodied agents. In the second task, where the map was not available, a very clear preference was shown for the more realistic agent. It is interesting to note that in this task, the participants were working off-screen, representing a realistic cognitive vision scenario, where they hardly looked at the screen, but still had a clear preference.

In the first task, the gaze of the eyes of the embodied agents pointed in the direction of the location of the hidden objects as they spoke, but two participants did not notice this subtle movement. In the cartoon-like condition, this movement was very subtle, and therefore less noticed than in the more realistic condition (only two participants indicated that they saw the gaze change in the direction of the objects). Some participants mentioned that the gaze of the realistic agent was even too strong and too clear, and said it did not feel natural, it felt arrogant, or like the agent was giving orders instead of reminders.

The participants mentioned many reasons why they preferred the map-style interface over the two agents: mentioned were that the map was unobtrusive, helpful, clear, practical, precise, and that it helped you when you did not exactly understand the voice, which implies that agents are not preferred in all kinds of tasks, or not as the only representation a cognitive vision should have.

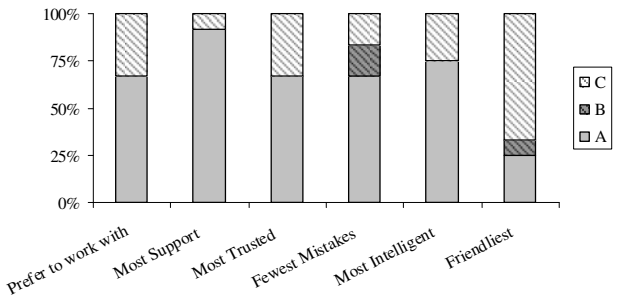

Figure 7: Results of the comparison questions in Task 1.

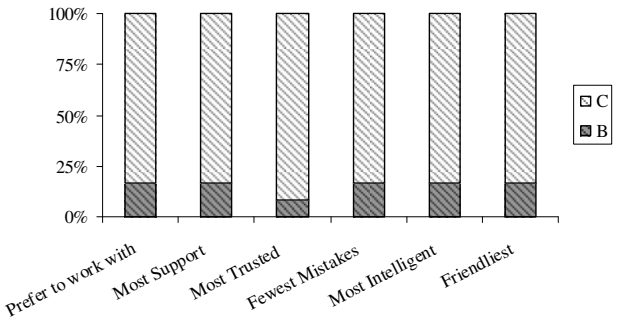

Figure 8: Results of the comparison questions in Task 2.

More than one user mentioned that the voice of the system was a point for improvement; the male voice was a bit harder to understand for many than the female voice, and both voices made some pronunciation mistakes, which makes it difficult for users to understand what the system wanted to communicate to them. Especially in the office environment, where tasks are performed away from the computer screen, users do not always look at the face that is displayed on the screen making interpretation a bit more difficult. One user suggested that the voice also reduced the credibility of the system and would prefer a system that would not speak but just display the text. Another participant mentioned that the voices of both agents were rather inanimate, as it did not reveal any emotion of the agent (because of the computer voice used). This decreased the realism, credibility and joy of working with an otherwise realistic agent.

As observed before, participants placed a lot of trust in the system. In the interview however, one participant did mention not wanting to rely completely on the system, because it could always make some mistakes and listened critically to the hints from the system. The other participants did not mention any of this doubt. One participant mentioned that she felt a bit uncomfortable during the whole experiment, because of the feeling of being watched all the time. For her, this would be a fundamental reason not to use a cognitive vision system because she couldn't get rid of that uncomfortable feeling. The other participants did not say anything about the four cameras that were positioned quite obviously in the lab and

refocused, zoomed and moved around in the lab throughout the session.

DISCUSSIONS

User Performance with the Cognitive Vision System

Based on our experiment, what can we say about the three different styles of interaction? We posed a number of questions in the beginning regarding the way users could interact with a cognitive vision system. As expected, we did not find a difference in terms of efficiency (question 1): each system guided the user approximately equally and we should at other pros and cons of the three interaction styles for the system. We also asked the participants which system they thought best supported their performance, and interestingly, all but one of them pointed to the GUI-style interface (without embodiment). This style of interface gave the users the most precise information about the location of the object, supporting the particular task the best. In this condition, very precise location information was given, more precise than could be given in the other conditions. This system might not be the best system to create a personal atmosphere, but it is efficient and to the point, and that is a more useful feature than the personality of the system.

In terms of intelligence, trust, and comfort (question 2), our participants did not see one system better than the others, which is reassuring in the sense that mere embodiment or personification does not automatically lead to a higher trust in or a higher interpreted intelligence of the system. Trust in general was very high in the three conditions of the first task (with average scores of up to 90 out of 100). This means that users do have confidence in that the system will give them the desired results and attribute certain trustworthiness to a system that might not be completely justified. In the second task, trust ratings were a bit lower. This can be explained because feedback about errors was much more direct in the first task compared to the second task (not finding a hidden object vs. solving the cube in an other way than the system had suggested), as well as because of the deliberate errors the system made in the second task.

The system did not always offer constructive help in both tasks (question 3). In the first task, it would look one time in the wrong direction or show a map with another location than what it said. When users recognized this wrong behaviour, they seemed to be shortly confused but then were still able to locate the item based on the audible information they heard. In task 2, the system would occasionally make a wrong observation and deduce a hint that was plainly wrong, such as “piece A is in the right position” when it actually was not. Then, after the user moved that piece anyway, ignoring the system, it could say again “piece A is in the right position”. This was done on purpose to see how users responded to a system that was very obviously not always correct (of which participants were warned in the beginning with the words that this new technology is not one hundred per cent fool-proof yet). Interestingly, despite these errors, users continued to rely on the hints that were given by the system and moving pieces as suggested. This is a danger in general for using cognitive system to assist users in problem-solving, where over-reliance on a system that is known to make mistakes can not only result in suboptimal performance (in terms of task time), it can also lead to errors that would not have occurred if the system would not have been used. This is in agreement with the point [15] made that a responsibility shift away from the user might be seen.

Presence and Realism

We found differences in the ratings users gave to the questions about the social presence of the three systems (question 4). The most embodied agent was seen as more socially present, more ‘real’ than the non-embodied agent. The interaction style *is* more human and personal to them, giving a personal touch to the system. In the case of cognitive vision systems, such a personal touch is not unimportant, as the application scenario of a personal assistant would mean that a personal ‘coach’ would always watch over your shoulder and accompany you in your work for years. When such a system is attributed human features, possibilities for emotional bonding are created. This can be positive in the sense that it increases the motivation to learn or work with such a system and can create a better atmosphere in the office. On the other hand, when a system is perceived as a social being and emotions are attributed to the system, this might be reason for other issues, for example the feeling that somebody is watching you, or a negative relation to an agent, similar to what happened to the embodied agents in older versions of Microsoft Office.

The questionnaires tell us that the great majority of the participants did prefer the most realistic agent, scoring higher on all questions comparing the two agents with each other, which is surprising in comparison to Mori’s uncanny-valley-theory and the idea that a too high degree of realism becomes creepy. This might be explained by the fact that our most realistic interface was still not perceived as “too realistic”. It could also mean that users are used to better graphical computer performances since e.g. [22] and that the perception of “realism” shifted or even partly disappeared.

The best scores in terms of performance were given to the interface without embodiment. The realistic agent was chosen as the second-best alternative: she gave more sense of direction with her gaze than the cartoon-like agent, even if it was a bit “overdone” according to the participants it was at least clear in which direction she was looking. The best results would be obtained when the two interaction-styles would be combined into a display showing both a (realistic) embodied agent and a map that shows more precise information about the location of objects. This would give the user the best of both worlds: a system giving precise location information through the map, and

providing an interface that provides more human-like interaction and increases the feeling of social presence.

The experiment showed that social presence scores were influenced by type of personality: where the social presence scores from introverts stayed the same for the three styles of interaction, extravert ratings showed a significant increase in social presence experience for the embodied agents. In this respect, the extraverts were solely responsible for the main effect in social presence. Extraverts generally prefer socially engaging activities than introverts and it is therefore logical to find a difference between these personality types. The relation between social presence and personality type is not new, [15] found that an extraverted computer voice can lead to higher social presence ratings from users than an introverted computer voice. The inverse of this relation, as we investigated, has not been reported before to our knowledge, although [4] mentions a difference in user trust for extroverts but not for introverts in a similar situation. [24] suggest that individuals who are introverted are more inclined to experience presence, but their findings were not significant.

In a cognitively very demanding task, like the second task we used, the participant hardly looks at the screen and only the quality of the voice was seen as much more important than the kind of agent that was used. One participant did mention that he would have liked to receive some more visual information on how two pieces of the puzzle fit together, in which case he would have looked more often at the screen. In the first task, participants did look at the screen, but preferred the interface that gave them the most information over the interface with the most personal identity.

CONCLUSIONS

We started this paper with the description of cognitive vision systems as the next step in the computer world. A cognitive vision system acquires more human capabilities. In the context of a cognitive vision system, interaction changes from traditional interaction to a more elaborate interaction, as the system "invades" users' space and can interact in the real world. The concepts surrounding vision-enabled computing are very different from traditional concepts of interaction, as the interaction moves from on-screen-interaction to off-screen interaction and becomes more personal. In this exploratory study, we presented users with such a partly off-screen interaction with three kinds of interfaces to see how such an interaction can take place. Do users prefer a more human interface for a more human-like form of interaction, or do they prefer the traditional GUI? What are the consequences of using embodied interfaces in a cognitive vision setting?

In the experiment, the GUI style interface provided a graphical overview and thus could give a bit more information about the office setting than the embodied agent: users had fewer problems interpreting where to find hidden objects with the assistance of this system, but the embodied agents scored better on social presence, giving the users a more 'real' interaction. A GUI-interface has more descriptive power, being able to show figures, graphs, text, etc., whereas a talking character only has the power of voice and some rudimentary gestures by changing the direction in which the eyes look, possibly extended by adding body gestures to the repertoire of the agent. Therefore, in tasks that do not rely on quick overview information, the advantages of a more personified representation could be more of use than the descriptive non-embodied interface, for example in cognitive tasks such as (spatial) learning [e.g. 6, 17].

In both tasks, the last comparison question we asked was which system was the friendliest, where most participants answered condition C in both tasks. It might not be the most supportive system (which was the non-embodied system), but it was the one they liked the most. A combination of both embodiment and overview information in the form of graphics or text would combine the best of the two interfaces.

We also saw that introverts are not as interested in the embodiment of an agent as extraverts are. This is not unexpected, as extraverts prefer activities involving interactions with other people whereas introverts tend to prefer solitary activities [10]. This suggests the use of different agents for different kinds of personalities, or using an adaptable version of an embodied agent, as suggested by [23].

This research has shown possible interactions with a cognitive vision system. Although we did not find significant differences between trust and intelligence ratings, we could show that users generally react positively to the more personal interaction style but also appreciate a map-like overview. Additionally, a more personal embodied representation increased social presence, especially with extraverted individuals.

FUTURE WORK

In task 1, we found that users prefer a combination of agent and GUI; in a follow-up study, we want to explore possibilities to combine these two modes in a meaningful way, e.g. like a TV weather forecast, where the presenter shows information on a map. This would yield additional insight into personal representation in addition to effective communication and could lead to very positive experiences with cognitive vision system.

In addition, we want to further explore the use of the agents gaze direction to point towards objects of interest in off-screen interaction and whether this can help to improve the interaction and guide the users' attention, in addition, more more attention can be directed to the findings regarding personality type in relation to social presence and how this influences interactions.

ACKNOWLEDGMENTS

This work was supported by the Austrian Science Foundation (FWF, project S9107-N04).

REFERENCES

1. Andre, E., Rist, T., and Muller, J., 1998. Integrating Reactive and Scripted Behaviors in a Life-Like Presentation Agent, *Proc. 2nd International Conference on Autonomous Agents (Agents '98)*, 261-268
2. G. Ball and J. Breese, 2000. Emotion and personality in a conversational agent. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, Eds., *Embodied conversational agents*, Cambridge, MA: MIT Press, 189-219.
3. Bickmore, T.W., 2003. Relational Agents: Effecting Change through Human-Computer Relationships, *Ph.D. Thesis*, MIT.
4. Bickmore, T.W. and Picard, R.W., 2005, Establishing and Maintaining Long-Term Human-Computer Relationships, *ACM Transactions*, 12 (2), 293-327.
5. Brennan, S.E. and Ohaeri, J.O., 1994. Effects of Message Style on Users' Attributions toward Agents, *Conference Companion Proc. CHI '94*, 281-282.
6. Cole, R., Vuuren, S.v., Pellom, B., Hacioglu, K., Ma, J., Movellan, J., Schwartz, S., Wadestein, D., Ward, W. and Yan, J., 2003. Perceptive Animated Interfaces: First Steps Toward a New Paradigm for Human–Computer Interaction, *Proc. IEEE,* 91(9), 1391-1405.
7. Cassell, J., 2001. Embodied Conversational Agents, *AI Magazine,* 22(4), 67-84.
8. Dahlbäck, N., Jänsson, A. and Ahrenberg, L., 1993. Wizard of Oz Studies - Why and How, *Proc. ACM International Workshop on Intelligent User Interfaces '93, 193-200.*
9. ECVision, online at http://www.ecvision.org. Last accessed: January 2006.
10. Furnham, A., 1997. *The psychology of behaviour at work*, Psychology Press, Publishers, East Sussex, UK.
11. de Greef, H.P. and IJsselsteijn, W.A., 2000. Social Presence in the PhotoShare Tele-Application. *Proc. PRESENCE 2000*, Delft, The Netherlands
12. Keirsey Online Temperament Sorter (German), http://www.keirsey.com/ Last accessed: January 2006.
13. King, W.J., and Ohya, J., 1995. The representation of agents: a study of phenomena in virtual environments, *Proc. RO-MAN'95.*
14. Koda, T. and Maes, P., 1996. Agents with Faces: The Effects of Personification of Agents, *Proc. HCI '96 UK*, 98-103.
15. Lanier, J., 1995. Agents of alienation, *ACM Interactions*, 2 (3), 66-72.
16. Lee, K.M. and Nass, C., 2003. Designing Social Presence of Social Actors in Human Computer Interaction, *Proc. CHI '03*, 289-296.
17. Lester, J.C., Converse, S.A., Kahler, S.E., Barlow, S.T., Stone, B.A. & Bhogal, R.S., 1997. The Persona Effect: Affective Impact of Animated Pedagogical Agents, *Proc. CHI '97*, 359-366.
18. Mori, M., 1970. The Uncanny Valley. *Energy*, 7, 33–35.
19. Mori, M., 2005. On the Uncanny Valley. *Proc. workshop Humanoids-2005.*
20. Nass, C., Steuer, J. and Tauber, E.R., 1994. Computers are Social Actors, *Proc. CHI '94*, 72-78.
21. Picard, R. W., 1997. *Affective Computing*. Cambridge, MA: MIT Press.
22. Power, G., Wills, G. and Hall, W., 2002. User Perception of Anthropomorphic Characters with Varying Levels of Interaction. *Proc. HCI '02 UK*, 37-51.
23. Reeves, B. and Nass, C., 1996. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press, Cambridge, UK.
24. Sas, C., Hare, G.M.P.O., and Reilly, R., 2004. Presence and task performance: an approach in the light of cognitive style, *Cognition, Technology & Work*, 6 (1), 53-56.
25. Short, J.A., Williams, E., & Christie, B., 1976. *The social psychology of telecommunications,* New York: John Wiley & Sons.
26. Sproull, L., Walker, J., Subramani, R., Kiesler, S., and Waters, K., 1996. When the Interface is a Face, *Human Computer Interaction*, 11, 97-124.
27. Vernon, D., 2004. Cognitive Vision – The Development of a Discipline, *Proc. IST 2004 Event 'Participate in your future'*.
28. Xiao, J., Catrambone, R. and Stasko, J., 2003. Be Quiet? Evaluating Proactive and Reactive User Interface Assistants, *Proc. INTERACT 2003*, 383-390.