OTTO-VON-GUERICKE UNIVERSITÄT MAGDEBURG
HUMAN INTERACTION TECHNOLOGY LABORATORY NEW ZEALAND

Digital Characters in the Real World

A review of embodied agents and augmented reality

Christian Graf
cgraf@cs.uni-magdeburg.de

July 12, 2004

Supervisors:
Dr. Knut Hartmann (Universität Magdeburg)
Mark Billinghurst, PhD (HITLab NZ)

ABSTRACT

Recent research has concentrated on the question how virtual characters (‚agents') and their set of reaction can be enhanced to better suit a certain purpose in a virtual reality environment (VRE). Another field of active research is augmented reality (AR), where computer generated graphics is superimposed over a real world image. Thus AR is able to add detail and enhance the underlying information.

Our idea is to create an augmented reality environment (ARE) with an agent for conversation centred areas such as presentations. Our goal is to find evidence in literature and research how agents are employed as effective means of communication. This includes a review of existing systems and a summary of user studies. We will a look at combining AR technology with agents, and what special issues arise. With this background we will develop design objectives and requirements of agents in AR. A prototype implementation of an agent architecture with a digital character as user front-end will be implemented. From our literature review and our practical implementation we will identify a list of fruitful areas of research and derive a list of challenging new questions.

CONTENTS

1. INTRODUCTION

1.1. Motivation

From early on mankind was fascinated by the idea of creating a human-like creature from inanimate material. Jewish legend tells us of a Golem made from clay and mystically filled with life (Idel 1990). It was taken up by modern authors in novels like “Frankenstein” and found their way in more recent productions like “Toy Story”. All are fuelled by the same thought: to ease human life by natural interaction with a human-like partner. Today’s technology does not rely on clay but on computers and electronic devices. With digital tools at hand research groups around the world focus again on an old idea – natural interaction with a human-like partner, this time it is computer generated.

Such interaction might not necessarily be ‘better’ in terms of efficiency than others but maybe in terms of quality. As Maes (1994) pointed out intelligent interactive agents are usefull to be employed in “fail-soft” systems. Fail-soft systems are such systems that preserve their essential operability even if parts of it fail. Applied to software agents that means that failure of communication with the agent must not mean the breakdown of the whole process. Agents enhance the quality of the application but they are not essential to it – a “nice to have” gimmick.Using advanced computer graphics it is possible to generate almost photo-realistic imagery of three dimensional virtual characters. In films it is now common to see digital characters interacting with real actors. Virtual characters become more and more indistinguishable from their real counterparts. However until recently such characters always inhabited the screen space separated from the real world, and they were not able to be generated in real time.

Augmented Reality (AR) is a research field concerned with overlaying computer graphics on the real world so that virtual imagery is seamlessly blended with the real surroundings. Unlike in film and television, these graphics are rendered in real time. With this technology virtual characters could co-exist in the same space with real humans. Despite the interesting possibilities that it offers, there has been little research work on Augmented Reality Agents. The goal of this paper is to review and summarise research work on embodied digital characters, to identify promising areas of application of augmented reality interfaces for the addition of embodied agents, to present some results of a prototype implementation, and to envisage further research directions.

1.2. Scenarios

Let us image we were in a future meeting of the senior managing board of a big car manufacturer. The issue is about a new car model and how its engine design should look like. All participants are immersed in an *Augmented Reality Environment* (ARE), where the computer generated representation of the car’s engine compartment is superimposed over the real world (imagine a virtual three dimensional model that each participant can see in correct perspective from his individual position). The participants can discuss freely with each other and have an almost hands-on impression of the engine’s location and constrains it has to match. Each of them can interact with it and everybody else will see the change. As the research & development (R&D) department is thousands of miles away, it has sent a proxy: a virtual agent. It has knowledge of the application domain and problems related to it, is capable of following the discussion, gives sound explanation, and is sensitive to the conversational context. Now the managers will consult the agent for details and background information and a conversation will develop. In the end, the group will have decided for an

innovative car design, even if nobody of the R&D experts was physically present, solely the agent and its abilities provided sound support. Artificial Intelligence has strived for such omnipotent agents for many years but a sound result has not emerged yet. We might shift our focus to other areas of application, for example education or entertainment.

Transferring the presentation scenario from the business world to the world of, for example, a museum means to shift from decision support to process support. In this area the skills of the agent are helpfull complementing other means, but not crucial. Nonetheless it is important to the joy and pleasure of all participants what skills the agent has and what actions it is able to perform. Presenting interesting stories in a museum involves a great deal of connecting facts to stories that relate back to the exposed artefacts and the audience. If a museum's guide additionally reacts flexibly on the audience's questions and suggestions, people will most likely regard the tour as successful and interesting. The audience's contentment can be clearly attributed to the convincing presentation of the agent.

Having envisaged such applications, we want to give some insight to the reader concerning the problems, the approaches, the application domains, the key questions to be answered and the possible outcome of this paper in the following chapters.

1.3. Problem, Goal & Approach

1.3.1. Problem & Research Question

Presentation through and collaboration with an agent in AR is a relatively new field of research. First, in recent years research in collaboration focused on *Virtual Reality* (VR) and its improvements. The main question was how representations of other users could be given more lifelikeness in a *VR Environment* (VRE). Little was done to improve the quality and thoroughness of interaction. Second, augmented reality is a very young field of research and the chosen domain of presentation has only been explored to a small degree before. Third, computers are used to present complex information to the user. The mismatch between humans ('analogue': senses, thoughts, emotions) and computers (digital: numbers, computation, complexity) suggests the idea to level the both sides by using human-like interfaces on the computer. At the time of conducting this research in the Human Interface Technology Laboratory, nobody had done anything like that before, and there was no expertise in the area of digital characters. The task and the questions behind were:

"What is the relation between embodied agents, augmented reality and collaboration? Review the literature and compile a comprehensive report with appropriate references about your findings! Implement and possibly test a prototype! What are your recommended directions for further research?“

1.3.2. Goal Definition

From the given hypothetical scenarios above, we can learn that agents might be a good idea for interfacing between human and computers. Our goal is to find evidence in literature and research how agents were employed as effective means of communication in conversation centred areas such as presentations. This includes a review of implemented systems and of the test results. Then, we will have a look at combining AR technology with agents, and what special issues might arise. With this background we will develop design objectives and requirements of agents in AR. A prototype implementation of an agent architecture with a digital character as user front-end will be implemented. This will conclude our feasibility study on agents as presenters in augmented reality. From our consideration during the literature review and our practical implementation we want to identify a list of fruitful areas of research and derive a list of challenging new questions.

1.3.3. Solution Strategy

Knowing very little about agents, their incorporation into AREs and how they might affect the interaction between human and computers, we need to ask some basic but essential questions. What constitutes natural communication among humans, how does it relate to human-computer interaction and how may we ease this interaction? Which communication cues are essential, and what means can mediate those? How can human behaviour be explained in conversational setting, and what implications can we draw from theories? We will mainly look into appropriate literature in the field of psychology and sociology to become clear about these questions. The findings are presented in Section 2.

Having set our focus on embodied agents, we will review the literature. What technologies were used to mediate what interaction? We are especially interested in the question what kind of systems had been proven to be effective in the sense of providing natural interaction to the user. The architecture of the beneficial systems will be of interest. A review of papers, articles and reports of other research groups will be the main source of our conclusion. We expect that the vague guess, an AR agent, might be a good idea. Read Section 3 and 4 to learn more about the findings.

Having learned about other approaches, we need to define our own. But what is different in an augmented reality setting, and is there any special condition we must be cautious about when building an agent for an ARE? What features and abilities constitute an agent and what measurement is to be taken to evaluate it? What areas of requirements can be identified when designing an agent? We have reviewed some proposed sets of criteria for agents, extended them and build our own. Read about it in Section 5 and 6.

Concerning the implementation, some questions arise. Should we build a new architecture, or use an existing one? What specific advantages and disadvantages do both approaches have? Outcomes are mostly based on our own practical tests. The experience during this stage of research greatly contributed to identifying further directions of research. You will find the results in Section 7 and 8.

1.3.4. Expected Outcome

We want to understand how human communication works, what means effects conversation in what way and how agents can be designed to use these effects. We expect to build an thorough understanding to what extent an agent can be useful in shared spaces and how it has to be build to engage humans in an effective or pleasant social interaction. Ideally a user study shall be done to confirm our expectation towards the effectiveness of agents.

1.4. Overview of this Paper

In order to develop a compelling virtual character there are many research problems that need to be addressed - language processing and generation, reasoning and machine learning, computer graphics, artificial intelligence, cognitive and social psychology, philosophy and sociology. Although interesting, most of these topics are beyond the scope of this paper. We will introduce and discuss topics from these fields as far as they contribute to our considerations. For further in-depth elaboration see the appropriate literature mentioned in the related section and the subsequent chapters.

In the remainder of this work we will give an introduction to user interfaces and some background on human interaction theories. Then the notion of *Embodied Agents* (EA) will be introduced and some properties explained. Next, we will summarise research on previous EAs and discuss several user studies conducted with such agents. Defining Augmented Reality and how agents relate to it will precede our work on design requirements for agents in the

presentation domain. After that, we will present our prototype implementation of an agent in AR and what consideration let to the design. Finally we will describe our findings about promising directions for future research.

As we could not include all the material we have produced for this work into this paper, there is a complementing website. Look at http://www.cs.uni-magdeburg.de/~cgraf/NZ/HITLab/Report/ to find out more about adjacent topics, detailed background information and pointers to appropriate literature.

2. HUMAN-COMPUTER INTERACTION

2.1. Computers in everyday life

Fifty years ago, when the UNIVAC (see Figure 1) was introduced as first commercially available general purpose computer for civilian-use, the computer was utilised to take the burden of vast computations from the human, but it could only be handled by experts. Nowadays, "more than 72 million employed people in the US age 16 and over - about 54 percent of all workers - used a computer on the job" [Web1] and consumer products are commonly computerised. Computer chips can be found in almost any household today, be it in washing machines, mobile phones, VCRs or microwaves - everybody has to cope with computers and handle them.

Figure 1: The U.S. Census Bureau was the first client to order a mainframe electronic computer.
© U.S. Census Bureau

Back in the 50s technological constraints (memory, speed, input & output channels) of early computer systems forced a concentration on the functionality. Only few lines of the programming code were concerned with the user interface. The end user had to be an expert to run the system. Lifting the hardware limitations has freed resources for considerable efforts to improve the user interface. "The effect of this rapid increase in the number and availability of computers is that the computer interface, must be made for everybody instead of just the professional of computer hobbyist" (Eberts, 1994).

Our world today shows that computer technology runs every sort of processes and machinery. Trying to accommodate the user performing his task we must concern the users first (so called 'user-centred design'). Machines should work through routines, tedious procedures, error-prone tasks so that humans might concentrate on critical decisions, planning and coping with unexpected situations. Human judgement is necessary for unpredictable events in the world (the open system) in which actions must be taken to preserve safety, avoid expensive failures, increase product quality. To achieve a higher task effectiveness of humans we need to decrease their burden to use technology.

2.2. Human-Computer-Interaction

Computers are at the foremost position in infiltrating all levels of our daily life – they work in all kinds of machinery and equipment commonly used. They run office, home and entertainment applications, they work in the industry and commercial systems as well as in exploratory, creative and co-operative systems. Whenever humans get in contact with machines, they have to manage to get along with the rather emotionless world of modern technology. Addressing this area, *Human-Computer-Interaction* (HCI) is "a discipline concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them" [Web2]. As Faulkner (1998) puts it "Human-Computer Interaction is the study of the relationship that exist between human users and the computer systems they use in the performance of their various tasks". HCI is an interdisciplinary field, relating computer science, psychology, cognitive science, human factors engineering (ergonomics), sociology, design, engineering, art, anthropology, physiology, artificial intelligence and others. See Figure 4 for a schematic illustration of the framework. The focus on the human and his needs becomes even more clear when considering the term *Human Factors Design*.

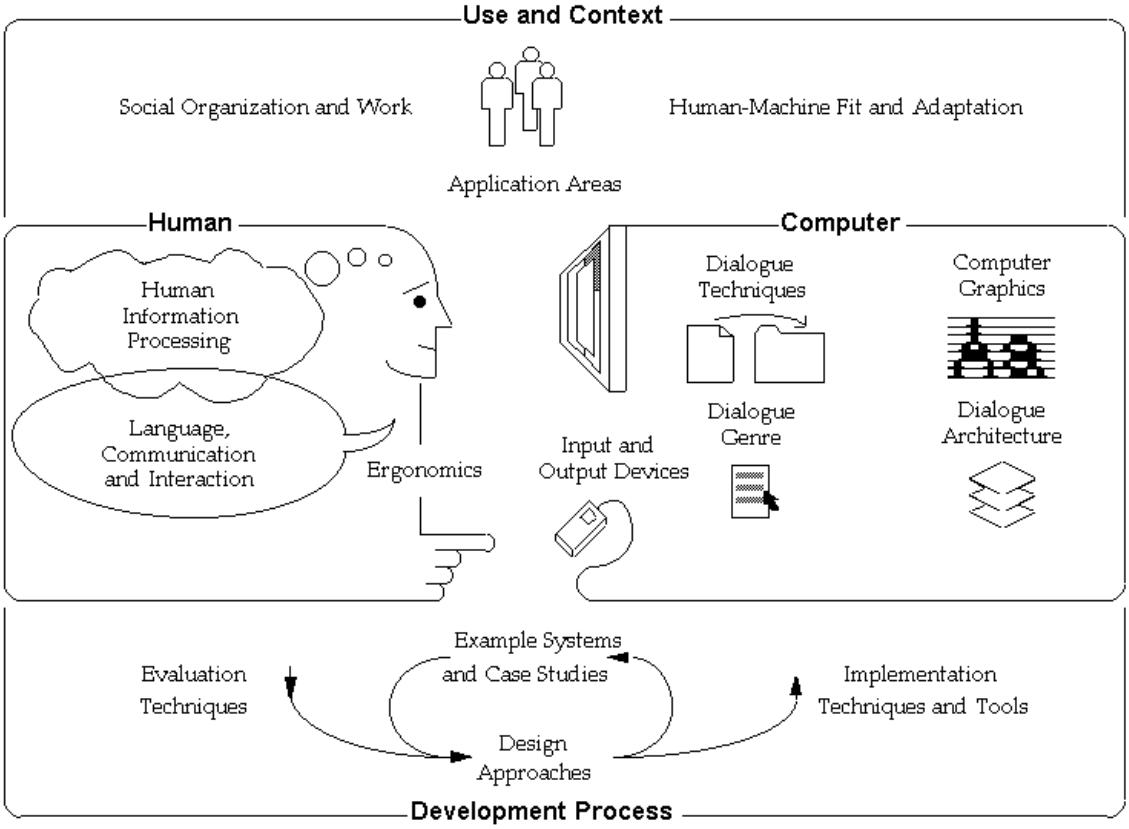


Figure 4: Interrelationship among topics in HCI (Courtesy SIGCHI)

Sometimes computers have an important place in life critical systems, e.g. in power plants or surgery support systems. Intuitively it becomes clear that such systems should be easy to use - that is the motivation behind HCI. If not, devastating consequences could result, as the public had to learn in the 1979 Three Mile Island nuclear power plant disaster [Web16][Web19]. The control panel did not give the operators any useful information to remedy to the situation when problems arose. Hundreds of alarms (audio and visual) went on at the same time. The system could not provide useful information about the exact current status of the system, was displaying to many uncoordinated warning without priority indicators and recovering actions to be taken were not obvious. As a result the core of the reactor almost melt down and the engineers could hardly stop it. Not only to avoid such fatal incidents, but to ease humans' need to work with technology in general, the aim of HCI is to develop or improve human-computer interfaces regarding (Shneiderman, 1998)

- safety (reduce errors generated by users, or offer better handling for such errors),
- utility (mapping of interface elements with user's tasks),
- effectiveness (does the system perform the tasks correctly?),
- efficiency (does that improve the user's efficiency/productivity?), and
- usability (can it be used?) of systems in general.

HCI aims on providing the users with interfaces that will make them more efficient in performing a task using the machine's abilities and advantages. The rating of efficiency is done in comparing the performance on the computer interface to the one on an equivalent manual system. This requirement is essential since "all to often computerized applications are produced that do not make the user's task easier and more satisfying, nor do they save time" (Faulkner 1998). He concludes: "The task of HCI is to design for people, for tasks and environments". The result of the design process is the user interface. It has to be task appropriate, efficient and suitable for the user. The UI connects the human to an environment that is filled with technology, mostly computer systems. The objective in designing the UI is to minimise the human's workload to handle the UI. Then he can invest as much as possible into solving the task. Thus the UI has to be the mediator between two different worlds: technology and humans.

During the last twenty years there has been significant effort to mend these two worlds and establish a mediating in-between. Research shows that the percentage of code for the UI and the percentage of money for its development has increased (Smith and Mosier 1984) (MacIntyre, Estep, and Sieburth, 1990; Rosenberg 1989). But merely computerising a formerly manual process will not guarantee an increase in efficiency (Eberts 1994). Findings from Hansen, Doring and Whitlock (1978), Kozar and Dickson (1978), and Gould (1981) show that inappropriate UI design can make tasks more difficult and time-consuming. This is contra-productive to what UI design strives for! Simply writing more code or investing more money seems not to be enough. In fact, concrete "evidence of improved usability is difficult to find", and experience shows that people still have significant problems with computers (Eberts, 1994). There was much hope in the *Object-Action Interface* (OAI) model to overcome these problems.

2.2.1. Object-Action-Interfaces

In the OAI model objects and actions are mapped from the real-world onto metaphorical objects and actions in the interface. Successful OAI interfaces provide the following features [Web9]:

- high visibility of interface objects;
- high visibility of actions of interest;
- rapid,
- reversible,
- and incremental actions.

The Direct Manipulation (DM) approach is the most prominent representative of OAI. DM is an interface concept that offers, among other things [Web9]:

- Display of current status of tools (overall view of system's status);
- Be as close to reality as possible (e.g. provide an appropriate representation or model of reality);
- Allow maximum control to the user;
- Display the result of an action immediately;
- Provide rapid response and display.

Compared to former UI concepts like command line interfaces, DM is more suitable to the user and his task. The user can find a solution through a succession of various tasks on the interface. He is supported in finding this solution by suitable representations of problems. See it as counting on your fingers: it gives you a physical real representation of numbers. The advantages of an DM object is that it represents the problem in a more intuitive way, that it combines both the data entry and data display in the same physical location, and that it gives immediate feedback.

Direct manipulation object are nice to add to an interface, but they have their limitations too. The following list provides some hints on key problems [Web9].

- Many objects with many possible actions could be present which is potentially confusing.
- Being bound to a finite screen, DM objects may consume valuable space that makes it necessary to hide some of the information off-screen and that requires scrolling and multiple actions. When more detailed information is needed the display is cluttered quickly.
- Menus hide their content at first. Not all options are displayed at the same time and the user has to search and retrace them. This takes time and guided attention - it is tedious.
- The meaningfulness of visual components is essential. Users must learn the semantics of the components (e.g. slider) that may lead to errors. The visual representation (e.g. icon) itself may be misleading.
- The choice of proper objects (including icons, menus, labels, buttons, other interface elements) and proper actions is difficult. The metaphor should be understood instantly without much attention. It is advisable to use simple metaphors and models with an associated minimal set of concepts.

Summarising these limitations, the DM approach seems to have reached its limitations concerning the problem of bridging the conceptual gap when humans communicate with or through computers.

2.2.2. A new medium evolves

There has been the tendency over the last ten years that computers have gained another meaning. The growing demand of information and global connectivity pushed the popularity of world-wide computer networks. The Internet has turned computers into a means of communication. They connect people, transport content and thus influence people. Being tools for humans in former days they have transformed into a medium now. Today's new purpose of computers might be a reason to re-consider the old DM metaphor. Such an interface might have been good for computers as tools to perform certain tasks. But the nature of the computer as medium today is quite different to the nature as instrument of action - think of a radio in contrast to a hammer. Thus the applicability of DM has to be reconsidered. One consequence could be to think about other metaphors or at least alternatives to choose from [Web20].

As Reeves and Nass could demonstrate in several studies (Reeves 1996), human-computer-interaction follows the same principles as human-human interaction. That means, we have an inherent tendency to respond to media and technical systems in ways that are social and common among humans. The question is, how do we allow the user to behave and communicate naturally when interaction with technology? One would expect that it probably lessens the cognitive workload for the user and contribute to the user satisfaction. With a proper design the task performance would increase as well. In the following chapter we conceptually suggest such a more human-like way of computing.

2.3. Affective Computing

The usual way humans interact with each other is face-to-face and by language. In such a communication at least two humans exchange meaning verbally. But words and sentences are not the sole part of the communication. Potentially humans can communicate with all five senses: sight, sound, touch, taste and smell. Using these channels, many other cues, verbal and non-verbal, are incorporated when transferring meaning. The sender employs these cues subconsciously and the receiver understands it in the same fashion - subconsciously. A vital feature is that our everyday interaction with each other and the world around us is a multi-sensory one, each sense providing different information that builds up the whole.

A key issue in interaction is speech (including listening). It is not a one-channel but always a two-channel process. The sender utters some words complemented by gaze, The receiver gives back some kind of acknowledgement to the sender, poses questions or takes turn, i.e. there is feed-back channel (see Figure 6).

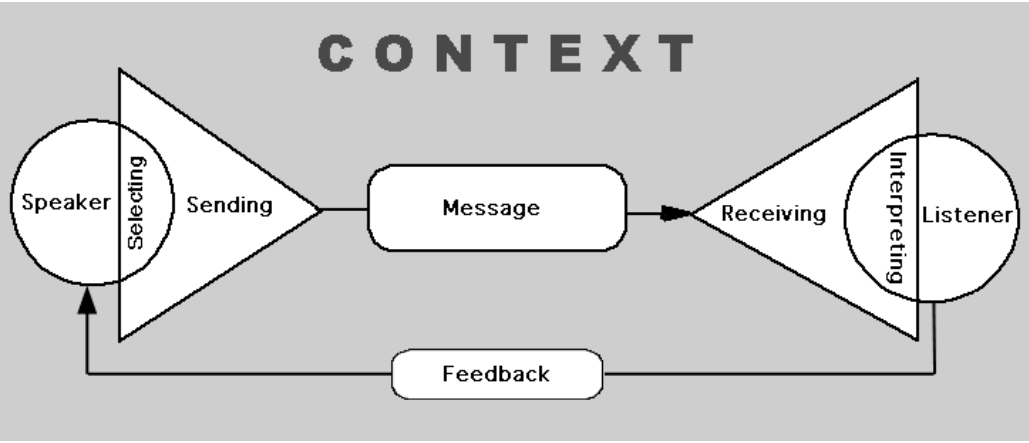


Figure 6: The sender-receiver model (a.k.a. discourse model) of human communication

The fluency, prosody and intonation of the sentences and words tell us important things about the speaker, e.g. about his inner state, about the circumstances he is speaking of or to give or emphasise a certain meaning.

Aside from the language centred cues there are multiple sources of sensory information, the so called non-verbal cues. These cues could support or complement the verbal utterances, e.g. pointing somewhere and saying 'there!'. It includes gestures like hand movements, head and eye movements as well as body movement and body orientation. Non-verbal cues can be expressions on their own, i.e. the body language. Turning your back on someone and crossing the arms in front of your chest will tell anyone that you don't want to interact with this person even if you don't say a word. On the other hand, when someone identifies a known person, the typical reaction is tracking with the eyes and a body orientation towards the person in question. That will be understood as a sign of openness and the disposition to start a conversation. Non-verbal cues play an important role for both, the speaker and the listener. For the speaker they are necessary means to accentuate his story. They support the listener to understand what the speakers wants to tell. Without non-verbal cues the conversation would be 'stiff' and quite unnatural.

An important aspect is the context of the conversation are the emotional states of the participants. They play a key role to truly understand what each one wants to tell. Thus the ability to recognise, interpret and express emotions - commonly referred to as "*emotional intelligence*" (Goleman, 1995) - plays a key role in human communication. Related is empathy, another typical property of human-human interaction. It means that someone shows understanding of the other's situation and is feeling with the individual, e.g. after a friend has suffered from a great loss one could express his sympathy by simply hugging him. Without showing some kind of empathy the human-human conversation is cold and distant, solely consisting of facts - not any social.

On the other hand, the use of computer technology often has unpleasant side effects, some of which are strong, negative emotional states that arise in humans during interaction with computers. Frustration, confusion, anger, anxiety and similar emotional states can affect not only the interaction itself, but also productivity, learning, social relationships, and overall well-being. Affective Computing takes up the social side of human-computer interaction and tries to actively support human users in their ability to regulate, manage, and recover from their own negative emotional states, particularly frustration. Thriving for this goal affective computing tries to resemble the human-human communication, especially on the emotional side of the interaction.

Emotion affects judgement, preference, and decision-making in a powerful, yet elusive way. Therefore it is an indispensable element in any interaction with technologies. Since emotions are an integral part of communication, computers with affect-recognition and expression skills would allow a more natural and thus improved human-computer interaction. An affect-recognising computer can "learn" during an interaction by associating emotional expressions (like pleasure or displeasure) with its own behaviour, as a kind of reward and punishment. The software or system would automatically adapt to the users needs. An affective computer-system could detect user frustration and show sympathy for the user and offer assistance, encouragement or comfort. Klein (1999) could show that user frustration levels were significantly lower if assistance was provided.

Since human interaction is improved by multi-sensory input, it makes sense to ask whether multi-sensory information would benefit to this endeavour. We can argue that only a multimodal system is a sound foundation for subsequent (re-)actions of an affective UI, because only the sum of multiple cues from different channels indicates what an utterance

truly means or how a person feels. If a system knows how its users feel, it can appropriately react to these emotions. It can guide, help, change its appearance or simply be as unobtrusive as possible.

Bringing affective computing into a more human form led to the development of *Social Agents*, resembling humans i.e. 'anthropomorphic'. They adapt to the user's context and system status, constructing intelligent responses to users and interacting verbally complemented by non-verbal cues. Social Agents are not intended to be a clone of humans, rather, we are applying human intelligence principles relevant to the specific technological and usage situations.

Equipping an interface with such agents will most likely increase the subjective pleasure of the interaction. Additionally it might positively influence the users' task performance. A functioning system will make human-computer interaction more intuitive, i.e. communicating with a computer will not require special technical skills anymore. We simply can use the same communication skills as in everyday life. The cognitive load for using the interface would vanish and we could free more time for concentrating on and solving the task.

Before concerning about how to implement a socially able agent we need to know what a general agents consists of.

2.4. **The Essence of Agency**

The essential attributes constituting an agent shall be defined here. A note from a popular textbook on artificial intelligence shows that we should not expect a mathematically sound definition: “The notion of an agent is meant to be a tool for analysing systems, not an absolute characterisation that divides the world into agents and non-agents” (Russell and Norvig 1995).

An common requirement for an agent is that it acts, or can act independently. In contrast to real world agent we deal with agents that ‘live’ inside computers: software agents. Sánchez (1997) identifies a hierarchy of eight types of software agents (see Figure 7). In his taxonomy he stresses the importance of the entity that perceives and benefits from the notion of agent. For example, the software developers use programmer and network agents as abstractions to cope with the complexity of system design and computer network. In the same matter, the end user shall benefit from the user agents. From findings by Friedman (1995) he concludes that end users can better deal with system complexity by viewing programs as animistic entities. The term *interface agents* or *user interface agents* has frequently been associated with this view of agency (see Laurel 1990, Kozierok and Maes 1993, and Wooldridge and Jennings 1995). But "interfaces" also exist between software modules and communicating computers (or between any independent systems). For clarity and to prevent ambiguity we will use "synthetic" to distinguish this class of agents.

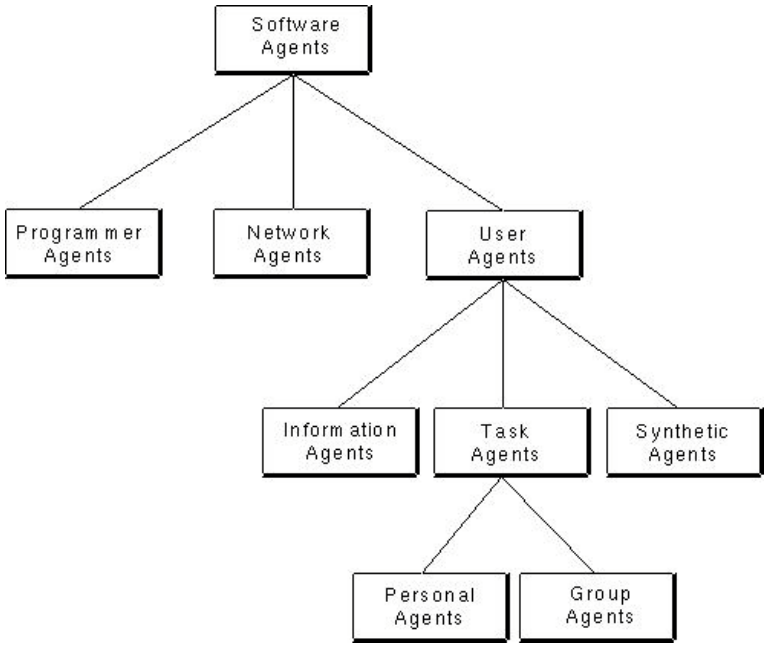


Figure 7: A Taxonomy of Agents (Sánchez 1997)

With this diversity given, one may ask what is common for every general agent? Franklin and Graesser (1996) identify the basic properties of a general agent as:

- reactive (sensing and acting), i.e. responds in a timely fashion to changes in the environment;
- autonomous, i.e. exercises control over its own actions;
- goal-oriented (pro-active, purposeful), i.e. does not simply act in response to the environment;
- temporally continuous, i.e. is a continuously running process.

In short, "an autonomous agent is a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future." (Franklin and Graesser 1996)

A further classification might be helpful to relate a wide variety of agents to each other, e.g. according to (a) the tasks they perform, (b) the range and sensitivity of their senses, (c) the range and effectiveness of their actions, or (d) how much internal states they possess. Other researchers suggest different taxonomies, e.g. (Brustoloni 1991). Focusing on synthetic agents, different researches provide criteria that programs must pass to be considered as "true agents" (e.g. Foner 1993).

Our defining criterion will be if the agent is communicative and socially able, i.e. communicates with people. Then we speak of a *Social Agents* (see last chapter), *Communicative Agent* (Franklin and Graesser 1996), *Synthetic Agent* (Sánchez 1997), or *Embodied Agent* (Cassell et al. 2000). Communication is a multi-party, multi-modal process that includes not only the subjects but their history, affective states and the context of the conversation. For more information on human communication see the accompanying web-design site.

Agents designed to be perceived directly by end users and to perform tasks on their behalf are an innovative class of agents that can facilitate human-computer interaction. The potential of agent-based user interfaces has been discussed extensively (Kay 1984, 1990; Laurel 1990, 1991; Negroponte 1990, 1995). On the other hand, as we have seen from earlier chapters,

applications that are primarily output orientated might not gain much from such an agent. Agents simple need to much time to interact with the human. But applications that rely on user satisfaction, that have a positive relationship between the interacting entities as central element or that aim on the user's pleasure can benefit from Social Interaction with agents. We normally find such applications in learning environments, museums and entertainment. This is exactly the area this paper wants to focus on to unwrap the potential of social interaction between humans and computers.

Conclusion

We could see during this section that human's interaction with technology was governed by an efficient but not necessarily enjoyable and self-explaining interface. Affective computing tries to employ concept known from human-human communication to ease the human's effort of interacting with technology. The subclass of affective interfaces this paper wants to focus on is the class of embodied agents. The next section will have a closer look on those.

3. **Embodied Agents**

Humans naturally (if unconsciously) attempt to use "a certain system of interaction with other intelligences – a system of *social interaction*" (Doyle 1999) when acting with computers. Humans have both evolved and learned that system and we "implicitly have expectations about the social and emotional behavior of machines, and even treat them in social ways" (Doyle 1999). However these expectations remain largely neglected in commercially available software and hardware computer system. Although users may speak and gesture at their computers, the machines do not gesture back, or engage in human-like communication. This lead to the idea to 'wrap' the agent into a 'hull' of appropriate visualisation and behaviours – the *Embodied Agent* (EA).

In this section we describe research that attempts to develop embodied characters that can engage more naturally with the user.

3.1. Delimitation

When considering human-like virtual characters in two terms that are often used and confused: that is "avatar" and "agent". The first one is a graphical representation of the user as he is immersed in a virtual reality or three dimensional graphics environment. An avatar is not autonomous – it totally relies on the human input for speech, gestures, movements etc. In contrast, an agent is an autonomous being, producing appropriate (re)action and utterances itself – the user is not obliged to steer it. Even if avatars and agents resemble each other in the outer appearance, in the animations they exhibit and the actions they perform, this key feature distinguishes them. Consequently avatars and agents are typically used for different types of applications. Avatars are often employed in chat systems or online communities to represent a human being. Agents can serve a number of roles, such as *pedagogical agents* (Johnson 1995; Rickel and Johnson 1999; Rickel and Johnson 2000) that are utilised in educational environments to train users on specific tasks, mostly manual work; or *conversational agents* that are designed to and capable of engaging the user in a conversation to accomplish or facilitate the desired task. They are employed mostly in information delivery. Agents and avatars could be implemented in a wide variety of different forms, but as we have already seen from the findings of Reeves and Nass (1996) human-like representations would possible be better for social interaction. Therefore we want to focus on EAs in this paper, especially on the conversational ones, the *Embodied Conversational Agents* (ECA).

An embodied conversational agent may be defined as a virtual graphical character that understand natural human communication cues and responds with it's own speech and gesture output. Developing such an agent is an extremely complex task that encompasses research from a number of different areas, including:

- Multimodal interfaces
- Anthropomorphism
- Application Area
- *Conversational models*
- *Ergonomics*
- *Psychology*

We will discuss the first three topics in the following chapters in more detail. The last three topics are interesting in the broader context but the are not essential and thus not elaborated on exhaustingly in this report. Some concepts from those areas might be used in the text later on with a short explanation. If interested the reader could browse the complementing web-site for further detailed information.

3.2. Multimodal Interfaces

Among the strengths of social communication are is the use of multiple modes and multiple information types and it's inherent flexibility. An ECA should mimic human behaviour and responses in a natural way to be believable to its conversation partner. Moreover, its perceptual mechanisms need to support interpretations of real-world events that can result in real-time action of the type that people produce effortlessly when interacting with their environment. These properties constitute a *multi-modal system*, a system that is able to integrate multiple modalities and thus is a higher-bandwidth communication interface. As interface to computers it means that other channels could be used at the same time providing additional or complementing information.

Considering the five senses, language (as a form of sound) is the ability that distinguishes humans from animals. Recognising speech has the advantage that it happens almost automatically with little attention. While producing speech the human body can move freely possibly engaged in some task, e.g. pointing at something. Think of disabled people who cannot use other means of interaction.

3.2.1. Speech as input

In the English language we can identify 40 *phonemes*, the atomic elements of speech (Dix et al. 1999). But language is more than simple sounds. Emphasis, stress, pauses and pitch can all be used to alter the meaning and nature of an utterance. The alteration in tone and quality of phonemes is termed *prosody*, it conveys a great deal of the actual meaning and emotion within a sentence. Prosodic information gives language its richness and texture, but is very difficult to quantify and thus to resemble it within the computational framework of computers. Another problem is that phonemes sound differently when preceded of followed by different phonemes, a phenomenon termed *co-articulation*. We need these distinctions for later elaboration, and as we can see using language is nothing but easy.

People who do not regard themselves as computer literate are appealed by the idea to converse naturally with the computer. Indeed, sometimes synthesised speech becomes the primary communication channel, e.g. for disabled people with visual impairment. Today single-user, limited vocabulary systems can work satisfactorily, but there are still problems when employing speech. On the input side, speech can only be applied to very specialised task because recognising complex vocabulary is hard to learn for computers. The speech

recognition process poses problems itself: accents, dialects, different intonation of words and 'continuation' noise such as 'Umm' to fill gaps in the usual speech confuse the machines. Interference from background noise hardens the extraction of meaningful sound.

Even if all the technical problems had overcome it still would be a problem for computer to interpret natural language. It is full of arbitrary meaning, sometimes meaning that is contrary to what the speaker wants to express (think of sarcasm, irony etc.). The computer had to have a huge world knowledge to correctly interpret all the human diversity in language. And, there is a fundamental difference between humans and computers: we concentrate on extracting the meaning from the whole sentence we hear rather than decomposing sounds into their constituent parts, analysing the structure (i.e. syntax), assigning meaning (i.e. semantics) and intension (i.e. pragmatics).

3.2.2. Speech as output

Using synthesised speech as output meets significant challenges as much as on the input side. Humans are “highly sensitive to variations and intonation on speech” (Dix et al. 1998). Listening to synthesised speech they are often intolerant to its imperfections. Speech synthesisers hardly produce natural sounding speech – they present to us mostly monotonic, non-prosodic tones. Human find it hard to adjust to that impartial and emotionless presentation. Some of today’s synthesisers can deliver a degree of prosody, but “in order to decide what intonation to give to a word the system must have an understanding of the domain” (Dix et al. 1998). If the feedback to the user is only a relatively small set of non-changing messages, a human speaker could be recorded and the messages played back at choice yielding in much more acceptable speech. But the dynamic production of speech still requires huge efforts and is not about to be solved satisfyingly in the near future.

We can conclude that using speech for interfacing between humans and technology is ambivalent but seems to be the best channel we can effectively use. On the one hand we can rely on the highly elaborated functions of the human brain to instantly convey messages. On the other hand, interface designers have to cope with the humans’ expectations and implicit reasoning that are not even conscious to them. Non-attentive behaviour displayed through body movement, posture, facial expression, and gestures all contribute to what the persons in a conversation perceive from each other. Information transferred through the unconscious behaviour conveys a great deal of meaning, personal feelings and context.

3.2.3. Vision

For social interaction, aside from language, humans use the visual system as the predominant channel for information transfer. Thus the computer interface should actively use sight as well. Observations from the surrounding world may complement the information gathered by sound or it may add entirely new aspects. Looking at a person who shrugs the shoulders saying "Hmmmm" in a monotonous way may tell that this person is probably disappointed or disinterested. The same person would be considered to by satisfied and happy when making the same sound while being observed eating some ice-cream with a smile on the face. Consequently the agent itself should have the ability to make some movements and exhibit certain displays of emotion. This would help others to understand how the character feels and what some words really mean on a certain 'emotional' background.

3.2.4. Visual Speech

Speech as a multimodal phenomenon is supported by experiments indicating that our perception and understanding are influenced by a speaker's face and accompanying gestures, as well as the actual sound of the speech. Many communication environments involve a noisy

auditory channel, which degrades speech perception and recognition. Visible speech from the talker's face (or from a reasonably accurate synthetic talking head) improves intelligibility in these situations. Visible speech also is an important communication channel for individuals with hearing loss and others with specific deficits in processing auditory information.

The number of words understood from a degraded auditory message can often be doubled by pairing the message with visible speech from the talker's face. The combination of auditory and visual speech has been called super-additive because their combination can lead to accuracy that is much greater than accuracy on either modality alone. Furthermore, the strong influence of visible speech is not limited to situations with degraded auditory input. A perceiver's recognition of an auditory-visual syllable reflects the contribution of both sound and sight. For example, if the ambiguous auditory sentence "My bab pop me poo brive" is paired with the visible sentence "My gag kok me koo grive", the perceiver is likely to hear, "My dad taught me to drive". Two ambiguous sources of information are combined to create a meaningful interpretation, the McGurk-effect (McGurk and MacDonald 1976).

There are several reasons why the use of auditory and visual information together is so successful. These include a) robustness of visual speech, b) complementarity of auditory and visual speech, and c) optimal integration of these two sources of information. Speechreading, or the ability to obtain speech information from the face, is robust in that perceivers are fairly good at speech reading even when they are not looking directly at the talker's lips. Furthermore, accuracy is not dramatically reduced when the facial image is blurred (because of poor vision, for example), when the face is viewed from above, below, or in profile, or when there is a large distance between the talker and the viewer (Massaro 1998).

Complement of auditory and visual information simply means that one of the sources is strong when the other is weak. A distinction between two segments robustly conveyed in one modality is relatively ambiguous in the other modality. For example, the place difference between /ba/ and /da/ is easy to see but relatively difficult to hear. On the other hand, the voicing difference between /ba/ and /pa/ is relatively easy to hear but very difficult to discriminate visually. Two complementary sources of information make their combined use much more informative than would be the case if the two sources were non-complementary, or redundant (McGurk and MacDonald 1976).

The final reason is that perceivers combine or integrate the auditory and visual sources of information in an optimally efficient manner. There are many possible ways to treat two sources of information: use only the most informative source, average the two sources together, or integrate them in such a fashion in which both sources are used but that the least ambiguous source has the most influence.

Perceivers in fact integrate the information available from each modality to perform as efficiently as possible. Many different empirical results have been accurately predicted by a model that describes an optimally efficient process of combination (Massaro 1998).

3.2.5. Other senses

We have seen that sight and sound are the dominant senses that detect and transmit most of the information. Tactile feedback is also important in improving interactivity. Without the hands humans would not have been able to form utilities that were central to his development. Tactile feedback forms an intrinsic part of their operation, and even today in the electronic office we handle many things that require holding, e.g. pens. On the other hand taste and smell are the least used of our senses. Their use is more for receiving information than for communicating it, and they have been difficult to implement into computer systems up to now. The secondary nature of those senses "tends to suggest that their incorporation, if it were

possible, would lead to only a marginal improvement" (Dix et al. 1998). With this reasoning we do not consider taste and smell in this paper any further.

3.3. Anthropomorphism

During the last chapter, we could see that non-verbal behaviour convey a great deal of meaning and context. If we don't use the non-attentive cues we probably loose the opportunity to set up a efficient communication channel between human and computer. Today's DM interfaces can't fulfil our demands concerning a natural interaction. We could learn throughout the last 50 years that technology has given massive computational power to us. With this advantage, we now can development interfaces that were not possible some years ago – human like interfaces. With anthropomorphic interfaces we use the natural ability of the human to communicate and co-operate with his species. Their is no interfacing needed anymore to translate between technology and human user, given that the interface follows the rules and conventions of human-human conversation. The human is freed from learning the interface and can instead concentrate on the task or simply enjoy interaction.

But research on human-like characters is seen critical by some researchers. They question if anthropomorphism is appropriate as user interfaces paradigm and what function that should serve (Maes and Shneiderman 1997; Shneiderman 1998; Shneiderman 2002). Any new technology should be significantly better than the existing solutions. Thus we have to evaluate this new type of human-computer interface.

One argument brought against anthropomorphic interfaces is the confusion they induce in users. With most of them, the appearance promises human-like interaction whereas their actual behaviour today is very much predictable and seems scripted. Additionally they lack the whole palette of non-attentive conversation cues like gaze behaviour, turn-taking and turn-giving, adaptation to different situations and different interaction partners. Another argument is that they cause slower response time of the user, take control away from the user in applications where it is essential (e.g. the Microsoft Helper Agent in former MS Office products) and thus have never been successful in the past.

Cassell argues in favour of giving the interface a human-like appearance, stating that "only conversational embodiment … will allow us to evaluate the function of embodiment in the interface" (Cassell, Vilhjálmsson et al. 2001). In their opinion well-designed agents have a human-like appearance and thus might address particular needs that are not met in current interfaces. Possible outcomes would include ways to make dialogue systems robust despite imperfect speech recognition, to increase bandwidth at low cost, and to support efficient collaboration between human and machines, and between humans mediated by machines. "This is exactly what bodies bring to conversation" (Cassell, Vilhjálmsson et al. 2001).

Although we will use the term *Embodied Conversational Agent* that was coined by the MIT Media Lab group around Justine Cassell (Cassell, Bickmore et al. 1999), different researches suggested other terms like *Social Agents* (Parise, Kiessler et al. 1996) and *Interactive Virtual Humans* (Gratch et al. 2002).

3.4. Application Areas

A misconception of social interaction applied to technology is that we could improve all interfaces by making them explicitly social. The failure of the Office Assistant in Microsoft's Office package is a supporting example. It was meant to providing task-specific help, status information and suggestions through a small animated character and dialog bubbles in screen space. The communicative agent should increase comfort of the system, ease of use,

efficiency in accomplishing one's task, or simply the pleasure in performing them. Instead of delivering some benefits to the user; "most users find these agents annoying, intrusive and distractive" (Doyle 1999). Consequently they disable them far more often than not.

But there have been promising results as well. In the domains of entertainment, information access and education (Badler et al. 2002, Bouras et al. 2000, Cassell et al. 2001, Fabri et al. 1999, Johnson et al.) researches found positive effects of social computing. This literature sheds some light on the consideration where and when a communicative agent is a good idea. (Doyle 1999) suggested that it is not a question of bad design or a bad approach in general but that the agents have to be chosen for the appropriate domain.

Especially with mechanical tasks, like building a spreadsheet, where the user knows the nature of the task and how to accomplish it, he is not likely to want a discussion about it. He needs a direct manipulation interface that supports a fast, painless and transparent operation. Social interfaces are nothing like that; they involve explanations, clarifications, redundancy in instructions, good for teaching somebody but intensely interfering when doing it yourself.

3.4.1. Conversational Agents

Doyle (1999) defines that "conversation implies an exchange of information and ideas, is literally 'keeping accompany with' one another". This holds for tasks where agents accompany us, like entertainment (Yoon, Blumberg et al. 2000; Laird 2001), therapy (Marsella, Johnson et al. 2000) and marketing (André, Rist et al. 2000; Cassell, Bickmore et al. 2000).

ECAs could be beneficial in other areas as well, but that depends on the user's demand and willingness for social interaction. For example customer support by telephone aims for maximum users' satisfaction through a transition from endless lists of numeric menus to a more social interface. In *computer supported collaborative work* (CSCW) agents could help co-workers with retrieving shared information, deliver information in digestible form, or integrate information from diverse sources – agents as mediators and facilitators of computer support.

There is a certain category of tasks for which it is more important to produce a satisfying result than to have a correct result in the end. In fact there might by no 'right' result but one that satisfies the user. (Doyle 1999) argues that communicative agents can have their share in producing a positive outcome by changing the nature of the task to make it more pleasant, and by encouraging the user's belief that the final choice is acceptable. Commercial applications arise in areas where preference is more important than substance. Findings in the field of real estates for information access (Cassell et al. 2001) support this direction of research on applications with main focus on user satisfaction.

It is reasoned that agents do not satisfy when employed to help with the pure decision making processes in commercial environments. They might be analytically strong but are too limited in their knowledge of how to achieve reasonable compromise between contrary constraints. Support comes from psychology that claims that human decision making is perceptual-recognitional (based on earlier experience) not analytical (Klein 1987). This is a matter of the different frameworks humans and computers exist in: the perceptual-cognitive and the computational one (see chapter 2.2). Thus it is hard to tackle such problems with computer logic.

However communicative agents can be valuable as guides or advisors. Dialog is the central factor for these tasks: "The value of conversation lies in synthesis; through conversation we clarify out thoughts, we receive criticism and suggestion, we learn new information, our perception changes" (Doyle 1999). An agent can make suggestions, offer advice, and present

facts to assist users who are attempting to make choices about which they are unsure. With personality and displays of emotions the agent can sympathise with and even reassure the user, especially if it takes on an authoritative role in the domain, e.g. as teacher. Consequently applications of this type are of educational purpose.

3.4.2. Pedagogical Agents

The Generic Pedagogical Agent would perform three main functions. As a facilitator, it helps direct the student through the learning environment in the manner best suited to each individual. As a tutor, it promotes of active learning by offering facilities and exercises which help the student learn to teach her- or himself. As an advisor, it displays some emotional responsiveness and problem solving capability. Our field of interest is presentation and information delivery, not testing what was learned, therefore tutoring is neglected.

As a facilitator, the pedagogical agent would act as a interactive gateway to the many features of the presentation system such as instructional presentations, paced exercises, examples, demonstrations, a database of common misconceptions. It will be able to respond to direct student queries via voice as well as offer suggestions to related issues and areas with adjacent topics - both reactive and proactive behaviour. As an advisor, the pedagogical agent displays a degree of emotional intelligence. Using a combination of reflective or active listening and offering suggestions which resources may help answering the customer's question, the agent would support motivation. Combinations of both modes are possible, for example by mixing the abilities or by switching between the roles.

As a computer controlled agent is very adaptable it can take different roles, not only the one of the teacher, but also the one of a team-mate, a student etc. Thus different users can benefit from agents in tasks normally requiring a human partner. The complexity of tasks is raising, research and production facilities of global corporations are not co-located anymore and collaborative work gains momentum – and virtual environments for joint training can be constructed for remote and local team members alike. Members of a team and the necessary instructors would be expected to be human-like resembling the situation in reality and could be implemented as agents.

Recent studies have produced results which indicate a variety of advantages in implementing interface agents as part of educational applications. A study by (Klein 1999) has shown that an interface agent designed to support users in managing and recovering from negative emotions can encourage user persistence at a difficult task. Presenting humorous or interesting facts or having a likeable outfit, pedagogical agents help to lift the pressure from the user and to keep up the engagement in the learning experience that may yield in more interest in the subject to learn (Lester, Towns et al. 2000). Learning environments become more interactive and engaging if agents can employ their social quality (Johnson, Rickel et al. 2000). An animated character as support for training purposes, e.g. to validate maintenance procedures, was successfully employed by (Badler et al. 2002).

The technical sophistication of animated pedagogical agents has progressed rapidly. STEVE, a 3D animated agent, can interact with learners in individual and team scenarios (Rickel and Johnson 1998). PPP Persona is able to generate tutorial presentations of Web-based learning materials (André et al 1998). Cosmo is able to generate critiques and explanations using a combination of speech and emotive gestures (Towns et al 1998). Early empirical results show that these agents can enhance the learning experience and improve its effectiveness (Lester et al. 1997). They demonstrated that the presence of a pedagogical agent by itself had a strong positive influence on the learning rate of the participants - the *persona effect*. The representation of computer generated characters as teacher/co-student/team-mate and not as

impersonal text display can make learning more pleasant and memorable. As agents with rich expressions facilitated the highest rates the authors postulated the *corrolar of expressiveness*. Especially the agents with rich expressions have a high positive influence on the learning experience. Expressiveness through embodiment in pedagogical agents can be employed to guide the user's attention to relevant features, e.g. through body language, gaze direction or simple pointing.

A major advantage of computer controlled virtual environments with an agent is that the user can observe the demonstration from different angles and that it can be repeated with different parameters. With the simulation of workspaces the user is not restricted through the limited affordances in the real world anymore. Complex systems that are not only a copy of the reality can be created with additional objects, situations etc. and could exceed the possibilities of actions in real world. These virtual or augmented environments provide the opportunity to simulate critical situations without the danger of loosing precious technology or even lives. With pedagogical agents the teacher – student relation can be transferred to a secure and task relevant environment without giving up human-like interaction.

Conclusion

The distinction between conversational agents and pedagogical agents is not as clear as it might seem, because agents of both types must have some means of communicating with the user and must be specialised in their field of application. Today, conversational agents have more sophisticated abilities to communicate naturally with the user whereas pedagogical agents have advantages in the structured and planed presentation of information to accomplish a specific goal and in controlling and correcting the user's actions. Findings combined from both areas of research may ultimately yield in a respected and supportive pedagogical agent as well as a truly believable and helpful conversational agent.

In the last chapters of this section we have learned what types of agents can be distinguished and in what application areas they are likely to be used. The section should have given an insight into the complex multi-disciplinary approach we have to take when considering and developing agents. The next section will shed more light on the what has been done to incorporate the different modalities into human-computer interfaces. We will look at a short history of multimodal systems and will point out the specific novelties and raised issues. The following section is not purely a technological view on multimodal system but should sensitise the reader to general problems of and approaches to multimodal systems. It will be a good background when looking on implementations of multimodal systems in the next but one section.

4. History of Multimodal Systems

In face-to-face conversation people generate and understand a wide variety of speech, gesture and non-verbal communication cues. There has be a long history of research attempting to give computers the same capability. The first systems were command and control type interfaces that had little dialog and literally no conversational behaviour. Communication with technology that happened to use natural language cannot be regarded as a conversation, i.e. social communication, because the social dimension is absent. Example are information booth or kiosk systems that provide timetables, weather information or the latest news in response to speech and touch screen gestures.

Bolt's "Put that there" (Bolt 1980) can be regarded as the first multimodal interface. The user points at some object in screen space and commands the computer by voice to move the desired object to another location (see Figure 8).

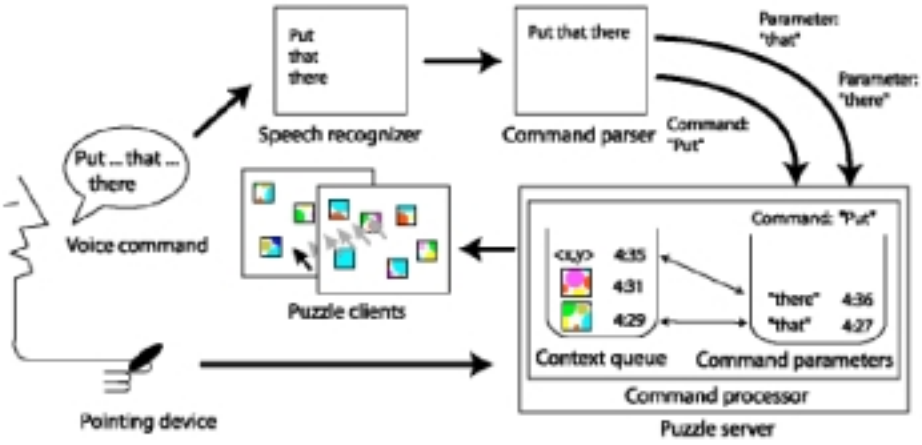


Figure 8: Principle of 'Put-That-There' applied to puzzle solving (Courtesy Harada et al. 2003)

The systems introduced cross-modal co-ordination in multimodal user input. It afforded multi-modal references (speech and *deictic gestures*, i.e. directing your attention to a spatial location) to single objects. A deictic gesture may in the simplest case produce a single point on the screen where the person pointed. This gesture could be accompanied by speech or a key press that is predefined to mean "multiple reference." A perceptual grouping algorithm can make interpretation of the gesture both independent of the gesture's form and of the input method used: references can be made with a mouse, touch screen, data glove or even gaze (Koons & Thórisson, 1993) - these will all look equivalent to the computer. A recurring problem in such interaction, however, is the inability of the computer to "see" the world in the same way as people do. For example, groups and groupings of objects that are obvious to users are invisible to the machine.

A key issue in 'Put-that-there' is binding the utterances "That" and "There" to pointing locations. This is done by simply picking out the location of the pointer at the instant moment when the relevant word is uttered. This approach is a specific type of what is referred to as "late fusion" or "semantic level fusion" (Oviat et al. 1996). A general problem in the system is, what Oviat et al. (1996) describe in their work: users engaging in pen/voice multimodal interaction prefer to use speech input as a descriptive tool, e.g. referring to objects by name or property, not only by deixis. In this way, people naturally refer to sets of objects and out of view objects, which is not possible in 'Put-That-There' neither.

Extending the work on deictic gestures, Bolt & Herranz (1992a) describe a system that allows a user to manipulate graphics with semi-iconic gestures. *Iconic gestures* are the kinds of gestures in which a body part, often the hands, play the part of another object for the purpose of demonstration. An example would be moving your hand forward, palm down and saying "The car drove like this" meaning that the car moved in the same way your hand does. Supplementing their research, they focused at two-handed input and gaze in a later work (Bolt & Herranz 1992b).

Koons et al. (1993) extended the work of Sparrell (1993) to integrate gaze, iconic gestures and speech. In the *Iconic* system, utterances can be mixed with freeform gestures: if a user speaks the words *"Move the chair"* while showing direction and amount of motion with a hand gesture, the system can execute the action without any further input. As drawback the system employs an e-mail style of interaction (construct and send command, wait for response) to minimise failures in the interaction sequence (Sparrell & Koons 1994). This is a major disadvantage in terms of intuitive user interfaces.

The presented work up to now concerned the input channel to computer systems. But multi-modality is both ways: input and output. Consequently Waters (1987) concentrated on the output mode. He proposes a muscle model for faces that is not specific to facial topology and that is more general for modifying the primary facial expression. In his model, the facial display is realised by local deformation of polygons representing the face. The process generally involves determining a mapping from text (orthographic or phonetic) onto visemes by means of vector quantisation (Morishima & Harashima 1991). Waters (1987) could show that the simulation of actions on muscles underlying the face looks more natural than manipulating the vertices directly. With this work it became possible to create lip movements synchronised to the synthesised speech signal. Implementing the basic functionality of human faces into computer representations brought computer generated faces closer to their natural counterparts, e.g. to simulate emotional display like disgust or joy.

Synthesised speech is normally generated by speech synthesisers from some text input. Text-driven facial animation like with Waters' face model was not smooth and rather rigid until Waters & Levergood (1993) introduced 'DECFace'. "The unique feature of DECface is the ability to generate speech and graphics at real-time rates, where the audio and the graphics are tightly coupled to generate expressive synthetic facial characters." (Waters & Levergood 1993) In order to achieve real time performance (15 frames per second with texture mapping), this approach compresses Waters 3D tissue model to a simple 2D wire frame representation of the frontal view (Figure 9). To animate lip movement a set of 55 viseme mouth shape key positions is utilised with a physically motivated non-linear interpolation. There is no eye or head movement. To add a level of dynamic realism, the eyelids are animated. But they are neither synchronised to any mode that is currently used to convey or complement meaning nor have any function to direct the discourse (e.g. initiate turn-taking).

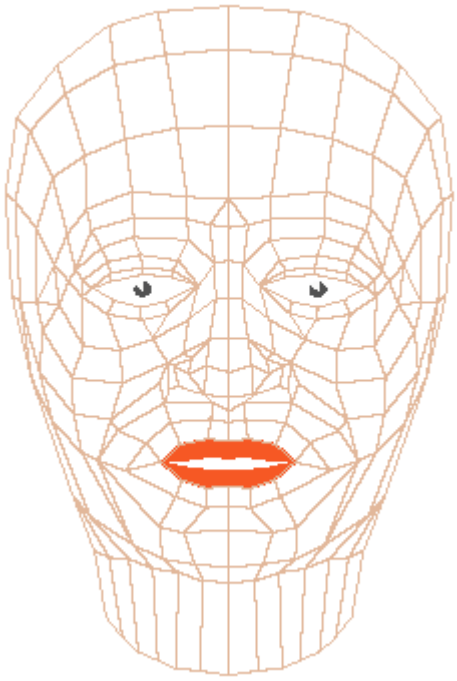


Figure 9: 2D polygonal representation of the face (Waters 1993)

As Oviatt (1996) found out multimodal interfaces have clear task performance and user preference advantages over speech only interfaces, in particular for spatial tasks such as those

involving maps. He could show that when users are free to interact with any combination of speech and pen, a single spoken utterance may be associated with more than one gesture (iconic and deictic). For example, a number of deictic pointing gestures may be associated with a single spoken utterance: ‘calculate distance from here to here’, ‘move this team to here and prepare to rescue residents from this building’. While users are not necessarily prone to make multimodal inputs, they can still integrate complementary output or use redundant output in noisy situations (Oviatt 1999). This clearly points out that multimodality is advantageous in any case, even if people cannot employ them as input.

The linguistic unification approach to multimodal integration proposed by Johnston et al. (1997) overcomes the limitations of previous approaches in that it allows for a full range of gestural input. Unlike speech-driven systems (like Bolts 1980 and Koons 1993), it is fully multimodal in that all elements of the content of a command can be in either mode. However, while this approach provides an efficient solution for a broad class of systems, there are significant limitations on the expressiveness and generality as a wide range of potential multimodal utterances fall outside the expressive potential of the architecture. Johnston (1998) show how unification-based approaches can be scaled up to provide a full multimodal grammar formalism and thus tackle the problem of synchronised integration of modes. These linguistic based systems have the drawback of depending on top-down grammatical parsing, which is not robust to disfluency, and require the user to learn their grammar and vocabulary.

Being just a short introduction to the development of multimodal systems we could demonstrate what kind of potential problems the researchers have to deal with. Solutions to some are proposed in new architectures and systems that are presented in the next chapter and its subsequent paragraphs.

4.1. Previous Embodied Agents

4.1.1. DECface

Probably the first ‘talking head’ mimicking human like interaction was DECface developed by Waters (1995) which uses DECtalk (Bruckert et al. 1983), a TTS (Text-to-Speech) System developed by Digital Equipment Cooperation. Interaction is mediated by a colour computer monitor on whom a face represents the agent. To achieve pictorial realism the face was made by constructing a geometric wire-frame model (see Figure 9), and mapping a digitised picture of a real human onto it. The gaze is fixed. During speech acts, the mouth is animated by applying the mouth posture (viseme) corresponded to the current linguistic unit (phoneme) and transitions between successive mouth shapes are interpolated and thus smoothed. Only language provides a channel for communication, either transferred through typed text (as input from the user) or through synthesised speech (as output from the talking head) with an acceptable rate of comprehension. The on-screen graphical representation does not expose any listening behaviour.

4.1.2. Gandalf

An early character with multimodal input and multimodal behaviour output was Gandalf (Thórisson 1996). Gandalf is embodied as a virtual hand and a face that appears 2D on a monitor beside the big screen for presentation (Figure 10). It verbally explains the solar system to the user how may ask questions and point to a desired location using a body tracker, an eye tracker and a microphone. Output is in the form of natural speech, deitic gestures, and beat gestures (short formless wave of the hand). Gandalf’s behaviour is not scripted or selected from a set of stock responses. The behaviour rules are based on research findings for human face-to-face interaction in the psychological literature on human-human interaction.

Figure 10: Gandalf – early character with multimodal input
© Media Lab, MIT, Boston

In the same way, Gandalf's dialog skills are modelled with data from the psychological and linguistic literature. In particular the following multimodal input behaviour is recognised: (a) Eyes: Attentional and deictic functions during speaking and listening, (b) Hands: Deictic gestures (c) Voice: Prosody (timing of partner's speech-related sounds, and intonation) and speech content, (d) Body: Direction of head and trunk and position of hands in body space, (e) Turn-taking signals: co-dependent and/or co-occurring multimodal events, such as intonation, hand position and head direction, and combinations thereof.

Gandalf's architecture produces real-time multimodal output in all dimensions of the input as well as: (a) Hands: Emblematic gesture (hand motion synchronised with speech production), (b) Face: Emotional emblems (e.g. raising eyebrows when greeting, facing speaker when listening), (c) Body: Emblematic body language (e.g. nodding, shaking head), (d) Speech: Back channel feedback. Input is processed in real-time and output timed with the ongoing conversation. This results in behaviour coherent with the spoken words and highly relevant to the user's actions, even under variability and individual differences. Thus Gandalf is capable of real-time multimodal, face-to-face interaction with a user (Thorisson 1996).

4.1.3. Cosmo – The Internet Advisor

The INTERNET ADVISOR is an interactive, screen based learning environment in which the agent Cosmo helps the user to accomplish a learning task. Cosmo is present in the scene constantly, observes the action of the user and provides hints, explanations, and help to him/her (see Figure 11). The agent can answer questions, can jump in if the user does not know any further, and can point out false steps in a solution attempt. Cosmo's interaction with the environment consists of three components: (a) Movements, (b) Language, and (c) References. The relationship between all three constitutes Cosmo's behaviours.

Cosmo can move around in the learning environment but he cannot actively manipulate any object even if it seems to be able to. The animations shown are not calculated in real-time instead they consist of pre-rendered and manually optimised sequences that can be combined to complex movements. Two types of sequences are distinguished: full body behaviours, e.g. greetings and farewell, and compositional behaviours to support explanations given, e.g. through head nods.

Analogous to the body behaviours Cosmo's speech acts are composed from predefined speech components. The agent 'knows' two hundred forty utterances from different length that were professional recorded (Lester, Towns et al. 2000). Animation of the lips is achieved through transition between pre-rendered lip-shapes. A heuristic working on the frequency distribution synchronises utterances and lip-shapes, e.g. opening the mouth if amplitude is high i.e. loud parts in speech acts (Lester, Voerman et al. 1997).

Figure 11: The Internet Advisor present on the user's screen
© North Carolina State University

Cosmo tracks the user's status and development in the learning environment to be able to build a history of known objects and how was referenced to them. By using a world model spatial knowledge is incorporated into the reference system, i.e. Cosmo determines when to say e.g. 'these' and 'those'. Such the agent can make use of references in speech acts (e.g. 'This' instead of saying again 'The subnet') enriched or complemented by deictic gestures (pointing to the object). In case of ambiguity, e.g. two objects are too close together to guarantee a clear identification, Cosmo would get nearer and point directly to the referenced one. It is important that the 'beat' of a gesture corresponds to the reference in the speech act. The aim of correctly synchronised deictic references in speech or through gestures is to produce natural and most of all believable communicative behaviour.

To resemble natural behaviour the display of emotions plays an essential role. Cosmo's embodiment of head, body and arms with hands implements the features known by professional actors and animators to express emotions very effectively: the face with eyes, eye brows, and mouth, the position of the head, the postures of body, arms, and hands. Believing in the corrolar of expressiveness the agent exhibits slightly exaggerated behaviours. Cosmo maps emotional behaviour onto eight different classes of speech acts (Lester, Towns et al. 2000), e.g. congratulatory act or assistance. When uttering a specific sentence one corresponding behaviour from the associated class is randomly selected.

Cosmo is able to construct emotionally augmented speech acts with correct deictic references and can give appropriate help in situations of need. Interaction with the agent is solely

dependent of the user's manipulation of the learning environment and a predefined catalogue of questions. The agent would not react on the user's questions while it gives an explanation. Cosmo is not able to have a natural conversation over the task as the agent lacks speech recognition and language processing. It simply explains facts and poses questions to the user to think about that are then answered in the agent's ongoing explanation. These features distinguish Cosmo as classical pedagogical agent with some conversational abilities.

4.1.4. Olga

The screen based assistant Olga is an agent enhancing the traditional consumer information booth through multimodal interaction with voice and visual interaction for navigation (Sundblad and Sundblad). Olga integrates spoken dialogue, 3d animated facial expressions, gestures, lip-synchronized speech synthesis and a graphical DM interface (Beskow, Elenius et al.).

Olga's representation is a cartoon-like anthropomorphic character (Figure 12) that is animated in real-time. It supports gesture, facial expressions and dialogue turn-taking, is able to visually refer to other on-screen graphics and may indicate the system's internal state (i.e. listening, understanding, uncertain, thinking/being occupied).

Figure 12: Olga's representation on the screen
© KTH Centre for Speech Technology, Stockholm

Regarding input, Olga affords clicking on buttons in its DM interface or using domain specific spoken language (according to a written scenario) that refers to objects already known by the agent. A knowledge database provides task related information and to some small extent world knowledge. Olga responses with natural language, pointing gestures (deixis) and gaze manipulation, potentially using all three to make references to graphical items visible on screen. Additionally the environment may display static icons for some particular actions.

Olga's dialog system composed of system goals and dialog strategies has the central aim to foster communication with the user and efficiently obtain information for database searches. Goals like minimising dialog breakdown and maximising dialog progress yield in behaviours like confirmation and requests for further information respectively. Through a history Olga

'remembers' what was said and done and can vary its behaviour over time - the dialogue strategies are dynamic (Beskow and McGlashan 1997).

All animations are done using combinations of basic non-rigid deformation on a parameterised polygon model. Articulation makes use of the parameterised model. It is controlled by a rule-based text-to-speech system framework (Carlson and Granström 1997) and is implemented as model that accounts for co-articulation effects (Beskow 1995). Having calculated the parameter trajectories the animation of lips, jaw, and tongue starts in synchronisation with audible speech output that is governed by the same framework.

Complex transitions like body movements, non-speech facial expressions and gestures can be scripted and grouped together (Beskow, Elenius et al.). Then these more complex movements (e.g. "shake head and shrug"), that are possibly to be displayed dynamically depending on the content of the conversation, can be triggered by a simple procedure call. To make scripting more flexible, especially with gestures, templates can be parameterised. "Template based handling of facial expressions and gestures has proven to be a simple, yet quite powerful way of managing non-speech movements" because it "makes it easy to experiment with new gestures and control schemes." (Beskow, Elenius et al.) Arguments supplied to the procedure change the realisation, e.g. the duration of the movement or which realisation is selected from a set of alternatives. Olga's 'idle loop' while it changes behaviour is a good example for latter.

4.1.5. Steve

The SOAR Trainings Expert for Virtual Environments ('Steve') is an embodied agent for teaching users how to accomplish manual tasks (Figure 13). The user and the agent are immersed together in a simulated VE, where the user can perform actions via multi-dimensional input devices and can see their effects on the environment (Elliot et al. 1997, Rickel et al. 1999).

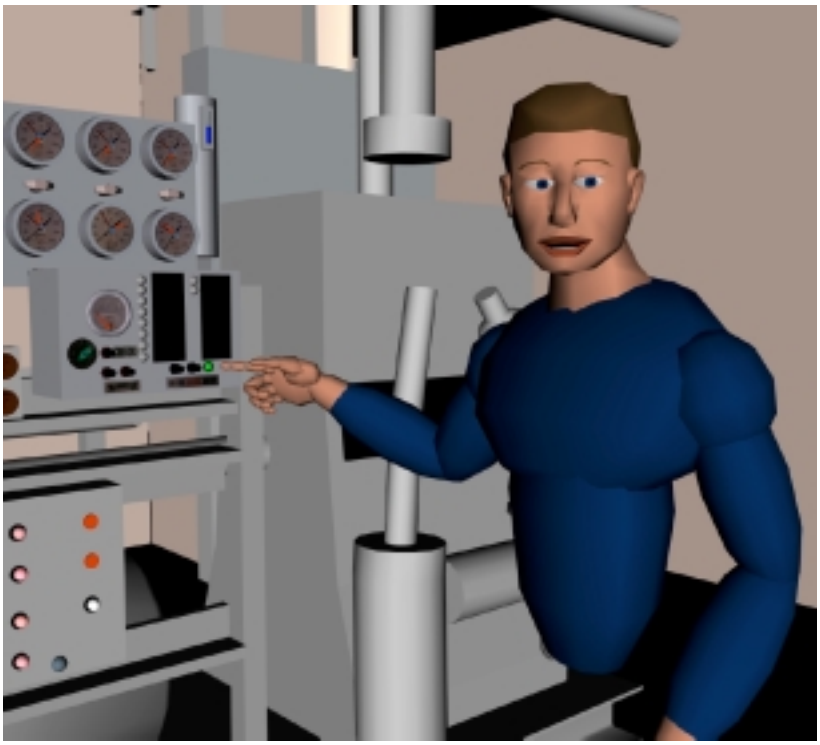


Figure 13: Steve pointing out a power light to the student
© Information Science Institute, University of Southern California

Steve's behaviour is not scripted. He has "domain-independent capabilities operating over a declarative representation of domain tasks" (Rickel et al. 2001). Steve is able to take the attributes of objects in the VE (e.g. if a switch has been turned on, or if a box has been moved), head position of the user in the simulation and which objects are in his view as input. He can move within the simulated environment and manipulate objects. Locomotion is realised through dynamically sequencing of pre-recorded snippets of motion captures. Communication with the user is possible through 'understanding' three predefined and simple questions (e.g. "why?") and the 'generation' of answers from pre-scripted fragments. The speech output via a text-to-speech system is displayed on the agent's face with a speak-like expression.

Steve's audio output is complemented with the following non-verbal behaviours: (a) Hands: neutral or pointing to an object, (b) Head: neutral or nods, (c) Eyes: Gaze direction to an object, the user or another agent. The gaze is influenced by the status of the agent itself: (a) Movements from one object to another, (b) Manipulation of an object, (c) Pointing to an object, (d) Direct gaze contact while approached by or speaking with the user, while waiting for him and while observing him during an (sensorial) action. These behaviours are employed during the demonstration period, and, if the end has reached or the user has pre-empted the demonstration, during the observation period. As goal-driven pedagogical agent Steve explains the rationale for his recommendations in terms of other relevant actions and goals. The student can ask follow-up questions about why these actions and goals are relevant until the rationale for Steve's initial recommendation becomes clear. In this way he is a good instructor.

On the other hand, Steve lacks a complementing lib-synch animation and non-verbal behaviour is literally non-existent. His architecture simply does not include any emotions. He does not recognise or exhibit non-verbal expressions or natural language. The dialogues are not rich in expressions and seem 'dry' even for the domain of task specific training. While this might be advantageous for a patient and tolerant collaboration, the teaching is emotionally flat – Steve cannot show any enthusiasm, cannot incorporated motivational cues and cannot distinguish between mundane and important instructions (e.g. put stress on some warning). Future implementation will acknowledges the human side of learning. It is planed to include Gratch's (2000) emotional model and the model for regulating emotional behaviour (Marsella et al. 2000) to animate expressive faces and human behaviour (Rickel et al. 2001). Eventually this may lead from a pedagogical agent to a more social one. Nowadays there is no such agent.

4.1.6. Towards Social Interaction

The 'Behavior Expression Animation Toolkit' (BEAT) is an architecture that enables an artificial animated human figure (for the procedure see Figure 14) to speak an animators' input typed text enriched by gestures, facial expressions and posture. The additional information is used to generate, control, and synchronise non-verbal behaviours with synthesised speech before it is sent to an animation engine. BEAT "extracts actual linguistic and contextual information from text in order to suggest appropriate gestures, eye gaze, and other nonverbal behaviors, and to synchronize those behaviors to one another" (Cassell, Vilhjálmsson et al. 2001). It generates appropriate non-verbal behaviours that are "assigned on the basis of actual linguistics and contextual analysis of the typed text, relying on rules derived from extensive research into human conversational behavior" (Cassell, Vilhjálmsson et al. 2001). When actually using BEAT. the author can specify some key words and phrases to look for in the conversation. That does not mean, that certain words are 'hard-wired' to certain gestures, but that the system 'knows', that it is a relevant word/expression and that it

should be stressed somehow. Through what behaviour this function is implemented is up to the system. This yields in natural appearing successions of animations. The toolkit is extensible in its rules (e.g. personality profiles, extra motion characteristics, scene constraints, and animation styles) and can be plugged into larger animation systems as the output format is generic XML.

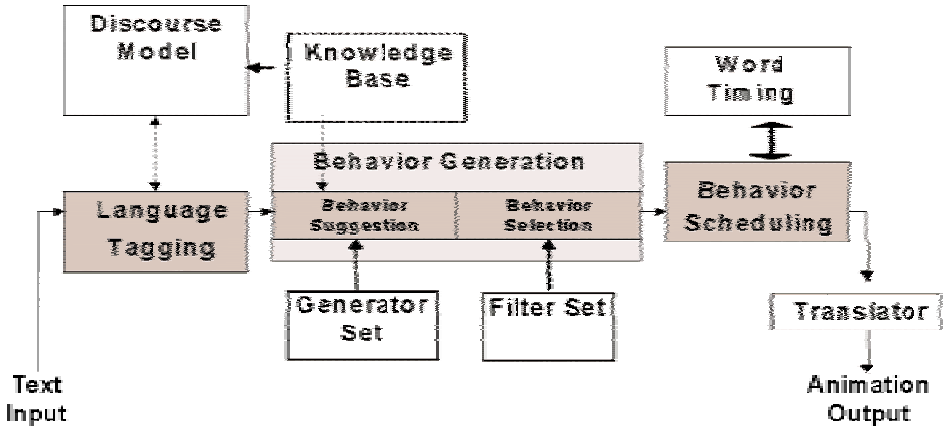


Figure 14: 'Behavior Expression Animation Toolkit': Text-to-Nonverbal Behaviour Module (Courtesy Cassell, Vilhjámsson et al. 2001)

BEAT was successfully used in follow-up projects, Rea and MACK. Similar to Gandalf these systems present a screen based virtual character that understands speech and gesture and answer in a multimodal manner. This time in a full 3d representation.

4.1.7. REA

With Rea – "Real Estate Agent" (Figure 15) – communication with agents even got closer to natural communication. The main focus of the system was to employ a functional conversational architecture that maps user input to conversational function, but it showed the importance of the social side of task-orientated dialog, e.g. small-talk. Small-talk is a common human behaviour for transitioning between otherwise disturbing and or confusing situations. Whenever solidarity with the user needs to be increased or the topic needs to be moved to a desired topic, Rea "implements the social, linguistic, and psychological conventions of conversation" (Cassell, Bickmore et al. 1999) to satisfy interpersonal goals. The system is designed to "conduct a mixed initiative conversation (…) while also responding to the user's verbal and non-verbal input that may lead in a new direction" (Cassell, Bickmore et al. 1999).

Figure 15: REA – a screen based virtual character that understands speech and gesture
© Gesture and Narrative Language Group, MIT Media Lab

Information contained in the speech acts of the conversational partner are compared to a dynamical knowledge base. If the fact is not known, i.e. it has not come up in the conversation yet, it is stored. If a certain object already exists, i.e. it was mentioned earlier in the conversation, Rea could reference this topic. The relation between dynamical knowledge extraction and the static domain knowledge makes it possible for Rea to bring new adjacent topics into the conversation. The ability of taking the initiative serves the expectation towards a natural conversation partner.

One main advantage of Rea is the ability to recognize and generate propositional and not only interactional information (see Chapter 3.5.1). Rea currently manages three main interactional discourse functions: (a) Acknowledgement of the user's present: posture, facing the user; (b) Multi-modal feedback: nodding, short statements ("I see"), paraverbals ("Hmm") or movements of the eyebrows; (c) Turn-taking and turn-giving function: tracking of who has the speaking turn, the possibility of interrupting verbally and through gestures, and the end of turn indication. Furthermore Rea employs functions that benefit both interactional and propositional components: (a) Greeting and farewell functions; (b) Emphasis function: recognition of accentuation of specific terms through beat gestures or pronunciation. In conjunction with the human-like, natural sized representation on a screen and the 'intelligent' knowledge representation the generation of propositional as well as interactional function is the basis to engage the user in an informative but interesting and pleasant conversation.

4.1.8. MACK

A more recent conversational agent, the "Medialab Autonomous Conversational Kiosk" (MACK) (Cassell and Stocky 2002) has additional spatial intelligence. The character has a life-size, knees- upward representation on a high resolution display (see Figure 16). MACK knows where it is situated in the environment and how this environment is structured. Thus the agent can relate directions or pointing movements to its own position or the position of other elements. ECA kiosks such as MACK engages the user in a shared environment, capable of interaction with shared object that are enhanced with computer generated overlays, i.e. MACK can highlight areas and draw paths on a physical map in front of the user. Interaction within the shared physical and information space that can be referenced provides

immersion of both kiosk and user in the actual physical space (Cassell, Stocky et al. 2002). Research indicates that kiosks serve a wide range of individuals, save time needed for physical staff to provide advice, and improve knowledge transfer (Steiger & Suter 1994).

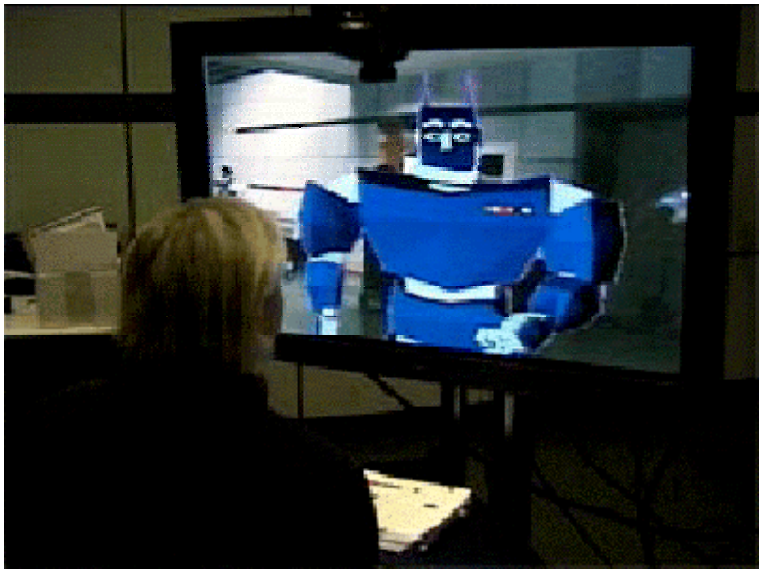


Figure 16: User interacting with MACK. (Courtesy Cassell, Stocky et al. 2002)

4.1.9. Summary

During this chapter we have investigated properties and characteristics of former and current EAs and the underlying architecture. It has become clear that the implementations heavily depend on the application they are used for. Some incorporate discourse functions (e.g. REA), others have strong abilities for training (e.g. STEVE). Nevertheless we can extract some higher order categories that can be used to assess the agents. Having all the details at hand, we chose the following abstract categories: (a) the type of agent, i.e. the main purpose of the system where "Ped" means pedagogical and "Con" means conversational; (b) the embodiment, i.e. with what parts of the human body; (c) the visualisation, i.e. how the agent is presented to the user; (d) which input and (e) which output channels the agents supports; and (e) which other features might by interesting. The reader will find the compiled list in Table 2.

System	Type	Embodiment	Visualisation	Input	Output	Other features
DECface	Con	Face only	Human-like, half life-sized, 2D on monitor	Arbitrary text	Lip-synch speech	
Gandalf	Con	Face & one hand	Comic-like 2D on monitor	Natural speech, simple pointing & beat gestures	Lip-synch speech, facial expressions, beat & deictic gestures, posture	
COSMO	Ped	Face, hands & arms, upper body	Comic-like 3D on monitor	Mouse	Speech, deictic gestures	
OLGA	Ped	Full-body	Comic-like	Natural	Lip-synch	

			3D on monitor	speech	speech, facial expressions, gestures	
Steve	Ped	Face, hands & arms, upper body	Human-like, 3D life-size in VE	Specific sentences through speech	Scripted speech, gaze, simple gestures, simple body language	Multi agent world, inter-agent communication
Rea	Con	Face, hands & arms, upper & lower body	Human-like, 3D life-sized on projector screen	Natural speech & non-verbal behaviour	Lip-synch speech, facial & body expressions, gestures	Turn-taking and giving
MACK	Con	Face, hands & arms, upper & lower body	Abstract, 3D life-sized, video based AR on projector screen	Natural speech & non-verbal behaviour	Lip-synch speech, facial & body expressions, gestures	Spatial knowledge, shares physical environment

Table 2: Embodied Agents - Overview and comparison of features

We have illustrated how Embodied Agents were designed, what specific properties they have got, what functionality they can offer and implicitly how the compare to each other. Some advantages and disadvantages of particular architectures and approaches were mentioned, but we have not assessed the usability of such systems and in what respect they have been tested and have proven to be effective. We will make up for this gap in the next chapter.

4.2. User Testing of Agents

As mentioned in the introduction and motivation, there is no formal guideline of how to assess interface agents or in our case the subcategory of embodied agents. This partly accounts for a lack of user testing in this new field of interface design. Another reason is that testing costs a lot of time and man power as it has to be set up all from scratch (especially for formal tests). From this background it is truly hard to find any sound user testing in literature. Most researchers simply do an informal assessment by asking the subjects to describe their impressions of and feelings towards the technology presented. This is not bad as such but it does not clearly answer the question if the UI under consideration is better as the existing ones. Keeping this in mind we will now look at the results of user testing.

4.2.1. DECface

In user test with the DECface by (Parise, Kiessler et al. 1996) researchers "observed that people's responses to the computer agent could be predicted from known principles of human behavior" reassuring the findings by (Nass & Steuer 1993). Research has found that personality attributes are influenced by people's physical appearance and voice (Warner & Sugarman 1986). The users exhibited more *impression management concerns,* i.e. giving less personal information to the talking face than through the text interface. On the other hand, they attributed more positive personality characteristics to the more pleasant looking agent as well.

In comparison to user-testing of former versions (Kiesler et al. 1996), co-operation with the realistic agent was much higher in the newer version of DECface. From analysis of the two

studies (Parise, Kiessler et al. 1996) infer that the newer agent "elicited more co-operation, not because it was more pictorially-real than the previous agent but instead because it was more human-like – better resembled, talked, and moved more like a normal human being".

4.2.2. Gandalf

Researches who used DECface as basis of their work found very early that it would be essential to add believable listening behaviour including head tilt, head nod, and paraverbals or 'vocalisations' ('hmm') to the agent (Parise, Kiessler et al. 1996).

Such a system was Gandalf, the early agent capable of multi-modal input and output. Tests showed that its dialogue performance has been rated high (Thorisson and Cassell 1996b). Users "preferred such a system to another embodied character capable of only emotional expression" (Cassell, Bickmore et al. 1999). They relied on the "interactional competency" of the system to negotiate turn-taking that was not present in the compared agent.

4.2.3. COSMO – The Internet Advisor

An informal study was designed to investigate (1) how well the spatial deixis approach produces explanations that are clear and helpful and (2) how an agent-based approach to deixis in learning environments compares with a non-agent-based approach. Results suggest that the spatial deixis framework produces clear explanations. "Most participants understood the agents advice most of the time." (Towns et al.) Although some wished it were less dramatic and suggested alternative organisations for the advice, its clarity was positively received. Subjects unanimously preferred the agent-based version over the agent-less version. Some suggested that a combination of agent gestures with other cues of guidance might be more effective than either in isolation. In general, the agent's motivating role was surprisingly strong, and subjects rated Cosmo as to be enjoyable and helpful (Lester et al. 1997).

4.2.4. Rea

In contrast to Gandalf, Rea (see Figure 17) can not only recognise and produce propositional information in a more elaborated form, but emphasises the interactional component of the conversational model. This way, Rea can handle repairs gracefully after misunderstandings. User testing showed that conversation with the newer system were comparatively more fluent (Cassell & Thorisson). Rea is more capable of making an intelligent content oriented contribution to the conversation. It is also „sensitive to the regulatory function of verbal and non-verbal conversational behaviors" and is "capable of producing regulatory behaviors to improve the interaction by helping the user remain aware of the state of the conversation" (Cassell, Bickmore et al. 1999). The engagement in social chit-chat is a way of reducing interpersonal distance and increasing trust between the user and the system.

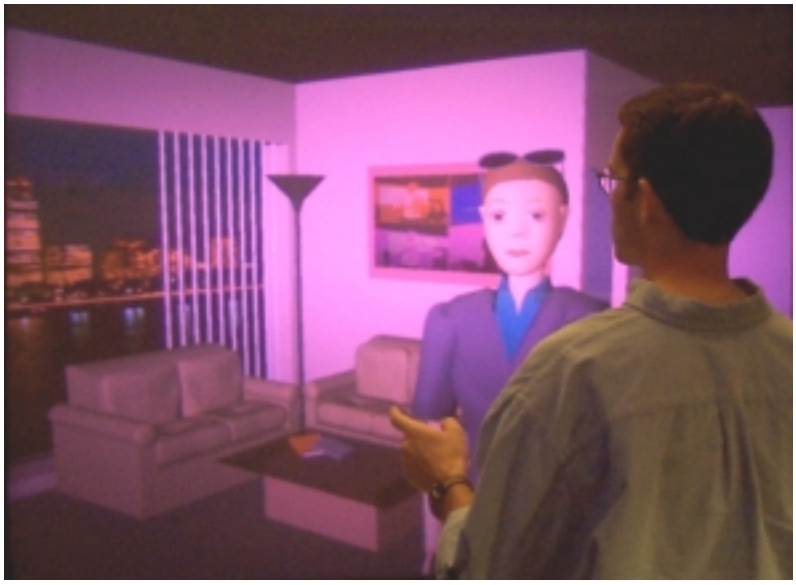


Figure 17: A human user in conversation with REA
© Gesture and Narrative Language Group, MIT Media Lab

4.2.5. MACK

During the practical use of MACK users interacted with it like with another person (i.e. when pointing in a direction, users turned around for a look and gave a nod as feedback). They trusted in information the system presented to them (Cassell et al. 2002). MACK also succeeded in engaging and entertaining users, but a formal evaluation was not found.

4.2.6. Results

Parise, Kiesler et al. (1996) and Welch et al. test users to determine what influence the factors pictorial realism, human-likeness and charm have on co-operation. They found out that pictorial realism positively affects the sense of presence in VR worlds and increases involvement. Involvement and presence may in-turn increase the own commitment and the impression of the other's commitment. Kerr & Kaufmann-Gillard (1994) have established commitment as an important mediator of co-operation in face-to-face discussions. If pictorial realism is essential to co-operation, than co-operation with less human-like characters should considerably drop off. Inexpertly the test results contradict these assumptions. The participants did not report a greater sense of presence in the more pictorial real condition. Neither did they feel that the environment was more natural or realistic. We can conclude that (pictorial) "realism … is not a condition for co-operation" (Parise, Kiessler et al. 1996). This is supported by former studies with even less realistically looking facial representations (Kiesler et al. 1996).

Social identity theorists claim that discussions amplify the feeling of being members of the same group (Hogg & McGarty 1990), and that group identity is increased among members of the same social category (Kramer & Brewer 1984). Thus our natural expectation that humans identify themselves more easily with other humans than with animals or lifeless object is theoretically supported.

Contrasting the former claim, one might consider another hypothesis: it could turn out that users tend to co-operate more with the charming but not necessarily human-like characters. Human-like agents might encourage social behaviour but might not be able to response adequately to it. User frustration about unmet expectations towards anthropomorphic agents on the one hand and a more playful, forgiving attitude towards non human-like characters on

the other hand could lead to favouring latter ones. Charm might favour co-operation even if the partner is not human-like.

Test results on interfaces confirm the expectations on human-likeness. Commitment and co-operation with anthropomorphic agents were significantly greater than with non-human like agents (Parise, Kiessler et al. 1996). Interestingly the hypothesis about the aspect of charm turned out to be true as well. Participants rated the non-human like agents as most ‘cute’ and trustworthy. The connection of both these findings yields an interesting result: high commitment (co-operation) does not correlate with social evaluation rating (liking). But “human-likeness moderates (is a conditions that influences) co-operation” (Parise, Kiessler et al. 1996).

Similar to the systems we have examined, most of the agents in the field are confined to the windows of computer monitor, to share a 3D space with their users. If an agent is implemented with virtual reality technology (like STEVE), it can share a 3D space with users. It shares, however, only a cyberspace and cannot get in touch with the real space. On the other hand, MACK can share physical space with it users. This agent, however, shares the space in a large computer monitor similar to a window.

We have seen from this section that agents may be beneficial in some applications and even may give joy to users doing that. As all the systems presented until now are screen based and not registered in 3D as we want to aim for, we will have a closer look on augmented reality and its special properties before we come to the implementation.

5. AUGMENTED REALITY AND AGENTS

5.1. The Essence of Augmented Reality

First, we want to introduce the notion of *Virtual Reality* (VR), then it becomes easier to explain what AR is. Virtual Reality, as given by (Aukstakalnis and Blatner 1992), is defined as “a computer generated, interactive, three-dimensional environment in which a person is immersed." There are many other definitions, essentially defined in terms of human experience VR is a mediated environment which creates the sensation in a user of being present in a (physical) surrounding. One of the identifying marks of a virtual reality system is the head mounted display worn by users. These displays block out all the external world and present to the wearer a view that is under the complete control of the computer.

Figure 18: A virtual object is superimposed on the real world

In contrast to virtual reality, augmented reality overlays the reality with synthetically images (see Figure 18) . Hence, the user can see the real world around him and in addition, computer-generated graphics that are composited with the real world. Mixing virtuality and reality happens using a *Head Mounted Display* (HMD) that is either see-through or overlays graphics on video of the surrounding environment (Figure 19). Instead of replacing the real world, we supplement it. Ideally, it would seem to the user that the real and virtual objects coexisted. The following list defines AR in short:

1. AR combines real and virtual.
2. AR is interactive in real time.
3. AR is registered in three dimensions.

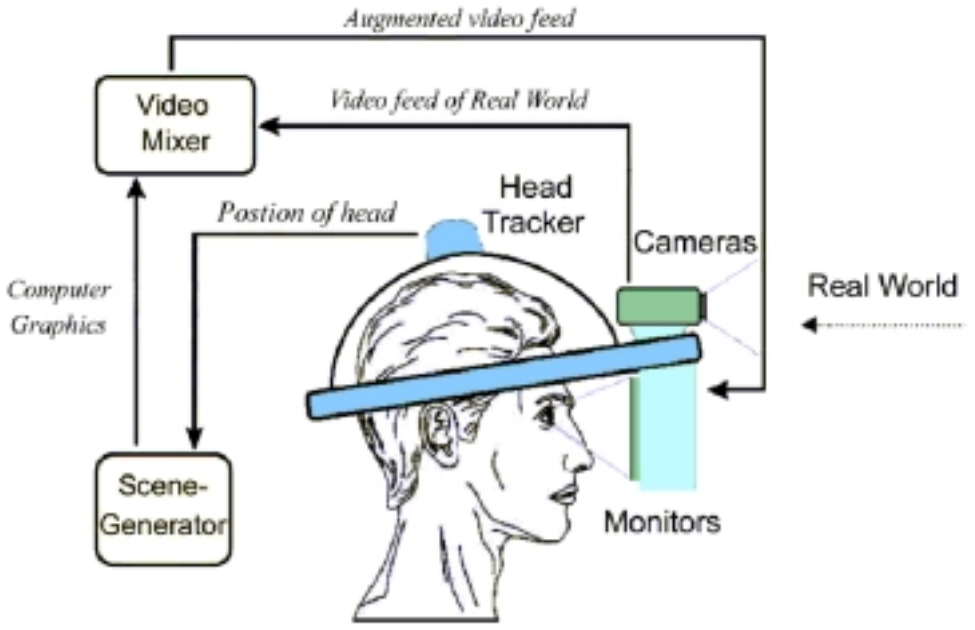


Figure 19: Principle of a video see-through HMD

AR and VR are in some respect quite similar but quite different in others. In Table 3, we have identified and compared the properties of VR and AR, respectively, along some key criteria.

Characteristics	AR	VR
Three-dimensional	Yes	Yes
Virtual immersion	Partly	Total
Sense control	Partly	Visual total others total/partly
Interactivity	Yes	Yes
Real world objects presence	Yes	No
Real time	Yes	Yes

Table 3: Properties of AR and VR environments

A very noticeable difference between these two types of systems is the immersiveness of the system. Virtual reality strives for a totally immersive environment. The visual, and in some systems aural and proprioceptive, sense are under control of the system. In contrast, an augmented reality system only enriches the real world scene. The user maintains a sense of presence in the real world. Otherwise he would not be able to co-ordinate his interactions with

objects and other humans. Similarities of AR and VR lie in the real time interactive style and the three-dimensional environment.

5.2. Augmented Reality as Interface

Augmented Reality offers a new quality of experience to the user. Single user Augmented Reality interfaces have been developed for computer-aided instruction (Feiner 1993), manufacturing (Caudell 1992) and medical visualisation (Bajura 1992). These applications have shown that Augmented Reality interfaces can enable a person to interact with the real world in ways never before possible. For example, Bajura et al. (1992) have developed a medical interface that overlays virtual ultrasound images onto a patient's body, allowing doctors to have "X-Ray" vision in a needle biopsy task (Figure 20). In Feiner's work (1993), users can see virtual annotations appearing over a laser printer, showing them how to repair the machine (Figure 21). In both of these cases the user can move around the three-dimensional virtual image and view it from any vantage point, just like a real object. For further detail, Azuma (1997) provides an exhaustive review of current and past AR technology and applications.

Figure 20: Biopsy supported by AR
(Courtesy Bajura 1992)

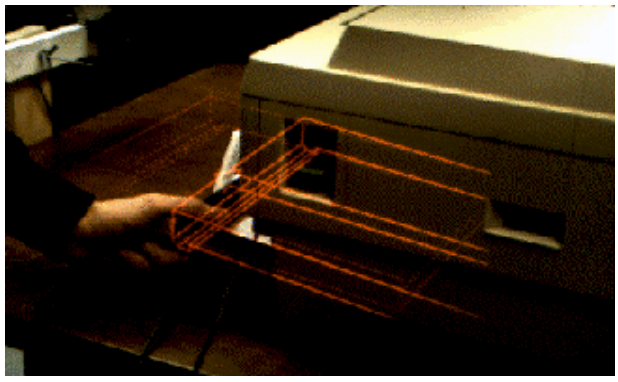


Figure 21: Repair support by AR
(Courtesy Feiner 1993)

Augmented Reality can also be used to enhance collaborative tasks. A good example of this is the Studierstube project of Schmalsteig et al. (1996). They use see-through head mounted displays to allow users to collaboratively view 3D models of scientific data superimposed on the real world (Figure 22). They report users finding the interface very intuitive and conducive to real world collaboration, because the groupware support can be mostly left to social protocols. The AR2 Hockey work of Ohshima et al. (1998) is very similar. In this case two users wear see-through head mounted displays to play an AR version of the classic game of air hockey. As they move a real mallet over a real table, they send a virtual puck towards each other's goals.

Figure 22: AR Interaction in Studierstube with scientific data
(Courtesy Vienna University of Technology)

AR technology has matured to the point where it can be applied to a much wider range of application domains, and education is an area where this technology could be especially valuable. The educational experience offered by Augmented Reality is different for a number of reasons, including:

· Support of seamless interaction between real and virtual environments
· The use of a tangible interface metaphor for object manipulation
· The ability to transition smoothly between reality and virtuality

The third property in this list makes up a new type of user interfaces: *transitional interfaces*.

5.2.1. Transitional Interfaces

Milgram (1994) points out that computer interfaces can be placed on a continuum according to how much of the user's world is generated by the computer (Figure 23). Moving from left to right the amount of virtual imagery increases and the connection with reality weakens. AR technology can be used to transition users smoothly along this continuum.

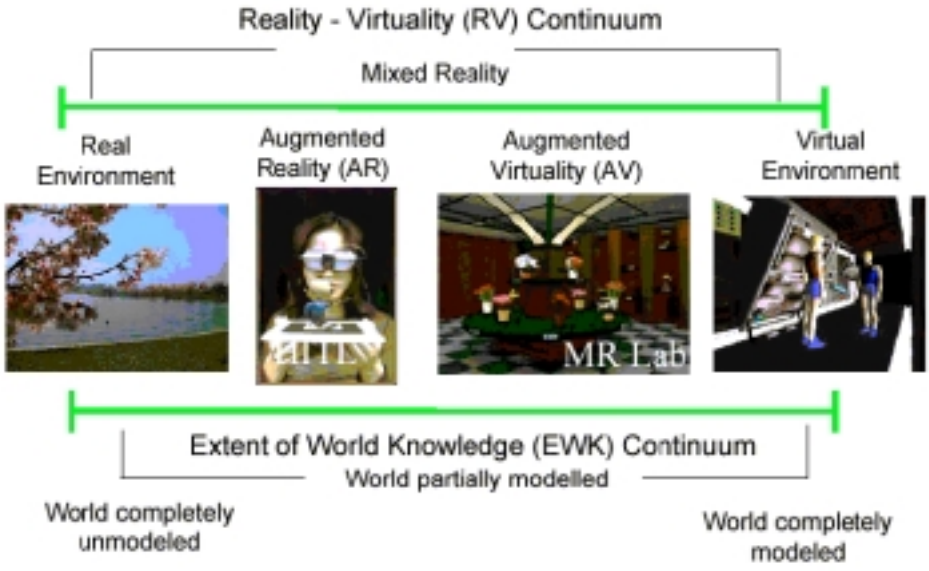
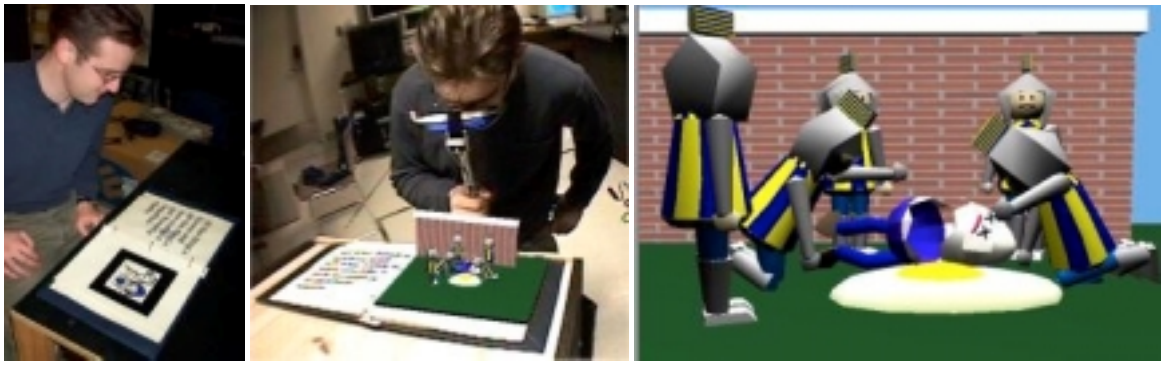


Figure 23: Reality-Virtuality Continuum (Adapted from Milgram et al. 1994)

A good example for transitional interfaces is the MagicBook (Billinghurst 2001). Young children often fantasise about being swallowed up into the pages of a fairy tale and becoming part of the story. The MagicBook makes this fantasy a reality by using a normal book as the main interface object. People can turn the pages of the book, look at the pictures, and read the text without any additional technology (Figure 24a). However, if they look at the pages through a handheld Augmented Reality display, they see three-dimensional virtual models appearing out of the pages (Figure 24b). The models appear attached to the real page, so users can see the AR scene from any perspective simply by moving themselves or the book. The models can be any size and are also animated, so the AR view is an enhanced version of a traditional three-dimensional "pop-up" book. Users can change the virtual models simply by turning the book pages. When they see a scene they particularly like, they can fly into the page and experience the story as an immersive virtual environment (Figure 24c). In the VR view, they are free to move about the scene at will and interact with the characters in the story. Thus, users can experience the full Reality-Virtuality continuum.



a: Reality b: Augmented Reality c: Immersive Virtual Reality

Figure 24: Using the MagicBook to move between Reality and Virtual Reality
(Courtesy Billinghurst 2001)

5.2.2. Tangible User Interface

In Augmented Reality there is an intimate relationship between virtual and physical objects. The physical objects can be enhanced in ways not normally possible such as by providing dynamic information overlay, private and public data display, context sensitive visual appearance, and physically based interactions. AR applications based on a *tangible user interface* (TUI) metaphor use physical objects to manipulate virtual information in an intuitive manner. In this way people with no computer background can still have a rich interactive experience. For example, in the Shared Space interface (Poupyrev 2000) users could manipulate three-dimensional virtual objects simply by moving real cards that the virtual models appeared attached (similar to Figure 19). There is no mouse or keyboard in sight. This property enables even very young children to have a rich educational experience. In educational settings physical objects or props are commonly used to convey meaning.

Aside from premature users and education, TUI have proven to be supportive for cognitive tasks, i.e. intuitive and spatial handling. It "comes close to the physical world in terms of trial time and number of user operations" (Fjeld et al. 2002). As Gav (1997) points out, speakers use the resources of the physical world to establish a socially shared meaning. Physical objects support common understanding both by their appearance, the physical affordances they have (see Gibson 1979 for information on affordances), their use as semantic representations, their spatial relationships, and their ability to help focus attention.

Looking through the literature we can identify other areas where augmented reality applications are currently in use to gain from the described advantages:

1. Medical applications
2. Entertainment
3. Military training
4. Engineering design
5. Robotics and tele-robotics
6. Manufacturing, maintenance and repair
7. Consumer design

In the following chapter we will look at some of these application areas and how agents in AR can benefit to them.

5.3. Agents in Augmented Reality Interfaces

Our motivation of this work was grounded in the belief that AR agents might be a good idea to help and guide humans. We had to learn that just a few researchers have worked practically on mixing virtual agents with reality, and thus little literature has been published. Up to now, synthetic characters are almost not employed in the real world. Some exceptions are given in Figure 24 and 25, showing an AR agent explaining the handling of a machine and a character playing checkers with a human counterpart, respectively.

Figure 25: A virtual human demonstrates the use of a real machine
(Courtesy L. Vacchetti et al. 2003)

Figure 26: AR Agent playing checkers against a real human
(Courtesy R. Torre, École Polytechnique Fédérale de Lausanne)

One of the key questions when designing and employing AR agents is if there are any special condition when using agents in AR and what problems can be expected. Aside from the technical question of registration (see Chapter 6.3.4.), there are concerns about how the

artificial human is perceived and rated by the human user. One AR character for that a qualified evaluation was done is 'Welbo' (Anabuki et al. 2000). This agent is a robot type interface agent (Figure 27) that guides, helps, and serve the user in an mixed reality Living Room, where the user can visually simulate the location of virtual furniture and articles in the physically half-equipped living room (Figure 28).

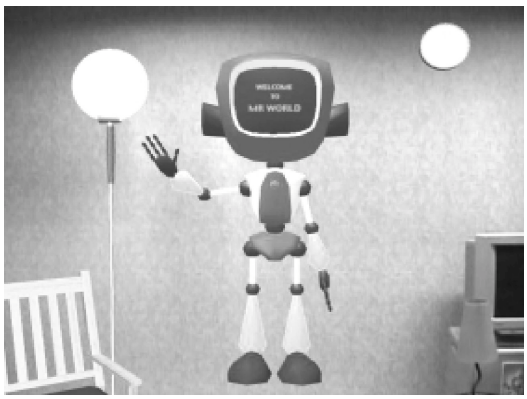


Figure 27: Welbo – a conversational character in the real world

Figure 28: A scene of the physical space (left) and the augmented scene (right) (Courtesy Anabuki et al. 2000)

We can learn from that work that the spatial factors of the design have a large impact on the impression. For example, change in the distance from the agent to the user significantly affect Welbo's appearance. The size and location also has similar effects. Through the experiment, the people preferred a size such that they could see Welbo's whole body in their field of view. Similarly, people liked it to stay some distance away from them. As people feel uncomfortable when other look down on them, Welbo gave an unfavourable impression when it float over them. Thus, it is sure that spatial factors affect the impression Welbo give to user.

Today the technical side of agents as AR interfaces might seem difficult to handle. The equipment needed for doing the first steps in AR is complex (HMD, cameras, tracking, fast computers etc.). On the hardware side, quite bulky and expensive apparatus is needed to make a basic interaction possible. On the software side, several programs have to work together (animation engine, speech engine, planer etc.). Today there exists no common framework or guideline to implement such a complex system. This is why we have to become clear about what requirements our agent and its environments should adhere to. After having specified our objectives we will be able to model and implement our system.

6. DESIGN OBJECTIVES FOR EMBODIED AGENTS

We could see in Section 4 that research groups around the globe and in some instances commercial suppliers have been developing embodied agents for a while now. The objectives that governed the work were most likely accustomed to their particular research interest/application and constrained by the available resources. The results have been very diverse and little compared, possibly due to the lack of objective measurements. Only in some cases an evaluation of benefits to the end user has taken place, e.g. Lester, Converse et al. (1997). Thus we have developed an own taxonomy of design objectives that are presented hereafter.

Preliminary thoughts

To clarify how a presentation agents should be designed, we first define what functionality we can expect and second through what concept this functionality might be implemented.

We have seen in Section 2 that the machine-like metaphor of a DM interface is not a good match to the communication needs of a computer presenter and its human audience. In order to be successful, an assistant-like interface will need to

- Support interactive give and take. Presenters do not speak only when asked a direct question. They themselves ask questions to clarify their understanding of what the user is interested in, describe their plans (e.g. how to proceed with the presentation) and anticipated problems, negotiate presentation plans to fit the skills of the audience and resources available, and present information as they fit into the context.
- Make adaptable plans and implement them. After being acquainted with the audience the particular interests shall become clear within some question in form of a warm-up chit-chat. Then the assistant has to make a first plan how to show around the audience and what artefacts, displays etc. to explain. This plan must not be static but be adaptable to the users interests which he may utter later on when the presentation already had started.
- Recognise the costs of omission or inclusion. Having the presentation plan formed, an presentation interface must model the significance of its decisions and the potential costs of leaving out details or include it, so that it can choose to avoid bothering the user with details that aren't important or include detail when there is a high probability that it might become important later on.
- Acknowledge of the social and emotional aspects of interaction. A human assistant should quickly learn that "appropriate behaviour" depends on the setting. To become a comfortable presenter, a computer assistant will need to vary its behaviour depending on the context. Social user interfaces have tremendous potential to enliven the interface and make the computing experience more enjoyable for the user, but they must be able to quickly recognise cues that might indicate the audience interest in particular artefacts, facts or displays.

Having these general thoughts for how an interface with a presentation agents should (re)act, we want to look at different concept through which these requirements can be implemented.

The goal is a common taxonomy that serves as design guideline as well as criterion for the evaluation of agents or components of agents. We reviewed taxonomies presented in Isbister and Doyle (2002) who examined several definition proposed for conversational agents and those suggested in Gratch, Rickel et al. (2002) summarising the outcome of an international workshop on virtual humans. Based on that work we have developed a new taxonomy of objectives to build believable embodied agents in augmented reality. The following section will explain the different components and how they contribute to the whole.

6.1. Agency

To convey a believable representation of an agent computational issues of agency have to be concerned. Properties like autonomy, reactivity, responsiveness, reliability, completeness, ability for parallel activities, efficiency, goal-directness, and optimality are perceived through the interface but have to be modelled and assessed in this layer.

To construct virtual humans that may effectively participate in a face-to-face conversation, it requires a control architecture as basis. The following list describes the features of such an architecture (Gratch et al. 2002).

1. *Multimodal input and output*: As human use all channels of communication in conversation, the architecture must support receiving and transmitting this information. Speech, gesture, intonation, and gaze should be supported.
2. *Real-Time Feedback*: The system must let the speaker watch for feedback and turn requests, while the listener can send these at any time through various modalities. It has to be sensitive to different threads of modalities and their appropriate response-time requirements, e.g. feedback and interruption occur on a sub-second time scale.
3. *Understanding and synthesis of the communication content and the processes of conversation*: For a meaningful multi-sentence output to the user, the system must include a planner that selects and manages the order of the presentation of interdependent facts. These facts come from a knowledge base that is divided into a static part with the fact knowledge and into a dynamic part with the discourse knowledge (e.g., What has the user been told already?). To behave accordingly to the rules of human communication, the system has to build a model of the current state of the conversational process, e.g. to determine who the current speaker is, if the speaker need to repeat the information as the listener has not understood it etc.)
4. *Conversational function model*: A range of different behaviours can be employed to achieve conversational functions, such as turn-giving (giving up the floor) or turn-taking (requesting the floor). Representing conversational functions eases the mapping on and combination of different modalities. At the input and output interfaces, behaviour is translated into functions or functions are translated into behaviour, respectively. This is symmetric as the same functions are present in input and output.

Regarding the muldimodal input and output, very few comparable systems have been evaluated using a corpus of unconstrained speech and gesture. There is only one exception (Quek et al. 2002), which includes an informal evaluation on unconstrained human-to-human communication, but no quantitative results.

To naturally model verbal and non-verbal behaviour that are influenced by both emotions and personality, researchers have developed computational models to mimic these phenomena. The models can be split into communication-driven and simulation-based approaches.

With communication-driven approaches the emotional expression is chosen on the basis of the desired impact on the user. The agent intentionally plans whether or not to convey a certain emotion and intentionally uses the expression with a specific goal, e.g. a question combined with a sorrowful face is more likely to evoke a affirmative response (Pelauchaud, Carofiglio et al. 2002). In contrast to this approach, the ones from the simulation-based category try to reproduce ‘true’ emotions. They build on appraisal theories of emotion, mostly the OCC model (Ortony, Clore et al. 1988) that “views the emotions as arising from valenced reaction to events and objects in the light of agent goals, standards, and attitudes” (Gratch, Rickel et al. 2002).

Coping theories that explain how people handle strong emotions have been incorporated in the OCC model (Marsella and Gratch 2002) and resulted in more diverse strategies: (a) problem-focuses coping, i.e. the selection of actions according to possible improvements of the agent's emotional state, (b) emotion-focused coping, i.e. altering of the agent's mental state to improve the agent's mental state (e.g. dealing with guilt by blaming somebody else). Simulation approaches have done further progress with the introduction of learning mechanisms, like AI-based planning (Marsella and Gratch 2001) that results in agents that are able to dynamically adapt emotions through own experience (El-Nasr, Ioerger et al. 1999; Yoon, Blumberg et al. 2000). Other researches derive an emotion's intensity from the importance of the goal and its probability of achievement (Sloman 1990; Gratch 2000).

Modelling the relationship between emotions and the resulting behaviours involves uncertainty. As Bayesian networks explicitly address this problem. They are often used by researchers (a) to determine the likelihood of different postures or gestures for individuals depended on their personalities and emotions, e.g. (Ball and Breese 2000); (b) to integrate different communicative functions on different channels, e.g. (Pelauchaud, Carofiglio et al. 2002); and possibly (c) to model how emotions vary over time. To effectively convey emotions, humans use body gestures, facial expressions, and acoustic realisations. For an overview of studies on emotive expressions see (Collier 1985).

6.2. Human-figure animation

There is vast evidence that an agent's embodiment may significantly contribute to the likeability of agents to users (McBreen et al., 2000). Without an appropriate display technique, emotive behaviour as well as speech and body animation for believable human-like actors would not be possible. In established methods from movie productions, animators design or script movements, or alternatively actor performances are brought into the computer system. Either data is then mapped onto the virtual character, its body and face creating locomotion, gestures and facial expressions.

With interactive agents that work simultaneously in many instances and over a limited bandwidth, there is no way to rely on animators to create their responses. They have to be autonomous in creating appropriate animations in real-time. "It is creating novel, contextually sensitive movements in real time that matters." (Gratch et al. 2002) Several parts of the body system need to be addressed: locomotion, body pose, gestures and hand movements, the face with all its components, and other physiological requirements of real humans, such as breathing, eye blinking, and perspiring. Traditionally research on animation has either focused on complete body animation or on facial animation. We will follow this distinction for now and will integrate both later.

6.2.1. Body animation

In body animation, there are two approaches to gain interactivity: motion capture with additional techniques to rapidly assign motion sequences according to immediate needs, or procedural code that allows controlling important movement parameters.

The motion capture approach produces natural looking movements but has problems with maintaining environmental constraints (e.g. solid foot contact to the underground, proper grasp). To bypass this problem, procedural approaches work through the parameterisation of target locations, motion qualities, and other movement constraints to form a plausible movement directly. Kinematics as the one technique of procedural approaches is best for goal-directed and slower (controlled) activities. Dynamic techniques on the other hand are

most suitable for movements directed by application of forces, impacts, or high-speed behaviours.

Thus the animation system can apply human behaviour on computer-generated models. With a diversity of body movements involved, we may build more consistent agents now. Multiple body communication channels are affected, and procedural animations control and co-ordinate them. “The particular challenge is constructing computer graphics human models that balance sufficient articulation, detail, and motion generators to effect both gross and subtle movements with realism, real-time responsiveness, and visual acceptability.” (Gratch et al. 2002) Beside all the good news there is still some chance for improvement. Today’s virtual humans animated with either of the introduced approaches do not display any kind of individualism, which is a characteristic property for humans. Artificial intelligence or machine learning might be a good start to challenge this issue.

With consideration of our application we will not elaborate on topics like movement strategies, soft deformable surfaces, or clothing. Even if they add considerable human qualities to the character, they are beyond the scope of our work. For references in these areas see Gratch et al. (2002). For an in depth investigation into the field of animated speech see [Web3].

6.2.2. Facial animation

Being central to human-human interaction, facial expressions can easily convey information about the emotional state of a person. This information is understood instantly, without further conscious processing. Thus we need to be cautious when creating artificial faces for communication purposes. Modelling and rendering the ‘wrong’ artefact might transport unintended meaning or evoke negative feelings in the receiver. “The great complexity and psychological depth of the human response to faces causes difficulty in predicting the response to a given animated face model.” (Gratch et al. 2002)

The choice of modelling and rendering technologies ranges from 2D line drawings to physics-based 3D models with muscles, skin, and bone. Textured polygons (e.g. non-uniform rational b-splines and subdivision surfaces) are by far the most common. A variety of surface deformation schemes exist that attempt to simulate the natural deformations of the human face while driven by external parameters. Among these techniques, we can identify three major categories of facial animation methods.

The first is ‘keyframing’, i.e. generate keyframes and interpolate between two adjacent frames. This methods provides complete artistic control and is very flexible. The animator can decide how could the output will be by inserting more keyframes, but on the downside it can be time consuming to be perfect. The second method to simulate faces is to apply motion captured data to a face model. In this approach human facial movements are measured directly, then post-processed (to filter artefacts or extract information, e.g. the position of markers) and finally mapped to the face model which changes accordingly to the input. The third method is to synthesise facial movements from text. A text-to-speech algorithm translates the textual information into phonemes, which are then mapped to visemes. A speech articulation model takes these visemes and accordingly animates the face. Basis of this approach is the Facial Action Coding System (FACS) by Ekman and Friesen (1978) that describes all “visually distinguishable facial movements”. In FACS, the facial movements are based on the combination of action units (group of muscles whose action is distinguishable for the human observer) that control facial expressions. This method can provide real-time animation with understandable acoustics and facial expression. An example is Perlin’s Responsive Face (Perlin) that demonstrates a subset of the full range of the facial expressions.

The output of all three approaches can be optimised by using a realistic 3D head model and updating a dynamic video texture map (Leung et al. 2000).

When deforming a face to produce speech animation we need to have a suitable representation of the face to apply deformation. The H-Anim 1.1 specification in [Web18] was the first and for a long time only standard that commonly defined how human bodies (including the face) are modelled and animated. Now MPEG-4 has overtaken most of the VRML specification and extended certain concept for face and body animation of virtual humans [Web15]. It defines *Face Definition Parameter* (FDP) feature points and locates them on the face (see Figure 29). Some of these points only serve to help defining the face's shape. The rest of them are displaced by *Facial Animation Parameters* (FAP), that specify feature point displacements from the neutral face position. Some FAPs are descriptors for visemes and emotional expressions. Most remaining FAPs are normalised to be proportional to neutral face mouth width, mouth-nose distance, eye separation, iris diameter, or eye-nose distance.

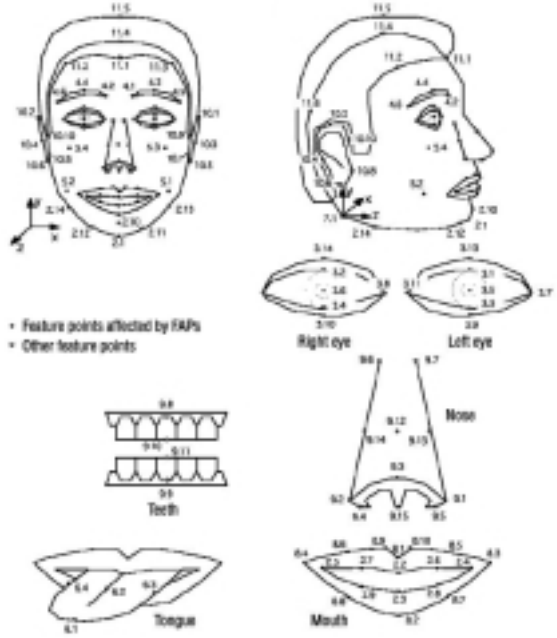


Figure 29: Facial Definition Parameter (FDP) feature points in the MPEG-4 standard (Courtesy Gratch et al. 2002)

6.2.3. Integration

MPEG-4 and its forerunner H-Anim do not only specify the face and its features. The Face and Body Animation (FBA) object describes the geometry representation and animation of the whole body, including the face. Using the FDP and the Body Definition Parameters (BDP) the MPEG decoder can create a FBA model with specified shape and texture. This is a hierarchical model of segments connected to each other via joints. The animation of body and face is done by interpreting the Body Animation Parameters (BAP), respectively the FAP. To animate a H-Anim/MPEG-4 character, the script obtains access to the joints and alters the orientation angles (yaw, roll, pitch) to match those defined in a BAP stream. A schematic illustration of the architecture can be found in Figure 30.

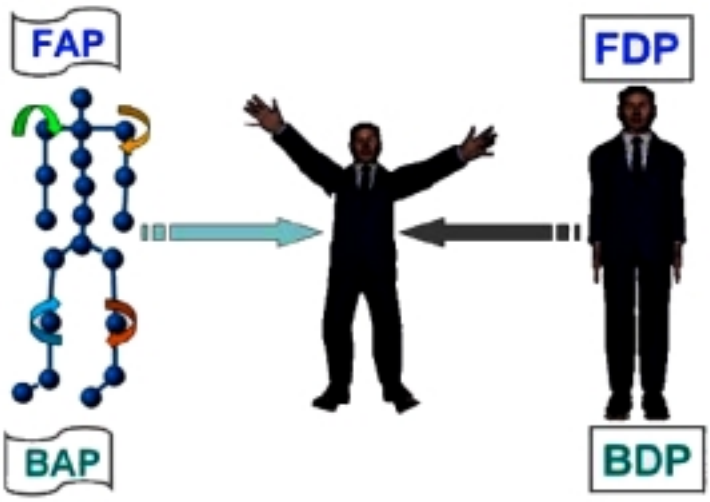


Figure 30: Components of MPEG-4 compliant figure definition
(Courtesy Thalman and Vexo)

6.3. Social Interface

Humans are considered as social animals. They communicate with each other extensively to build up social relationships. The Media Equation (Reeves & Nass 1996) states that "media = real life" and "computers are social actors". In a series of experiments, Reeves and Nass (1996) demonstrate that individuals' interactions with computers and other media are fundamentally social and natural and follow the same rules as individuals' interactions with other people. Similarly, other researchers argue that autonomous agents, in their interaction with people, must be governed by the same principles that underlie human collaboration findings from human-human interaction (Rich & Sidner 1998). Thus, knowledge about interpersonal interaction can give good indications for human-computer interaction.

Researchers have learned that human face-to-face communication involves language and non-verbal behaviour that function in parallel and independently. For in-detail information about human-human interaction and the psychological functions behind it see the accompanying website to this work.

Words support the interpretation of e.g. a gesture and vice-versa (Gratch et al. 2002). In human communication, the different channels operate at different time scales (a quick nod to stress the meaning of a word is quicker than speaking the word whereas a descriptive gesture may take longer than speaking a short sentence). Meaning conveys through the patterns of co-occurrence in different channels. Thus integrating verbal and non-verbal conversational behaviour like speech, intonation, gaze, and body movement is a interrelated challenge. Cassell et al. 2000 could show that “speech and nonverbal behaviors do not always manifest the same information, but what they convey is virtually always compatible” (Gratch et al. 2002).

We can distinguish three cases of co-occurrence in different modalities: reinforcement, complementation and process simplification. When people accentuate important words by speaking more forcefully or illustrate their words with an iconic gesture, two modalities convey the same meaning i.e. they are redundant but reinforce each other. When people produce a gesture that complements what they mean while speaking, e.g. “Let’s go there!”, it adds important information which is essential for the understanding (i.e. a semantic attribute of the message). Speakers turning their eyes towards the listener when coming to the end of a

thought and listeners nodding within a few hundred milliseconds when the speaker's gaze shifts, ease the communication process (i.e. a pragmatic attribute of the message).

Every realisation of modalities can be seen with respect to its goal, either to increase the conversation content or to advance the conversation process (Gratch et al. 2002). Some of the goals can be equally met by one modality or the other. Thus the interface we design must support multimodal communication through the choice of appropriate functions to resemble human interaction.

6.4. Personality

An embodied agent's appearance and (re-)action to the world define the agent's 'personality'. Personality accounts for the joy people feel when interacting with animate characters (Hayes-Roth and Doyle 1998). It is related to such properties as display of broad, well-integrated capabilities, behaviours, goals, and emotions. Personality aims to avoid dullness, to engage the user to stay with the presenter, and to enjoy the interaction.

An agent's personality influences locomotion and movements (e.g. a specific gait), visual appearance, intonation and fluency of speech, and non-verbal behaviour. The base for these functions was build in the previous chapters, but the individual modelling and implementation will be done at this level. Defining personality is a heavily manual process, for example creating characteristics of the walking or sets of facial expressions.

In their study of the effects of the personification of agents, Koda & Maes (1999) found that in an interactive application, an embodied interface agent (as opposed to a disembodied dialogue box) was more engaging to the user. The creators of WebPersona (Andre et al. 1998) discovered that their subjects rated learning tasks presented by the Persona agent as less difficult than the presentations viewed without such an animated, graphically depicted interface agent. Initial studies of a plant biology tutoring agent by (Lester et al. 1997) revealed that lifelike, personable interface agents were perceived by students as being very 'helpful, credible and entertaining' and that agents which offer a range of levels of advice can increase learning performance.

6.5. Production

Integrating all the different objectives, technologies and design requirements into one agent is challenging. The different components discussed so far need to work to together to create a overall natural artificial human. Of special interest in this matter is consistency, timing and registration.

6.5.1. Consistency

One problem when combining a variety of components is to maintain consistency between the internal state of the agent (e.g. goals and emotions) and the multiple channels of behaviour (e.g. speech and body movement). When real people interact, they present multiple behaviour channels, and we interpret them for consistency, honesty, and sincerity, and for social roles, relationships, power, and intention. Gratch et al. (2002) conclude "When these channels conflict, the agent might simply look clumsy or awkward, but it could appear insincere, confused, conflicted, emotionally detached, repetitious, or simply fake." Practically they found out that

- Arm gestures without facial expressions look odd;
- Facial expressions with neutral gestures look artificial;
- Arm gestures without torso involvement look insincere;

- Attempts at emotions in gait variations look funny without concomitant body and facial affect;
- Otherwise carefully timed gestures and speech fail to register with gesture performance and facial expressions
- Repetitious actions become irritating because they appear unconcerned about our changing (more negative) feelings

One approach to remedying this problem is to explicitly co-ordinate the agent's internal state with the expression of body movements in all possible channels. The co-ordination of the agent's internal state with the expression of the body movement in all possible channels should remedy this problem. For example, the Emote system is used to modify the execution of given behaviour and thus change its movement quality or character (Kovar et al. 2002). If the endeavour is successful in the end, we will not only have a modular architecture for building agents, but a guideline how to assess others work and compare it to the own.

6.5.2. Timing

Problematic is that the virtual human's behaviour develops over time, and is subject to a variety of temporal constraints. These timing constraints are tuned according to the nature of the architecture. Most architectures for agents focused on a specific aspect of behaviour (e.g. speech, reactivity, or emotion), only a subset of the whole spectrum.

In speech centred systems behaviour is a slave to the timing constraints of the speech synthesis tool (e.g. in BEAT, see Section 4). In Emotion centred systems behaviour is a slave to the constraints of emotional dynamics (e.g. Emote, see Kovar et al. 2002). The third class of systems focus an conditions in the environment. Here, behaviour is a slave to environmental dynamics. These are competing constraints. One solution could be to arrange the components of the architecture in a pipeline and share more information between them.

The well timed succession or synchronicity of elements from verbal or non-verbal behaviour is a major concern. For example, speech-related gestures must closely follow the voice cadence. Synchrony is essential to the meaning of conversation. When it is destroyed, as in low bandwidth videoconferencing, satisfaction and trust in the outcome of a conversation diminishes (O'Conaill and Whittaker 1997). Therefore multimodal input to a virtual human must be incremental and time-stamped to allow interpretations of different input events. With the time stamps they can be fused to understand what behaviours act together to convey meaning. To align words and non-verbal behaviour, the speech recogniser has to provide word onset times. For the output the speech synthesiser needs to maintain synchrony with the body animation to produce co-occurrence of speech and e.g. gestures (for details see Gratch et al. 2002). Tight synchronisation along the modalities is essential for conveying meaning!

6.5.3. Registration

Special to augmented reality is the problem of registration. As we render virtual objects into the real world scene, they must be drawn at the correct position, with the correct dimensions and the correct alignment according to the current viewpoint of the user. They are dynamically altered when the user's viewpoint changes, e.g. when he moves or looks somewhere else. Quality of registration depends on what system is being used and what latency (negatively proportional to the update rate) the system provides. Several different approaches are possible, depending on the application and the accuracy needed.

Cheapest and mostly used is the vision-based approach where markers displaying a distinctive and unique pattern represent some virtual object to be displayed. When such a marker comes into the visual field of the camera, an algorithms finds that marker, extract the relative

position and relative orientation from the viewpoint. Then it renders the virtual object according to its relative position and orientation (see Figure 31 for more details). This approach depends heavily on the quality of the camera and the distance of the camera to the markers. When no marker can be distinguished in the video feed anymore, registration is lost and all virtual objects disappear. There are techniques with multiple markers and natural feature tracking that promise some relief but the basic problems remains.

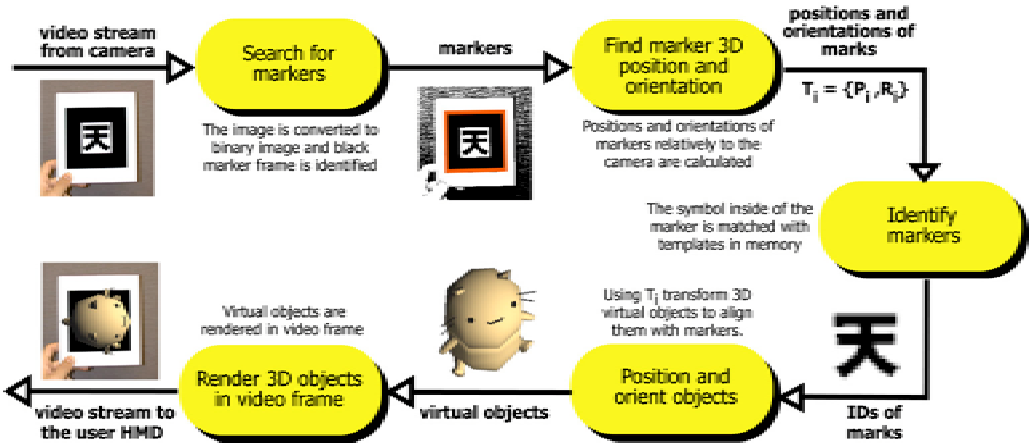


Figure 31: Vision-based tracking with markers (Courtesy HITLab, University of Washington)

An approach that is not dependent upon a clear camera view is inertia tracking. Starting at a fix point in space an inertia sensor registers every movement with velocity and time stamps. That data is recalculated into changes of position and the new view point is calculated. Alignment is less exact with this method then with marker tracking, because small errors accumulate which cannot be adjusted easily because no point of reference is giving during the tracking.

The second non-vision approach relies on tracking through the Global Position System (GPS) or direct field sensing. When installing a GPS receiver directly at the HMD and knowing the global co-ordinates of the virtual objects, one can easily calculate the distance and hence the size of the virtual objects. Inertia and GPS tracking only gain three degrees of freedom each at maximum. They have to supplemented with other methods to allow the complete calculation of the viewing direction. Direct field sensing measures a natural or synthetic magnetic or electric field and calculates all data through physical and electrical phenomena such as the Hall-effect or the magneto-inductive effect.

Tracking methods can be distinguish according to the way information is presented in the HMD:

- Head-stabilised - information is fixed to the user's viewpoint and doesn't change as the user changes viewpoint orientation or position.
- Body-stabilised - information is fixed relative to the user's body position and varies as the user changes viewpoint orientation, but not position.
- World-stabilised - information is fixed to real world locations and varies as the user changes viewpoint orientation and position.

Each of these methods require increasingly complex head tracking technologies; no head tracking is required for head-stabilised information, viewpoint orientation tracking is needed for body-stabilised information, while position and orientation tracking is required for world-stabilised. The registration requirements also become more difficult: none are required for

head-stabilised images, while complex calibration techniques are required for world stabilisation. World-stabilised information display is usually attractive for a number of reasons, one of which is that you may annotate the real world with context. But in the context of presenter agents it poses the problem, that when registered in the world we automatically expect that the virtual human will move with us (e.g. walking). This ability is not central to such an agent and affords additions to (in worst cases) the whole architecture. This issue might be circumvented by introducing an agent without legs that is just hovering through the space.

We could see in this chapter what technical challenges have to be met when building an agent in AR. For an exhaustive introduction to tracking in AR have a look into Rolland et al. 2001. The next one will shed some light on how to specialise the general agent to its application domain.

6.6. Application domain

Being the most specialised area when conceptualising an agents, we need to give some attention to the application domain and its specific requirements. It is evident that the application must not be considered one time in the end. It is essential to be clear about the purpose of the complete system at the beginning of the assembly to have clear objectives for later decisions.

When all the other steps of implementation have been successfully taken, the agent should be ready for training in his later field of expertise or work or training. Therefore we need to apply some specific knowledge, behaviours, emotional display etc. The ideal agent should allow to be trained on these areas without implementing additional routines or changing the existing system layers. Seeing the other components as bricks and the whole architecture as house, the training on the application domains is the roof.

Training for the later application involves speech (e.g. the user's language, and special terms specific to the domain), special behaviours (e.g. typical greeting and farewell gestures), domain knowledge (what the agents needs to know and world knowledge), spatial knowledge (where am I? what is around me? etc.) and the setting (e.g. in a museum for mostly families, or in a managing board meeting).

6.7. Believability

The criterion for success is if the agent conveys the ‘illusion of life’ to the user. This can be evaluated through subjective measures such as questionnaires or surveys. Bates adds, that “to our knowledge, whether an agent’s behaviour produces a successful suspension of disbelief can be determined only empirically” (Bates 1992). Some researchers proposed objective measures for a subset of the mentioned properties that involve formal settings on a specific task (Doyle 1999). Consequently, the research area in ‘Social Interactions’ has been developed during the last years and aims to describe the mechanisms underlying the process of establishing and maintaining social relationships between human agents and artificial agents (Dautenhahn 1998). Canamero (2001) offers a good overview of issues relating to emotional agents in social interactions (known as ‘Socially Intelligent Agents’ in her work).

We are convinced that *believability* Bates (1994) is the measure all research results have to be assessed against. It’s not a single category but an overall objective within building agents. This implies by no means that other category might be neglected, but instead that each of them has to contribute its share to the common goal. Ideally this will result in a socially and situational aware, strong autonomous character with self-motivation and emotions. With such

a virtual human it should be possible to engage humans in a joyful, pleasant and at the best effective interaction.

Conclusion

We have specified requirements for agents that may serve as design guideline for agents. This bottom-up approach guarantees a solid basis before building the higher, dependent layers. A representation of the complete model is given in Figure 32.

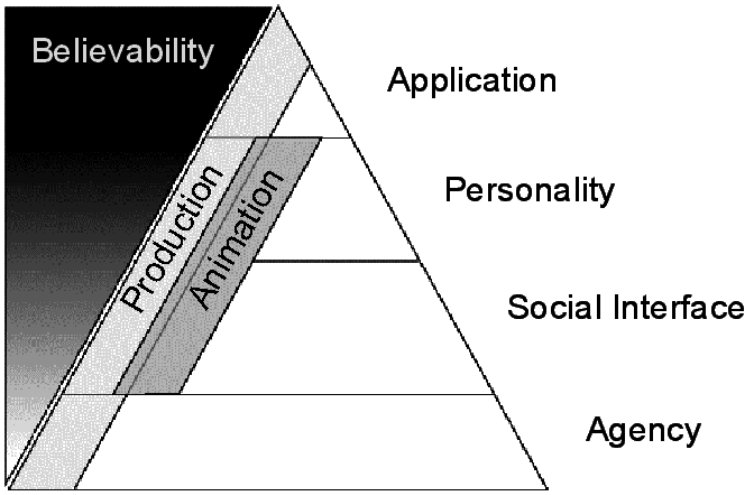


Figure 31: Relation of components when building AR agents measured against their contribution towards believability

The categories proposed here are not as independent as it might seem, especially the border between personality and social interface might be fuzzy. But this can be understood as result from successful agent design: producing believable agents with appropriate behaviours is a highly integrative process that transcends set boundaries and needs to be interdisciplinary.

A set of common objectives for design and evaluation of embodied agents and components of them would provide benefits to all researchers. Understanding of each others' results and setting them in context would be eased, the importance of work in certain areas identified and acknowledged, and inter-operatability of components were in safe hands. It would yield in a modular architecture and standardised interfaces while preserving the standards for excellence and guaranteeing the extensibility of to-be-developed agents, as others could easily relate their development to and build it on existent ones.

We could learn from this section what parts an agent in augmented reality consists of and what requirements it has to meet to be useful to the user. We have developed a taxonomy of design objectives assessed against their particular impact to the level of believability. The next section will concern the second part of our own work, the implementation of a prototype presentation agent.

7. THE IMPLEMENTATION

Having focused on a review of embodied agents and the definition of areas for agent design during the last sections, we now want to apply our knowledge and introduce the reader to our prototype implementation. In this section we explain what considerations led to the final design, what problems we were confronted with, and how we solved them. In the end, we present a complete synthetic embodied agent for presentations that lives in the user's augmented reality space and can engage him in natural conversation.

7.1. Basic Considerations

When implementing an agent in AR we need to define a certain architecture which is able to integrate the aspects of agency, social interface, personality, animation and production. Additionally it must be flexible to guarantee the custom adaptation to the application domain.

From the previous chapters the basic architecture of an AR agent was introduced in terms of necessary functionality that build on top of each other or that need to be considered throughout some stages of the development. At this point, we want to interpret the model in terms of engineering, i.e. what kind of hardware and software is appropriate to implement the desired functionality.

The goal to create a functioning synthetic character that converses with the human user in his personal augmented space involves two big issues: the input to the agent and the output from the agent. During the last sections we have concentrated on the output channels and we will go on in the same manner in the subsequent chapters. Incorporating all the different input channels in our model would at least two-folds the complexity and increases the burden of integration. Thus we assume that understanding the human user and preparing meaningful answers was done in a processing step before our agent system takes over, e.g. by a language parser to process the textual input, and a reasoning unit in combination with a knowledge base to create matching answers. The agent only gets textual information of what should be answered. This black box approach is a simplification of the whole problem but allows us to focus on the generation of visual and audio realisations and its effect on the user. Input parsing and answer generation are left unconsidered here. We will see in the further work that this is no disadvantage but meets the necessary requirements.

Concentrating on the agent's output, we need to have a parser to extract information from the sentence to be said for the subsequent steps, a sophisticated mechanism to create non-verbal behaviour, a text-to-speech engine that transforms the text into audible voice, and an animation engine to realise the movements of the face and the body into display. This is a processing pipeline with text (the answer to the human) as input form the knowledge base and meaningful speech and accompanying behaviour as output to the user.

We will go through the steps of the processing shortly. The parser enriches the textual input from the knowledge base with meta-information about the syntactical information. It analyses the structure of the utterance and to identify relationships among the single words and marks typical words or constructions in sentences. We cannot expect semantic 'understanding' as the arbitrariness of human language is to complex. Then, the non-verbal-behaviour generation sets in. It analyses the meta-information and the output according to a knowledge base and adds animation parameters (for example, to realise supportive beat gestures). The knowledge base is trained to a specific vocabulary and dynamically adds transitional information. Gaining a high level of appropriateness between the gestures and the speech is the main difficulty in this step. Near to the end of the pipeline, the text-to-speech engine forms speech from the textual output and sets additional parameters to control the subsequent lip animation (so called 'word-onset times'). The very end of the processing pipeline, the animation engine takes all the animation information and applies it to the representation of the agent with the correct timing. Throughout each phase of the process, the output of modules contributing to the animated character directly must be well synchronised (see Chapter 6.5.2) to ensure consistency and believability of the agent.

7.1.1. Alternative Approaches

Implementing such an architecture might happen in two ways: to completely design a system from bottom-up using existent or newly developed components or employing an existent

framework and make certain changes to adopt it to the given requirements. We will shortly elaborate on both approaches and present the conclusion that guided our work.

When beginning with nothing, we can specify all our needs and be confident that they will be met in the end. Nothing else constrains the development, except from technological boundaries. There is the high chance to build a perfectly suited system according to the special needs of the customer. It will be very specific and centred at one unique application. Problematic with this approach is the amount of work one has to invest to see first results. The creation of several modules would be very complex, time and labour intensive. It has to be conceptually designed, implemented and tested. The result might be a working system, put if it is good in terms of software-engineering (e.g. error-rates, reliability and robustness) or the application (completeness, soundness) cannot be decided without an extensive evaluation. This approach is appropriate for large and well equipped development teams in highly specialised areas that can ensure good software-engineering and thorough testing.

In the other approach, we look for existing modules or whole architectures to meet our specifications. This usually introduce some kind of compromise because mostly you do not find the exact realisation of the functionality you need. It might be less efficient, equipped with a terrible documentation etc. Using others' software means overhead. Functionality you don't want but which causes errors forces you to build a workaround. And functionality that is missing has to be added according to the design of the existing system, even if that is inefficient. We have assumed that the source code of the system is available, but in fact this is very optimistic hope. Many modular programmes are written using compiled libraries whose input/output parameters are properly documented only. If the source code is available as, for example, in Open Source[1] you might not understand the code because the implementation is in an unusual programming language you don't know, or the concept behind the code stays unclear. Either this means to learn a new language, to rewrite the code or to choose another software. But there is a good side to modules and recombination of existing software. The programmer can rely on basic functionality that is already implemented. A system or architecture that was successfully used somewhere else has proven to be a mature and stable piece of software. By no means, this guarantees a flawless operation. But the chance that bugs have been eradicated is high. Moreover people have already worked with such a system and gained some expertise that might be helpful when looking for solutions to particular problems. This approach is suitable for endeavours that can handle some flexibility in the requirements and that have limited resources available.

In the last chapter we have presented some pros and cons of two approaches to software development and we could see that under some circumstances one approach might be better than the other.

7.1.2. Research situation

Now we will cast light onto the situation I found myself in when joining the Human Technology Lab in Christchurch, New Zealand (HITLab NZ) and what results were effected by that. The subsequent chapters will be a technical report about my internship at the HITLab, not a common research paper.

My stay at the HITLab NZ was characterised by several constrains. First of all, the research lab had just begun its operation, meaning that there was no big working research group established. Therefore I hardly could rely on structural or personal support and seeking help was an endeavour. Basically, I was a 'research group' on my own. Secondary, the work of the lab aligned along certain projects that were mostly funded by third party companies. The area

[1] For more information see http://en.wikipedia.org/wiki/Opensource (07.07.04)

of research I chose was appealing to some of them but did not match their core interests. I wrote a project proposal to motivate the incorporation of my work as 'edgy' supplement to a remote collaboration project, but I didn't work out. Thus I had no money to invest in some high quality hardware or commercial software. Fortunately the lab had common equipment for projects in augmented reality that I could use, all on basis of common Intel PC technology. A last constrain was time, usually five months with the option to extent it up to the time my visa was valid (that was one month more).

7.2. Approach

Under the described conditions I had to decide for an approach to realise a life-like animated character in augmented reality. Because of the limited resources, especially time and man-power, I opted for the second approach - to build my agent upon an existing system. It is illusionary to create a new architecture in half a year's time with a staff of one, if other groups need several years with many people involved. Thus, I had to find some piece of software which could serve as basis for my system. Believing in the Pareto Principle[2] this basis should provide 80% of the necessary functionality with a little effort of 20% of the total time, whereas implementing the 20% of extra functionality would take 80% of the total time. The basic functionality includes such issues as setting up the modules and get them working together. The extra functionality covers the adoption to augmented reality and the training to the application domain. The site for the practical tests is the HITLab itself. There are plenty of experiments and displays to explain to frequently coming visitors.

Before we look at the particular implementation of the agent prototype in AR, we will re-consider the some issues outlined earlier. Section 6 will be a helpful background for this chapter.

7.2.1. Preliminary considerations

Believability

Our social presenter will be the servant of the user. As such, it must be clear to the customer that he is the master. Clarifying this role model is a question of UI design. The agent must not be perceived as omnipotent figure that guides the conversation effortlessly and that knows everything. The expectations of the human would (unconsciously) increase tremendously. And if the agent cannot keep up with these expectations frustration would take over. Thus it is better to design a human-like but not a photo-realistic conversational AR agent. The behaviours will be the same social ones, the communication not diminished, only the outer appearance changes. Then the human will more easily forgive small flaws in the realisation (representation and behaviour) of the agent or conversational break-down (e.g. misunderstanding). We clearly look for a non-photorealistic agent to be front-end of our architecture.

File Formats & Body Description

Implementing an animated human-like presenter is complicated by a relative lack of generally available tools. Body models tend to be proprietary (e.g. Extempo Expert Agents, www.extempo.com), optimised for real time and thus limited in body structure and features (e.g. DI-Guy, www.bdi.com). Constructions built with standard commercial 3D modellers such as Poser, Maya, or 3DSMax that save in proprietary formats can hardly be used in other animation tools. Translating between the different programs is possible and file format converters are available. But on the one hand our tests showed that there is no 100%

[2] For more information see http://en.wikipedia.org/wiki/Pareto_principle (07.07.04)

correspondence of the output with the input (textures are missing, dimensions become distorted etc.). On the other hand, modellers and converters whose results only need a few manual steps of post-processing are commercial. We financially cannot afford using these programs and their respective formats.

As described in Section 6.2, the best way to design avatar is the Web3D Consortium's H-Anim standardisation (www.h-anim.org) or the MPEG-4 Industry Forum's effort[3] (www.m4if.org). With well-defined body structure and animation capabilities, the H-Anim specification inside *VRML* ('Virtual Reality Modelling Language') engenders model sharing and testing not possible with proprietary approaches. The great advantage in contrast to MPEG-4 is that VRML is open source and free. With this property VRML would benefit to all researchers active in the area of conversation or pedagogical agents. On the other side, obtaining or using the MPEG-4 specification costs money. This clashes with our constrains. Preferring H-Anim over the more recent specification does not introduce a disadvantage as MPEG-4 incorporates VRML for body and face description, exactly the area we focus on. We would need no other features of the MPEG standard, hence VRML is entirely suitable. Thus, the animation engine in our architecture should support VRML data.

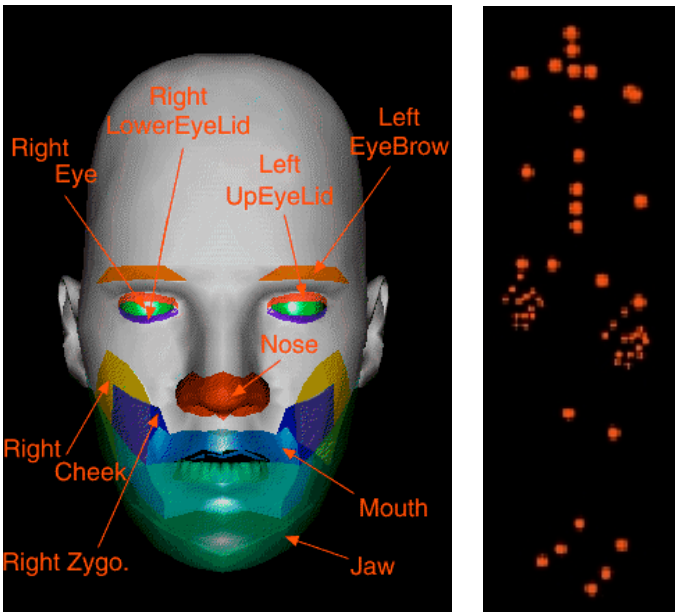


Figure 32: (a) Regions of deformation with SMILE[4] (a subset of VRML97) and (b) VRML97 body joints (Courtesy of Christian Babski, École Polytechnique Fédérale de Lausanne)

Rendering & Animation

There are commercially available tools for rendering and animating virtual agents, such as the 3D Game engines, but due to the fact that these engines are built especially for games, they also have a number of limitations. The first is that they use specific file formats (see the chapter above). Furthermore, the agents cannot use an arbitrary body hierarchy (they usually have simple skeletons) and there are few actions that they can implement, such as running, jumping and picking objects. It is very hard (if not impossible due to the simplicity of the skeleton) to design more complex actions using game engines, e.g. to have an agent gesture to you, or make a facial expression. Finally, game engines tend to be implemented in very

[3] An overview of the MPEG-4 standard from March 2002 can be found at: http://www.chiariglione.org/mpeg/standards/mpeg-4/mpeg-4.htm (08.04.04)

[4] P. Kalra, A. Mangili, N.M. Thalmann, D. Thalmann, "SMILE : A Multilayered Facial Animation System", Proc. of Conference of Modeling in Computer Graphics, Springer, Tokyo, 1991, pp 189-198

sophisticated, highly optimised code that makes the integration with other software difficult. Therefore, one cannot use such an engine to have an agent as a part of a larger application.

Registration

In Chapter 6.4.3 we were sensitised towards the problem of registration. It mainly concerned how the virtual objects are mixed with the real world to be in the right place. In this short paragraph the viewpoint is specifically on how we could present our agent to the user.

We need to visualise the presentational agent somewhere. This brings up the problem of occlusion. As the agent occupies some space in the world, the original content of that space will be occluded. As bigger as the agent becomes the less the user will see from the real scene. The question is where to register the agent. Should it be bound to a fix position in 3d ('world stabilised') or should it be displayed directly in the field of view of the user ('HMD stabilised')?

The first method would let more freedom to the user but introduce the challenge of modelling the topology of the surrounding space. Otherwise the agent would not be (partly of totally) occlude when being behind a desk, shelf etc. Registered somewhere in the world rather than in the field of view, the user's interaction with the environment will stay the same but the agent does not have such close connection to its master anymore. That might not contribute to the information delivery process and the narrative experience.

The second method would present the agent as truly personal guide, always present, always ready to give some explanation or hints. The down-side is that the user has to adjust his position and the field of view more frequently as parts of it are occupied and thus permanently occluded by the agent. On the positive side, we do not have to worry about walking animation or the user's impression of a floating, leg-less body. Both methods seem feasible and each one has its advantages. The further direction for our implementation depends on the visual result of each methods. We will try both!

During the last chapter we have considered some special issues of the agent's design that will lead us to a successful prototype. Questions that were already taken up in the more generic, preceding chapters about agent design (Section 6) are understood to be included in the considerations. The next chapter will be about particular software architectures that are evaluated against their suitability as basis for our presenter agent in AR.

7.2.2. Software Candidates

During the course of this paper we have introduced different software systems that implemented a diversity of agents. In this chapter we will cast view on existing architectures to find one as good basis for our AR agent. The candidates should implement an synthetic, animate agent that shows natural behaviour. The systems should be extensible, documented and PC based, at best supporting VRML and OpenGL (OGL). The following list can also serve as short overview of the history and the development in the field of animate agents. None of the systems under consideration has been converted or tested in augmented reality.

JACK (Badler et al. 1993) is targeting at the fields of industrial simulation, human factor analysis, and similar environments that require accurate biological and mechanical modelling of the human body. Not much attention is paid on appearance, nor to the display of any social actions solely basic movements. It only runs on workstations and thus not applicable for us.

Humanoid (Boulic et al. 1995), a system for realistic representation of the skin and muscles. Focusing more on appearance, it features realistic skin deformation for the body and hands,

and facial animation. It uses its own type of human models and works only on workstations. With this properties it is not suitable for us.

IMPROV (Perlin & Goldberg 1996) consists of an animation engine that uses procedural techniques to generate layered, continuous motions and transitions between them, and a behaviour engine that is based on rules governing how actors communicate and make decisions. The combined system provides an integrated set of tools for authoring the 'minds' and 'bodies' of interactive actors. It seems to suit well for interactive storytelling, but it is not available.

STEVE (Rickel & Johnson 1999), that was introduced earlier in this report, uses the JACK system and is commercial software. Thus it is not affordable.

Virtual Teletubbies (Aylett et al. 1999) uses a novel approach which applies a robot architecture to virtual agents and their behaviours. It cannot be considered as a good tool to start, as there are no hints on how to construct other applications using this method.

Synthetic Actor model connects emotions and social attitudes to personality, providing long-term coherent behaviour in games (Silva et al 1999). Implemented in VRML it meets one of our requirements. But unfortunately, there is no indication on the system's architecture and how to obtain the software.

EXCALIBUR (Nareyek 2000) is a system that uses a generic architecture for autonomously operating agents that can act in a complex computer-game environment. The project has been built for computer games, and the focus was information gathering and handling of incomplete information. With this project it is not clear how to interface the proposed architecture to other modules, e.g. a speech synthesisers that is lacking.

SimHuman (Vosinakis & Panayiotopoulus 2000) is a sophisticated simulation architecture for virtual humans that relies on VRML and OpenGL/C++. It cleary defines callback functions as interfaces to modular extensions. Unfortunately the agent architecture only contains perceptual mechanisms for physical properties of the environment, neither language nor non-verbal-behaviour as input or output. Promising in the beginning, this project turns out to be for physical simulation only, similar to JACK and STEVE, and it does not work on PCs.

The Rutgers University Talking Head - RUTH - (DeCarlo 2002) is a methodological tool for developing and testing psycholinguistic theories of the functions and behaviours of natural face-to-face conversation. Following some tagged textual input, it speaks with lip synchronisation, does limited facial expressions and head movements. Being just a talking head it is impossible with RUTH to display some kind of gestures, postures etc. EMOTE (Chi et al 2000), the Expressive MOTion Engine, on the other hand solely focuses on non-facial movements and the independently defined underlying movements. There was the idea to employ both in conjunction - an extended version of RUTH for visual speech and an adopted version of EMOTE to drive the body language. RUTH is freely available as source code in C++, EMOTE is not - which is why this approach fails.

The CSLU Toolkit (Cole et al 1999) presents a talking head that does language training, e.g. for deaf children (Stone 1999), or educates its audience, e.g. as vocabulary tutor (McTear 1999). The toolkit integrates different pieces of software to offer speech-recognition, text-to-speech engine, audio and display tools. The modular architecture is extendable and customisable, documentation and tutorials are available, it is PC-based, and it is free of charge for academic use. As with RUTH the idea is to extent the toolkit with EMOTE and further embodiment. Being highly promising, the sole problem is the download procedure. To control licensing, it is rather complicated. Long lasting efforts to get the software transferred and working failed. This is the practical reason for no further evaluation.

Studierstube (Schmalstieg et al. 1996; Schmalstieg et al. 2002), the workbench for AR application is mentioned here for reasons of completeness. In contrast to the others it supports tracking, displays and input devices of various kinds and the augmentation of the video stream. The source is open, well documented and actively developed by the University of Technology in Vienna. The problem with the program bundle is that the rendering is done with *OpenSource OpenInventor* (OSOIV) by SGI. But OSOIV cannot display VRML data! Thus we can't include our virtual humans and equivalent detailed body descriptions are not offer in Studierstube. It's an AR framework, not designed to support virtual humans. The second and main problem is connected with the this. There is no functionality to control and animate virtual humans, let alone the issue of implementing autonomous verbal and non-verbal behaviour. Thus Studierstube is a powerful architecture for the development of a wide range of AR applications, but it is not suitable for imbedding social agents.

Today, BEAT (see Chapter 4.1.6) is the most advanced architecture to drive social actors. The representation is a VMRL character on top of *OpenGL* (OGL - a class library for graphics operation) that offers visual speech, and non-verbal behaviours like gestures or body postures. The modules are all well integrated, satisfyingly documented and run on a standard PC. Problematic with this system is the licensing of certain modules. Both, the part-of-speech tagger ("Machinese" by Conexxor[5]) to interpret the input and the library to render the character ("OpenInventor" by TGS[6]), are commercial software. The animation engine ("Phantomime") is a in-house solution from the MIT MediaLab done as Master Thesis by a student some years ago. A lack of available technical detail is to be noticed with all the components involved, and the author of the animation engine cannot be consulted any longer. Apart from the particular drawbacks mentioned, BEAT offers the best starting point to implement a presenter agent in AR.

7.3. Details of the Implementation

Considering the software candidates we have presented in the last chapter, only BEAT meets the criteria we have announced. Visual speech and non-verbal behaviours are realised and can be adapted to the user's needs. This chapter will show how to realise an social agent with BEAT in AR.

Having identified BEAT as good start, there are still problems to overcome, most off all the licensing. The MIT needed some months to decide if they can give away the source code of their animation engine, the main part for my further implementation. This was due to substantial personal changes in the research group. TGS and Conexxor responded more quickly but contributed a time limited license. As the other process took such a long time, I had to reapply two times because the licenses had expired. And from time to time it got harder to explain to the business people that I have no budget but need their software anyway. Substituting the commercial software by non-commercial ones was not possible. The rendering library by TGS offers VRML rendering in contrast to the non-commercial OSOIV. Experience has shown[7] that the POS Tagger by Conexxor is essential for the purpose of BEAT, as it offers a level of tagging not reach by other products.

Reviewing literature, setting the topic of my work, searching and evaluating software, acquiring licenses and waiting for other researchers to respond took a long time, approximately 4-5 months. During this time the idea to extend BEAT (that had not been officially released until that time) was born, but I could not start immediately because of the

[5] http://www.conexxor.com
[6] http://www.tgs.com/
[7] Personal communication with the general developer of BEAT, Hannes Vilhjamson.

aforementioned problems. Thus I decide to test some other approaches not covered by the evaluation of existing systems. For example, to directly build an animated human into an AR world. I could expect that this approach will be far too difficult (see the considerations at the beginning of Chapter 7.2), but gaining some insight in AR was helpful. One of the results is displayed in Figure 33. The figure is registered in the real world to a marker.

Figure 33: H-Anim figure displayed in Augmented Reality registered to a marker

The main problems with this approach is, besides the lack of an behaviour animation engine, that the figure is really small. To get a feeling for the perceives size, compare the agent to the pencil in the background. This agent has the dimension of a mouse, not a human. One could scale the whole figure to an appropriate size, but the field of view in a HMD is limited and applying to this scene we would only see the feet and legs. While searching for a solution, another property of marker based registration was noticeable during the test runs. Due to the low resolution of the camera (max. 320x240 pixels) extracting the markers was limited to a maximal distance of 1.5-2.0meters under best light conditions. Thus, if we scaled up the agent to an appropriate human-like size and would go back to see it in our field of view, the marker would eventually be too far away, pattern recognition lost and the agent would vanish. We have seen that simple scaling is not a solution, and world-aligned visualisation seems not to be proper for our AR agent. Thus we propose the registration in the field of view of the user. This implies the advantages and disadvantages mentioned in Chapter 7.2.1.

Eventually the desired software in source code and all necessary licenses were available. The first task was to extend BEAT to display the agent in the real world. Second would be to train the agent into a particular field (in our case the application displays of the HITLabNZ), and third to evaluate the usefulness of and the pleasure with the presenter in demo situation.

As found out through experimentation before, the presenter will be aligned in the user's field of view. Display-wise, this is equivalent to a human figure at a fixed position on a dynamic background. The dynamic background is the video-feed from the camera (see Figure 19). Phantomime, the TGS *OpenInventor* (OIV) based animation engine, was identified as the right place to hook into the graphics pipeline. The new animation engine has the name 'PantAR' - Phantomime in AR.

In OpenGL it is fairly easy to insert a video from a camera or file as background. As Inventor is a high class library of OpenGL (meaning that it abstracts more and eases programming) it should be fairly easy as well by applying the same concepts as in OpenGL. There are three main problems to this approach.

First, the inner concepts of OGL and OIV. OGL is a state machine and has a strict pipelining concept. Whenever a state (=parameter) is changed in the rendering pipeline, for example, the width of lines, this change will persist until it is changed again. Pipeline means that an object will perform all the operation defined there (e.g. rotations, colouring etc.) in succession until it is rendered to the canvas (=screen). In OGL, objects are rendered in the order they pass through the pipeline. OIV in contrast builds a scene database (i.e. a hierarchical 3D scene graph) which defines the objects to be used in an application. Objects to be rendered come from this database that sets the order of rendition as well. In OIV, states are not defined for the whole pipeline, but for single objects. The two fundamental differences between OGL and OIV can result in fairly awkward renditions when both concepts are used at the same time and interfere with each other.

Second, the processes in OpenInventor are event-based, meaning that only if something happens in the scene the scene is rendered again. Compared to that, OpenGL runs continuously, rendering the scene as often as possible. The difference becomes evident when implementing a video texture in OIV and in OpenGL. In OpenGL you define a rectangle in the view port, you initialise a texture map on the rectangle, and you update the content of the texture map with a pointer to the camera feed in each rendering call. In OIV you can almost do the same, the programming is quicken then in OpenGL but the results looks quite odd - there is no video feed in the drawing area, just a static image. Then, if the mouse is moved or a key pressed, the canvas will update with the current video image from the camera, a single static frame. That is because the mouse and the keyboard trigger events that force the scene to be re-rendered, and the canvas is updated.

Third, knowing about the first two problem we have to solve the question how OIV and OGL can work together without interference. Three solutions are possible. When we want to use OIV without any troubles from OGL, we should implement all functionality in OIV without using native OGL code. This circumvents the problem of integration but might not be possible in some cases. Then, we could embed OGL code into OIV applications or exactly the other way around, OIV code in OGL. In both cases we need to be aware the states. They can be changed by both, independently from each other, and 'confuse' the running rendering process with inconsistencies. Whenever OIV changes states it must get sure that the state in the end is coherent to the states at the start for OGL. Programming a mix of OGL and OIV can be tricky and efforts lots of attention by the developer.

Having learned about OpenInventor and having tested several methods, I found that realising the video texture background is the easiest when implementing the video canvas solely in OIV. Unconstrained by that, I used the *AR ToolKit*[8] (ART), a native OpenGL C library, to get the processed video feed from the USB camera via DirectShow. Contradicting my intention, I mixed the concepts of OIV and OGL. But this was no matter, as the one was for display only, the other to get and pre-process the video-feed only. There was no interference.

My `SoMyBackground` node was added to the OIV hierarchy as top node to ensure that every transformation is done after this node has been traversed. Thus the video background stays at the same position in the viewing frustrum. It takes some tests to set the rectangle on which the images are mapped to the right dimensions. Later on in my work I realised that the

[8] The AR ToolKit is OpenSource and freely available from http://www.hitl.washington.edu

people at TGS must have had the same idea as I did. They have programmed a general `SoBackgroundImage` node that can handle all kind of exceptions and special cases. I opted for this new, far better version and discarded mine. Essentially they both do the same: setting up a `SoTexture`, loading the video image into the texture, stretching the texture to fit the viewport, and transfer it to the canvas via `glDrawPixels`. The small problem of different colour coding in the incoming video and the outgoing graphics stream turned out to be more serious than expected but was eventually solved.

Ensuring that there is not only one static image but a running video feed efforts some digging in the OIV hierarchy. The `SoTimerSensor` node triggers itself after a set interval of seconds and schedules a function `tickFunc()`. In our case this functions reloads the texture memory with new data from a pointer to the camera stream and triggers traversing the rest of the object tree (=the 'world'). This effects a re-rendering of the whole scene many times per second including our canvas for the background. Hence we can see the camera images in a fast succession - a video! PantAR was successfully implemented[9].

BEAT with PantAR as animation engine was successful tested. No problems were expected and none turned up. The speech output of the agent was driven by a hidden human operator who observed the user's interactions with and movement in the environment. The BEAT framework was trained to some specific words of its application field. They were from the context of AR research and typical for the applications that were on display in the HITLab at that time.

7.4. Results

The extended animation engine works in the given framework without problems, and the AR agent was successfully implemented, see Figure 34. It is non-photorealistic, comic-like to level the expectation of the users. And the representation is only knee upward, as the legs are static anyway, but the arms and hands hanging down along the body must be visible because they may convey meaning through gestures.

[9] For code snippets see the complementing web-site: http://www.cs.uni-magdeburg.de/~cgraf/NZ/HITLab/

Figure 34: Our AR agent explains the HITLab New Zealand

Because of time constraints, the thorough evaluation could not take place. The preparation, implementation and evaluation would have taken too much time. Promising were the preliminary informal tests that have shown a high excitement and commitment of the users. Eventually there shall be informal test with questionnaires and formal tests in a social dilemma situation to assess if and how good is the social interaction with the agent. Key concern will be its ability to flexibly handle customer needs. Through presenting an individual narrative experience enriched by available displays user satisfaction and information delivery in educational settings, e.g. a museum, should be improved. This is planned for a follow-up research work.

The drawback of this approach is, that non-verbal input, gestures and postures are not considered as inputs. Our agent is deaf and blind - not a good approximation of the real world. This disadvantage can be levelled by introducing ‘ears’ and ‘eyes’ to the computational model. Adding channels of communication requires the implementation of input processing, either automatic or manual. Automatic input processing would include speech recognition, natural language parsing, gesture recognition etc. – it is similar to the output pipeline, just the other way around, from visual speech to text. A reasoning unit than takes the input, processes it with some algorithm to find appropriate answers and sends it to the output modules (that we have described before). Manual input processing means that a human watches the user and his actions, and manually inputs his usual reactions to the output module. This ‘direct’ processing substitutes the language processing and the reasoning unit. Naturally, the human offers both. The second approach may look a bit ‘fake’, but indeed it is used in UI research quite frequently. So called ‘Wizard of Oz’ studies explore the usability of potential new interfaces whose functionality cannot be implemented with today’s technologies or whose complexity is to high to be realised in considerable time. The planed evaluation might be such a ‘Wizard of Oz’ study. For an introduction to and overview of WoZ studies see Dahlbäck et al. 1993.

Considering the implementation one could argue that the use of ART introduces a avoidable overhead, because only a small fraction of its broad functionality is used. Essentially, PantAR employs ART methods to initialise, open and grab video frames. We could do the same directly using DirectShow. The advantage of keeping ART as mediator between the camera

and the rendering engine is that we could easily implement other versions of the presenter, for example one that is world-aligned and thus needs marker recognition and calculation of positions. This is all done inside ART and we would not have to worry about the processing.

In Section 7 we have introduced PantAR, the extended animation and rendering engine in BEAT. It realises the synthetic human-like agent that gives the impression of a personal presenter in the user's field of view. Having begun with elaboration about basic considerations and different approaches to the problem, we summarised earlier work in this area to identify candidates for our further development. On the background of the context of our research and some design considerations, we presented our implementation and provided details of the realisation. After concluding with on overview of the results, we identified some potential problems and their solution. In the next section we will provide ideas for further research, both practical and theoretical, that originate from questions or problems that came up during our work.

8. DIRECTIONS FOR RESEARCH

Testing

PantAR works as an initial prototype and suits well for use as a testbed to determine some of the following questions.

1. Do students learn the content of the displays (more quickly, more easily, more thoroughly) in the ARE compared to real world experience?
2. Do they learn the facts presented in single displays, or the relation among the different displays?
3. Does the interaction with the agent increases enjoyment or comprehension of the presented material?
4. How likely are students to engage with the presenter (e.g. interruptions, clarification etc.)?
5. Do students show enthusiasm is the presented topic, e.g. are they more likely to ask additional questions?
6. Do students find the emotional side of interaction useful? Does it increase motivation to continue though difficult areas?
7. Do students find presentation more enjoyable with PantAR in contrast to classic displays and multi-media only exhibition?
8. Do students experience less frustration or confusion with PantAR in comparison to the usual displays?
9. Is motivation positively affected? Do students spend more time on the presentation run with the agent? Do they indicate a stronger desire to continue or a higher chance of success? Does the system increase self-confidence?
10. What is the minimum and maximum level of agent representation that people like and find usefully?
11. What is the influence of different parameters in the representation of the agent (e.g. style, size, behaviour, lifelikeness etc.) on the believability and likeness scale?

Implementation

Considering practical issues like licensing problems because of commercial software, it would be really advantageous to have the whole package as OpenSource. That would mean a migration from TGS OpenInventor to a free version of OpenInventor with the same functionality, e.g. COIN. And an equivalent substitute for the Conexxor's POS tagger has to be found. At the time of the implementation of PantAR, COIN lacked support for VRML but that feature was announce for future releases. Not essential but to ensure compatibility into the future and as more and more tools handle that format, MPEG-4 compatible body and face descriptions should be introduced.

Conceptual Design

The modular architecture of BEAT should be extended by an emotion module that can truly simulate emotional states. The emotional state can then affect the choice of utterance or animation in a particular situation. Working together the architecture could simulate behaviour like being happy because of successful completion of a task, or sad because of repeated misrecognitions. The incorporation of input channels should be designed. Extraction of facial features and body language to determine the emotional state of the conversational partner. Listening, body pose and gesture recognition are some essential pieces to make the presenter really autonomous. This could be another input channel into the behaviour management and language generation module that could generate more appropriate responses of the agent. It would yield in a more social agent, i.e. 'it not only understands me verbally but shows empathy'. But as much as we would wish this feature, it is hard to realise.

Theory

The concept of presence is important in virtual reality and mixed realities. It describes the feeling of the user to be 'there', in the synthetic environment, and feel the presents of the virtual characters or objects. In VR there is usually a distinct point in time when the feeling of presence swaps, the moment when closing the HMD and the moment when taking it off. Does the question of being present in either the real or the virtual world have any meaning anymore with AR that makes this distinction obsolete? AR smoothes the seam between reality and virtuality - they co-exist in one space! But how can the individual distinguish between them? And what consequences does the existence in the one realm have for the other? Which opportunities offer combined affordances (Gibson 1979) from real and virtual world to the person in terms of his effectivities? Maybe humans cannot imagine that today as the supporting technology is not widely available and other inventions have not been made yet. As these questions are more philosophical than touching computer science and user interface design we would let them to the appropriate people, most likely psychologists and sociologists.

Research

From research we know that there are culture invariant facial expression that almost anybody can produce and read. As most ECAs and their behaviour are made for English speaking people we ask if the non-verbal behaviour is appropriate for other cultures. Other societies with other languages might have different behaviours.

Our test system is designed for one user only. Presentations or other kind of collaboration experience is most likely a multi-user settings. The need for spatial audio becomes essential to distinguish between multiple simultaneous speakers. Another environmental aspect would be the rendering of virtual objects according to the real world light conditions at the position of the object. This would require the knowledge of light sources an the topology of the world.

Then you would even be able to draw natural looking shadows from virtual objects or overlay real world shadows to virtual objects.

Essential for researchers in the field of user interface agents is the development of a standard user test to assess the quality of the interface. Even if there are many applications and a diversity of areas where agents are employed, it should be possible. We all consider ourselves as humans even if we have different backgrounds, different jobs, and different experience. Thus any truly social able agent should be measured up to that expectation. To assess the quality, we could to set up a catalogue of questions to be answered by the designer or the human tester. The answers should not only aim on formal properties but on informal ones as well.

9. CONCLUSION

In the beginning of this paper we gave an introduction to Human-Computer Interaction and why agent could be a good idea to be employed there. We assessed a wide field of literature and elaborated on embodied agents, one half of the subject in this work. Then we reviewed multimodal systems developed over the last 20 years and presented results from user tests with them. Augmented reality, the second half of the subject of this work was introduced and related to agents. After that long part of introduction and necessary background, we have developed and structured design objectives for embodied agents in augmented reality. Some of these objectives and thoughts went into our prototype implementation that was explained in the last but one section. In the end we presented a diversity of questions and thoughts that might be interesting for other researchers as well.

Social Agents are advantageous in education, presentation and other soft skills focused areas. From the presented findings we can learn that certain properties in agents have ambivalent consequences. On the one hand humans treat computer and its interfaces in a human-like way, regardless of their appearance. Implications are manifold: By creating e.g. comic like characters with proportionally larger heads, clearly distinguishable facial features (especially the lips) and lip-synchronised speech designers could increase understanding. Typical for comic figures are exaggerated gestures that could improve the display of non-verbal behaviour. On the other hand the weak performance of non anthropomorphic agents in collaborative tasks suggests that we should be cautious about the application domain for which we design lovable or charming agents. Agents should be human-like not for its own sake but to increase the user's performance. Secondly augmented reality environments are a promising area of research and have a good chance to be accepted as intuitive interfaces. The prototype implementation of the communicative agent in AR was successful. A naturally behaving presenter engaged out guests in a conversation to deliver some information. The preliminary test was positive and suggests a further thorough evaluation. In the end of this paper we have identified promising directions for research. Originating from different areas the suggestions and questions could guide the reader to further, thrilling findings.

In this last paragraph I want to draw my personal conclusion of this project. I got to know how great it can be to be part of the research community. Without borders and in the best case without limits these people that are scattered all around the world can make the world go round. Not everything went right, but somebody help me out when I had a problem. They fostered my interest in doing research, of course in the area of user interfaces. AR is great! Interaction is fun!

REFERENCES

André, E., T. Rist, et al. (2000). The Automated Design of Believable Dialogues for Animated Presentation Teams. Embodied Conversational Agents. J. Cassell, J. Sullivan, S. Prevost and E. Churchill. Cambridge, CA, USA, MIT Press**:** 220-255.

Anabuki, M., Kabuka, H., Yamamoto, H. & Tamura, H. (2000), *Welbo: An Embodied Conversational Agent Living in Mixed Reality Space*, Videos of the 2000 Conference on Human Factors in Computing Systems (CHI'2000), pp.10-11. The Hague, The Netherlands.

Andre, E, Rist, T, Mueller, J. 1998. WebPersona: a lifelike presentation agent for the World-Wide Web, Knowledge Based Systems, 2 (1), 25-35.

Aukstakalnis, S. and D. Blatner (1992). Silicon Mirage - The Art and Science of Virtual Reality. Berkeley, CA, Peachpit Press.

Aylett, R., Horrobin, A., O'Hare, J., Osman, A. and M. Polshaw (1999), "Virtual Telebubbies: reapplying a robot architecture to virtual agents", in Proceedings of the Third International Conference on Autonomous Agents, New York, pp.514-515.

Azuma, R. (1997). A Survey of Augmented Reality, Presence, 6, 4, pp.355-385

Badler, N., Phillips, C., Webber, B. (1993). Simulating Humans: Computer Graphics Animation and Control, Oxford University Press, 1993.

Bajura, M., Fuchs, H., Ohbuchi, R. (1992) "Merging Virtual Objects with the Real World: Seeing Ultrasound Imagery Within the Patient." In Proceedings of SIGGRAPH '92, New York: ACM Press, pp. 203-210.

Ball, G. and J. Breese (2000). Emotions and personality in a conversational character. Embodied Conversational Agents. J. Cassell, J. Sullivan, S. Prevost and E. Churchill. Cambridge MA, USA, MIT Press**:** 189-219.

Bates, J. (1994). The role of emotion in believable agents. Communications of the ACM, 37(7): 122-125.

Beskow, J. (1995). Rule-based Visual Speech Synthesis. Proceedings of Eurospeech '95, Madrid, Spain.

Beskow, J., K. Elenius, et al. Olga - A Dialogue System with an Animated Talking Agent. Stockholm, Sweden, Department of Speech, Music and Hearing.

Beskow, J. and S. McGlashan (1997). Olga - a conversational agent with gestures. Workshop on Animated Interface Agents: Making them Intelligent. Nagoya, Japan.

Bolt, R. A. (1980). "Put-That-There": Voice and Gesture at the Graphics Interface. Computer Graphics, 14(3), 262-70.

Bolt, R. A. & Herranz, E. (1992a). Giving Directions to Computers via Speech, Gesture and Gaze. *Proceedings of UIST '92*.

Bolt, R. A. & Herranz, E. (1992b) Two-Handed Gesture in Multi-Modal Natural Dialog. Proceedings of *UIST '92, Fifth Annual Symposium on User Interface Software and Technology*, Monterey, CA, November 15-18. New York: Academic Press.

Boulic, R., Huang, Z., Shen, J., Molet, T., Capin, T., Lintermann, B., Saar, K., Thalmann, D., Magnetat-Thalmann, N., Schmitt, A., Moccozet, L., Kalra, P., and Pandzic, I. (1995). A system for the parallel integrated motion of multiple deformable human characters with collision detection. Computer Graphics Forum, vol. 13(3), pp. 337-348.

Brustoloni, J. C. (1991). Autonomous Agents : Characterization and Requirements. Pittsburgh, USA, Carnegie Mellon University.

Bruckert, E., Minow, M., and Tetschner, W. (1983). Three-tiered software and VLSI aid developmental system to read text aloud. *Electronics*.

Buxton, W. (1997). Living in Augmented Reality: Ubiquitous Media and Reactive Environments. In K. Finn, A. Sellen & S. Wilber (Eds.). *Video Mediated Communicatio*n. Hillsdale, N.J.: Erlbaum, 363-384.

Canamero, L. (2001). Building emotional artifacts in social worlds: Challenges and perspectives. In *Proceedings of 2001 AAAI Fall Symposium in Emotional and Intelligent II: The Tangled Knot of Social Cognition*: 22–30. AAAI Press, Technical Report FS-01-02.

Card, S. K., Moran, T. P., and Newell, A. (1983). The Psychology of Human-Computer Interaction. Hillsdale, NJ: Lawrence Erlbaum Associates.

Carlson, R. and B. Granström (1997). Speech Synthesis. The Handbook of Phonetical Science. W. Hardcastle and J. Laver. Oxford, Blackwell Publisher Ltd**:** 768-788.

Cassell, J., T. Bickmore, et al. (1999). Embodiment in Conversational Interfaces : Rea. Proceedings of CHI '99, Pittsburg, PA, USA.

Cassell, J., T. Bickmore, et al. (2000). Human conversation as a system framework: Designing embodied conversational agents. Embodied Conversational Agents. J. Cassell, J. Sullivan, S. Prevost and E. Churchill. Boston, MIT Press.

Cassell, J. and T. Stocky (2002). Shared Reality: Spatial Intelligence in Intuitive User Interfaces. IUI '02, San Francisco CA, USA.

Cassell, J., O. Torres, et al. (1999). Turn Taking vs. Discourse Structure : how best to model multimodal conversation. Machine conversations. Y. Wilks. Boston, Kluwer Academic.

Cassell, J., H. Vilhjálmsson, et al. (2001). BEAT: The Behavioural Expression Animation Toolkit. SIGGRAPH '01, Los Angeles, CA, USA.

Cassell, J., Stocky, T., Bickmore, T., Gao, Y., Nakano, Y., Ryokai, K., Tversky, D., Vaucelle, C., Vilhjálmsson, H. (2002). MACK: Media lab Autonomous Conversational Kiosk. In *Proceedings of IMAGINA'02*, Monte Carlo

Caudell, T.P., and Mizell, D.W. (1992)"Augmented Reality: an application of heads-up display technology to manual manufacturing processes." In Proceedings of the Twenty-Fifth Hawaii International Conference on Systems Science, Kauai, Hawaii, 7th-10th Jan. 1992, Vol. 2, pp. 659-669.

Chi, D., Costa, M., Zhao, L., Badler, N. (2000). The EMOTE Model for Effort and Shape. In *Proceedings of SIGGRAPH 2000*, ACM Computer Graphics Annual Conference, New Orleans, Louisiana, 23-28 July, 2000, pp. 173-182.

Cole, R., Massaro, D.M., de Villiers, J., Rundle, B., Shobaki, K., Wouters, J., Cohen, M., Beskow, J., Stone, P., Connors, P., Tarachow, A., and Solcher, D. (1999). New tools for interactive speech and language training: Using animated conversational agents in the classrooms of profoundly deaf children. In *Proceedings of ESCA/SOCRATES Workshop on Method and Tool Innovations for Speech Science Education*, London, UK, Apr 1999.

Collier, G. (1985). Emotional Expressions. Hillsdale NJ, USA, Lawrence Erlbaum.

Darwin, C. (1859). The Origin of Species by Means of Natural Selection: Or, the Preservation of Favored Races in the Struggle for Life. London: John Murray.

Dautenhahn, K. (1998). The art of designing socially intelligent agents – science, fiction, and the human in the loop. *Applied Artificial Intelligence, Special Issue on Socially Intelligent Agents*, 12 (7–8): 573–617.

Dahlbäck, N., Jönsson, A., and Ahrenberg, L. (1993). Wizard of Oz Studies – Why and How. *Proceedings of the 1st International Conference on Intelligent User Interfaces*: 193-200

DeCarlo, D., Revilla, C., Stone, M., and Venditti J. (2002). Making discourse visible: Coding and animating conversational facial displays. In *Computer Animation 2002*: pp. 11-16

De Soussa, C. S. (1993). The Semiotic Engineering of User Interface Languages. *International Journal of Man-Machine Studies, 39*: 753-773.

Dix, A.J., Finlay, J.E., Abowd, G.D., Beale, R. (1998). Human-Computer Interaction (Second Edition): 555-561. Prentice Hall Europe.

Doyle, P. (1999). When is a Communicative Agent a Good Idea? Third International Conference on Autonomous Agents, Seattle.

Eberts, R. (1994). User Interface Design. Englewood Cliffs, NJ: Prentice Hall.

Ekman, P. and Friesen, W., The Facial Action Coding System, Consulting Psychologists Press, San Francisco, CA, 1978.

El-Nasr, M. S., T. R. Ioerger, et al. (1999). A Pet with Evolving Emotional Intelligence. Proceedings of the 3rd International Conference on Autonomous Agents, Seattle, Washington, ACM Press, New York NY, USA.

Elliott, C., Rickel, J., and Lester, J.C. (1997). Integrating affective computing into animated tutoring agents. In Proceedings of the IJCAI Workshop on Animated Interface Agents: Making Them Intelligent, pages 113--121, Nagoya, Japan

Faulkner (1998). The Essence of Human-Computer Interaction. NY: Prentice Hall

Feiner, S., MacIntyre, B., and Seligmann, D. (1993) "Knowledge-Based Augmented Reality." Communications of the ACM, Vol. 36(7), pp. 53-62.

Fiske, S. (2004). Social beings. Wiley.

M. Fjeld, S. Guttormsen Schär, D. Signorello, H. Krueger (2002): Alternative Tools for Tangible Interaction: A Usability Evaluation. In Proceedings of the IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR 2002), pp. 157-166.

Foner, L. (1993). What's an agent, anyway? A sociological case study. Agents Memo 93-01, MIT Media Lab, E15-305 20 Ames St., Cambridge, Mass.

Fosnot, C. T. (Ed.) (1996). Constructivism: Theory, Perspectives, and Practice. Teachers College Pr.

Franklin, S. and A. Graesser (1996). Is it a agent or just a program? A taxonomy for autonomous agents. Agent Theories, Architectures and Languages. Berlin, Germany, Springer-Verlag**:** 21-95.

Gav, G., Lentini, M. "Use of Communication Resources in a Networked Collaborative Design Environment." http://www.osu.edu/units/jcmc/IMG_JCMC/ResourceUse.html.

Gibson, J. J. (1979). The ecological approach to visual perception. Boston, Houghton Mifflin.

Goleman, D. (1995). Emotional Intelligence: Why it can matter more than IQ. New York: Bantam.

Gould, J.D. (1981). Composing letters with computer-based text editors. *Human Factors, 23*: 593-606.

Gratch, J. (2000). Emile : Marshalling Passions in Training and Education. Proceedings of the 4th International Conference on Autonomous Agents, Barcelona, Spain.

Gratch, J., J. Rickel, et al. (2002). "Creating Interactive Virtual Humans : Some Assembly Required." IEEE Intelligent Systems (July/August).

Griffin, Em (2002). A First Look at Communication Theory. 5th Edn. London, New York: McGraw-Hill.

Hansen, W.J., Doring, R., and Whitlock, L.R. (1978). Why an examination was slower on-line than on paper. *International Journal on Man-Machine Studies, 10*, 507-519.

Harada, S., Hwang, J., Lee, B., and Stone, M. (2003).“Put-That-There”: What, Where, How? Integrating Speech and Gesture in Interactive Workspaces. UBIHCISYS 2003 Online Proceedings. Stanford University. http://ubihcisys.stanford.edu/online-proceedings/Ubi03w7-Harada-final.pdf

Hayes-Roth, B. and Doyle, P. (1998). Animate characters. Autonomous Agents and Multi-Agent Systems, 1 (2):195–230.

Idel, M. (1990). Golem : Jewish Magical & Mystical Traditions on the Artificial Anthropoid. Albany, USA, State University of New York Press.

Isbister, K. and P. Doyle (2002). Design and Evaluation of Embodied Conversational Agents : A Proposed Taxonomy. AAMAS '02 Workshop on Embodied Conversational Agents.

Johnston, M., P. R. Cohen, D. McGee, S. L. Oviatt, J. A. Pittman, and I. Smith. 1997. Unification-based multimodal integration. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*: 281-288.

Johnston, M. 1998. Unification-based multimodal parsing. Proceedings of COLING–ACL 98: 624-630

Johnson, W. L. (1995). Pedagogical Agents in Virtual Learning Environments. International Conference on Computers and Education.

Johnson, W. L., J. W. Rickel, et al. (2000). "Animated Pedagogical Agents : Face-to-Face Interaction in Interactive Learning Environments." International Journal of Artificial Intelligence in Education **11**: 47-78.

Kato, H., Billinghurst, M., Asano, M., Tachibana, K. (1999) An Augmented Reality System and its Calibration based on Marker Tracking, *Transactions of the Virtual Reality Society of Japa*n, Vol.4, No.4, pp.607-616, 1999

Kay, A. 1984. Computer software. Scientific American 251, 3 (Sept.), 52-59.

Kay, A. 1990. User interface: A personal view. In The Art of Human-Computer Interface Design, Brenda Laurel, Ed. Addison-Wesley, Reading, Mass., 191-207.

Kendon, A. (1990). A negotiation of context in face-to-face interaction. Rethinking context : language as an interactive phenomenon. A. Duranti and C. Goodwin. Cambridge England ; New York, Cambridge University Press.

Klein, G. A. (1987). Analytical Versus Recognitional Approaches To Design Decision Making. System design : behavioral perspectives on designers, tools, and organizations. W. B. Rouse and K. R. Boff. New York, North-Holland**:** 175-186.

Klein, J T. 1999. Computer Response to User Frustration, MSc Thesis at Massachusetts Institute of Technology.

Koda, T and Maes, P. 1996. Agents with Faces: The Effects of Personification of Agents, Proceedings of HCI'96, London, 1996.

Koons, D. B., Sparrell, C. J. & Thorisson, K. R. (1993). Integrating Simultaneous Input from Speech, Gaze and Hand Gestures. Chapter 11 in M. T. Maybury (ed.), *Intelligent Multi-Media Interfaces*, 252-276. Cambridge, MA: AAAI Press/M.I.T. Press.

Koons, D. B. & Thorisson, K. R. (1993). Estimating Direction of Gaze in Multimodal Context. Presented at *3CYBERCONFÑThe Third International Conference on Cyberspace*, Austin TX, May 13-14.

Kozierok, R., and Maes, P. 1993. A learning interface agent for scheduling meetings. In *Proceedings of the ACM-SIGCHI International Workshop on Intelligent User Interfaces* (Orlando, Fla., Jan.). ACM Press, New York, N.Y., 81-88.

Kovar, L., Gleicher, M., and Pighin, F.(2002). Motion Graphs. Proceedings of ACM SIGGRAPH 2002.

Kozar, K.A., and Dickson, G.W. (1978). An experimental study of the effect of data display media on decision effectiveness. *International Journal of Man-Machine Studies, 10*, 494-505.

Laird, J. E. (2001). It Knows What You're Going To Do: Adding Anticipation to a Quakebot. Proceedings of the Fifth International Conference on Autonomous Agents.

Laurel, B. 1990. Interface agents: Metaphors with character. In The Art of Human-Computer Interface Design, Brenda Laurel, Ed. Addison-Wesley, Reading, Mass., 355-365.

Laurel, B. 1991. Computers as Theatre. Addison-Wesley, Reading, Mass.

Lester, J. C., S. Converse, et al. (1997). The Persona Effect : Affective Impact of Animated Pedagogical Agents. Proceedings of CHI '97, Atlanta, GA, USA.

Lester, J. C., S. G. Towns, et al. (2000). Deictic and Emotive Communication in Animated Pedagogical Agents. Embodied conversational agents. J. Cassell, J. Sullivan and S. Prevost. Cambridge, Mass., USA, MIT Press.

Lester, J. C., J. L. Voerman, et al. (1997). A Life-like Pedagogical Agent with Deictic Believability. Working Note of the IJCAI '97 Workshop on Animated Interface Agents: Making Them Intelligent, Nagoya, Japan.

Lester, J., Voerman, J.L., Towns, S.G., and Callaway, C.B. (1997). Cosmo: A life-like animated pedagogical agent with deictic believability. In *Proc. of the IJCAI97 Workshop on Animated Interface Agents: Making them Intelligent*, Nagoya.

Leung W. H., Tseng B., Shae Z.-Y., Hendriks F. and Chen T. "Realistic Video Avatar", IEEE International Conference on Multimedia and Exposition, New York, July 2000.

Littlejohn, S. W. (1996). Theories of human communication. Belmont, CA: Wadsworth Publishing Company.

Maes, P. and B. Shneiderman (1997). "Direct Manipulation vs. Interface Agents : a Debate." Interactions **5**(6).

Marsella, S. C. and J. Gratch (2001). Modeling the Interplay of Emotions and Plans in Multi-Agent Simulations. Proceedings of the 23rd Annual Conference of the Cognitive Science Society, Edinburgh, Scotland.

Marsella, S. C. and J. Gratch (2002). A Steps Towards Irrationality : Using Emotion to Change Belief. Proceedings of the 1st International Joint Conference on Autonomous Agents and Mulit-Agent Systems, Bologna, Italy.

Marsella, S. C., W. L. Johnson, et al. (2000). Interactive Pedagogical Drama. Proceedings of the Fourth International Conference on Autonomous Agents, ACM Press.

Massaro, D. (1998). Perceiving talking faces: From speech perception to a behavioral principle. Cambridge, Massachusetts: MIT Press.

McBreen, H., Shade, P., Jack, M., and Wyard, P. (2000). Experimental assessment of the effectiveness of synthetic personae for multi-modal e-retail applications. In Proceedings 4th International Conference on Autonomous Agents (Agents'2000), pages 39–45.

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264: 764-748.

McIntyre, F., Estep, K.W., Sieburth, J.M. (1990). Cost of user-friendly programming. *Journal of Forth Application and Research, 6 (2),* 103-115.

McTear, M. (1999). Using the CSLU Toolkit for Practicals in Spoken Dialogue Technology. In *Proceedings of ESCA/SOCRATES Workshop on Method and Tool Innovations for Speech Science Education,* London, UK, Apr 1999.

Morishima, S. and Harashima, H. (1991). *A Media Conversion from Speech to Facial Image for Intelligent Man-Machine Interface*. In IEEE Journal on Selected Areas in Communications 9(4), 594-600.

Nareyek, A. (2000) "Intelligent Agents for Computer Games", in Proceedings of the Second International Conference on Computers and Games.

Negroponte, N. 1990. Hospital Corners. In The Art of Human-Computer Interface Design, Brenda Laurel, Ed. Addison-Wesley, Reading, Mass., 347-353.

Negroponte, N. 1995. Being Digital. Alfred A. Knopf, New York, N.Y.

Nielsen, J. (1993). Usability Engineering. Boston, MA: Academic Press.

O'Conaill, B. and S. Whittaker (1997). Characterizing, predicting, and measuring video-mediated communication: A conversational approach. In Video-mediated communication: Computers, cognition, and work. K. E. Finn and A. J. Sellen. Mahwah, N.J., Lawrence Erlbaum Associates, Inc.: 107-131.

Ohshima, T., Satoh, K., Yamamoto, H. and Tamura, H.(1998) "AR2 Hockey: A Case Study of Collaborative Augmented Reality," Proc. IEEE VRAIS '98, pp.268-275 .

Ortony, A., G. Clore, et al. (1988). The Cognitive Structure of Emotions. Cambridge, Cambridge University Press.

Oviatt, S. (1999). Mutual Disambiguation of Recognition Errors in a Multimodal Architecture. *Proceeding of the CHI 99 conference on Human factors in computing systems*: 576 - 583

Oviatt, S. L. 1996. Multimodal interfaces for dynamic interactive maps. In *Proceedings of Conference on Human Factors in Computing Systems*, 95–102.

Owen, D. (2003). Cognitive Ergonomics. Organisational Psychology in Australia and New Zealand. M. O'Driscoll, P. Tylor and T. Kalliath, Oxford University Press**:** 239-261.

Parise, M., S. Kiessler, et al. (1996). My partner is a Real Dog : Cooperation with Social Agents. Proceedings of Computer Supported Cooperative Work '96, Cambridge MA, USA, ACM Press, New York, NY, USA.

Pelauchaud, C., V. Carofiglio, et al. (2002). Embodied Agent in Information Delivery Application. Proceedings of Autonomous Agents and Multiagent Systems.

Perlin, K. and A. Goldberg (1996). "Improv: A system for scripting interactive actors in virtual worlds". In *Proceedings of ACM Computer Graphics Annual Conference 1996*, 205-216.

Picard, Rosalind (1997) Affective computing. Cambridge, MA/London: The MIT Press.

Poupyrev, I., Billighurst, M. Kato, H., May, R.(2000) "Integrating Real and Virtual Worlds in Shared Space." In Proceedings of the 5th International Symposium on Artificial Life and Robotics (AROB 5th'00), Oita, Japan, 26-28 January 2000, Vol. 1, pp. 22-25.

Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S., and Carey, T. (1994). Human-Computer Interaction. Wokingham, UK: Addison Wesley.

Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X.-F., Kir-bas, C., McCullough, K. E., & Ansari, R. (2002). Multimodal human discourse: gesture and speech. *ACM Transactions on Computer-Human Interaction (TOCHI*), 9, 171–193.

Reeves, B. and C. I. Nass (1996). The media equation : how people treat computers, televisions, and new media like real people and places. Stanford, Calif., New York; Cambridge, Cambridge University Press.

Rickel, J. and W. L. Johnson (1999). "Animated Agents for Procedural Training in Virtual Reality: Perception, Cognition, and Motor Control." Applied Artificial Intelligence **13**: 343-382.

Rickel, J. and W. L. Johnson (2000). Task-Orientated Collaboration with Embodied Agents in Virtual Worlds. Embodied conversational agents. J. Cassell, J. Sullivan, S. Prevost and E. Churchill. Cambridge, Mass., USA, MIT Press.

Rickel, J., Gratch. J., Hill, R., Marsella, S., and Swartout, W. (2001). *Steve goes to Bosnia: Towards a new generation of virtual humans for interactive experiences*. In AAAI Spring Symposium on Artificial Intelligence and Interactive Entertainment

Rolland, J.P., Baillot, Y., and Goon, A.A. (2001): „A SURVEY OF TRACKING TECHNOLOGY FOR VIRTUAL ENVIRONMENTS", In Fundamentals of Wearable Computers and Augmented Reality, pp.67-112.

Rosenberg, D. (1989). Cost-benefit analysis for corporate user interface standards: What price to pay for a consistent "look and feel". In J. Nielsen (Ed.), *Coordinating user interfaces for consistency* (pp. 21-34). New York: Academic.

Russell, S. J. and P. Norvig (1995). Artificial Intelligence : A Modern Approach. Englewood Cliffs, NJ, USA, Prentice Hall.

Sánchez, J.A. (1997). A Taxonomy of Agents. Technical Report No. ICT-97-1, Interactive and Cooperative Technologies Lab, Universidad de las Américas-Puebla, Cholula

Santoro, G. M. (1995). What is Computer Mediated Education?. In Zane L. Berge and Mauri P. Collins (Eds.), *Computer Mediated Communication and the Online Classroom*, volume 1: Overview and Perspectives, Hampton Press.

Schmalsteig, D., Fuhrmann, A., Szalavari, Z., Gervautz, M., Studierstube - An Environment for Collaboration in Augmented Reality. In CVE '96 Workshop Proceedings, 19-20th September 1996, Nottingham, Great Britain.

Schmalstieg, Fuhrmann, Hesina, Szalavari, Encarnação, Gervautz, Purgathofer (2002). The Studierstube Augmented Reality Project. In PRESENCE - Teleoperators and Virtual Environments 11(1), MIT Press.

Shneiderman, B. (1998). Designing the user interface : strategies for effective human-computer-interaction. Reading, Mass, Addison Wesley Longman.

Shneiderman, B. (2002). Leonardo's laptop : human needs and the new computing technologies. Cambridge, Mass., MIT Press.

Silva, D., Siebra, C., Valadares, J., Almeida, A., Frery, A., and Ramalho, G. (1999). Personality-Centered Agents for Virtual Computer Games, In *Proceedings of Virtual Agents 99*, Workshop on Intelligent Virtual Agents, Salford, UK.

Sloman, A. (1990). Motives, Mechanisms and Emotions. Cognition and Emotion. The Philosophy of Artificial Intelligence. M. A. Boden, Oxford University Press**:** 231-247.

Smith, S.L., and Mosier, J.N. (1984). Design Guidelines for the user interface for computer-bases information systems. Bedford, MA: The MITRE Corp.

Sparrell, C. J. (1993). Coverbal Iconic Gesture in Human-Computer Interaction. M.S. Thesis. Cambridge, MA: Massachusetts Institute of Technology.

Sparrell, C. J. & Koons, D. B. (1994). Capturing and Interpreting Coverbal Depictive Gestures. AAAI 1994 Spring Symposium Series, Stanford, USA, March 21-23, 8-12.

Stone, P. (1999). Revolutionizing Language Instruction in Oral Deaf Education. In *Proceedings of the International Conference of Phonetic Sciences*, San Francisco, CA, Aug 1999.

Sundblad, O. and Y. Sundblad OLGA - a Multimodal Interactive Information Assistant. Stockholm, Schweden, CID - Center for User Orientated IT Design.

Thalmann, D. and Vexo, F. MPEG-4 Character Animation. Virtual Reality Laboratory, Swiss Federal Institute of technology, Lausanne, Switzerland

Towns, S.G., Callaway, C.B., Voerman, J.L., and Lester, J. Coherent Gestures, Locomotion, and Speech in Life-Like Pedagogical Agents. Multimedia Laboratory, Department of Computer Science, North Carolina State University

Thórisson, K. R. (1997). Gandalf: An Embodied Humanoid Capable of Real-Time Multimodal Dialogue with People**.** First ACM International Conference on Autonomous Agents, Mariott Hotel, Marina del Rey, California, February 5-8, 1997, 536-7.

Thórisson, K. R. (1996). Communicative Humanoids – A Computational Model of Psychosocial Dialogue Skills**.** Doctoral Dissertation. Massachusetts Institue of Technology. Cambridge, USA.

Thórisson, K. R., Cassell, J. (1996). Why Put An Agent In a Body : The Importance of Communicative Feedback in Human-Humanoid Dialoque. Presented at Lifelike Computer Characters, Utah, October 1996.

Torre, R., Balcisoy, S., Fua, P., Ponder, M., and Thalmann, D. (2000). Interaction Between Real and Virtual Humans: Playing Checkers, Eurographics Workshop on Virtual Environments}, Amsterdam, Netherlands, June 2000.

Turing, A. (1950). Computing machinery and intelligence. *Mind, 236*: 433-60.

Vacchetti, L., Lepetit, V.. Papagiannakis, G., Ponder, M., Fua, P., Magnenat-Thalmann, N., and Thalmann, D. (2003). Stable Real-Time Interaction Between Virtual Humans and

Real Scenes. In International Conference on 3-D Digital Imaging and Modeling, Banff, Alberta, Canada, October 2003.

Vosinakis, S, and Panayiotopoulos, T. (2000). A tool for constructing 3D Environments with Virtual Agents. Knowledge Engineering Laboratory, Department of Informatics, University of Piraeus, Piraeus, Greece.

Waters, K. (1987). A muscle model for animating three-dimensional facial expression. *Computer Graphics*, 21(4):17–24.

Waters, K. & Levergood, T. M. (1993) "DECface: an automatic lip synchronization algorithm for synthetic faces". Technical Report CRL 93/4, DEC Cambridge Research Laboratory, Cambridge, MA.

Wooldridge, M., and Jennings, N. (1995). Agent theories, architectures and languages: A survey. In *Intelligent Agents*, M. Wooldridge and N. Jennings, Eds. Springer-Verlag, New York, N.Y., 1-39.

Yoon, S., B. Blumberg, et al. (2000). Motivation Driven Learning for Interactive Synthetic Characters. Proceedings of the Fourth International Conference on Autonomous Agents, ACM Press.

WEB REFERENCES

(ordered alphabetically regarding their titles, all valid on 01.03.2004)

[Web1] 100 Years and Counting -- Census Bureau Celebrates Centennial. U.S. Census Bureau. http://www.census.gov/pubinfo/www/photos/centenial.html

[Web2] ACM SIGCHI Curricula for Human-Computer Interaction. Hewett, Baecker, Card, Carey, Gasen, Mantei, Perlman, Strong and Verplank (1996). http://sigchi.org/cdg/

[Web3] Animated Speech: Research Progress and Applications. Massaro, D., Cohen, M., Tabain, M., Beskow, J. & Clark, R. (2002). Perceptual Science Laboratory, University of California, Santa Cruz, CA, USA. http://mambo.ucsc.edu/pdf/massaroetal1.pdf

[Web4] COMM 103: Interpersonal Communication. Rowley, R. (1999). Communication Department, Mt. San Jacinto College. http://www.aligningaction.com/prgmodel.htm

[Web5] Computer Modell Katalog – Zeitlinie. http://home.t-online.de/home/cyrill.cmk/history.htm (01.03.2004)

[Web6] Ereignisse in der Geschichte des Computers. http://www.adp-gmbh.ch/personal/histoire/histoire.html

[Web7] Facts for Features Photos. U.S. Census Bureau. http://www.census.gov/pubinfo/www/photos/FFFPhotos/FFFPhotos.html

[Web8] Grundlagen der Informatik und der Numerik. Meiler, M. (2002). Institut für Informatik,Universität Leipzig. http://www.informatik.uni-leipzig.de/~meiler/GL.dir/GLWS02.dir/V01_Inf_Num.pdf

[Web9] Human Computer Interaction. Pelletier, G. (2000). Lulea Tekniska Universitet. http://www.cdt.luth.se/~pelle/smd045/

[Web10] Human Computer Interaction. Universita degli Studia di Roma La Sapienza. http://cesare.dsi.uniroma1.it/~ium/

[Web11] Interaction As a Topic of Its Own Right. Duncker, E.. School of Computing Science, University of Middlesex. http://www.cs.mdx.ac.uk/staffpages/elke/cmt3210/lectures%202002/CMT%203210%20lecture%209.PPT

[Web12] LANGUAGE AND CULTURE: An Introduction to Human Communication. O'Neil, D. (2004). http://anthro.palomar.edu/language/

[Web13] MagicBook website http://www.hitl.washington.edu/magicbook/

[Web14] MIT Media Lab Europe Homepage: http://www.medialabeurope.org

[Web15] MPEG-4: A Multimedia Standard for the Third Millennium, Part 1. IEEE Computer Society. http://www.computer.org/multimedia/articles/mpeg4_1.htm
[Web16] Meltdown at Three Mile Island. http://www.pbs.org/wgbh/amex/three/
[Web17] Responsive Face - Facial animation demo. Perlin, K.. http://www.mrl.nyu.edu/perlin/facedemo/
[Web18] Specification for a Standard Humanoid, Version 1.1. Web3D Working Group on Humanoid Animation (August 1999). http://h-anim.org/Specifications/H-Anim1.0/
[Web19] Three Mile Island. http://en.wikipedia.org/wiki/Three_Mile_Island/
[Web20] User Interface – Diskurs. http://diskurs.digital.udk-berlin.de/wiki/index.php/User-Interface