Proceedings of the 23rd IASTED International Multi-Conference
ARTIFICIAL INTELLIGENCE AND APPLICATIONS
February 14-16, 2005, Innsbruck, Austria

INTRODUCING AN EMBODIED VIRTUAL PRESENTER AGENT IN A VIRTUAL MEETING ROOM

Anton Nijholt, Herwin van Welbergen and Job Zwiers
Human Media Interaction Group
University of Twente
Enschede, the Netherlands
anijholt@cs.utwente.nl

ABSTRACT
In this paper we survey our research on modeling presentations in virtual environments. This research is performed in several of our research projects, in particular the European FP6 AMI (Augmented Multi-party Interaction) project. This project is about capturing and modeling of meetings. One of our aims in the project is to have real-time transformation of events during a meeting – hence, including presentations – to similar events in a virtual meeting room. Another aim is to model presenters and presentation making in order to make it possible to have remote presenters, anytime presenters, and modified presentations in the virtual meeting environment. During meetings presentations are often done using a data projector and PowerPoint sheets. These are the presentations that are mainly discussed in this paper. Timing and synchronizing of the multi-modal outputs displayed by the embodied virtual presenter is discussed.

KEY WORDS
Virtual reality, multimedia presentation, embodied agents.

1. Introduction

Information has to be presented. It should be possible to interact with media that allow the presentation of information. There are many ways to present information. When we ask directions on the street we can get quite detailed information, about how to go from the current location to a desired location. The explanation consists of verbal and nonverbal utterances, that is, sometimes the verbal utterances support pointing gestures or gestures that explain objects (landmarks) and situations that will be met, sometimes the gestures support the verbal utterances. A museum guide verbally and nonverbally (using gestures and gaze) interacts with his or her audience to explain the interesting parts of a sculpture or painting, addressing one, several or all persons in his or her audience. And, after having explained a piece of art, the guide will make it clear where to go or look next, again by verbal and nonverbal means of addressing the audience. Yet an other way of presenting information is to give a presentation, using overhead, data, or video projector. Clearly, also in these cases we can expect that the presenter will use deictic references to pictures, bullets, and texts fragments that appear during a presentation, e.g., a power-point presentation, on the screen. In this paper we introduce our research on presenting, where presenting means that - verbally and nonverbally - explanations are added by a virtual human-like presenter, to a scene or an object that is visible to this presenter's audience.

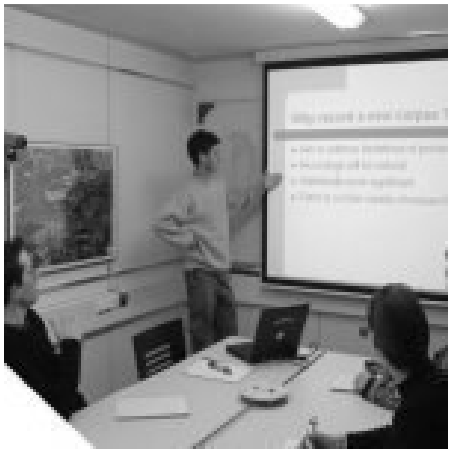

Figure 1, AMI Meeting Presentation

2. Background of our Research

Our research on presenting is – in principle - not tuned to one particular application. However, it is part of research in a FP6 European IP on Augmented Multi-party Interaction (AMI: http://www.amiproject.org/) [1]. This project is concerned with the modeling of interactions in a smart meeting environment, with the aim to provide real-time support to the meeting partners and to allow off-line multimedia retrieval and multimedia browsing of information obtained from a particular meeting. For that reason our examples are drawn from the domain of meetings. However, as explained in [2], the technology and the models that are being developed and designed can find their way in all kinds of applications of smart environments, including museum and exhibition environments that give real-time support to their visitors.

The technology that is being developed allows to detect, track and identify people in a particular environment and to interpret their activities and their interaction with other people or with objects and locations in the environment. Our main line of research is to translate our findings on meeting modeling into tools (meeting assistants, meeting browsers, meeting visualization, etc.) that can be used in real-time during a meeting or offline to browse through a previous meeting. Part of this research is how to model presentations during meetings (cf. Figure 1) and how to translate them to sufficiently realistic presentations in a virtual reality representation.

A FP6 European project that is related to our AMI project is CHIL (Computers in the Human Interaction Loop). This project is also on modeling multimodal interactions, but rather than to concentrate on meetings, the current activities concentrate on studying a presenter and its interaction with its audience in, e.g., a lecture room [3].

3. A Corpus of Presentations

In the AMI project a corpus of meetings is being collected. These meetings contain whiteboard and PowerPoint presentations. Especially in this latter case, we have a presenter explaining what is already visible on the screen, not essentially different (apart from the content) from a guide in a museum that explains a painting or a sculpture to a group of tourists visiting the museum. Gestures are made, there is pointing to the interesting parts, and there is some interaction with the audience, verbally and nonverbally. The corpus that was available at the start of the AMI project consisted of mock-up meetings, where, for example, during a meeting someone stands up to deliver a presentation. During the project new corpora will emerge, depending on the research interests of the different partners in this large-scale European project. One, fully unstructured corpus that has been added is a series of thirty videos of presentations during a workshop associated with the project. Designing models for multimodal human presentation (see among others [4]) goes together with designing annotation schemes and annotating presentations from the corpora. For this purpose some annotation tools have already been developed that make it possible to relate spoken content with gestures. New annotation tools are developed that also take into account pointing and other gestures that refer to parts of a scene, e.g. part of a painting, a sheet of a PowerPoint presentation or a location.

4. Towards Virtual Presenters

How can we put knowledge about presentations in use? That is, how can we model this knowledge in software and hardware and then give computer support in situations where presentations need to be delivered? Our aim is to have virtual presenters available on websites, 3D and virtual reality environments that are human-like (i.e., embodied conversational agents) and in which this presentation knowledge is modeled. Examples of such presenters have already been made available [5,6], either in 2D or 3D form. Related research for a physical robot presenter has also been performed [7]. Our work is related to that of Noma and Badler [8]. Their presenter can make presentations in a 3D virtual environment or on the WWW. It gets its input from speech texts with embedded commands that relate to the presenter's body language. This presenter behaves as a TV presenter, e.g., a weather report presenter that knows about the camera and the presentation screen, but does not have an audience in its direct environment.

Our presenter (see Figure 2) will deliver its presentation using a wide range of multi-modal channels, including speech, gesture and the use of sheets. The presentations are generated from a script describing the synchronization between those channels. For now, those scripts are generated by hand, based on annotation of existing presentations. In the future, we would like to be able to generate those scripts from the presentation text, and modify the presenters' behavior, based on his personality and emotional state.

To display the sheets, the virtual presenter uses a projector screen. This screen is a visual 3D entity in a virtual environment (a meeting room) displaying sheets. On these sheets, areas of interest are defined, at which the presenter can point.

5. Synchronizing Multi-modal Output

Timing and synchronizing of multi-modal outputs is still a challenging problem in embodied conversational agent

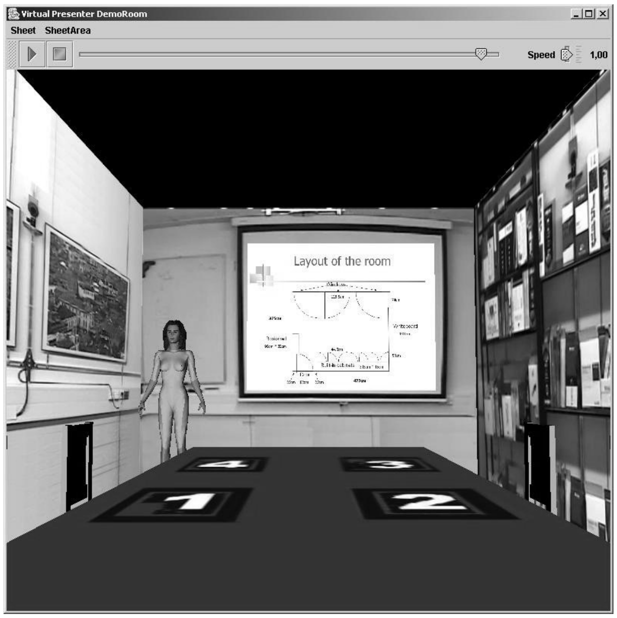

Figure 2, Presentation in the Virtual Meeting Room

research [9]. Some approaches tackle this problem by synchronizing the expressions on all modalities by using time stamps with constant values (NITE-XML, CoGest, etc). While this is fine for annotating multi-modal input, this limits the flexibility for multi-modal output, because in such an approach it is necessary to know the time needed for all actions beforehand.

Other approaches choose one modality to be the 'master' (usually the speech modality) and let this modality determine the execution of expressions on other modalities (MURML, etc.). However, during multi-modal interaction there is no single modality that always determines the synchronization of the other modalities. Even if there would be such a thing as a leading modality, then it would not always be the same modality. By example, while speaking hand gestures can be modeled as being guided by speech, but when taking a break to drink a glass of water during a talk, the hand movement becomes the leading modality.

SMIL-like languages are used for synchronization in multi-modal outputs in languages like CML, STEP and VHML. SMIL has a par, seq and wait tags to create respectively parallel actions, sequential actions and a wait action (used to let one of the parallel actions start later). With those three tags it is possible to express every possible way of synchronization. Using these kinds of constructions might work fine when the amount of different modalities is small. For a greater amount of modalities SMIL might become less manageable especially because different modalities are not clearly separated in SMIL. Another problem with those languages is that constructions (by example: a par wrapped around the complete script) can be realized that require a parser to walk through the complete script before it can be executed. This kind of constructions can possibly take a long time to plan and dynamically changing such a script will be hard.

To solve the issues with the different approaches in the languages described above, a synchronization language describing the generation of multi-modal expressions should satisfy the following constraints:

- The synchronization should not rely on constant time values;
- It should be possible to change the synchronizing modality over time;
- For easy parsing and reading, the modalities should be clearly separated in the language;
- It should be possible to read the language as a 'stream', that way the multi-modal actions can be executed before a script written in the language is fully parsed and new expressions can be added to the script dynamically.

We developed the MultimodalSync language in order to be able to use a language that satisfies the constraints mentioned above. The MultimodalSync language is based on SMIL and the BEHAVIOUR language used in our ANGELICA project [10,11] on embodied information presentations. In BEHAVIOUR, timestamps are set on a verbal channel. Gestures in the nonverbal channels can be synchronized with expressions in the verbal channel, using these timestamps.

In MultimodalSync, the multi-modal expressions are separated in channels. Different channels are executed in parallel. Expressions within one channel are executed in sequence. Each channel can define synchronization points. Expressions within a channel can also define synchronization points. Using the UseSynchronisationPoint tag, a channel can be synchronized with other channels, by waiting for a time value. This time value can be defined in a (to be defined) expression language, which makes use of synchronization points defined in other layers. It is also possible to use synchronization points within a channel, by example to synchronize the stroke of gestures with a certain word in a verbal expression.

MultiModalSync places no restrictions on the abstraction level of the expressions to be synchronized, or on the combination of abstraction depths in the different channels. The expressions can be on the level of intention, but also on the level of completely defined body animations.

Figure 3 shows an example of a MultimodalSync script, using a verbal, sheet-control and deictic channel. Figure 4 shows how the expressions in this script are synchronized in time.

```
<MultimodalSync>
  <Segment>
    <Channel name="verbal">
      <DefineSynchronisationPoint id="t0"/>
       ...
      <DefineSynchronisationPoint id="t1"/>
      <Verbal>Okay, let me just go through
       the constraints first and then we can
       continue this discussion if there's
       still need.
      </Verbal>
      <DefineSynchronisationPoint id="t2"/>
    </Channel>
    <Channel name="sheetcontrol">
      <UseSynchronisationPoint value="t2"/>
      <ChangeSheet name="sheet1">
    </Channel>
  </Segment>
  <Segment>
    <Channel name="verbal">
      <UseSynchronisationPoint
       value="t3+2"/>
      <Verbal>So, the
       <DefineSynchronisationPoint id="t4"/>
       bookshelf right now is sitting here.
      </Verbal>
    </Channel>
    <Channel name="deictic">
      <Point target="bookshelf"
       stroke=”t4”>
    </Channel>
  </Segment>
</MultiModalSync>
```

Figure 3, Example MultiModalSync Script

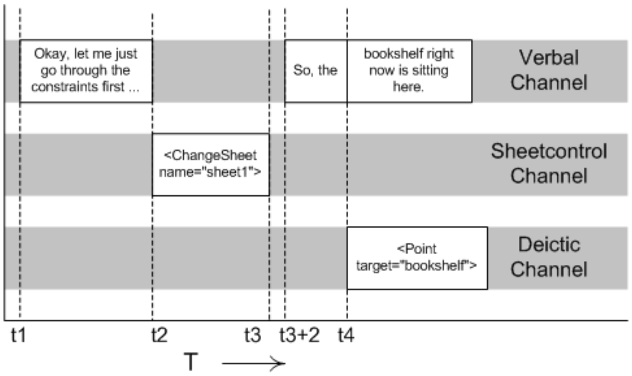
Okay, let me just go through the constraints first ...
So, the
bookshelf right now is sitting here.
Verbal Channel
<ChangeSheet name="sheet1">
Sheetcontrol Channel
<Point target="bookshelf">
Deictic Channel
t1
t2
t3
t3+2
t4
T ⟶

Figure 4, Timeline of the Expressions generated by the Example Script

Stream-ability is achieved by creating segments that contain a part of the expressions in each layer. In following segments, the expressions on the same layer are supposed to be executed in sequence; however, expressions in a different layer from following segments can overlap with a layer in the current segment.

More details about the technical implementation of the presenter can be found in [12].

6. Putting Virtual Presenters in Virtual Environments

In Figure 22 we already saw our virtual presenter in a virtual environment that allows presentations on a screen. Pictures, photographs, paintings, and sequences of these objects (e.g., a PowerPoint presentation) can be displayed on this virtual screen, and the virtual presenter can be provided with a script to explain and give comments, as explained in section 4. However, where is the audience? Clearly, we can introduce a fully generated virtual audience, as has been done in the experiments of Slater [13]. However, it is much more interesting to make a link with a real audience. We make a distinction between different types of audiences:

- An audience that is physically present during a presentation that is done by a human presenter in the same room;
- An audience that is provided with a presentation done by a virtual presenter without the ability to interact or to become aware of others that are accessing the same presentation;
- An audience that is physically present during a presentation that is done by a human presenter in the same room but that has also real-time access to a virtual reality representation of what is going on during the presentation, maybe allowing them to get additional information about the presentation;
- Audiences that in addition consist of one or more persons that are not physically present in the presentation location, but that can take a virtual position in this audience from a remote location and from that location have a real-time view on things that are going on in the presentation location. They can be represented visually as virtual humans, observable for other members of the audience. In more advanced applications they also can take part in discussions and interactions with the virtual presenter;
- An audience that is interested in what has been going on during a presentation in the past. In this case we assume that the audience has facilities to browse through the presentation event, to retrieve information about the presentation event, and to ask meta-questions about the presentation event (who were present?, who asked a question about what?, was there consensus?, was there someone who didn’t agree?, etc.).

In our research in the AMI project our first aim is to model a virtual presenter (as discussed in section 4, and displayed in Figure 2), to allow users access to presentations provided by a virtual presenter and, much more interesting, to put the presenter in a virtual environment allowing us to represent the presenter’s audience in various ways (off-line representation of the physically present audience, real-time representation of the physically present audience, allowing remote participation of the event, etc.). However, in addition, we have successfully looked at ways to translate the behavior of a human presenter and a human audience (assuming a presentation by a human presenter in a physical environment) to a virtual reality presentation event. This is shown in Figure 5 and Figure 6.

These figures show how the movements of the audience members and a human presenter are captured by cameras and mapped on an H-Anim standard representation of the human body and the limbs [14]. Using this representation, the members and presenter can be visualized as virtual humans in a 3D virtual environment. Presently, in the AMI project meeting participants are visualized in a virtual meeting room. Since we are able to make this translation it is also possible to have different viewpoints on a presentation event or to let an observer navigate in the virtual environment. A remote participant, whether represented and made visible to other participants or not, can take a reserved position during the event and an off-

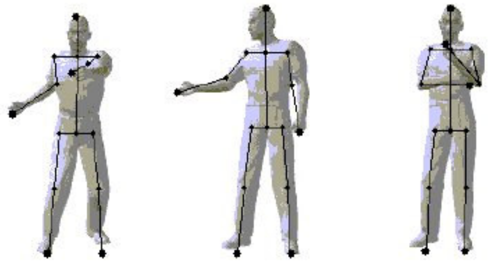

Figure 5, Translating Presentation Behavior to 3D Virtual Reality

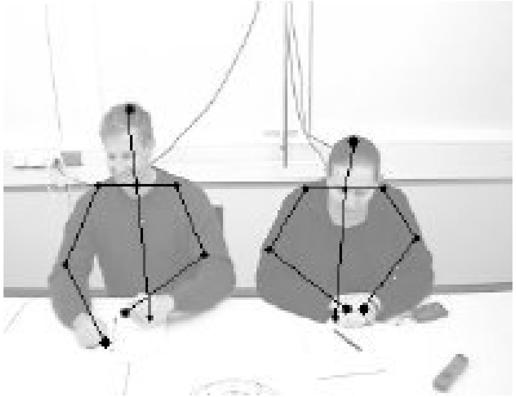

Figure 6, Translating Audience Behavior to 3D Virtual Reality

line user can experience a presentation event from the point of view of one of the real-time participants (see Figure 7).

7. Conclusions and Future Research

In this paper we have looked at research on presentation modeling. Although the research takes place in the context of a European project on meeting modeling, the results (models and tools) can be used in other domains as well. We can learn from real presenters. That is why we aim at introducing annotation tools and schemes to annotate real presentations and that is why we design models that describe useful human presentation behavior. Another way to make use of real presentations is to have them transformed in real-time to a presentation by a virtual presenter in a virtual environment that can be accessed by everyone who is invited or otherwise interested. Yet another way is to use motion capturing in order to obtain smooth animations of gestures, movements and pointing behavior of a virtual presenter. Presently we take two roads: (1) real-time translation of the behavior of a human presenter into the behavior of a virtual presenter in a virtual environment [4] and (2) have scripted presentations performed by a virtual presenter without having a human counterpart [12]. One important line of future research is to design virtual presenters that can be interrupted and are able to handle simple interruptions. Obviously, there are multiple possibilities of interruption (requesting additional information, requesting the rephrasing of information, going back or forward to another position in a PowerPoint presentation) and the presenter has many possibilities to react (elaborating on a specific element of the presentation, giving an alternative presentation of a specific topic, move forward or backward in a presentation, informing the audience that the requested information will be given later or that it is not available, requesting additional information from the audience). A model of the process of interrupting a presenter, based on a taxonomy of such interruptions, is in development. The model that is in development describes how the different interruptions lead to their respective responses [15].

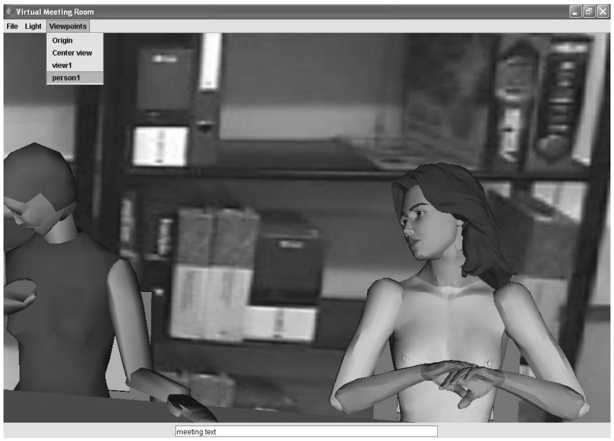

Figure 7, Participating from a Particular Viewpoint in the Virtual Meeting Room

Other topics of future research are realistic gaze behavior of the virtual presenter, in particular when the presenter is interrupted. It certainly is not difficult for the reader to come up with a lot of other interaction issues that need to be modeled in order to obtain more realistic virtual presenters. Fortunately, it is not always necessary to simulate all human characteristics in order to obtain an embodied agent that is sufficiently believable for its human partners

8. Acknowledgements

We are grateful to Dennis Reidsma for many useful remarks during the design of the virtual presenter tools. One of our students, Jaak Vlasveld, is presently working on the technical aspects of the design of an interruptible virtual presenter. This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication AMI-31).

References

[1] I. McCowan, D. Gatica-Perez, S. Bengio, D. Moore and H. Bourlard. Towards Computer Understanding of Human Interactions. In: *Ambient Intelligence*, E. Aarts, R. Collier, E. van Loenen & B. de Ruyter (eds.), Lecture Notes in Computer Science, Springer-Verlag Heidelberg, 235 - 251. See also: http://www.amiproject.org/.

[2] A. Nijholt, Smart Exposition Rooms: The Ambient Intelligence View. *Proc. Electronic Imaging & the Visual Arts (EVA 2004)*, V. Cappellini & J. Hemsley (eds.), Pitagora Editrice Bologna, March 2004, 100-105.

[3] A. Waibel, H. Steusloff, R. Stiefelhagen and the CHIL Project Consortium. CHIL - Computers in the Human Interaction Loop. 5th Intern. Workshop on *Image Analysis*

for Multimedia Interactive Services, April, 2004, Lisboa, Portugal. See also http://chil.server.de/.

[4] R. Poppe, D. Heylen, A. Nijholt & M. Poel, Towards real-time body pose estimation for presenters in meeting environments. Submitted for publication, University of Twente, December 2004.

[5] N. Magnenat-Thalmann & P. Kalra, The Simulation of a Virtual TV Presenter. *Proc. Pacific Graphics 95*, World Scientific, Singapore, 1995, 9-21.

[6] T. Rist, E. Andre & J. Muller, Adding Animated Presentation Agents to the Interface. *Proc. Intelligent User Interfaces*, 1997, 79-86.

[7] Y. Nozawa, H. Dohi, H. Iba, M. Ishizuka. Humanoid Robot Presentation Controlled by Multimodal Presentation Markup Language MPML. Proc. 13th IEEE Int'l Workshop on *Robot and Human Interactive Communication (RO-MAN2004)*, Kurashiki, Japan, No.026 (2004.9).

[8] T. Noma & N.I. Badler, A virtual human presenter. *Proc. IJCAI-97 Workshop on Animated Interface Agents*, 1997, 45-51.

[9] J. Gratch, J. Rickel, E. André, E. Cassel, E. Petajan, and N. Badler, Creating Interactive Virtual Humans: Some Assembly Required. Workshop report of the *Virtual Humans Workshop*, Marina del Rey, 2002.

[10] M. Theune. ANGELICA: choice of output modality in an embodied agent. Proc. Intern. Workshop on *Information Presentation and Natural Multimodal Dialogue (IPNMD-2001)*, Verona, Italy, 89-94.

[11] M. Theune, D. Heylen & A. Nijholt, Generating embodied information presentations. Chapter 3 in: *Multimodal Intelligent Information Presentation.* O. Stock & M. Zancanaro (eds.), Kluwer Series on “Text, Speech and Language Technology”, Vol. 27, O. Stock & M. Zancanaro (Eds.), Kluwer Academic Publishers, 2004, 47-70.

[12] H. van Welbergen, A virtual human presenter in a 3D meeting room. Manuscript in preparation, University of Twente, December 2004.

[13] D.-P. Pertaub, M. Slater & C. Barker, An experiment on public speaking anxiety in response to three different types of virtual audience. *Presence: Teleoperators and Virtual Environments 11*(1), 68-78.

[14] Humanoid Animation Working Group: http://www.h-anim.org/

[15] J. Vlasveld, Interacting with a virtual human presenter. Internal report, University of Twente, December 2004.