Presenting in Virtual Worlds: Towards an Architecture for a 3D Presenter Explaining 2D-Presented Information

Herwin van Welbergen, Anton Nijholt, Dennis Reidsma, and Job Zwiers

Human Media Interaction Group, University of Twente Enschede, The Netherlands
anijholt@cs.utwente.nl

Abstract. Entertainment, education and training are changing because of multi-party interaction technology. In the past we have seen the introduction of embodied agents and robots that take the role of a museum guide, a news presenter, a teacher, a receptionist, or someone who is trying to sell you insurances, houses or tickets. In all these cases the embodied agent needs to explain and describe. In this paper we contribute the design of a 3D virtual presenter that uses different output channels to present and explain. Speech and animation (posture, pointing and involuntary movements) are among these channels. The behavior is scripted and synchronized with the display of a 2D presentation with associated text and regions that can be pointed at (sheets, drawings, and paintings). In this paper the emphasis is on the interaction between 3D presenter and the 2D presentation.

1 Introduction

A lot of meeting and lecture room technology has been developed in previous years. This technology allows real-time support to physically present lecturers, audiences and meeting participants, on-line remote participation of meetings and lectures, and off-line access to lectures and meetings [1,2,3]. Whether it is for participants that are physically present (e.g. while being in the lecture room and looking back on part of the presentation or previous related presentations), for remote audience members or for off-line participants, multi-media presentation of captured information needs a lot of attention.

In previous research we looked at the possibilities to include in these multimedia presentations a regeneration of meeting events and interactions in virtual reality. We developed technology to have a translation from captured meeting activities to a virtual reality regeneration of these activities that allows adding and manipulation of information. We looked at translating meeting participant activities [4], and at translating presenter activities [5].

While in the papers just mentioned our starting point was the *human* presenter or meeting participant, in the research reported here our starting point is a *semi-autonomous, virtual* presenter (Fig. 1) that is designed to perform in a virtual reality environment. The audience of the presenter will consist of humans, humans represented by embodied agents in the virtual world, autonomous

M. Maybury et al. (Eds.): INTETAIN 2005, LNAI 3814, pp. 203–212, 2005.

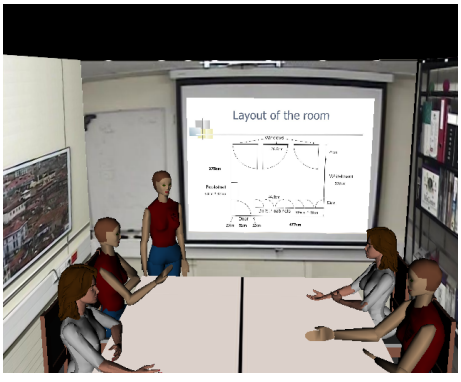

Fig. 1. The virtual presenter

agents that decided to visit the virtual lecture room or have been assigned roles in this room, or any combination of these humans and agents.

In this paper we confine ourselves to models and associated algorithms that steer the presentation animations of a virtual presenter. The presentations are generated from a script describing the synchronization of speech, gestures and movements. The script has also a channel devoted to slides and slide changes; they are assumed to be an essential part of the presentation. Instead of slides, this channel may be used for the presentation of other material on a screen or wall.

1.1 Organization of This Paper

In section 2 of this paper we highlight previous research on presentation agents that can interact with visual aids. Section 3 of this paper introduces our architecture of a virtual presenter. In the sections 4, 5, 6 and 7 the separate parts of this design are further discussed. Section 8 concludes this paper and discusses possible further work on the virtual presenter.

2 Presentations by Embodied Agents

A great amount of projects about presenting and virtual agents can be found. In this section we highlight a few projects projects featuring human-like presenters that use visual aids.

Prendinger, Descamps and Ishizuka worked on specifying presentations related to web pages, which are executed by MS-agents, a robot, or a 3D presenter. Their main focus has been the development of a multi-modal presentation language (MPML) with which non-expert (average) users can build web-based interactive presentations [6]. MPML allows one to specify behaviour in several modalities including gestures, speech and emotions.

The work of Noma, Zhao and Badler aims at simulating a professional presenter such as a weather reporter on TV [7]. The presenter can interact with a visual aid (usually a 2D screen). The animation model is rather simple. It implements two posture shifts, namely those needed to look at the screen and then back in the camera. The arm movement is determined by pointing actions specified in the animation script and by the affirmation level (neutral, warm, enthusiastic). For the hand movement, canned animations for grasping, indicating, pointing and reaching were used.

André, Rist and Müller [8] describe a web agent drawn in a cartoon style to present information on web pages. Most of their work focuses on planning such a presentation script. The agent is capable of performing pointing acts with different shapes (e.g. punctual pointing, underscoring and encircling). It can express emotions such as anger or tiredness. "Idle animations" are performed to span pauses. The character is displayed in 2D using completely predefined animations. Pointing acts are displayed by drawing a pointing stick from the hand to the pointing target.

3 An Architecture for a Virtual Presenter

Building a virtual presenter brings together many different techniques, including facial animation, speech, emotion/style generation and body animation. The main challenge is to integrate those different elements in a single virtual human. The two major concerns for this integration are consistency and timing [9].

Consistency. When an agents internal state (e.g. goals, plans and emotions) as well as the various channels of outward behavior (like speech, body movement and facial animation) are in conflict, inconsistency arises. The agent might then look clumsy or awkward, or, even worse, it could appear confused, conflicted, emotionally detached, repetitious, or simply fake. Since in the current version of the virtual presenter its behavior is derived from the annotated script of a real presentation, consistency conflicts arise mostly between channels that are implemented and those that are not. When the presenter gets extended to dynamically generate its behaviour, consistency will become a more important issue.

Timing. The timing is currently a more crucial concern. The different output channels of the agent should be properly synchronized. When an agent can express itself through many different channels the question arises what should be the leading modality in determining the timing of behavior. For example, BEAT [10], a toolkit used to generate non-verbal animation from typed text, schedules body movements as conforming to the time line generated by the text-to-speech system. Essentially, behavior is a slave to the timing constraints of the speech synthesis tool. In contrast, EMOTE takes a previously generated gesture and shortens it or draws it out for emotional effect. Here, behavior is a slave of the

constraints of emotional dynamics. Other systems focus on making a character highly reactive and embedded in the synthetic environment. In such a system, behavior is a slave to the environmental dynamics.

To combine these types of behavior you need at least two things. In the first place, the architecture should not fix beforehand which modality is the "leading modality" wrt the synchronization. In the second place, different components must be able to share information. If BEAT has information about the timing constraints generated by EMOTE, it could do a better job scheduling the behavior. Another option is to design an animation system that is flexible enough to handle all constraints at once. Norman Badler suggests a pipeline architecture that consists of 'fat' pipes with weak up-links. Modules would send down considerably more information (and possibly multiple options) and could pull downstream modules for relevant information (for example, how long it would take to point to a certain target, or how long it would take to speak a word).

3.1 The Architecture

The architecture of the virtual presenter (Fig. 2) is inspired by Norman Badler's pipeline and the work described in [9]. Expressions on separate channels are specified in the presentation script. The planners for different channels determine how those expressions are executed. This can be done using information from another module (for example, the text-to-speech engine can be asked how long it takes to speak out a certain sentence, or the sheet planner can be asked which sheet is visible at what time) or from human behavior models. The planner could even decide not to execute a certain expression, because it is physically impossible to do so, or because it would not fit the style of the presenter.

We choose to implement a selected set of dimensions of the behavior of a presenter in a theoretically sound way. To be able to insert new behavior later on, the presenter is designed to be very extensible. The rest of this paper discusses the different implemented aspects in more detail.

4 Presentation Script

As can be seen in Figure 2, the script is the starting point for (re)visualization of presentations. A script can be created from annotation of behavior observed in a real presentation, or generated from an intention to convey certain information. The multi-modal channels used by the presenter are scripted at different abstraction levels. Gestures are specified in an abstract manner, mentioning only their 'type' or communicative intent (deictic reference, stressing a word in speech, indicating a metaphor, etc.), leaving the exact visualization (e.g. which body-part, hand shape or movement path to use) to the planners. Speech is annotated at a word level. Resting poses and pose shifts are annotated at a very low level, the joint rotations in the skeleton of the presenter are specified for every pose. Sheet changes are specified whenever they should occur.

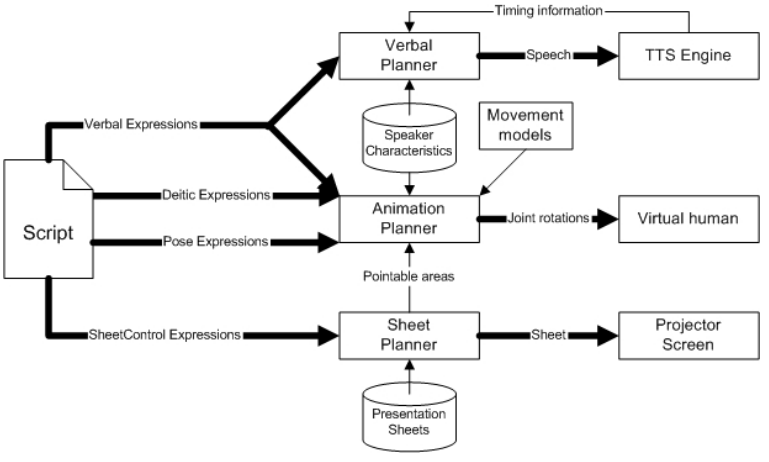

Fig. 2. Full system architecture of the virtual presenter

4.1 MultiModalSync

For the synchronization and timing of the presentation in its different channels we developed the MultiModalSync language. The synchronization is realized by setting synchronization points in one modality and using these synchronization points in another modality. For example, a synchronization point can be set before a word in the verbal modality. This point can then be used in the pointing modality to define a pointing action that co-occurs with the spoken word. Synchronization points can be set and used on all modalities so that the "leading" modality can be changed over time. [11] describes the constraints and synchronization definitions of the MultiModalSync language in greater detail and explains why it was necessary to develop a new script language.

5 Presentation Planning

The animation planner (Fig. 3) is responsible for the planning and playback of body animation. It makes use of movement models derived from neurophysics and behaviorial science to perform this task. Details on the exact use of these models can be found below. Currently, the animation planner is capable of playing deictic gestures, pose shifts and speech (mouth movement) specified in the script. Static speaker characteristics influence how this behavior is executed. The architecture can easily be extended to execute other gesture types. The verbal planner and the sheet planner regulate the Text To Speech generation and the sheet changes, respectively.

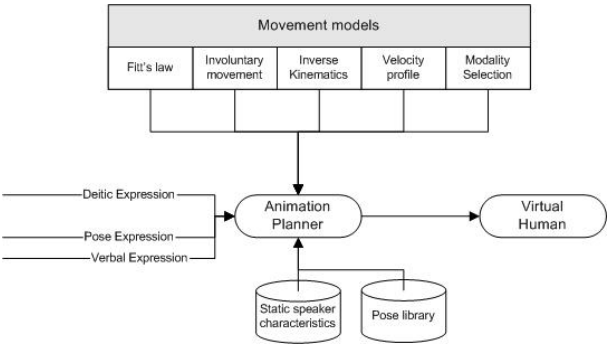

Fig. 3. The animation planner

5.1 Speech Planning

Loquendo's Text-To-Speech engine is used to generate speech, lipsync and speech timing information from the verbal text. A very simple form of lip synchronization is used by the Animation Planner: the rotation of the jaw is proportional with the volume of the speech, averaged over a short time period.

5.2 Involuntary Movement

Even while standing still, our body still moves in very subtle ways: we try to maintain balance, our eyes blink and our chest moves when we breath in and out. An avatar that does not perform such subtle motion will look stiff and static.

Our presenter uses an involuntary movement method described in [12]. Involuntary movement is simulated by creating noise on some of the joints in the skeleton of the avatar. This method is chosen to avoid the repetitiveness of predefined scripted idle animation and because the presenters' model is not detailed enough to use realistic involuntary movement models. The choice of which joints to move is made ad-hoc. For example, the two acromioclavicular joints (the joints between the neck and the shoulder) can be moved to simulate small shoulder movement that occurs when breathing. Small rotations of the vl1 joint are used to simulate subtle swaying of the upper body.

5.3 Posture

A pose is defined as the resting position of the body. Poses are used as start and end positions for the limbs within a gesture unit. In "real" monologues, posture shifts frequently occur at discourse segments [13]. In our presenter system, each pose is specified separately in a library containing the joint positions. In the current scripts, references to these poses are included based on the human presenter's pose in the real presentation (Fig. 4).

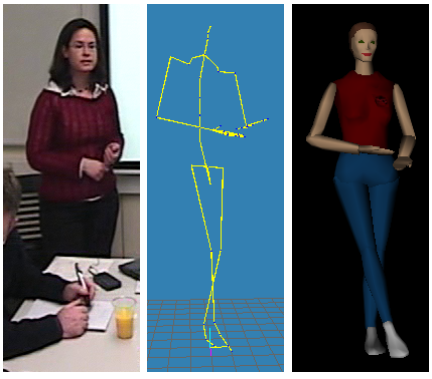

Fig. 4. Annotating and simulating poses. From left to right: the pose in the meeting room, a manually created representation in the Milkshape modelling tool and the virtual presenter showing that pose.

5.4 Pointing

During presentation, a presenter can refer to areas of interest on the sheet, by using a gesture with a pointing component. Our pointing model takes several aspects of pointing movement in consideration, so that the pointing movement can be generated given only the intention to point and a pointing target. Like Noma, Zhao and Badlers presenter [7], the presenter points to the right using its right hand and to the left using his left hand, to keep an open posture. When this preferred hand is occupied the presenter will point using gaze.

Timing. Fitts' law predicts the time required to move from a certain start point to a target area. It is used to model rapid, aimed pointing actions. Fitt's law could thus give a minimum value for the duration of the preparation phase of a pointing action in a presentation. The virtual presenter uses a 2D derivation of Fitt's law described in [14]: $T = a + b \cdot log_2(\frac{D}{min(W,H)} + 1)$, where T is the time necessary to perform the pointing action, a and b are empirically determined constants, D is the distance to the object to point to, W is the width and H the height of the object.

Movement in the Retraction Phase. Gestures are executed in three phases. In the (optional) preparation phase the limb moves away from the resting position to a position in gesture space where the stroke begins. In the (obligatory) stroke phase we find the peak of effort in a gesture. In this phase, the meaning of the gesture is expressed. In the (optional) retraction phase the hand returns to a rest position. Preparation (or retraction) will only occur if a gesture is at the beginning (or end) of a gesture unit.

According to [15], gesture movement is symmetric. We conducted a small experiment to find out if this is also true for more precise pointing actions. This was done by creating and looking at videos of pointing gestures, the same

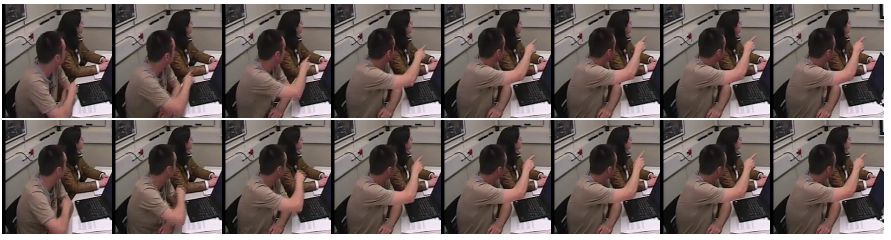

Fig. 5. Screen capture of the preparation (upper half) and retraction phase (lower half, backward) of a movie with a pointing action

pointing gestures played backward and a single video with both the forward and backward played gestures. Figure 5 shows screen captures of such a video. The following was found out:

- Pointing gestures that form a complete gesture unit by themselves are rare.
- Those gestures that do form a gesture unit by themselves have symmetric looking preparation and retraction phases.
- As Kendon noted, it is hard to tell if such a gesture is played backward or forward

Based on these findings, the retraction phase is defined by moving the arm back to the resting position in the same way it would be moved to the position of the stroke from the resting position (but backward).

Pointing Velocity. The velocity profile of a pointing movement is bell shaped [16]. This bell can be asymmetric. The relative position-time diagram is sigmoid-shaped. We use the sigmoid $f(t) = 0.5(1+\tanh(a(t^p-0.5)))$ to define the relative position of the wrist. In this function t represents the relative movement time ($t = 0$ is the start time of the pointing movement, at $t = 1$ the wrist reached the desired position). $f(t)$ describes the relative distance from the start position: $f(0) = 0$ is the start position, $f(1) = 1$ represents the end position. The steepness of this sigmoid can be adjusted using a. p can be used to set the length of the accelatory and decelatory phases.

Pointing with Gaze. Gaze behaviour during pointing movements is implemented on the basis of Donders Law [17], which defines both the necessary movements and end orientations for the eyes and for the head, given the fact that the presenter will look at its pointing target.

Shoulder and Elbow Rotation. If the wrist position is determined by the location and size of the pointing target, the rotations of the elbow and shoulder joints can calculated analytically using the inverse kinematics strategy described in [18]. The elbow though is still free to swivel on an circular arc, whose normal is parallel to the axis from the shoulder to the wrist. To create reasonably good looking movements, the presenter always rotates the elbow downward.

5.5 Sheet Planning

To display the sheets, the virtual presenter uses a virtual projector screen. On these sheets, areas of interest are defined, at which the presenter can point. The sheets for a presentation are described in an XML presentation sheets library. The sheet planner is responsible for the display and planning of sheet changes. It can provide other planners with planning information (e.g. what sheet is visible at what time).

6 Digital Entertainment in a Virtual Museum: Moving to a Different Domain

In order to demonstrate the broader applicability of the technology that was developed for the Virtual Presenter we are currently investigating a different domain, namely that of a Virtual Museum Guide. A corpus of annotated paintings such as the Rijksmuseum database used in [19] shares many characteristics with the presentation sheets. The information about the painting covers general aspects as well as remarks about specific subareas of the painting (e.g. relative composition, details in a corner of the painting, etc). The multimedia presentations they generated automatically from this content could easily be made interactive by using the virtual presenter as a museum guide who talks about the paintings while pointing out interesting details. In those cases where the relations between the text and areas in the paintings is only implicitly encoded in the text, techniques could be developed for automatic extraction of those relations from the text.

7 Conclusions and Future Research

Using the architecture described in here, we were able to create a speaking, involuntary moving, posture shifting and pointing presenter. The behavior on these different output channels is synchronized using a script. The designed architecture has already been adapted in two student projects with respectively new gesture modalities and possibilities to allow interruption of the presenter by an audience.

Further work could broaden the presenters abilities to express itself. This can be done by adding additional gesture types (like beats, iconics or metaphoric gestures). It is also possible to raise the virtual presenting process to a higher abstraction level. Currently, the script determines what part of the presentation is expressed in speech and what part is expressed by gestures. The next logical abstraction step would be to implement a process that determines what to say and what gestures to make, based on the content of what the presenters wants to tell. Such a selection process can be guided by the presenters style and emotional state.

References

1. Rui, Y., Gupta, A., Grudin, J.: Videography for telepresentations. CHI Letters **5**(1) (2003) 457–464
2. Rogina, I., Schaaf, T.: Lecture and presentation tracking in an intelligent meeting room. In: Proc. IEEE Intern. Conf. on Multimodal Interfaces. (2002) 14–16
3. Waibel, A., Steusloff, H., Stiefelhagen, R., the CHIL Project Consortium.: CHIL - computers in the human interaction loop. In: 5th Intern. Workshop on Image Analysis for Multimedia Interactive Services, Lisbon, Portugal (2004)
4. Reidsma, D., op den Akker, H., Rienks, R., Poppe, R., Nijholt, A., Heylen, D., Zwiers, J.: Virtual meeting rooms: From observation to simulation. In: Proc. Social Intelligence Design, Stanford University, CA (2005)
5. Poppe, R., Heylen, D., Nijholt, A., Poel, M.: Towards real-time body pose estimation for presenters in meeting environments. In: Proc. of the Intern. Conf. in Central Europe on Computer Graphics, Visualization and Computer Vision. (2005) 41–44
6. Prendinger, H., Descamps, S., Ishizuka, M.: MPML: a markup language for controlling the behavior of life-like characters. J. Vis. Lang. Comput. **15**(2) (2004)
7. Noma, T., Zhao, L., Badler, N.I.: Design of a virtual human presenter. IEEE Computer Graphics and Applications **20**(4) (2000) 79–85
8. André, E., Rist, T., Müller, J.: Webpersona: A life-like presentation agent for the world-wide web. Knowledge-Based Systems **11**(1) (1998) 25–36
9. Gratch, J., Rickel, J., André, E., Cassell, J., Petajan, E., Badler, N.I.: Creating Interactive Virtual Humans: Some Assembly Required. IEEE Intelligent Systems **17**(4) (2002) 54–63
10. Cassell, J., Vilhjálmsson, H.H., Bickmore, T.: BEAT: the Behavior Expression Animation Toolkit. In: SIGGRAPH '01: Proceedings of the 28th annual conf. on Computer graphics and interactive techniques. (2001) 477–486
11. Nijholt, A., van Welbergen, H., Zwiers, J.: Introducing an Embodied Virtual Presenter Agent in a Virtual Meeting Room. In: Proc. of the IASTED Intern. Conf. on Artificial Intelligence and Applications. (2005) 579–584
12. Perlin, K.: Real time responsive animation with personality. IEEE Transactions on Visualization and Computer Graphics **1**(1) (1995) 5–15
13. Cassell, J., Nakano, Y., Bickmore, T., Sidner, C., Rich, C.: Annotating and Generating Posture from Discourse Structure in Embodied Conversational Agents. In: Workshop on Representing, Annotating, and Evaluating Non-Verbal and Verbal Communicative Acts to Achieve Contextual Embodied Agents, Autonomous Agents 2001 Conference, Montreal, Quebec (2001)
14. MacKenzie, I.S., Buxton, W.: Extending Fitts' law to two-dimensional tasks. In: Proc. of the SIGCHI conf. on human factors in computing systems. (1992) 219–226
15. Kendon, A.: An Agenda for Gesture Studies. The Semiotic Review of Books **7**(3) (1996)
16. Zhang, X., Chaffin, D.: The effects of speed variation on joint kinematics during multi-segment reaching movements. Human Movement Science (18) (1999)
17. Donders, F.C.: Beiträge zur Lehre von den Bewegungen des menschlichen Auges. Holländische Beträge Anat. Physiol. Wiss. (1) (1848) 104–145
18. Tolani, D., Goswami, A., Badler, N.I.: Real-time inverse kinematics techniques for anthropomorphic limbs. Graph. Models Image Process. **62**(5) (2000) 353–388
19. Smeulders, A., Hardman, L., Schreiber, G., Geusebroek, J.: An integrated multimedia approach to cultural heritage e-documents. In: Proc. 4th Intl. Workshop on Multimedia Information Retrieval. (2002)