INTERACTION APPROACH FOR DIGITAL VIDEO BASED STORYTELLING

Norbert Braun

Department of Digital Storytelling
Computer Graphics Center (ZGDV e.V.), Rundeturmstrasse 6
64283 Darmstadt
Germany
Norbert.Braun@zgdv.de http://www.zgdv.de/distel

ABSTRACT

This paper shows an approach for a storytelling oriented interaction on digital video. All the interaction capabilities of the system are driven by the video context and therefore media centric. Interaction possibilities are given to the audience by conversation (on topics of the video content) or by classical Direct Manipulation (of video objects). Conversations can be done by the user with a personalized assistance or directly with the video. The implementation of the approach is shown by a discussion about an application architecture on the basis of Real Media Server, SMIL and Java 3D. The application runs as a Video On Demand system, accessible through the world wide web.

Keywords: Interaction, Conversation, Media Centred, Digital Video, Digital Storytelling.

1. INTRODUCTION

User interfaces surprise the every day user with permanently new methods of interaction on their computer systems. Operating systems please their audience with approaches like Direct Manipulation WIMP-Interfaces, drag-and-drop, conversational interfaces for delegation of tasks etc. . Anyhow - generally the interaction metaphors are built as interaction with the system - not interaction with the information that users are looking for and systems are delivering. The interaction is system- centred, but not centred on the needs of the user. If users shift from one system to the other, the interaction opportunities on these systems change too. Users have to adapt to the system's interaction, therefore spend a lot of time learning interaction methods. This disturb the users concentration on the information they need to manage within every day work.

The change to User Centred Interaction is nevertheless tricky. What is User Centred Interaction exactly?! If someone suppose that one major task of every day user's work is the collection of information and the information is given through media - so the work with media is important for the user. If now the interaction capabilities of the user are focused on the interaction with the media and the information context, the interaction is centred on the user's needs.

How should the user interact Media Centred - this question directs to another paradigm of User Centred Interaction. The user should interact with the media in a somehow natural, *human like* style. This implies that the user should be able to directly manipulate objects he recognizes in the media, as well as he should be able to delegate some tasks to the media. Those tasks like for example an altered orientation of information delivery through the media or the request for background information onto information objects presented by the media. Users should be able to phrase such delegations in a human like way, i.e. through a conversation. Even the information delivery through the media should be in an conversational way, because humans are trained on information handling through conversations. For the input channels of a Media Centred Information System this implies the use of natural speech. The output channels of a Media Centred Information System should use mimic and gesture as well as speech to adapt to user's needs of natural like conversation and natural like information assimilation.

So Media Centred Interaction is on the one hand specific dependent, on the other hand should be designed and adjusted, in relation to the type of media that is used to deliver information. Continuous media types like audio or video (audiovisual) are time dependent. For the interaction capabilities on this media, it is important that users are able to interact synchronous to the media - this means onto the information objects that are actually experienced in the media - as well as asynchronous to the media, what means on information objects that are actually not experienced within the media, but that have been experienced or will be experienced while the media is giving information to the user. Like in the user's real live it should be possible to interact onto actions of the present or of the past - or to prepare for actions in the future.

Last but not least the user interaction is strongly related to the type of story that is narrated through the media. The approach should not interfere with suspension and immersion of the story. The story context should limit user's interaction demands to a level that is fulfilled by the interaction capabilities of the system.

The approach of this paper has shortly been described in this section. This paper is organized as follows: The next paragraph will give an overview about related work in this research area. Conversational interfaces on video will be discussed in Section 3, Direct Manipulation for video in section 4. The approach of Media Centred Interaction will be shown in section 5. A Media Centred Interaction system on video will be introduced in section 6. The paper will end with a conclusion and some words onto future work.

2. RELATED WORK

Design-based approaches to interaction within storys are closely related to the narration of the storys. At MIT Media Lab [Davenport96] exists a lot of work based on Automated Storytelling with avatars within an virtual environment. The interaction capabilities of their systems are strongly related onto the aesthetic of the narrated story. Both Direct Manipulation and Conversational Interaction is used in this concepts. At Carnegie Melon University [Bates96] a story engine had been developed that is based on virtual reality and comes with object manipulation and Conversational Interaction too.

A video based approach of storytelling is discussed in [Sawhney96]. The video is used to give coffee house impressions to the user. Interaction is done by Direct Manipulation of several video streams and text-hyperlinks. Like most video based systems, a conversational access to the video information is not developed.

3. CONVERSATIONAL INTERACTION

Conversational Interaction is often combined with a humanlike avatar because users need someone to talk to character - instead of talking to a video or to some audio stream. The vision, see [Spierling00], is a very realistic conversation within an immersive environment like shown in figure 1.

Figure 1: Idealistic view of Conversational Interaction

The use of conversations in HCI is based on the idea of task delegation. Users should be able to delegate their work to the computer – in a somewhat humanlike, easy to learn way of interaction. The computer on the other hand should be able to present its results in the same humanlike and intuitive way to the users. Multimedia and Multimodality are often used words in this context – but these technologies are not really relevant to a conversation – text based interfaces, used years ago in AI (see Eliza [Weizenbaum65]) showed that conversations are able without any picture, animation or sound-interfaces. But the use of multimedia, e.g. video presentations, has an other mayor advantage: They give context to a conversation.

Conversational Interaction lacks often on a absent restriction on the conversation topics. Users are allowed often to talk about anything that's in their mind. The use of a context to restrict conversation topics is a nice approach to give the user the impression that the Conversational Interface, used within the HCI-component, has really *something to say* within the conversation. Users do not have the impression to the of another fancy stuff that annoys with the message 'I didn't understand – please repeat your sentence'. User Tests showed that video presentations are a method to focus the conversation goal of the user to topics which are within the range of the conversational practice of the

system. In addition, nonlinear video presentation can be used to apply a narrative structure to the conversation. This enables the system to conduct the user talk to a direction that is understandable for the system.

If context restriction based on video presentations is used in conjunction with speech recognition, the speech recognition component can be optimised on the given context for a more advanced understanding of the user's words.

Beside the input and output of speech there is a another factor that gives Conversational Interaction a humanlike style: mimics and gesture, see [Buxton90]. With mimic and gesture the conversation range can be expanded to emotional and social signals that are common to the average user and therefore easy to understand. Beside that, emotional and social behaviour of an avatar can expand the selective story narration of a video presentation in a generic way. Figure 2 gives an example of an avatar using its hands for gesture and its face for mimic information transfer. The user has the imagination of an humanlike opponent talking with his hands.

Figure 2: Avatar using its hand and face for mimik and gesture information

If a system is expanding the narrative approach of a video presentation with an humanlike avatar, the avatar can be used in several ways, see [Braun00]:

- Part of the story: The avatar is used as a narrative part of the story. The restrictions on the video clip narration (fixed characters, action and suspense) are expanded by the generated behaviour of the avatar. As a result of this story structure the user has the impression of the avatar behaving 'inside' of the story world.
- Conferencier: The avatar is used as a kind of show master. Different to the previous point the avatar is acting outside of the story and therefore noticeable (for the user) not as a part of the story.
- Audience: The avatar is behaving and talking as if it was a part of the audience. This is different to the previous points as the avatar has no (for the user) noticeable impact on the story's narration. Certainly the avatar is forcing user's emotions and reactions and therefore its behaviour and speech is planned on authoring time as a part of the narration of the story.

Figure 3 shows an avatar as a conferencier within a Virtual Trade Show. Its task is to present videos about some topics – therefore to moderate the presentation and to entertain the audience of the show. As promised by results in psychological research [Gleich97] users accept the avatar as a conversational partner.

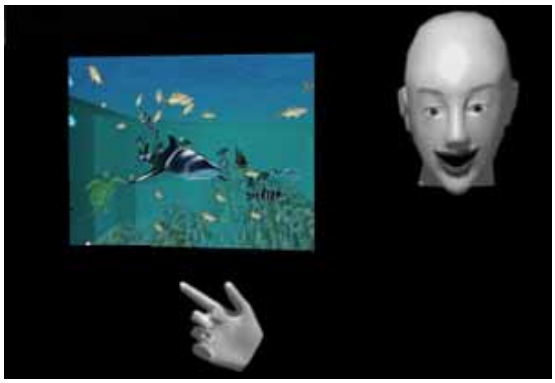

Figure 3: Avatar is working as a conferencier

The three proposed usages of the avatar are driven by the content of the video presentation – and therefore the usage is Media Centric. As a benefit of the Media Centric approach the conversation context is reduced in a way that it can be handled by speech and behaviour engines.

4. DIRECT MANIPULATION FOR VIDEO

The Direct Manipulation approach on video is split into Direct Manipulation on whole video clips, e.g. treatment of the video as a BLOB (binary large object) and the treatment of the video as a continuous media object that can be subdivided into several information objects. Those objects have a temporal and spatial appearance within the video. As the BLOB -based manipulation of a video is not that interesting for storytelling (e.g. it's primitive vcr-functionallity), this section will focus on the manipulation of the information objects within the video.

All information objects within a video have a spatial and temporal characteristic – as well as a media specific characteristic based upon the presentation within the graphics or acoustics of the video clip. Beside the media presentation, the interaction possibilities on these objects are not that versatile: On can annotate the objects in a hypermedia style to give a interaction access to the user. This annotation can be done within the media stream that is containing the annotated object (intramedia annotation) – for instance by drawing a polygon around the annotated object if it is within the visual stream – or by showing a signal within a separated media synchronous to the presence of the annotated object (intermedia annotation). Intermedia annotation could be a texthyperlink or a button that is active synchronous to the presentation time of the annotated object, as seen at [Martel96]. Intermedia annotation has several problems, e.g. the problem of associating the annotation to the annotated object. The approach of this paper is a intramedia annotation due to the better association of object and annotation.

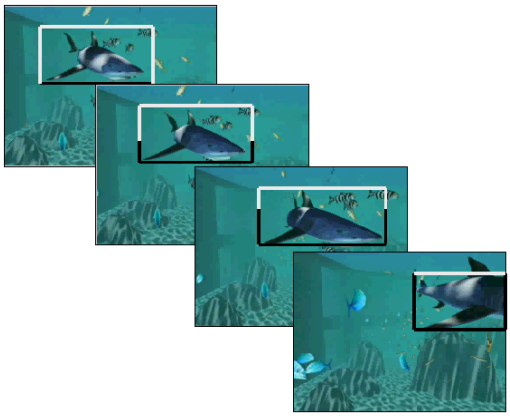

Figure 3: Concept of Temporal Video Hyperlink within the videos graphic

The problematic of the spatial and temporal presentation of objects within the visual media stream can be solved by using Temporal Video Hyperlinks, see [Braun99]. Those Temporal Video Hyperlinks are presenting the temporal structure of a hyperlink to the user by giving him an explicit notion of the following points:

- The start of a hyperlink: With the start of the annotation, the user notices an sign that is explicitly assigned to the annotated object.
- The duration of the hyperlink: While the annotation is accessible, the hyperlink reflects the remaining time, which is given to access the hyperlink.
- The end of the hyperlink: The end of the hyperlink is shown to the user by an explicit end symbol. Users know after the ending of the hyperlink that there is no more access to the object.

Figure 3 shows this approach for a visual annotation. A white rectangle is drawn around the annotated object. This rectangle is changing its colour to black while the hyperlink is elapsing in time. Figure 4 shows the same approach with a circle drawn around the annotated object. The circle's colour starts with white and is by degrees changing its colour to red. User Tests show that users interaction is more comfortable with this approach: the stress of clicking onto an annotation as fast as it is possible for the user is reduced to a minimum.

Figure 4: Temporal Video Hyperlink showed by filling ring around the target object

The same approach holds for the acoustic channel of a video presentation. Acoustic objects within the audio channel can be annotated by a sound, played parallel to the object. Since Auditory Icons (sound that has an natural source, see [Brewster97]) don't work well with this approach one can use Earcons (synthetical sounds, see [Brewster97]) for sonification of the duration of the hyperlink. Figure 5 shows how two converging tones, played synchronous to the annotated audio object, are showing the duration of the hyperlink by converging to one tone.

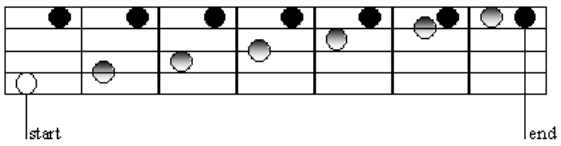

Figure 5: Temporal Audio Hyperlink by converging tones

Figure 6 shows how the same concept works with increased repeats of a tone. Figure 7 shows how the duration of the hyperlink is shown with tones that shorten their duration.

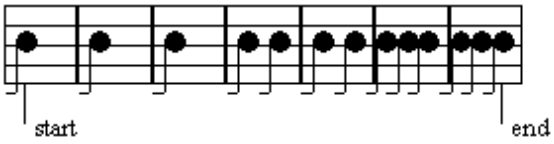

Figure 6: Temporal Audio Hyperlink by increased repeats of tone

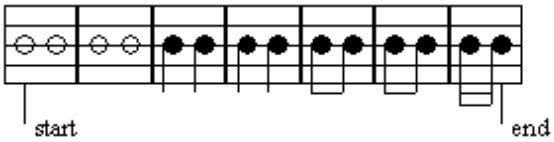

Figure 7: Temporal Audio Hyperlink by shortened tone duration

User Tests show that users can make a distinction between annotation and audio object (known as the cocktail party effect, see [Arons92]) on the one hand and that users can predict the duration of a temporal audio hyperlink by the start of the annotation on the other hand.

5. MEDIA CENTRED INTERACTION

Having Conversational Interaction and Direct Manipulation of videos discussed in section 3 and section 4, the question is how this can be combined to a Media Centric Interaction metaphor. If the conversational component, showed as an avatar in section 3, is driven by the content of the video (therefore the behaviour and speech of the avatar is shown synchronous to the video and its context) one needs a processing unit that takes into account the user interaction on both interaction components.

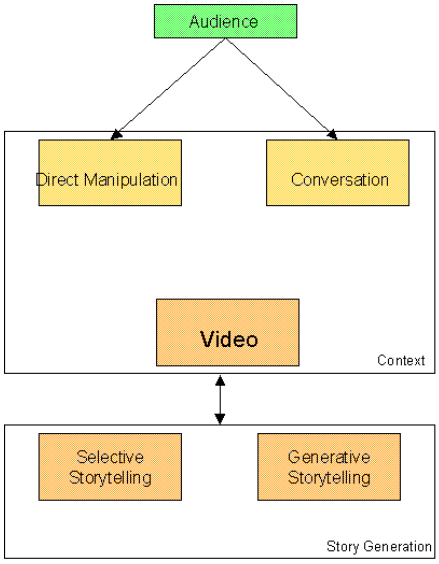

Figure 8: Concept of Media Centric Interaction on video, based on selective/generative storytelling

This unit has to calculate an adapted system reaction on every user interaction based on the story requirements. These requirements are based on the narration of the story and therefore on the intentions of the author for the story processing. Conversations have to be done as a part of the story dramaturgy, allowing immersion and suspense of the user within the story space. The approach of this paper is a combined selective and generative narration of story. This combination allows a playout of video clips and avatar behaviour and speech, based on the plots that are processed by the story and presented as a nonlinear video. As the selective and the generative components are combined, every information output is based on the story processing and every user interaction is based on the context given with the video/avatar presentation. User interaction is therefore Media Centric. Figure 8 shows a sketch of the approach with the video as the central information unit, flanked by the Direct Manipulation and Conversational Interaction capabilities, the story generated by a combined selective/generative storyengine.

6. ARCHITECTURE

The approach of this paper, nonlinear storytelling combined with Direct Manipulation/Conversational Interaction, has a widespread use of state of the art technologies to allow users a Media Centric access to storys presented by video. The architecture of the system has to considerate the following components:

- A conversational approach based on a humanlike avatar
- A Direct Manipulation approach, based on video annotation
- Story generation based on selective and generative elements
- Speech generation
- Speech recognition
- Video service
- Animation of avatar

Figure 9 shows an architectural overview of the system. The whole system is based upon the real media video server, see [HeftaGaub97].

Server

As one can see a database is used to store the selective and generative components of story, like video clips and avatar behaviour primitives. This database is used by the story engine to calculate the actual story narration – dependant on the authoring (pre authored video, combined with video-synchronous avatar animation and text) and on user reply, send by the real media client. These story components are given via the real media server (using plug-ins) through the world wide web to a real media client, running within a Netscape browser.

The animation control is synchronized to the video stream by SMIL [SMIL98].

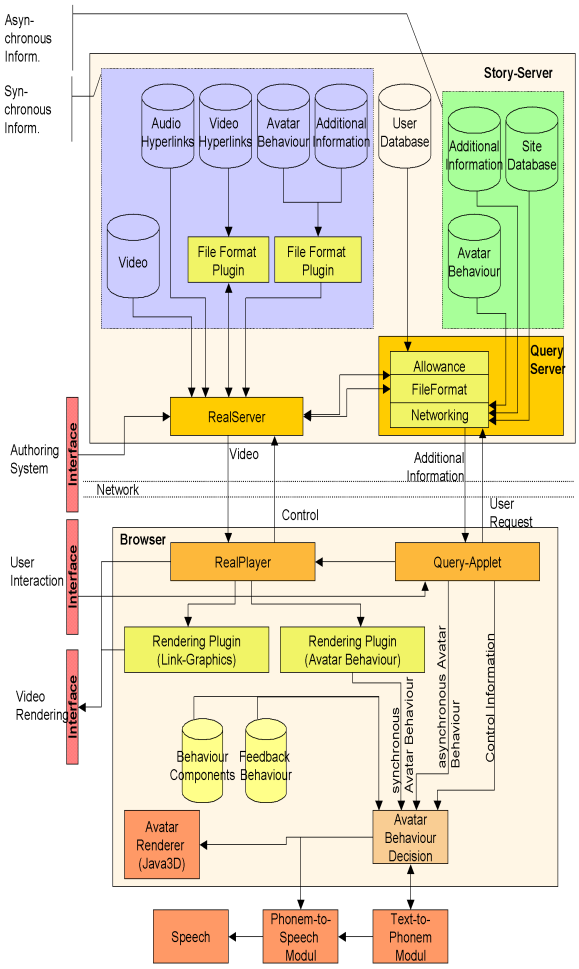

Figure 9: Architecture of the Media Centric system

Client

The real media client distributes the presentation streams to

- the video rendering unit
- avatar behaviour decision unit
    - Java-3D based feature morphing renderer, see [Alexa99]
    - phonem generator [Portele92], combined with a speech generator [Dutoit96]
    - viseme generator

The avatar decision unit decides which avatar behaviour to show to the user. The avatar behaviour consists of three components:

- the behaviour streamed by the real server. Since this behaviour is streamed synchronous to the video (see figure 10) it is called synchronous behaviour.
- the asynchronous conversation reply generated by the story engine. Since this behaviour is streamed asynchronous to the video it is called asynchronous behaviour.
- feedback behaviour generated on the client due to direct feedback on user inputs.

The avatar behaviour, combined with the viseme-animation, is on the fly generated on the client side of the system. The viseme-animation is synchronized lipsync to the speech rendered by the phoneme to speech-generator.

The Conversational Interaction of the user is processed via speech recognition. The speech recognition of the system is done within a separate application (using Java speech and IBM Via Voice) due to Java runtime limitations within the Netscape browser. The speech recognition unit processes the user speech to a symbolic query reply. This query reply is given via the real media client to the real media server.

The speech recognition too is processing the Direct Manipulation interaction on the acoustic part of the video, since the interaction on the acoustic part of the video is done by user's voice. Every user interaction on a temporal audio hyperlink is given as a reply to the real media server.

The Direct Manipulation interaction on the visual part of the video, based on annotations of the visual objects within the video, is done by the real media client. Every user interaction is given as a reply to the real media server.

On the server side a reply is calculated by the story engine (in form of a video clip and/or avatar behaviour and text) and streamed back to the real media client.

Behaviour Streaming

The streaming of the avatar behaviour is done by using high level behaviour instructions instead of preanimation of the behaviour on the server side. The avatar behaviour is hierarchically divided into four levels:

- Motivation: Commands that trigger a pro active high level behaviour like 'angry' or 'sad'.
- Task: Commands that trigger direct actions like 'salutation' or 'illustration'.
- Feature: Commands that trigger a modification of high level body components like 'left eyebrow up' or 'mouth smile'.
- Geometric: Commands that trigger low level changes on the avatar geometry.

With those commands one can reduce net traffic and maintain an avatar animation and rendering on the client side that responds quickly on user interaction. As a nice side effect the streamed avatar

behaviour is comfortable synchronized by SMIL, see figure 10.

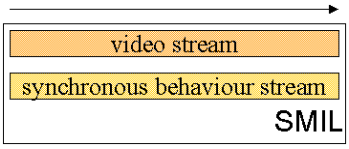

Figure 10: Avatar behaviour streamed synchronous to video

Due to the feature morphing, animation sequences have to be defined only once. If the geometry of the avatar is changed the avatar behaviour can still be used by the system. Figure 11 shows two morph targets based on the geometry of a video cassette. The animation defined with the morph targets can be used by any other avatar geometry.

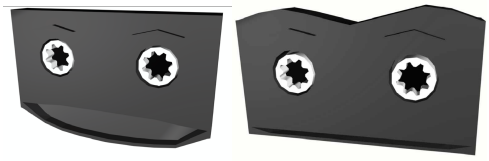

Figure 11: Morph targets (left: smile, right: brow up)

Prototype

The whole system is implemented as a prototype with the client running on a 900 mhz Windows-NT Computer.

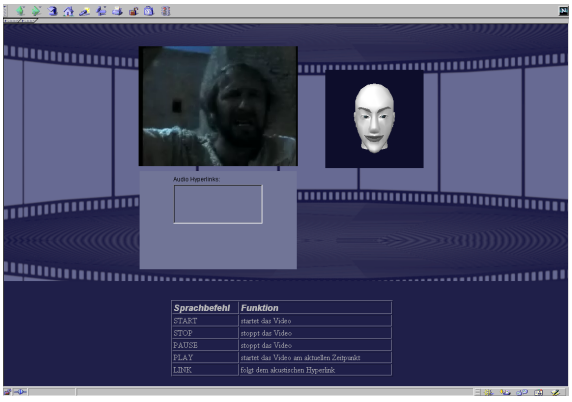

Figure 12: System prototype with humanlike avatar and video

Figure 12 shows the prototype's interface, presenting an avatar and a video. The avatar is acting synchronous to the video, talking about the movie. It is changing its behaviour if a user interaction happens. Then a direct feedback to the user is given, followed by a (optional) change of the video clip and/or the avatar behaviour and speech. Figure 13 shows another avatar, this time not humanlike, that acts in the same way as the avatar described above.

Figure 14 shows a user interacting by voice with a video. The avatar is not visible, the conversation is done to the video stream directly. Tests show that user prefer the avatar shown on the screen.

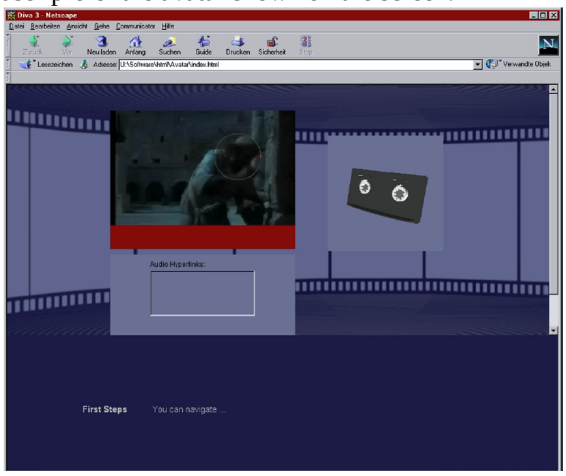

Figure 13: System with video cassette avatar and video

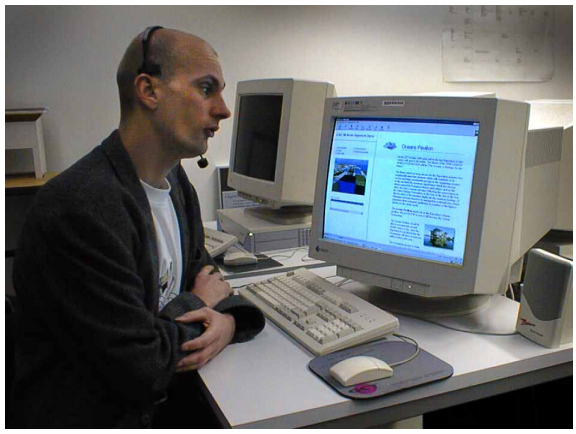

Figure 14: Media Centric user interaction by voice

The prototype is used for user testing. Users gave a very positive feedback on the system, they interacted with the story shown by the system in a very intuitive way. The users get use to the approach without a longer training phase. The test results underlay earlier tests based on video hyperlinks [Zahn00] and assistance systems [Nietschke00].

7. CONCLUSION

The approach discussed within this paper showed a storytelling system based on a story engine that combines selective and generative components for story narration. The story is presented using video clips and avatar animation. User interaction is possible via a hypermedia metaphor based on temporal acoustic and visual videohyperlinks as well as a conversational metaphor. The interaction

facilities are given to the user in a Media Centric way – the context of the conversation and hypernavigation is always based on the story narrated through the video.

The story narration of the system is done using a plot based story engine, the story management is located in event triggers combined with hyper navigation. Future work will expand the story engine to a generative storytelling model with a larger degree of freedom in story generation, granularity of control and a global location of the story management.

Another interesting extension of the approach to the web based interactive television is a multiple user connection based on avatars. The avatar behaviour and speech should be driven by the story impressions of several users, based on the non linear story narration presented by the video.

REFERENCES

[Alexa99] Alexa, M and Müller W.: The Morphing Space, *Proceedings of The 8-th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*, 1999

[Arons92] Arons, B., A Review of the Cocktail Party Effect, *Journal of the American Voice I/O Society*. 1992

[Bates96] Bates, J. and Smith, S.: Towards a Theory of Narrative for Interactive Fiction, *Technical Report*, School of Computer Science, Carnegie Mellon University, 1989

[Braun00] Braun, N. and Finke, M.: Interaction of Video on Demand Systems with Human-like Avatars and Hypermedia, In Scholten, H. and van Sinderen, M.J. (Ed.) *Interactive Distributed Multimedia Systems and Telecommunication Services, Springer*, 172 – 186, 2000

[Braun99] Braun, N. and Dörner R.: Temporal Hypermedia for Multimedia Applications in the World Wide Web, *Procceedings of the 3rd International Conference on Computational Intelligence and Multimedia, Applications, World Scientific*, 1999

[Brewster97] Brewster, S.A. and C.V.Clarke: The Design and Evaluation of a Sonically-Enhanced Tool Palette, *Proceedings of the International Conference on Auditory Displays ICAD*, 1997

[Buxton90] Buxton, W.: The Natural Language of Interaction: A Perspective on Non-Verbal Dialogues. In Laurel, B. (Ed.). *The Art of Human-Computer Interface Design, Reading, MA: Addison-Wesley*. 405-416, 1990

[Davenport96] Davenport, G.: 1001 Electronic Story Nights: Interactivity and the Language of Storytelling, *Proccedings of the Australian Film Commission's 1996 Language of Interactivity Conference*, 1996

[Dutoit96] Dutoit, T. and Pagel, V. and Pierret, N. and Van Der Vreken O. and Bataille, F.: MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes, *Proceedings of the ICSLP 96*, Philadelphia, 1996

[Gleich97] Gleich, U.: Parasocial interaction with people on the screen, *New Horizons In Media Psychologie*, 1997

[HeftaGaub97] Hefta-Gaub, B.: The Real Media Platform Architectural Overview, *Real Networks Conference*, San Francisco, USA, 1997.

[Martel96] Martel, R.: A Distributed Network Approach and Evolution to HTML for New Media, *Real Time Multimedia and the World Wide Web, No. 25*, France, 1996.

[Nietschke00] Nitschke J. and Wandke H.: Human Support as a Model for Assistive Technology, *Proceedings of the OZCHI*, 2000

[Portele92] Portele, T. and Steffan, B. and Preuss, R. and Sendlmeier, W.F. and Hess, W.: HADIFIX - a speech synthesis system for German, *Proceedings of the ICSLP '92, Banff*, 1227-1230, 1992

[Sawhney96] Sawhney, N., Balcom, D. and I. Smith: HyperCafe: Narrative and Aesthetic Properties of Hypervideo, *Proccedings of Hypertext '96*, 1996

[Spierling00] Spierling, U. and Gerfelder, N. and Mueller, W.: Novel User Interface Technologies and Conversational User Interafces for Information Appliances, Proccedings of *ACM Conference on CHI, Development Consortium*, 2000

[SMIL98] World Wide Web Consortium W3C. Synchronized Multimedia Integration Language (SMIL) 1.0 Specification *http://www.w3.org/TR/REC-smil/* , 1998.

[Weizenbaum65] Weizenbaum, J.: ELIZA - A Computer Program for the Study of Natural Language Communication Between Man and Machine, *Communications of the Association for Computing Machinery*, *9*, S. 36 – 45, 1965

[Zahn00] Zahn C., Schwan S., Barquero B.: Authoring and navigating video: Design strategies and user's needs for hyperlinks in educational films, *Report on an explorative study*, University of Tuebingen, 2000