DESIGN ISSUES FOR CONVERSATIONAL USER INTERFACES: ANIMATING AND CONTROLLING 3D FACES

W. Müller[1], U. Spierling[2], M. Alexa[1], I. Iurgel[2]
[1] *Technische Universität Darmstadt, Interactive Graphics Systems Group, Darmstadt, Germany*
[2] *ZGDV Computer Graphics Center, Darmstadt, Germany*

Keywords: Avatars, anthropomorphic conversational interfaces, facial animation, behavioral animation.

Abstract: Software agents and assistants, together with their adequate visual representations, lead to so-called social user interfaces, incorporating natural language interaction, context awareness and anthropomorphic avatars. Today's challenge is to build a suitable visualization architecture for anthropomorphic conversational user interfaces, and to design believable and appropriate face-to-face interactions, including human attributes, such as emotions. An integrated approach to these tasks is presented.

1. INTRODUCTION

With the end of the 20th century, the vision of a new human-computer interaction paradigm of "assistance" seemed destined to overtake the as yet still valid paradigm of the "computer as a tool" (Maes 94). However, recent discussions in research and development for human-computer interaction have lead to the following agreements:

1. The introduction of task delegation to software assistants will not replace, but complement, the direct manipulation of software tools. (Maes et. al. 97)

N. Magnenat-Thalmann et al. (eds.), *Deformable Avatars*

2. The delegation of tasks to an assistance software and their monitoring claims for special and social interfaces, which resemble a human-human relationship rather than tool usage. (Nass et. al. 94)
3. Consequently, so-called conversational interfaces evolve, not only relying on natural speech interaction, but also on non-verbal behavior, such as facial expressions and gestures. (Cassell et. al. 00)

Especially in the home area, a convergence between TV, VCR, and household appliances with desktop computer and web interfaces is apparent. In this context, a unique approach is undertaken to suggest and evaluate prototypes and solutions for face-to-face interaction with virtual characters (here called "user interface agents" or "avatars") integrated in traditional interaction concepts. This approach is interdisciplinary and addresses the following challenges:

1. Human factors research towards an academic basis for non-verbal communication: Evaluation of human-human interaction as a starting point for the design and generation of human-computer interaction.
2. Technological platform for animated behavior: A lean behavior animation platform with emphasis on real-time interaction and flexibility.
3. Design for appropriate usage and acceptance: Integrated design of face-to-face scenarios regarding human and technical requirements, as well as context of interfaces and contents.

In this contribution, we present current results of our ongoing projects with respect to the last two points, technology and design.

2. STATE OF THE ART

In the area of facial animation, the first synthetic faces based on a parametric model were created by Parke in 1972 (Parke 72). Psychological studies from Ekman and Friesen (Ekman and Friesen 69) built the bases for most of today's approaches, allowing for a control on a higher level of abstraction. Their Facial Action Coding System (FACS) describes the facial muscle activities based on 58 action units. A current adaptation of this work can be found in Facial Animation Parameters (FAPs) in MPEG-4 (Ostermann 98). A good overview of FACS and FACS-based approaches can be found in (Parke 98). Though MPEG-4 also targets small and medium platforms, FAPs implementations are usually relatively complex and real-time performance can hardly be achieved. An alternative approach based on a modification of standard morph targets has currently been introduced by Alexa et al (Alexa et. al. 00), providing good performance, scalability, and better control during authoring.

There have been several approaches to control the emotional appearance of a synthetic character. Bates (Bates 94) introduced virtual characters with their own personalities based on individual goals and emotions. Similarly, Perlin and Goldberg (Perlin and Goldberg 96) developed an interactive animation system with hierarchical goal descriptions and artificial personalities for automated choreography. André et al (André et. al. 98) worked on a virtual presenter with internal models of emotion. However, all these approaches target the control of a virtual character's general behavior and have not yet proven to provide sufficient modeling for believable, complex facial expressions. Rule-based models have been introduced for this purpose by Cassell and Pelachaud (Cassell et. al. 94). Another example in this context is the work of Beskow (Beskow 95). Our work described in this paper borrows from Cassell and Pelachaud. However, we are using greatly simplified models due to a different application context. Here, a sound and believable appearance on small systems is a sufficient criterion for success.

3. AVATAR ANIMATION PLATFORM

We target at a UI control module appropriate for rendering animated characters with speech output and lip sync on standard PC platforms while communicating over low bandwidth connections. These technological constraints, as well as considerations for the future design of successful conversations, have been considered for the realization of a new and flexible Avatar Platform (see Figure 1):

The overall human-machine dialogue is controlled by a preceding dialogue manager, which manages all user-interface components and modalities. It also decides on explicit sentences to be spoken.

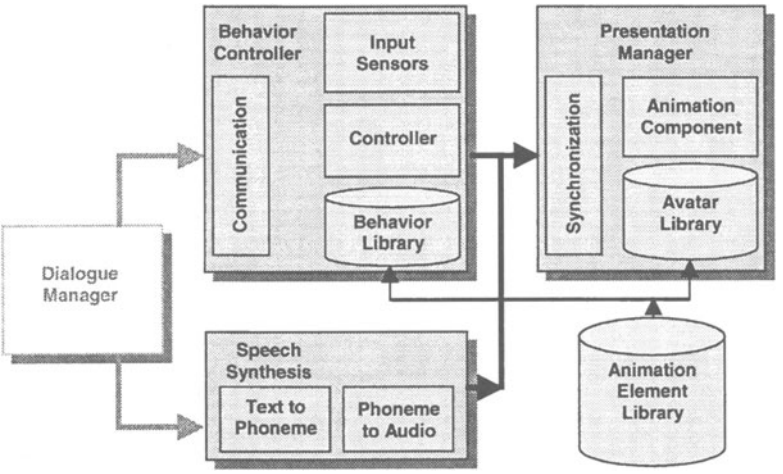

Figure 1: *Architecture of the Avatar Animation Platform (with preceding Dialogue Manager)*

3.1 PRESENTATION MANAGER

This module provides the functionality to present animated artificial characters, to perform facial animations, and to achieve lip sync. Animated characters are represented in a structure conforming to H-ANIM (H-ANIM 99). H-ANIM joints are augmented by a facial structure based on Morph Targets (see Figure 2) and efficiently realized as a Morph Node (Alexa et. al. 00). Hereby, a much broader range of facial expressions can be provided than in known real-time systems, while the parametrization of the face is kept simple.

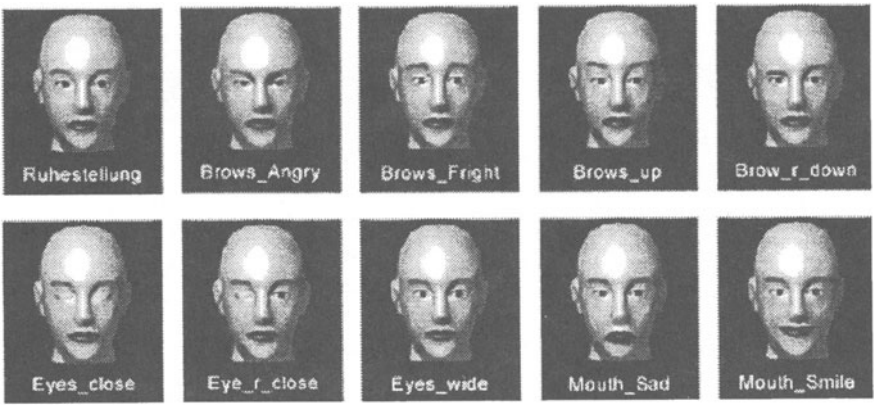

Figure 2: *Example set for basic morph targets*

Key-frames defining the state of the animated character consist of only a small number of values and playback is possible even over low-bandwidth networks and on small, portable devices. Moreover, facial animations can be easily mapped to different even very cartoon-like faces. Figure 3 depicts a different avatar representation, able to present the same behaviors, though partly with different expressions (e.g., using ear movements to express certain emotional states). Even within the same topology of a generic avatar face the expression range could be extended to morph targets that lead to different character representations (see Figure 4).

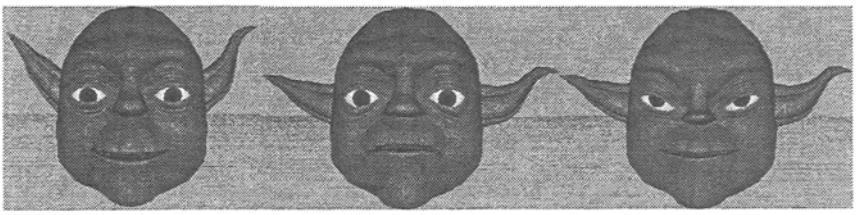

Figure 3: *Alternative avatar representation based on a Yoda geometry (Platinum 2000)*

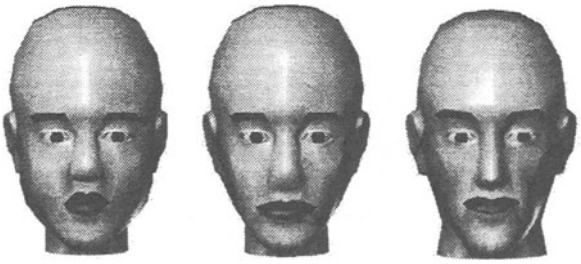

Figure 4: *Alternative character representations (baby, young, elder) within consistent topology*

3.2 BEHAVIOR CONTROLLER

While the Presentation Manager allows for the control of the avatar on the fundamental geometric levels, the Behavior Controller provides an interface on a more abstract level. Tasks and even motivations can be specified and the corresponding actions are performed automatically. Examples of such actions are gestures and movements of the avatar. In addition, behavior patterns for specific motivations or moods can be activated. This corresponds to a rule-based (possibly stochastic) activation of behavior elements. Applications for this are simple Avatar actions (e.g. accidental looking around or smiling) to avoid repetitive behavior. Another

possible application is to provide emotional expressions to emphasize system states, such as a puzzled look in case of an unexpected user input.

As base behaviors, we provide emotional states corresponding to the human universal prototype emotions, which are: fear, happiness, sadness, surprise, anger, disgust (Ekman and Friesen 69), and embarrassment (Castelfranchi and Poggi 90). Typically, the first four of these emotions are more relevant for our application field, since user interface agents tend to be polite.

The Avatar Platform is controlled by the dialogue manager, which is responsible for managing the multimodal and multimedia dialogue with the user. The application context is provided by the assistance functionality behind the dialogue manager.

3.3 SPEECH SYNTHESIS

In order to achieve a realistic appearance, an important requirement for the Avatar is a realistic synchronization of speech output with lip sync motions, facial expressions, gestures, and head movements. Since the dialogue with the user is not known in advance, prerecorded animation sequences with lip and facial animation cannot serve as a solution. Instead, text segments provided by the dialogue manager have to be converted automatically by a text-to-speech (TTS) engine. The speech synthesis system used is expected to generate phoneme information with appropriate timing information, which will be mapped to corresponding visemes and facial animation. This concept offers several advantages:

- Lip sync is realized automatically assuming that speech synthesis works in real time.
- Interactive applications with speech generation during and based on the interaction are possible.
- It is easy to implement a transparent system providing speech synthesis for different languages.
- New developments in the area of speech synthesis systems are readily usable. Even new features of such systems appear to be easy to include.

Our concept is based on a Hadifix-based text-to-phoneme conversion (Portele 97, Portele 96) and MBROLA (MBROLA 99). Here, phonemes are communicated using the international standard SAMPA. The mapping to visemes or sequences of visemes is based on timing and frequency information available in SAMPA.

In addition, heuristics are used to generate non-verbal facial expressions from phoneme information. Similar to (Poggi and Pelachaud 2000), we animated the eyebrows, eyelids, eye movement, and head movement based on typical structures in the phoneme stream. The rules applied include:

- Tone pitch attendance: In human communication, a tone pitch increase is usually connected with a raising of the eyebrows. The eyebrows are raised if the tone pitch exceeds a specified level by an amount dependent on the tone pitch.
- Semantic accentuation: When the tone pitch increases over a longer time frame, we assume an accentuation in the spoken sentence. The animation is accentuated by an raising of the eyebrows and a nod with the head. In addition, we direct the gaze of the avatar towards the user.
- Pause attendance: A longer pause between some words or sentences is accompanied by the avatar closing its eyes.
- Speech Rhythm attendance: We achieved good results by not blinking completely at random, but in coordination with the duration of a long vowel and of the following phoneme. This seems to support accordance to speech rhythm.
- Turn signal support: After finishing a speech act, we automatically supply some non-verbal behavior to signal a switch in the speaker role, that is, the user may interact with the system now. Here, we again supply a nod of the head. As a default, we accompany this nodding with a smile.

3.4 PSYCHOLOGICAL COHERENCE

An additional module ensures that the facial display will always appear to be psychologically coherent, giving the impression of the avatar undergoing psychological processes. The coherence module is derived from psychological literature, but all dependencies are simplified and adapted to the avatar's function of assistance (e.g. Smith and Lazarus 90, Ekman and Friesen 69, Kemper 84). The avatar is not an agent that makes choices on its own, but it does show considerable autonomy in the ways commands are displayed (comparable to IMPROV, Perlin and Goldberg 96). Arousal and mood are two important parameters governing the display of emotions.

Our goal is a fully parametrizable set of rules, supplemented by databases of idiosyncratic facial expressions, so that many different "actors" can easily be defined, each one interpreting the same directions according to its "personality". For example, an avatar that reads news would show almost no arousal or mood changes, while an assistant that is part of the extended family should show deep concern for problems. An assistant for the elderly will always be serene, while another for children will never show aggression.

3.5 IMPLEMENTATION

The implementation has to fulfill the following requirements:

- **Short response time.** The delay between request and display of the processed animation of a single sentence should be less than a second.
- **Expressive dynamics.** Mainly for movements of the head, varieties of acceleration are important for an appealing impression.
- **Resolution of conflicts in rules.** Conflicts between rules are common, e.g. between a rule to blink and a rule to open the eyes widely.
- **Continuous behavior.** Even if the avatar is not displaying commands of the dialogue manager, it should continuously show some simple behavior patterns – yawning, for instance.

We met these requirements by decomposing the module that determines morph weights and transformations. The resulting bundle of concurrent animations is coordinated by a blackboard and a multiplexer. Figure 5 depicts the architecture.

We use two distinct types of animations to set weights and positions:

- Computing a key frame sequence in advance. For example, expensive prosody-dependent animation parts are all preprocessed and stored as key frames.
- A real time animation, which does only a minimum of preprocessing and relies on a finite state machine to determine in real time weights or transforms. Eye blinks, for example, are produced this way.

Thus, the first three requirements mentioned above are met. Continuous behavior is assured by the generation of specific animations in the background.

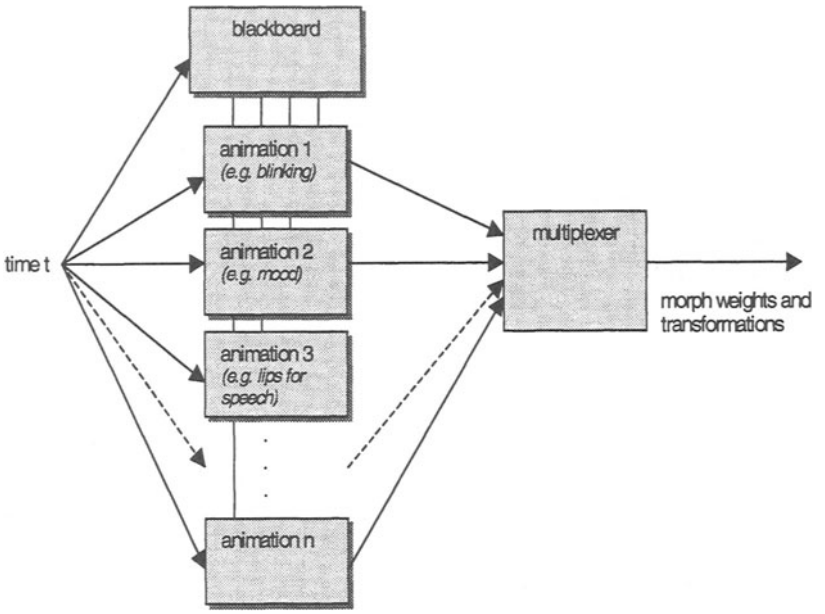

Figure 5: *Determination of morph weights and transforms by a coordinated bundle of animations*

4. FACIAL EXPRESSIONS FOR CONVERSATIONAL INTERFACES

The avatar platform is conceived as part of the rendering pipeline in the multimodal system and covers several anthropomorphic output modalities. User interface designers of the future must utilize these possibilities in an integrated way along with traditional output media such as graphics (Spierling 2000).

Up to now, there have been no known user interface design tools or common methods that allow the integrated design of anthropomorphic interfaces, but input can be taken from several existing concepts, such as UI design, character design, 3D animation, TV show design, and film grammar. The key aspects are:

- Believable animation of the user interface agent, making use of an autonomous behavior engine: creating a character, assigning characterized behavior and a role of assistance, animation and dialogue design.

- Integration of the avatar into context with other interface elements: giving stage directions for screen layout, and synchronizing camera and screen elements in time.

It is to be expected that with regard to converging systems, not only rules of building interfaces, but also influences from the realms of entertainment and storytelling will shape the future platform. The challenge is to provide authoring possibilities on top of the rather autonomously working user interface agent software. The separation of the geometry database and the behavioral animation library is the key concept for our solution.

One goal, for example, is to allow an intuitive scripting of avatar behavior by interface designers, comparable to stage directions. Stage directions tell an actor what to do, but give certain degrees of freedom to the performer. This can be done on various levels of abstraction, from a precise instruction up to a more improvising level. In our system, this is reflected by four hierarchical layers (direct, feature, task, motivation), that can be employed for different scenarios.

Facial expressions and the expression of emotions, as well as distinct character, are important aspects along with believable storytelling. Measurement of effectiveness and efficiency of conversational user interfaces has to be expanded to include a measurement of acceptance. The function and semantics of single features in facial expressions have not yet been analyzed or evaluated by human factors research and the effort is expected to be enormous. At this point, a potential risk of non-acceptance is obvious, just as with every other media that includes emotional attributes.

Our proposed solution is to give responsibility not exclusively to the engine, but to writers and storytellers, by opening the system for entertainment designers. Instead of completely building upon research results, designers can make suggestions based on intuition, or on experience in traditional animation (Thomas and Johnson 81) and experimental problem solving strategies. Figure 6 shows an environment for experiments.

5. APPLICATION EXAMPLE

We used the system in various test scenarios. One application of the Avatar Platform is a Virtual News Reader. Here, a web-based news service is polled and new messages are presented by the Avatar. In addition, the Avatar's emotional behavior is based on the content of the news message. For this, the news text is scanned for keywords stored in a database, and corresponding emotional expressions are automatically selected.

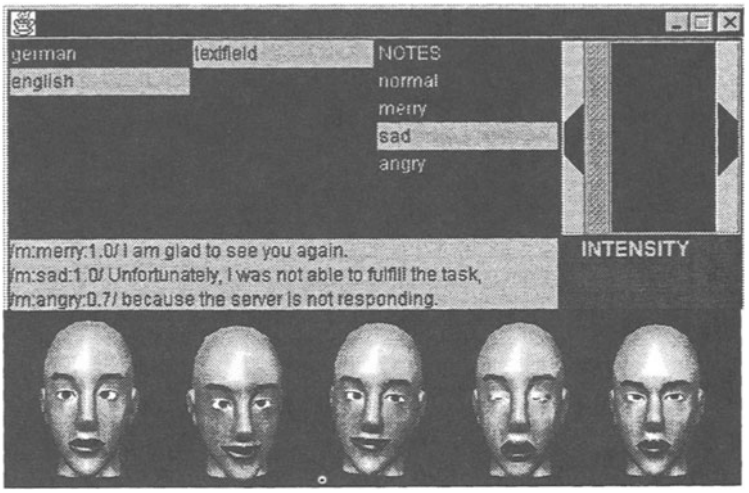

Figure 6: *Animation sequence automatically produced based on speech information*

6. CONCLUSIONS

In this paper, we presented the conception and realization of a presentation engine for conversational user interface agents. This software engine distinguishes itself from other solutions in using a novel approach for the representation and animation of facial expressions, enabling real-time lip-sync animations even on small machines. Furthermore, the system allows for an easy exchange of the animated face during runtime.

Based on a behavior controller and a library with animation elements and animation rules, the Avatar can be easily controlled at a task level and by making use of the motivation layer. Very complex animations can already be achieved automatically just from the phonetic information of speech output employing psychological rules of communication.

The Avatar Platform has been applied successfully to realize a Virtual News Reader. For broader application in more complex scenarios, however, usability studies and a better understanding of human communication rules and the use of non-verbal cues are needed.

7. ACKNOWLEDGEMENTS

This work has been partially funded by the German "Bundesministerium für Bildung und Forschung," BMB+F through the Focus Project EMBASSI (BMB+F-No. FKZ 01 IL 904 U8) [http://www.embassi.de]).

8. REFERENCES

Alexa, M., Behr, J., and Müller, W.: The Morph Node. In: Proc. Web3d/VRML 2000, Monterey, CA., 2000, pp. 29-34.

André, E., Rist, T., and Müller, J.: Integrating Reactive and Scripted Behaviours in a Life-Like Presentation Agent. Proc. 2nd Int. Conf. on Autonomous Agents '98, 1998, pp. 261-268.

Bates, J.: The Role of Emotion in Believable Agents. Communication of the ACM, Vol. 37, No. 7, 1994, pp. 122-125.

Beskow, J.: Rule-based Visual Speech Synthesis. in: Proc. of Eurospeech '95, Madrid, 1995

Castelfranchi, C. and Poggi, I.: Blushing as a discourse: Was Darwin wrong? in: Crozier, R. (ed.), Shyness and Embarrassment: Perspectives from Social Psychology, 1990, pp. 230-251, Cambridge Univ. Press, Cambridge, MA.

Cassell, J., Pelachaud, C., Badler, N.I., Steedman, M., Achorn, B., Beckett, T., Douville, B., Prevost, S., and Stone, M.: Animated conversation: rule-based generation of facial display, gesture and spoken intonation for multiple conversational agents. Computer Graphics (Proc. SIGGRAPH '94), 28 (4), 1994, pp. 413-420.

Cassell, J., Sullivan, J., Prevost, S., and Churchill, E. (eds.): Embodied Conversational Agents. MIT Press, Cambridge, MA, 2000.

Ekman, P. and Friesen, W.: The repertoire of nonverbal behavior: Categories, origins, usage, and coding. Semiotica 1, 1969.

Humanoid Animation Working Group (H-ANIM): http:// ece.uwaterloo.ca:80/~h-anim/ 1999.

Kemper, T. D.: Power, Status, and Emotions: A Sociological Contribution to a Psychophysiological Domain. in: K. Sherer and P. Ekman (eds.), Approaches to Emotion, Hillsdale 1984, pp. 369-383.

Maes, Pattie: Agents that Reduce Work and Information Overload. Communications of the ACM Vol.7/7, July 1994.

Maes, Pattie, Shneiderman, Ben, and Miller, Jim (Mod.): Intelligent Software Agents vs. User-Controlled Direct Manipulation: A Debate. Panel Description, ACM CHI '97, 1997

The MBROLA Project: Towards a Freely Available Multilingual Synthesizer. http://tcts.fpms.ac.be/synthesis/mbrola.html, 1999

Nass C., Steuer J., and Tauber, E.: Computers are Social Actors. Proceedings of ACM CHI '94, Boston. MA, 1994.

Ostermann, J.: Animation of Synthetic Faces in MPEG-4. Computer Animation, Philadelphia, Pennsylvania, June 1998, pp. 49-51.

Parke, Frederic I., and Waters, Keith: Computer Facial Animation. 1998.

Perlin, Ken and Goldberg, Athomas: Improv: A system for scripting interactive actors in virtual worlds. Proc. SIGGRAPH '96, 1996, pp. 205-216, http://www.mrl.nyu.edu/improv/

Platinum Multimedia Pictures Inc.: 3dCafe, http://www.3dcafe.com/asp/anatomy.asp, 2000

Poggi, Isabella and Pelachaud, Catherine: Performative Facial Expressions in Animated Faces. in: Cassell et al (ed.): Embodied Conversational Agents, MIT Press, Cambridge, 2000.

Portele, Thomas: Hadifix. http://www.ikp.uni-bonn.de/~tpo/Hadifix.html, 1996.

Portele, Thomas: Txt2pho - German TTS front end for the MBROLA synthesizer. http://tcts.fpms.ac.be/synthesis/, 1997.

Smith, C.A., & Lazarus, R.S.: Emotion and Adaptation. In: Pervin (ed.), Handbook of Personality: theory & research, Guilford Press, NY, 1990, pp. 609-637.

Spierling, U.: Conversational Integration of Multimedia and Multimodal Interaction. Development Consortium "Beyond the Desktop", Ext. Abst. of ACM CHI 2000, Den Haag.

Thomas, F. and Johnson, O.: The Illusion of Life. New York Abbeville Press, 1981.