Gesture Generation by Imitation:
From Human Behavior to Computer Character Animation

by

Michael Kipp

ISBN: 1-58112-255-1

DISSERTATION.COM

Boca Raton, Florida
USA • 2004

Gesture Generation by Imitation:
From Human Behavior to Computer Character Animation

Dissertation.com
Boca Raton, Florida
USA • 2004

ISBN: 1-58112-255-1

Gesture Generation by Imitation

From Human Behavior to Computer Character Animation

by
Michael Kipp

Dissertation

Faculties of Natural Sciences and Technology
Saarland University

Saarbrücken 2003

To my parents

Abstract

In an effort to extend traditional human-computer interfaces research has introduced embodied agents to utilize the modalities of everyday human-human communication, like facial expression, gestures and body postures. However, giving computer agents a human-like body introduces new challenges. Since human users are very sensitive and critical concerning bodily behavior the agents must act naturally and individually in order to be believable.

This dissertation focuses on conversational gestures. It shows how to generate conversational gestures for an animated embodied agent based on annotated text input. The central idea is to *imitate* the gestural behavior of a human individual. Using TV show recordings as empirical data, gestural key parameters are extracted for the generation of natural and individual gestures. The gesture generation task is solved in three stages: observation, modeling and generation. For each stage, a software module was developed.

For observation, the video annotation research tool ANVIL was created. It allows the efficient transcription of gesture, speech and other modalities on multiple layers. ANVIL is application-independent by allowing users to define their own annotation schemes, it provides various import/export facilities and it is extensible via its plug-in interface. Therefore, the tool is suitable for a wide variety of research fields. For this work, selected clips of the TV talk show "Das Literarische Quartett" were transcribed and analyzed, arriving at a total of 1,056 gestures. For the modeling stage, the NOVALIS module was created to compute individual gesture profiles from these transcriptions with statistical methods. A gesture profile models the aspects handedness, timing and function of gestures for a single human individual using estimated conditional probabilities. The profiles are based on a shared lexicon of 68 gestures, assembled from the data. Finally, for generation, the NOVA generator was devised to create gestures based on gesture profiles in an overgenerate-and-filter approach. Annotated text input is processed in a graph-based representation in multiple stages where semantic data is added, the location of potential gestures is determined by heuristic rules, and gestures are added and filtered based on a gesture profile. NOVA outputs a linear, player-independent action script in XML.

Acknowledgements

I want to thank the DFG (German Research Foundation) for the two-year full scholarship that made this research possible, and also the associated "Graduate College for Cognitive Science".

Most of all, I thank Prof. Wolfgang Wahlster for his continual support during my time at Saarland University and at the DFKI (German Research Center for Artificial Intelligence), for accepting and encouraging this project, and for his intense supervision in the final stages of this work. I also want to thank Prof. Elisabeth André for helping me getting this project off the ground in the first year and for joining the doctoral committee.

I would also like to thank the DFKI for providing office, equipment and a creative research environment. Special thanks to the CrossTalk project team for the great spirit, inspiration and motivation. I am especially indebted to Dr. Norbert Reithinger and Prof. Thomas Rist for inviting me to work on various DFKI projects in my post-scholarship days.

Thanks to Prof. Marcel Reich-Ranicki, Prof. Hellmuth Karasek, and the ZDF broadcast company for giving their permission to publish the video stills used in this dissertation which were all taken from the TV show "Das Literarische Quartett".

Some individuals deserve special mention. Thanks to Martin Klesen and Ralf Engel for patiently annotating gestures, to Dr. Christian König and Dr. Kerstin Seiler for a decisive mid-thesis discussion, to Stefan Baumann and Dr. Jürgen Trouvain for personalized advise on Phonetics and to Patrick Gebhard for nerve-wrecking debates and continual life-saving technical support. I would also like to thank all ANVIL users who provided bug reports, suggestions and encouragement. Special thanks to Cornelia Messing for lightning speed proofreading.

Contents

Chapter 1

Introduction

> For some presumptuous reason, man feels the need to create something of his own that appears to be living, that has inner strength, a vitality, a separate identity – something that speaks out with authority – a creation that gives the illusion of life.
>
> — Thomas and Johnston (1981: 13)

1.1 Computer Animated Characters

A new star has stepped onto the computer screen: the human body. Computer animated characters have always populated computer games before establishing themselves firmly in the movie industry in 1995 with *Toy Story*, the first completely computer animated feature film. Progress in computer graphics and character animation[1] refueled the ideas of early Artificial Intelligence (AI) to create artificial humans. However, not as physically present robots but as virtual beings living in a computer-generated graphical environment. While traditional AI research focused on the thought processes of human beings, now that virtual bodies are possible another issue is coming to the foreground: communication. In terms of communication, the human body has much more to offer than text or spoken language. The research area of human-computer interaction (HCI) is concerned with applying AI methods to make the complex software and hardware systems of today more accessible to human users by offering interfaces that go beyond the customary keyboard/mouse input and windows/text output[2]. The human body is considered a potentially powerful interface where the hidden and overt channels of everyday human-human communication can be exploited, such as gestures, facial

[1] cf. Witkin and Kass (1988), Badler et al. (1993) and Magnenat-Thalmann and Moccozet (1998)

[2] The traditional computer interface is sometimes referred to as WIMP: windows, icons, mouse, and pointer.

expression, gaze, posture and posture change. Such an interface consists of one or many human faces or bodies that interact with the human user. These computer animated characters are called anthropomorphic agents, embodied agents or life-like characters (Prendinger and Ishizuka, 2003, Cassell et al., 2000b). The term *avatar* refers to a special kind of embodied agents. An avatar is a puppet that is fully or partially controlled by the human user. It is meant to represent the user in a virtual space. Notwithstanding this specific meaning, the notion of avatar is sometimes used synonymously with embodied agent (Lindner 2003).

1.1.1 Embodied Agents Systems

Embodied agents are the focus of several research projects. The famous simulations of Marilyn Monroe and Humphrey Bogart in the short film *Rendez-vous in Montreal* by MIRALab in 1987 anticipated the task-oriented systems of today. Monroe was later put into an application for virtual tennis matches as a referee and to announce game results (Molet et al. 1999). This kind of task, information presentation, appears to be a natural application for embodied agents. For instance, in the PPP[3] system an anthropomorphic agent called *Persona* uses speech and gesture to explain technical devices (André et al. 1996). Pointing gestures are used to disambiguate references in speech and to focus user attention (see Figure 1.1). The Persona agent is animated by keyframe-based animation (Müller 2000). It relies on a library of animations in the form of keyframe sequences. The keyframes can be concatenated or merged to a single frame to give the illusion of continuous movement. The more sophisticated approach, called model-based animation, is based on an internal 3D bone model that is used to compute the animation's frames at runtime. Gestures are produced with the help of a library of pre-fabricated motion patterns that is accessed at runtime to animate the 3D model. An internal 3D model offers much more flexibility in animation. Pointing gestures and manipulative actions can be adapted to arbitrary situations, i.e. varying locations, shapes, dimensions of objects, people and places. Movements can be modified along various dimensions such as abruptness, smoothness, force etc. (Chi et al. 2000). Also, parallel motions can be merged in a single motion (e.g., a smile and a gesture) and sequential motions can be connected by smooth transitions (Perlin and Goldberg 1996). The *Virtual Human Presenter* (Noma and Badler 1997) is such a model-based system based on the *Jack* engine, a 3D character animation software that is controlled by a script of text and commands. Beyond libraries of predefined gestures, the feature-based animation approach aims at creating each new gesture on the fly from single form or motion features (Kopp and Wachsmuth 2000).

Presentation agents like Jack and PPP Persona can be used in arbitrary information systems, for instance to read the news, present tourist information or

[3] **P**ersonalized **P**lan-Based **P**resenter

report book reviews. They can also be used in e-commerce applications to advertise and sell products, or in e-learning environments to teach and supervise. Cassell et al. (2000a) developed REA[4], a 3D agent who presents houses to potential buyers. The agent coordinates gesture and speech with respect to both semantics and pragmatics. For instance, REA makes a circular gesture to semantically express "surrounding", and in terms of pragmatics she places gestures on *new* items in the speech stream. For another pragmatic function, signalling beginning and end of discourse segments, REA has been extended to utilize posture shifts (Cassell et al. 2001a).

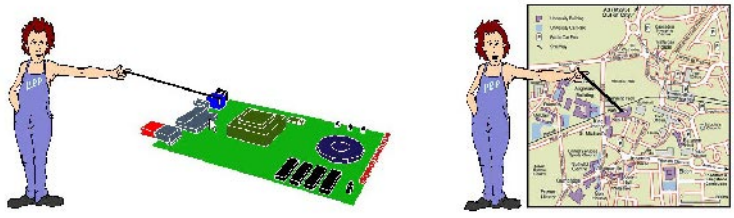

Figure 1.1: *Two applications of the PPP Persona system which automatically generates presentations, coordinating gesture and speech. On the left, Persona explains technical details of a modem. To the right, Persona acts as a city guide using a map of Portsmouth. Pointing gestures are used for focusing user attention on regions and for referencing concrete objects. (Taken from Müller, 2000.)*

While most presentation systems consists of a single embodied agent, André and Rist (2000) argued for a *team* of presentation agents to exploit the benefits of dialogue (see also Rist et al., 2003). Dialogue is livelier and easier to follow than monologue. Different agents can represent different viewpoints or degrees of expertise. This can even be used to manipulate the opinion of the listener. André et al. (2000) implemented this vision in a scenario called the Inhabited Marketplace where embodied customer and sales agents engage in an automatically generated dialogue about a product. The viewer is thus informed about the product's various properties. Selectable agent profiles of personality and interest guide the dialogue generation. Gaze behavior is used to focus the viewer's attention on the current speaker. The CrossTalk project is based on the same paradigm of team presentation (Gebhard et al. 2003). It is a self-explaining interactive system where a separate agent welcomes the user, explains the system and starts the actual presentation: a car sales dialogue. The agents use conversational gestures to make their interactions more life-like. Even if no user actively interacts with the systems the agents give

[4] **R**eal **E**state **A**gent

the impression of "living on" by engaging in smalltalk amongst themselves. This is to show that the system is permanently on stand-by, never to be turned off, never "freezing" or becoming inactive as electronic devices usually do. In the nonverbal behavior of the agents this is reflected in idle-time actions like scratching the forehead or blinking and breathing (Müller 2000). A similar idea is followed in the PEACH[5] project where continuous assistence is to be guaranteed in the form of a museum guide that jumps to different end devices, e.g., from a projected painting to a mobile palm top (Kruppa et al. 2003). The illusion of a continuous life is central to these systems and must be backed by believable nonverbal behavior by the agents.

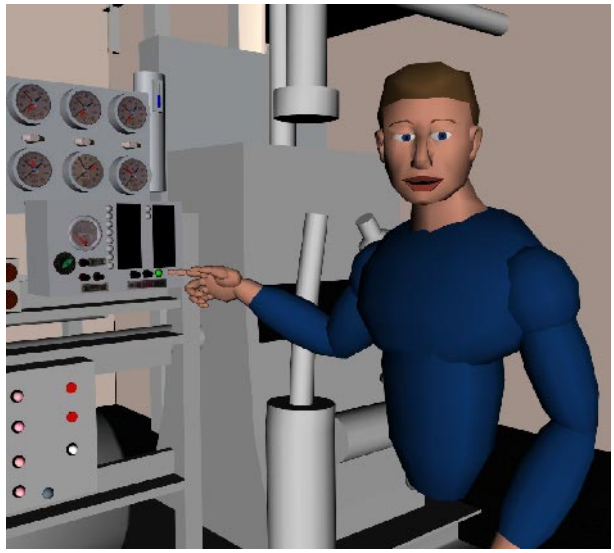

Figure 1.2: *The Steve agent describing an indicator light (figure taken from Rickel and Johnson, 1999). His pointing gestures help to resolve speech references to the currently explained object. Steve uses gaze behavior for pointing (looking at objects) and regulating the interaction with the user (looking at the user when expecting input).*

Besides the presentation of facts and products, agents can be employed in educational and training systems to convey knowledge and skills. Examples are Cosmo, who teaches how the Internet works (Lester et al. 1997c), and Herman the Bug,

[5] **P**ersonal **E**xperience with **A**ctive **C**ultural **H**eritage

a system to explain plants (Lester et al. 1999). The Steve[6] system was developed to accompany a human trainee in his/her hands-on experience of operating complex machinery in a virtual reality environment (Rickel and Johnson 1999). It interacts with the user by answering questions and demonstrating procedures. Steve uses pointing gestures to indicate the explained object and gaze behavior to show that Steve is listening to the user (Figure 1.2)[7]. Based on the Steve agent technology (Rickel et al. 2002), the Mission Rehearsal Exercise (MRE) project creates training simulations for a whole *team* of soldiers in a virtual reality theater with projections of 3D life-size embodied agents on a large curved screen with a 150 degree field of view (Swartout et al. 2001). MRE is supposed to prepare soldiers for critical situations on peacekeeping missions. For instance, faced with a wounded local inhabitant lying in the street next to his crying mother and with an urgent mission waiting somewhere else what decision must the platoon leader take? The system allows to realistically act out possible alternatives. Appropriate nonverbal behavior must be generated to recreate the social factors that lead to the above described stress situation. In a similarly immersive 3D environment, the VirtualHuman[8] project provides both a virtual teacher and a virtual student to give astronomy lessons to a human user (Figure 1.3). The teacher follows different paedagogical paradigms and behaves according to parametrized personality settings. The co-student extends the usual one-to-one (computer-human) setting to a classroom situation where students can help each other and compete with each other. Conversational gestures and facial expressions must be generated to make the experience as authentic as possible.

Complex applications like MRE and VirtualHuman demonstrate that embodied agents can inspire wholly new forms of interaction. Gottlieb (2002), co-creator of the highly popular computer game *You don't know Jack*, sees the potential of embodied agents in acting as guides, thus offering a new interaction style. It lies between a navigation-style communication (web-browser, newspaper) and a continually running show (TV, movie, lecture). The user can set his/her own *pacing* but the system controls the structure of the information which is important in educational scenarios. Pacing can be influenced by the agent's nonverbal behavior, for instance by yawning or tapping with one foot when the user pauses for too long.

To support the development of embodied agents applications, toolkits have been created that provide high-level scripting languages to control the agents. Examples are: Jack (Badler et al. 1993), IMPROV (Perlin and Goldberg 1996), Microsoft

[6] Steve is an acronym for **S**oar **T**raining **E**xpert for **V**irtual **E**nvironments. Soar is a general cognitive architecture for developing systems that exhibit intelligent behavior (Laird et al. 1987) and has been in use since 1983. For further information visit `http://www.eecs.umich.edu/~soar/main.html`

[7] Copyright by the University of Southern California

[8] `http://www.virtual-human.org`

Figure 1.3: *Screenshot detail of the 3D VirtualHuman system. The virtual student (left) listens to the virtual teacher (right). The teacher formulates a question that must be answered by virtual student or human user in direct competition.*

Agents and CharActor[9]. All tools are based on pre-fabricated motion patterns, some offering motion blending and online motion modifications with respect to form and tempo.

In summary, embodied agents are being developed for many application areas, including presentation/information, sales, assistance, education, training, and entertainment. They make possible new forms of interaction by bringing a new realism and social factors into computer applications.

1.1.2 Why Use a Body?

Using a body opens up new possibilities: broader and more efficient communication, expression of personality and emotion and the motivation resulting from the social presence of a life-like entity.

In terms of communication, hand and arm gestures play a major role in the body's communicative capabilities. Gestures can be used for pointing in order to resolve references to world objects, e.g. when asking "what's that?" while pointing to an expresso machine. The listener can resolve the anaphor "that" by following the pointing gesture. Gestures can also visually illustrate aspects of the message that are difficult to express verbally, e.g. by drawing the shape of an object into the air, by demonstrating a manual action or by recreating complex spatial arrangements with hands, fingers, arms. Consider the complex arrangements one would have to describe when retelling scenes from a Sylvester & Tweetie animation movie

[9] `http://www.charamel.de`

(see Figure 1.4 for examples from McNeill, 1992). With gestures both dynamic (speed, trajectory) and static (direction, distance, size) aspects can be expressed in a way that is simple to perform and quick to comprehend. In contrast to these highly context-dependent gestures that must be invented anew for each new situation there are gestures with standardized form and conventionalized meaning like the thumbs-up gesture, meaning "OK" or "good!", that can be used instead of speech where speaking is restricted by noise (construction site), convention (library) or taboo. Gesture can also be used to regulate a conversation, i.e. to assign, yield or claim the speaking turn using e.g. pointing or conventionalized signs like waving (Duncan and Fiske 1977). This is especially important since embodied agents systems strive to become more interactive and thus need to implement behavior that regulates agent-user as well as agent-agent dialogues. On the discourse level, gestures are used to segment the speech stream, to "highlight" parts of particular interest and to signify rhetorical relations (McNeill 1992). Politicians exploit such gestural devices to increase the intelligibility of their public speeches and even to control audience reactions like applause and laughter (Atkinson 1984). A major advantage of communication by gesture is that the signals are well-known to human users from everyday usage so that, when used in a computer interface, users do not have to learn new signs and behaviors.

Embodied agents have advantages beyond communication issues. With their social presence they can act as a guide, giving orientation, or as a trainer, demonstrating physical actions, but most importantly, they can motivate human users (Lester et al. 1997a). This motivation may stem from pure curiosity in the virtual "personality" or from the lowered technological barrier since human-agent interaction requires less expertise than interaction with traditional WIMP interfaces (McBreen 2001). For pedagogical applications, Lester et al. (1997a) conducted a formal empirical study suggesting that embodied agents can be pedagogically effective. Lester et al. (1997b) found that the students perceived the agent as being helpful, credible, and entertaining. McBreen (2001) and van Mulken et al. (1998) both found that an embodied agent makes an application more enjoyable and engaging but that user trust in the system is not necessarily enhanced. Reeves and Nass (1996) show how easily human users take technical equipment as living beings with a personality. They conducted two series of social-psychological experiments on social interaction, one with a human partner and one where this partner was substituted with a computer. Various aspects of human interaction were paralleled in human-computer interaction. For example, humans behaved politely when interacting with computers, they liked to be flattered by computers, and they judged computers that praised themselves lower than computers that praised other computers. Systems that aim at producing personality thus reinforce a natural tendency. However, while human users easily ascribe a personality to technical gadgets, they are at the same time highly sensitive to inconsistencies and

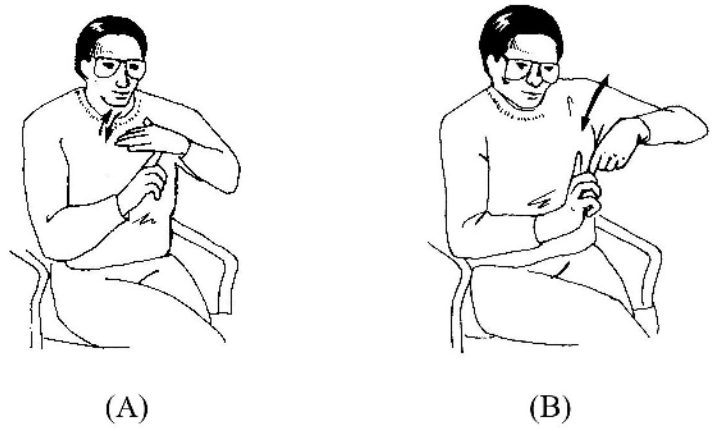

Figure 1.4: *Gestures of subjects retelling scenes from a Sylvester and Tweetie animation movie (drawings taken from McNeill, 1992). In (A) the speaker says: "he steps on the part where the street car's connecting". The gesture complements this by expressing aspects of direction and trajectory (lower hand) and shape (upper hand). In (B) the speaker says: "he swallows it". The gesture expresses relative locations, direction of movement and aspects of shape.*

mistakes in the agent's behavior (Nass et al., 2000, Paiva et al., 1999). A human body must always display a consistent picture of human behavior. The resulting challenge is to create *believability*.

1.1.3 Believability

Letting agents create an "illusion of life", making them *believable* and *like-life*, is a major goal of embodied agents research. Since DePaulo (1992: 234) found that it is impossible to regulate nonverbal behavior in such a way that no impression at all is conveyed, the agents' behavior must be carefully controlled to convey the intended impression. Speakers who actively suppress movement are perceived as being unexpressive, inhibited, withdrawn and uptight (DePaulo and Kirkendol 1989). Schaumburg (2001) found that designing an interface that takes advantage of the social bias of the user is difficult because users are easily annoyed by unsocial conduct.

Personality and emotions have been found to be key concepts to make an agent believable and can be used to guide speech and gesture generation. In speech, emotion was shown to correlate with intonation, tempo, intensity and voice quality (Schröder et al. 2001), and also personality has been shown to be marked in speech

(Scherer 1979). As far as the body is concerned, Ekman and Friesen (1975) claim that emotion is mainly expressed by the face[10]. However, other researchers found that gestures as well as postures say something about the speaker's emotional state, about his or her personality and status (Collier, 1985, Scheflen, 1964, Scherer et al., 1979). A number of popular science books exploit these insights to advise people on how to interpret and control "body language" (Fast, 1970, Molcho, 1983). As concerns posture, McGinley et al. (1975) showed that a speaker can achieve a higher degree of *opinion change* in his/her addressee when assuming an open posture as opposed to a closed one. In terms of status and liking, Mehrabian (1972) found evidence for two correlations: a more relaxed posture is perceived as low status, and a more *immediate* posture (forward lean, eye contact, body orientation) increases liking. In contrast to posture findings, the relation between gestures and emotion is still quite unexplored.

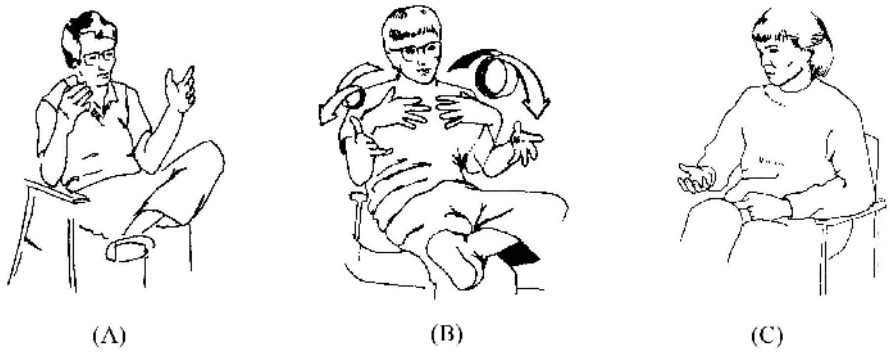

Figure 1.5: *Three instances of metaphoric gestures that frequently occur in normal conversation (drawings taken from McNeill, 1992). In (A) the speaker says: "it was a Sylvester and Tweety cartoon". The gesture indicates a substance held between the hands. The substance is taken as a metaphor for "cartoon". In (B) a circular gesture metaphorically illustrates a process or transition while the speaker says: "and now we get into the story proper". In (C) the speaker a variant of the gesture in (A). A virtual substance is presented on the open palm as a metaphor for something also expressed in speech.*

Most scenarios of embodied agents systems involve normal conversation with the user. Conversational gestures must not necessarily have an explicit function. McNeill (1992) explored a class of gestures he called *metaphorics* that illustrate the

[10]In fact, the correlation between emotions and facial expression is so strong that it works in both directions, that is not only does emotion affect the face but changing the facial expression affects the emotions, a phenomenon called *facial feedback* (Tomkins, 1962, Izard, 1990).

spoken content only via a metaphor as shown in Figure 1.5. According to Webb (1997), such gestures dominate most conversations, so automatically generating conversational gestures should become a research focus to let embodied agents act more life-like. Cassell and Thórisson (1999) show that users are more likely to consider agents life-like when they display *appropriate* nonverbal behavior. A small number of such gestures were integrated in a system by Cassell et al. (1994), using a functional approach (Figure 1.6)[11]. However, since these gestures' function is difficult to unearth and their benefits in terms of communication unclear, there should be an effort to implement a broad spectrum of conversational gestures in a shallow approach. Then, the generated gestures can not only be used to make a single agent believable but also, to make each agent acting in a team stand out as a distinct individual.

Figure 1.6: *In a functional approach the* Animated Conversation *system annotates utterances with how the content can be expressed in gesture, in this case: metaphorically. The agent says: "Will you help me get fifty dollars?". The open palms illustrate the readiness to receive a substance. This substance acts as a metaphor for the answer. (Figure taken from Cassell et al., 1994.)*

Making embodied agents believable still needs much interdisciplinary research (cf. Gratch et al., 2002). The research by Lee et al. (2002) shows how specialized yet important research topics for embodied agents have become. The authors implemented the simulation of saccadic eye movement based on empirical measure-

[11] Copyright by Justine Cassell, Northwestern University

ment with human subjects. An evaluation study showed that this added movement made the face look more natural, friendly and outgoing. In contrast, switching off eye movement led to attributions of lifelessness while random movement led to attributions of unstability. This demonstrates the task complexity of simulating humans: the blink of an eye may count as much as moving the whole body.

1.1.4 Multimodal Interfaces

In the past, research in HCI has primarily been concerned with understanding *input* from different modalities like keyboard, mouse, speech, gesture, touch or facial expression. Gestures were seen as a powerful modality to complement speech input for a more efficient human-computer communication. The Put-That-There system was one of the first systems that understood both speech and (pointing) gestures (Bolt 1980). The system used speech recognition and a 3D space sensing device to let the user manipulate virtual objects on a wall-sized display. The XTRA[12] system, designed as an interface to expert systems, allowed input by gesture and speech using empirical results from experiments on the functions of deixis (Wahlster 1991). The projects ICONIC (Koons et al. 1993), SGIM[13] (Latoschik et al. 1998) and IFP-GS[14] (Hofmann et al. 1998) added data gloves to recognize gestures. SignRec (Hienz et al. 1999), like IFP-GS a system for sign language recognition, relies on a video-based approach: subjects are fitted with colored marks that can be reliably located in image processing. Most of these approaches to gesture recognition[15] consist of three steps. First, the gesture must be segmented, i.e. it must be established where a single gesture starts and where it ends. Second, the gesture must be classified, i.e. in a list of predefined classes the current gesture must be assigned to one class. Third, the recognized gesture must be understood in conjunction with co-occurring speech input.

Whereas early multimodality research focused on *understanding* only, current research is pushing toward *symmetric* multimodality (Wahlster 2003). This means that not only input should be multimodal but that also output should be generated in multiple modalities (text, sound, diagrams, gesture, posture, facial expression). As part of the multimodal output, embodied conversational agents (ECA) are integrated in multimodality projects like SmartKom (Wahlster 2003). SmartKom is a mixed-initiative multimodal dialogue system with three applications as a communication, infotainment and mobile travel companion. The integrated embodied agent Smartakus uses speech, facial expression and gestures coordinated with graphical output to communicate with the user. How to coordinate speech and gesture thus

[12] e**X**pert **TRA**nslator

[13] **S**peech and **G**esture **I**nterfaces for **M**ultimedia

[14] Interdisziplinäres Forschungsprojekt "Gebärdenerkennung mit Sensorhandschuhen", German for: interdisciplinary research project "gestural sign recognition with sensory gloves"

[15] See Wachsmuth and Fröhlich (1998) for representative papers on gesture recognition.

becomes part of the more general question of how to coordinate different modalities. Since fully symmetric multimodal applications must process input as well as output representations, research strives for a single working representation that contains complex multimodal content as well as information about segmentation, synchronization and other processing data. In SmartKom this is called M3L[16] and can be thought of an interlingua for semantic and pragmatic aspects of a message.

A major and often neglected prerequisite for symmetric multimodal interfaces are empirical studies based on annotated corpora (Bunt et al. 2003). However, much is lacking in terms of software to aquire and manage the data as well as exchange of existing corpora. For the systematic study of nonverbal communication, body movements (arms, face, posture) must be recorded in actual communicative situations. While Efron (1941) had to rely on sketches and photographs, researchers have moved to VCRs and now, to digital video for their analysis (Loehr and Harper 2003). However, the move to digital video and computerized transcriptions is still in progress. Generic tools and standards of transcription are a matter of current research.

When human coders transcribe observed movements from video, they necessarily reduce the primary information in an interpretative process. For certain purposes more objective and exact methods are required. Therefore, some researchers work on the automated capturing of movement using image processing. Quek and McNeill (2000) developed a tool that computes hand position and head orientation from video frames. Grammer et al. (1997) point out the neglect of motion quality (speed, acceleration, spatial extension etc.) in behavior research and ascribe this deficit to the methods used. They developed a system of automatic movie analysis (AMA) where digitized video is analyzed using image filters. The motion energy detection (MED) works by computing the difference of a gray-scale video frame from the previous frame pixel by pixel. Alternatively, one could obtain exact data by using data gloves or other methods from motion capturing and gesture recognition.

1.2 Research Aims

The previous sections introduced embodied agents as a potentially beneficial interface between human and computer. However, to make these agents work the human user must perceive them as living beings without being distracted by unnatural gestural behavior. When a team of agents works together an additional requirement arises: that the agents display individual differences. Otherwise, the human user would perceive them as clones with identical gestural behavior. This is potentially distracting even if each single agent has believable behavior.

[16] **M**ulti**m**odal **M**arkup **L**anguage

1.2.1 Generation by Imitation

This dissertation deals with the problem of generating gestures for a team of computer-animated agents. The gestures must be believable, entertaining and individual. To generate gestures means to simulate an aspect of human behavior. The Oxford English Dictionary (OED) defines *to simulate* in the narrow sense as to "produce a computer model of (a process)" (Brown 1993). In Cognitive Science, simulations refer to *functional* simulations of cognitive processes that are created to test hypotheses on the original human processes. Other simulations recreate, according to the OED, "the conditions of (a situation or process), esp. for the purpose of training". For the gesture generation approach of this dissertation this definition of simulation appears to be too broad. Neither is the creation of a functional model of human gesture production nor that of a training environment simulation intended. Therefore, the more specific notion of *imitation* will be used here. The Chambers Science and Technology Dictionary gives the following definition (Walker 1991):

> **imitate** *(Behav.).* Learning through the observation of another individual (model) which is accomplished without practice or direct experience.

This definition contains some important concepts. It emphasizes that imitation usually refers to human individuals. One *imitates* a specific person, whereas one *simulates* more generally a human being. For three reasons it makes sense to take a single, especially selected individual as the basis for modeling as opposed to relying on a population of subjects. First, in a team of agents each agent must display individual behavior to avoid creating behavioral clones which degrades the believability of the team. Second, for the target applications of presentation, sales, education etc. the agents should be more regarded like actors on a stage instead of simulated humans (André and Rist 2000), actors who perform for an audience: the user(s). Consequently, the agents should display a certain proficiency with gestures or, in other words, they must not display monotonous or distracting gestures. Such a proficiency can be ascertained by selecting experienced public performers. Third, instead of focusing on a few specimen that are functionally modeled, the aim is to arrive at a broad range of output gestures. The focus lies on creating a rich gestural base behavior that can be complemented by functionally modeled gestures where necessary.

The dictionary description of imitation also states that the method of imitation is pure observation without "direct experience". Technically, this can be translated to a corpus-based approach to generation in three phases. First, the behavior of the target must be observed. The observed behavior is strongly context-dependent and has many degrees of freedom. Therefore, in the second phase, the observed behavior must be generalized from its specific context and those parameters must