Vis Comput
DOI 10.1007/s00371-009-0401-x

ORIGINAL ARTICLE

Video-realistic image-based eye animation via statistically driven state machines

Axel Weissenfeld · Kang Liu · Jörn Ostermann

Abstract In this work we elaborate on a novel image-based system for creating video-realistic eye animations to arbitrary spoken output. These animations are useful to give a face to multimedia applications such as virtual operators in dialog systems. Our eye animation system consists of two parts: eye control unit and rendering engine, which synthesizes eye animations by combining 3D and image-based models. The designed eye control unit is based on eye movement physiology and the statistical analysis of recorded human subjects. As already analyzed in previous publications, eye movements vary while listening and talking. We focus on the latter and are the first to design a new model which fully automatically couples eye blinks and movements with phonetic and prosodic information extracted from spoken language. We extended the already known simple gaze model by refining mutual gaze to better model human eye movements. Furthermore, we improved the eye movement models by considering head tilts, torsion, and eyelid movements. Mainly due to our integrated blink and gaze model and to the control of eye movements based on spoken language, subjective tests indicate that participants are not able to distinguish between real eye motions and our animations, which has not been achieved before.

Keywords Eye animation · Talking-heads · Sample-based image synthesis · Computer vision

1 Introduction

Computer-aided modeling of human faces usually requires a lot of manual control to achieve realistic animations and to prevent unrealistic or nonhuman-like results. Humans are very sensitive to any abnormal lineaments, so that facial animation remains a challenging task till today. Facial animation combined with spoken output, also known as talking-head, can be used as a modern human–machine interface [1].

Talking-heads give a face to spoken output, which is either generated by a text-to-speech synthesizer (TTS) or recorded from a human subject. Dialog systems, as used in e-commerce and e-care, can integrate facial animations with synthesized speech generated by a TTS synthesizer in web sites to improve human–machine communication. Instead of producing expensive TV and video productions, a talking-head can be animated by the spoken output of the human subject. In comparison to video, a voice recording is very easy to accomplish and inexpensive. Furthermore, talking-heads can be given as add-ons to audio productions. For instance, a talking-head can read audio books, so customers have the choice between audio only and a talking-head reading the story .

Today animation techniques range from animating 3D models to image-based rendering [1]. In order to animate a 3D model consisting of a 3D mesh, which defines the geo-

Electronic supplementary material The online version of this article (http://dx.doi.org/10.1007/s00371-009-0401-x) contains supplementary material, which is available to authorized users.

A. Weissenfeld (✉) · K. Liu · J. Ostermann
Institut für Informationsverarbeitung,
Leibniz Universität Hannover, Appelstr. 9A, 30167 Hannover, Germany
e-mail: aweissen@tnt.uni-hannover.de

K. Liu
e-mail: kang@tnt.uni-hannover.de

J. Ostermann
e-mail: ostermann@tnt.uni-hannover.de

Published online: 17 November 2009

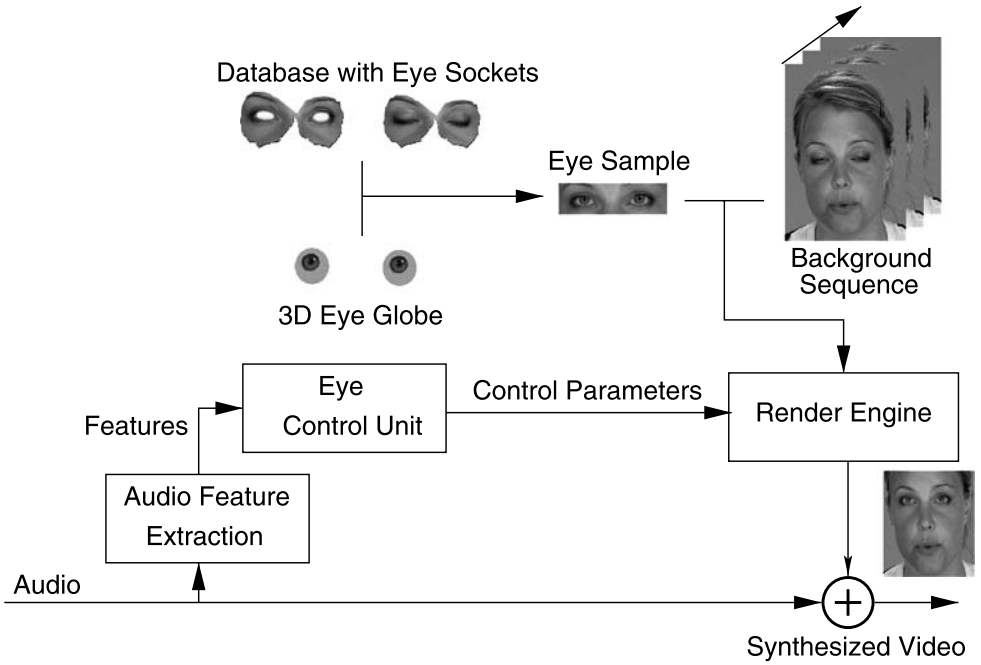


Fig. 1 Image-based eye animation system: Initially phonetic and prosodic information are extracted from the audio. The ECU generates eye blinks and movements and sends control parameters to the render engine

metric shape of the head, the vertices of the 3D mesh are moved [2, 3]. Image-based rendering processes only 2D images, so that new animations are generated by combining different facial parts of recorded image sequences. We use an image-based system since we believe that in the near future this type of system will allow the animation of avatars, which cannot be distinguished from a real person.

Image-based facial animation, which was introduced by [4–6], concentrates on synthesizing smooth mouth animations by replacing the mouth area in a background sequence with previously stored samples in a database. Background sequences are recorded video sequences of a human subject with typical short head movements. However, the proposed systems have several shortcomings, which are important for achieving video-realistic facial animations. We define video-realism as the synthesis of facial animations, which are indistinguishable from real recorded videos and at the same time correctly model the human-like behavior. Hence, facial animations need to appropriately model nonverbal communication to spoken output to appear human-like. However, facial expressions, and head and eye movements are mainly neglected in image-based systems. This work focuses on replacing the eye area to generate video-realistic eye animations to spoken output, since eyes play an essential role as a major channel of nonverbal communication.

Eye animation systems (Fig. 1) consist of an eye control unit (ECU) and a rendering engine to synthesize animations. The ECU consists of models controlling gaze patterns, blinks, and the dynamics of human eye movements and sends generated eye control parameters to the rendering engine. Optionally eye animation systems have a unit extracting audio features from the spoken output, which are sent to the ECU.

Eye animations are either controlled by conversational rules, statistical models, or a combination of both. Conversational rules take the spoken language into account in order to determine gaze. For instance, when the speaker wants to give her turn of speaking to the listener, she usually gazes at the listener at the end of the utterance. Statistical models are based on measurements of the human eyes. For example, a statistical model may be used to determine the magnitude of a saccade. Saccades are rapid eye movements repositioning the eye gaze to new locations in the visual environment. Rather simple models are designed in [7–13], which are only capable to model gaze patterns for exactly predefined conversational settings with a limited set of predefined sentences. Hence, if eye animations need to be automatically generated to arbitrary spoken output, e.g., as provided by an audio book, these models are not appropriate. These techniques included neither the dynamics of eye movements nor eye blinks.

Cosatto [18] implements eye blinks and eye globe motions with probabilistic state machines. Eye movements are generated by moving the gaze point between the interlocutor and an imaginary desk. The animation is rendered by combining a 3D eye globe model with image-based rendering of the eye area. However, the proposed method is not further evaluated, so that we cannot grade the quality of the modeling. In general, the proposed control models are very simple and do not take audio features into account. Hence, his models do not distinguish between listening and talking, although human gaze patterns strongly vary [19]. Thus, the animations will not appear video-realistic.

Lee et al. [15] propose a comprehensive statistical model to control eye motion developed from their own gaze tracking analysis of real people. The avatar can be in one of the three cognitive states: listening, talking, or thinking. For this, a human operator manually segments the original eye-tracking video. Each state has its own model and probabilities to perform saccades. Saccades are rapid eye movements repositioning the eye gaze to new locations in the visual environment. The gaze pattern consists of two states, looking at and away from the interlocutor. These states and the execution of saccades are modeled by measured probability distributions. In addition, their model considers head rotations. Thus, if the direction of the head rotation has changed and its amplitude is larger than a threshold, then a gaze shift accompanying the head motion is created. Subjective tests with a cartoon-like avatar show that the proposed statistical model achieves higher scores than a stationary and random eye movement model. We regard their work as a reference method, since, on the one hand, their model takes many details into account such as modeling the dynamics of saccades. On the other hand, their approach of designing a model based on measured statistics guarantees a great flexibility. However, their method has still several shortcomings. Manually distinguishing between talking and thinking mode is redundant, since the spoken language already contains this information. Their designed model does not distinguish between small gaze shifts and short fixations within the facial area of the interlocutor and looking away. Hence, a large saccade may be executed with a short duration in the look away state resulting in flickering eye motions, which look unnatural. They do not model eye blinks.

In the work of Deng et al. [14], an automated eye animation is present in which new eye motions and blinks are synthesized by texture synthesis. A database with information of the eye blink signal and eye gaze position is generated by analyzing recorded sequences. The initial eye gaze and eye blink position are randomly selected. Afterwards a number of similar samples of the database are determined, and one is randomly selected, which determines the new gaze and blink position. In this way an animation path for eye gaze and eye blinks is generated. Statistical dependencies between gaze and blinks are not explicitly considered. The animation is evaluated by subjective tests. The animations are synthesized by using the model derived in [15] for gaze and eye blinks generated by a Poisson distribution or the proposed model, which was favored. However, their designed eye animation is only generated for listening mode, while our main focus is on synthesizing realistic eye movements while talking.

Masuko and Hoshino [20] present a method to generate eye and head movements synchronized with the conversation of virtual actors. Their model is mainly based on works of [12] and [15]. Moreover, their model induces head rotations due to the distance between the current and subsequent gaze point. They subjectively evaluate their results by comparing video with only blinks, only eye movement without head movement and the proposed method, which achieved the best scores. A comparison with [15] is missing. Since the control of head rotation is challenging using image-based rendering, we do not take their extension into account.

The work of Ma and Deng [17] is based on training a Dynamic Coupled Component Analysis model by recording the gaze and head movements of six humans for 150 s each. Eye animations are generated by predicting eye movements from a head motion sequence as input. They evaluated their approach by subjective tests. For this, they created animations of the original recorded sequence, the model derived in [15], and their proposed model. The original was favored followed by their new approach, which is founded in [16]. In [16] the saccade kinematics consisting of the velocity and duration are investigated in experiments in which the saccade kinematics of three monkeys are measured. The behavior of eye movements of humans in a conversation, however, is not considered in their work. In addition, only a dependency between saccade kinematics and head movements are presented in [16] but not a model describing the coupling. In their evaluation they generated the video samples of the reference method [15] only by taking the simple head-gaze model into account. However, one part of the model of Lee et al. [15] generates gaze shifts to speech content, which is not considered in their results. Since a meaningful model of predicting the saccade kinematics from head movements is not available yet, we do not consider this relationship in our work.

In the previous paragraphs it has been shown that a number of issues are either not or only insufficiently addressed. Instead of manually adding the mode thinking as in [15], we want to automatically control gaze movements with spoken language. This approach has the advantage of generating eye movements to arbitrary spoken output without manual interference. Although, Cranach [21] already showed a correlation between eye movements and blinks in 1969, eye control models proposed in literature neglected this aspect. Our designed model will integrate possible statistical dependencies. In order to model eye blinks, we will explore whether eye blinks can be controlled by spoken output as investigated in a psychological study [22]. Since flickering eye motions result in unnatural-looking animations, we will refine the model of generating saccades proposed in [15] based on a careful analysis of mutual gaze. In order to overall improve the quality of the eye animation, we will integrate research results from studies about eye movement physiology carried out in the field of neurophysiology and ophthalmology. Furthermore, an appropriate image-based rendering engine needs to be developed.

Our designed ECU has the following innovations with respect to the state-of-the-art. Firstly, gaze shifts and eye

blinks are fully controlled by audio features in talking mode, which allows one to automatically create eye animations to arbitrary spoken output. Secondly, one integrated model steers eye blinks and gaze shifts in talking mode. Thirdly, saccadic eye movements are improved by considering Listing's law, head tilts, and the coupling between vertical saccades and eye blinks. Fourthly, the state in which the speaker is looking to the interlocutor is refined by a model taking short fixations within this state into account. Last but not least, the eye blink model considers temporal as well as audio features to generate eye blinks.

This paper is organized as follows. Section 2 gives a brief introduction to 3D motion of rigid objects. In Sect. 3 the recording and analysis of eye movements and blinks of human subjects is described. This data is further evaluated in Sect. 4 in order to investigate statistical dependencies between eye blinks, gaze patterns, and spoken language. Models controlling eye movements and blinks are designed in Sect. 5, and Sect. 6 explores the designed rendering engine. The results are presented in Sect. 7.

2 Motion of face and eye model

The pose of rigid objects (Fig. 2a) can be described by a translation vector $\vec{T}$ and rotation matrix $\underline{\mathrm{R}}$ with rotation angles $(\omega_x, \omega_y, \omega_z)$. The rotation matrix $\underline{\mathrm{R}}$ is defined by three consecutive rotations about the x-, y-, and z-axes:

$$\underline{\mathrm{R}} = \underline{\mathrm{R}}_z \cdot \underline{\mathrm{R}}_y \cdot \underline{\mathrm{R}}_x. \tag{1}$$

The rotation of the eye globe (Fig. 2b), which is modeled as a half sphere, can either be described by (1) or a quaternion. The latter approach is more intuitive and offers several

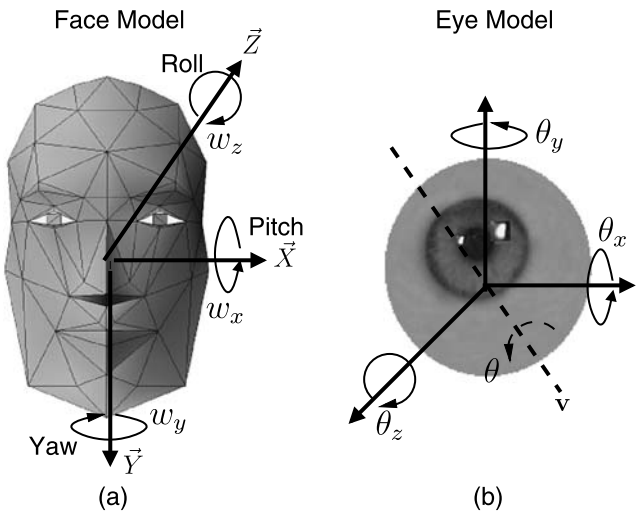


Fig. 2 (**a**) The pose of the face model is defined by its rotation matrix and translation. (**b**) Rotation of the eye globe, which is modeled as a sphere, can be either executed by a rotation matrix (1) or quaternion (2) with the rotation angle θ around the rotation axis $\mathbf{v}$

advantages over (1) [23]. A quaternion $q = (q_0, q_1, q_2, q_3)$ can be associated with a rotation by an angle $\theta \in [0, \pi]$ about the rotation axis $\mathbf{v}$ using quaternion notation [23]:

$$q = \cos\left(\frac{\theta}{2}\right) + \sin\left(\frac{\theta}{2}\right)\mathbf{v}. \tag{2}$$

3 Analysis of eye movements and blinks

We distinguish between two gaze states, mutual gaze and gaze away. We define mutual gaze as the state in which the direction of the gaze is located within the facial area of the interlocutor consisting of the mouth and eye area. If the gaze is not in this defined area, then the system is in gaze away. If the speaker switches from mutual gaze to gaze away, a gaze shift is performed. In order to analyze the statistical dependencies between gaze and blink patterns and spoken language, we conduct experiment 1 in Sect. 3.1.

The dynamics of saccadic eye movements cannot be analyzed in the first experiment, since the spatial and temporal sampling frequency of ordinary cameras is too low. Therefore, in the second setup, which is similar to that in [15], an eye tracker is used (Sect. 3.2). The significant drawback of eye trackers is either the necessity of wearing an eye tracking device or the restriction of a stationary head pose. Hence, we only use eye trackers to determine characteristics which cannot be analyzed by the first experiment, e.g., the kinematics of saccades.

3.1 Experiment 1

In the first setup, we record in two sessions a conversation of two persons who are interviewing each other and discussing current-affairs. In each session, which lasted for 30 minutes, the same moderator and a different human subject participated. They are sitting in front of a table and facing each other. The camera is located next to the head of the moderator, and a microphone is positioned on the table. Both human subjects are informed that not eye but mouth movements and facial expressions are investigated in this study in order to avoid potential change of eye movement behavior. In each session the beginning of the conversation is not recorded in order to let the human subject get used to the setup. After a while (5–7 minutes), we believe that the subjects acclimate and their subconscious controls the eye movements. While recording took place in a lab environment, subjects seemed to behave naturally. For the analysis, the recorded video is manually divided into segments in which the recorded human subject is listening and talking, since the gaze behavior varies [19]. All frames of the "listening segments" are only labeled with their gaze and blink patterns. All frames of the "talking segments" are additionally labeled with audio information. The audio information contains phonetic

and prosodic features. We perform phoneme labeling on the recorded sequences using the speech recognition software HTK [24]. Given an audio sequence and an associated text transcript of the speech being uttered, HTK uses forced Viterbi search to find the optimal phoneme durations for the given audio sequences. The most important characteristics of prosody are pause, which is already detected by the phonetic labeling, speech rate (or phoneme duration), and pitch variations [25]. We use an alignment-based method as proposed in [26] in order to classify the words of the spoken output as slow, normal, or fast. There are different definitions of the term prominence. We use Terken's definition [27], who defines prominence as "word or syllables that are perceived as standing out from the environment", because glances are used by speakers to emphasize particular words or phrases [19]. Our algorithm based on [28, 29] automatically detects areas of high pitch variability, which indicate an emphasis, and labels these as prominence. An example of a labeled video is presented in Fig. 3.

3.2 Experiment 2

In this setup we use a table-mounted eye tracker [30] which requires a stationary head pose. The eye tracker is used in a video conference system in which the communication takes place over a camera and microphone. This setup is most similar with respect to possible applications of talking-heads, e.g., in a dialog system. The conversational settings are the same as in the first experiment. The eye tracker processes the measured point of regard (POR) on the computer screen in order to automatically determine saccadic eye movements and fixations. Under POR we understand the direction of the gaze where a person is looking. The distance between the POR before and after a saccade is performed can be converted into a rotation with a direction and magnitude characterizing the saccade if the distance between the eyes and computer screen is known. Furthermore, this eye tracker gives the opportunity to analyze gaze shifts of humans while looking at the interlocutor (mutual gaze) in more detail.

4 Statistical dependencies

In this section important statistical dependencies and distributions of eye blinks and eye movements are investigated. These results will be incorporated in the designed models controlling eye blinks and eye movements. Note that extreme values of eye movements and blinks are eliminated in order to prevent unnatural animations. The distributions may highly vary between individuals, because each human has its own nonverbal behavior.

4.1 Gaze and blink patterns

Psychological studies show a significant difference of human gaze patterns while listening and talking [19]. Hence, for each mode, the relative frequency distributions of remaining in mutual gaze and gaze away as well as the duration between two consecutive eye blinks, which is denoted as nonblink duration, are calculated. While the sample mean $\bar{X}$ of the duration in gaze away increases from 0.63 s in listening to 0.89 s in talking mode, the eye blink distribution remains very similar. The relative frequency distributions of duration in mutual gaze and gaze away while listening are illustrated in Fig. 4a, b. Furthermore, the distributions of gaze away duration while talking and nonblink duration are depicted in Fig. 4c, d.

In order to describe these relative frequency distributions, it is useful to model them by a well-known probability distribution such as a Poisson or geometric distribution. In order to evaluate the equality of the frequency distribution with a reference probability distribution, the Kolmogorov–Smirnov test (KS test) is used [31], which quantifies a distance between the distribution function of the sample and the cumulative distribution function of the reference distribution. As null hypothesis H_0, the KS test states that both cumulative distribution functions are the same. In Table 1 the results of the KS test of the relative frequency distribution of Fig. 4c are exemplarily shown. As this test reveals,

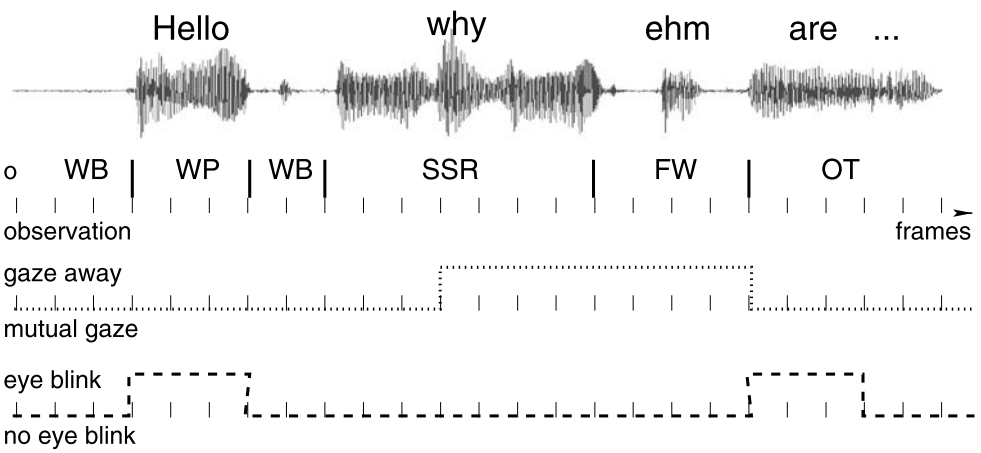


Fig. 3 The speech waveform of a segment, which is labeled with its corresponding observations, is depicted ("word boundary or pause" (WB), "slow speech rate" (SSR), "word prominence" (WP), "filling word" (FW), "other" (OT)). The speaker is either in mutual gaze or gaze away (illustrated by *the dotted line*). *The dashed line* illustrates when the speaker performs eye blinks

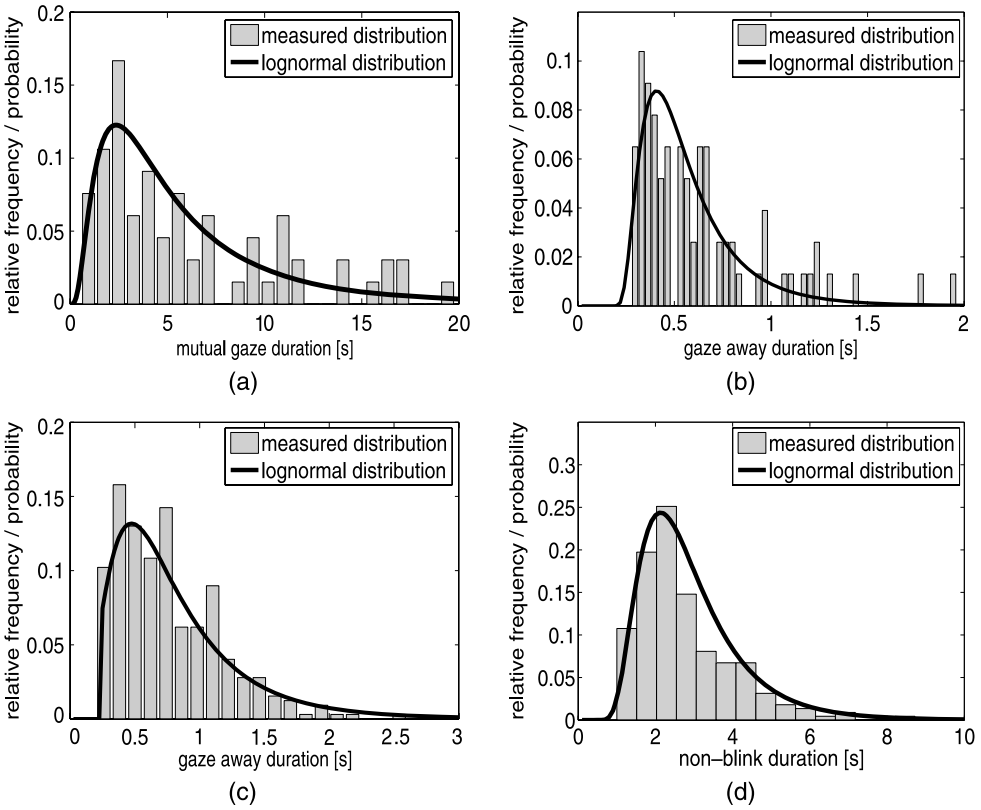


Fig. 4 Relative frequency and fitted lognormal distribution of the duration of (**a**) mutual gaze ($\bar{X} = 5.35$ s, $S = 5.63$ s), (**b**) gaze away ($\bar{X} = 0.63$ s, $S = 0.35$ s) while listening, (**c**) gaze away duration while talking ($\bar{X} = 0.89$ s, $S = 0.52$ s), (**d**) duration between two consecutive eye blinks ($\bar{X} = 5.29$ s, $S = 4.15$ s)

Table 1 Results of the Kolmogorov–Smirnov test by fitting different probability distributions to the relative frequency distribution of Fig. 4c. The corresponding p-values and whether the null hypothesis H_0 is rejected or retained using $\alpha = 0.05$ as level of significance are displayed

Distribution	p-value	H_0
Poisson	0.039	rejected
Negative binomial	0.008	rejected
Chi-Square	$<10^{-3}$	rejected
Lognormal	0.48	retained

only the lognormal distribution is retained, which is defined as

$$f(x) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} \cdot e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}, & x > 0, \\ 0.0, & x \leq 0, \end{cases} \tag{3}$$

with parameters μ and σ [32]. For this, the lognormal distribution $f(x)$ is fitted to the relative frequency distribution by a maximum likelihood estimate. Lognormal distributions are often used if measurements show a more or less skewed distribution. Skewed distributions are particularly common when mean values are low, variances large, and all values equal or larger than zero [33], as in our case. However, we evaluate $f(x)$ only at a discrete set of uniformly spaced points x_k. The endpoints of the interval are the smallest x_1 and largest x_K measured durations. Therefore, $f(x)$ is normalized, resulting in $\check{f}(x_k)$. For $\check{f}(x_k)$, it applies

$$\sum_{k=1}^{K} \check{f}(x_k) = 1. \tag{4}$$

We characterize the distributions by $\bar{X}$ and S (standard deviation), which is common for data with a lognormal distribution [33].

The analysis of eye tracking data indicates that humans do not constantly stare at the interlocutor in mutual gaze, but rather move the POR across the face, which is observed in both modes. In order to analyze the POR within mutual gaze in more detail, we define four regions of interest (ROIs) in the face: left eye, right eye, mouth area, and the rest of the face. The first three ROIs are defined as an elliptic region (Fig. 5a). In order to calculate the experimental probability that the POR is located inside an ROI, the measured fixations are assigned to the corresponding ROI (Fig. 5b) and modeled by a normalized exponential distribution. The measured probabilities are very similar for both modes.

4.2 Gaze patterns, eye blinks, and spoken language

In this section the statistical dependencies between eye blinks, eye movements, and spoken language are determined. For this, we calculate the experimental conditional

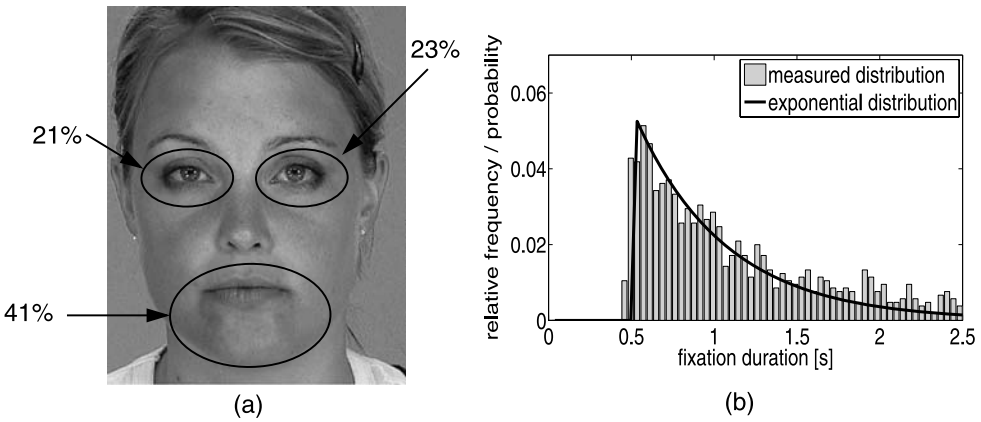


Fig. 5 (**a**) The ROI of the left and right eyes and of mouth area are marked by *an ellipse*. Each ROI is labeled with its corresponding probability that the human subject looks at this ROI. (**b**) The relative frequency distribution of the duration of fixations within the ROI "mouth area" is modeled by an exponential distribution

Table 2 For two recorded human subjects, the experimental conditional probabilities of gaze shifts and observations are presented (WB: word boundary, SSR: slow speech rate, FW: filling word, WP: word prominence, OT: other)

		WB	FW	SSR	WP	OT
Subj. 1	$p(o\|\text{GS})$	33.5%	9.0%	33.0%	3.7%	20.8%
	$p(\text{GS}\|o)$	7.6%	8.7%	4.4%	0.3%	1.3%
Subj. 2	$p(o\|\text{GS})$	41.2%	9.4%	14.6%	9.0%	29.6%
	$p(\text{GS}\|o)$	4.3%	3.3%	6.0%	0.7%	0.8%

probability of executing a gaze shift (GS) or blink given an observation o.

In Table 2 the experimental conditional probabilities $p(o|\text{GS})$ and $p(\text{GS}|o)$ are depicted for two human subjects, which significantly vary between both speakers. The probability $p(o|\text{GS})$ illustrates the distribution of observations while performing a gaze shift. For instance, subject 1 executes 33.5% of her gaze shift during a "word boundary".

The three observations "word boundary", "slow speech rate", and "filling word", which may indicate that the speaker is in thinking mode, have a high probability $p(\text{GS}|o)$ that a gaze shift is performed. During "word prominence" (WP), the speaker usually looks to the interlocutor, and therefore $p(\text{GS}|o \neq \text{WP}) \gg p(\text{GS}|o = \text{WP})$.

Moreover, we analyze the gaze behavior at the beginning and end of utterances. At the end of an utterance, the speaker always glances to the interlocutor as a turn-taking signal, which is consistent with [34]. Hence, we introduce the observation "end of sentence". Kendon et al. [34] observed that at the beginning of an utterance a gaze shift is performed. We analyzed this in more detail and determined that a gaze shift is executed if one of the observations "word boundary", "slow speech rate", and "filling word" occurs.

We did not observe a dependency between observations and gaze away duration. Therefore, we assume that the gaze away duration is not influenced by its corresponding observation.

Condon and Ogston [22] observed that eye blinks mainly occur during vocalization at the beginning of words or utterances, the initial vowel of a word, and following the termination of a word. Hence, we label each frame of an utterance with one of the following observations: "vowel", "consonant", and "word boundary". Since vowels already account for vocalization, we neglect vocalized consonants. First, we calculate the experimental conditional probability $p(o|\text{B})$ that the observation o occurs if a blink B is executed (Table 3). A large number of blinks are performed at word boundaries (WB), e.g., $p(o = \text{WB}|\text{B}) = 0.61$ of subject 1. Afterwards we determine the conditional probability $p(\text{B}|o)$ that a blink is performed given o. This conditional probability indicates a high statistical dependency between word boundary and blinks, while the other two observations have low conditional probabilities. Since $p(\text{B}|o = \text{consonant}) \approx p(\text{B}|o = \text{vowel})$, the system does not distinguish between vowels and consonant, and labels these simply as other (OT).

4.3 Characteristics of saccades

The characteristics of saccades which induce a shift from mutual gaze to gaze away or vice versa are only briefly analyzed, since in [15] a comprehensive investigation is already presented. The relative frequency distribution of the direction and magnitude A while talking are illustrated in Fig. 6. The relative frequency distribution of A is modeled by an exponential distribution.

4.4 Gaze shifts and head movements

Stahl [36] presents an attempt to describe the relation between head rotation and horizontal saccadic magnitudes in more detail. He shows this correlation only for horizontal saccades, whereas we assume that head rotations also accompany saccades in other directions. Within a certain magnitude range, head movements and gaze shifts are independent. If the saccadic magnitude A is larger than a threshold, e.g., 15°, then the head movement accompanies the saccade.

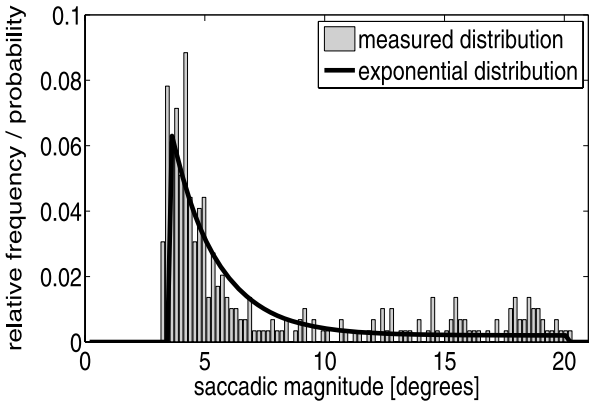
Table 3 For two recorded human subjects, the experimental conditional probabilities of eye blinks and observations are presented

	Word boundary		Consonant		Vowel	
	$p(o\|\mathrm{B})$	$p(\mathrm{B}\|o)$	$p(o\|\mathrm{B})$	$p(\mathrm{B}\|o)$	$p(o\|\mathrm{B})$	$p(\mathrm{B}\|o)$
Subj. 1	60.7%	20.1%	16.0%	7.7%	23.3%	6.1%
Subj. 2	52.1%	11.0%	18.5%	2.3%	29.4%	2.4%

Fig. 6 (**a**) Relative frequency distribution of saccade directions. For instance, the direction 0° defines a gaze shift to the right. (**b**) Relative frequency distribution of the magnitude of large saccades in talking mode is superimposed with a fitted exponential distribution. In general, the saccadic magnitude is larger while talking than listening

Direction	rel. frequency
0°	0.16
45°	0.09
90°	0.15
135°	0.08
180°	0.19
225°	0.09
270°	0.16
315°	0.08

(a)



(b)

Since the direction of the head movement is defined by the background sequence, we appropriately adapt the direction of the saccade.

Previous studies have related head movement amplitude θ_h to gaze shift amplitude [35]. Head movements of the background sequence are partitioned into segments. For each segment of the background sequence, the quaternion q_h is determined describing head rotation. The corresponding rotation angle θ_h from (2) gives the magnitude of the head rotation. If θ_h is larger than a threshold, a saccade accompanying the head rotation is generated.

4.5 Gaze shifts and eye blinks

Eye animations concentrate on gaze patterns and usually do not pay attention to eye blinks. In 1969 Cranach et al. [21] already investigated a correlation between gaze shifts and blinks. Namely the larger the gaze shift, the higher the probability of simultaneously executing a blink and gaze shift. A well-known hypothesis is that eye blinks due to gaze shifts are executed in order to prevent apparent motion [21].

In this section, we specify the statistical dependency between gaze shifts (GS) and eye blinks (B) in more quantitative terms. Note that eye blinks do not induce gaze shifts. While for listening, the experimental conditional probability $p(\mathrm{B}|\mathrm{GS})$ is only 2.7% of subject 1, whereas it increases to 12.6% of subject 2. This high variation is due to the different nonverbal behavior of both subjects. While subject 2 regularly executes large head rotations, e.g., while laughing, subject 1 exhibits less emotion. Since a large head rotation induces a large gaze shift and large gaze shift induces an eye blink again, $p(\mathrm{B}|\mathrm{GS})$ of subject 2 is much higher than of subject 1. However, since we do not get any feedback from the interlocutor in listening mode, the talking-head only performs typical short head movements in the background sequence. Hence, if we only consider eye blinks during small or normal head movements, $p(\mathrm{B}|\mathrm{GS})$ decreases to 2.5%. Then the events B and GS are statistically independent because of $p(\mathrm{B}|\mathrm{GS}) \approx p(\mathrm{B})$.

In talking mode, $p(\mathrm{B}|\mathrm{GS})$ significantly increases to 23.2% of subject 1 and even 64.5% of subject 2. This high probability is due to the higher number of large saccades performed while talking. As a result, we need to couple saccades and blinks in one control model. For this, we analyze the dependency in more detail by designing an experiment in which we measure the experimental conditional probability $p(\mathrm{B}|A)$ of executing a blink B, given the saccadic magnitude A. Note that A already implies a gaze shift. In the experiment we present single images on which a single green dot is displayed. We ask the participant to follow the green dot in the image sequence with the eyes while measuring the POR with an eye tracker. The spatial difference between consecutive dots is between 5° and 22°. The experimental results whether a blink is executed while changing the POR can be described by the following function:

$$p(\mathrm{B}|A) = \begin{cases} 0.02; & 5 \le A^s < 7.5, \\ 0.09; & 7.5 \le A^s < 12.5, \\ 0.24; & 12.5 \le A^s < 17.5, \\ 0.41; & 17.5 \le A^s < 22. \end{cases} \tag{5}$$

In general, $p(\mathrm{B}|A)$ increases by an increase of the saccadic magnitude A as proposed in [21]. Since the execution

of a blink during a large saccade is due to controlled by the subconscious and speech does not seem to be a dominant factor, we assume that the statistical relationship between a saccade and an eye blink in this experiment is the same as in a two-way conversation.

5 Eye control unit

First, the characteristics of eye globe rotations and models to generate the different types of eye movements are briefly explored. The ECU contains models controlling eye blinks and movements (Fig. 7). This unit also selects the appropriate model for listening and talking mode.

5.1 Eye movement physiology

Donders' law states that each time the eye looks in a particular direction, it only assumes one 3D orientation [37]. If the eye looks straight forward with the head straight and fixed, which is denoted as the primary position, torsion is not induced. All other gaze positions have their own unique torsion component (rotation along the z-axis in Fig. 2b). Listing's law (LL), which is considered to be one of the most important principles in eye movement physiology, states exactly what those torsion values are [38]. It states that when the head is upright and stationary with the eyes fixating a distant object, all rotation axes lie within the same plane, denoted as Listing's plane (LP). In our defined eye coordinate system, LP is orthogonal to the z-axis and intersects the globe center (Fig. 2b). Hence, the rotation axis $\mathbf{v}$ must be located in this plane, which is achieved by setting $q_3 = 0$ of the quaternion representation. If we use rotation matrices (1) instead to describe the rotation of the eye globe, then with simple algebraic manipulation the torsional component in (1) can be calculated. The rotation θ_z (Fig. 8a) is equal to

$$\theta_z = \operatorname{atan}\left[\frac{\sin(\theta_x)\sin(\theta_y)}{\cos(\theta_x)+\cos(\theta_y)}\right], \quad \theta_x, \theta_y \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right). \quad (6)$$

Although LL is only fulfilled under the mentioned conditions, violations like binocular convergence only induce small changes of the orientation of LP. Binocular convergence refers to the angle that results when the viewer turns both eyes to fixate a target. Eye movements, which neither start nor end in the primary position, fulfill LL only by a rule called half-angle [39], which we take into account. Then the plane of rotation is tilted by half the angle between the momentary and primary position.

While pitch and yaw head motions cause only a small change in the orientation of LP, head tilts induce ocular counter roll to keep the image upright. Schworm et al. [40] acquired data from five subjects using video oculography devices to measure ocular counter roll induced by head tilts. The ocular counter rolls induce torsion to the eye globe and

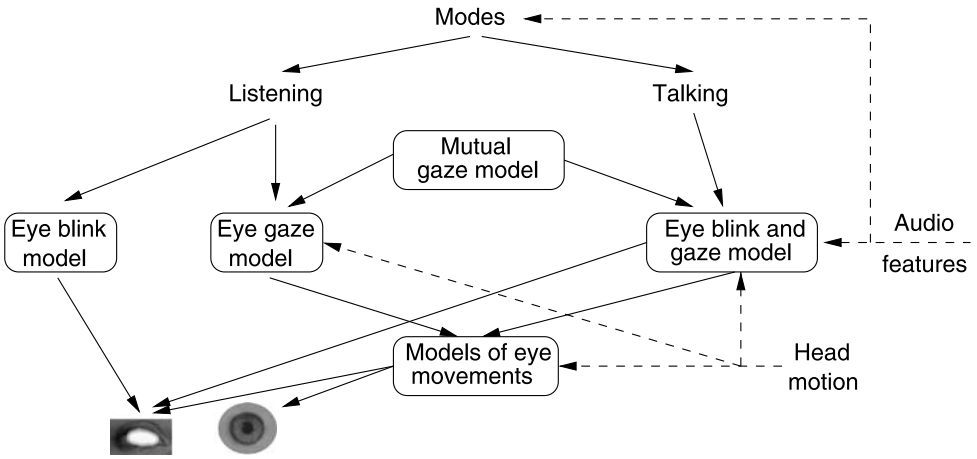


Fig. 7 Overview of the ECU: While in listening mode two independent models control the eyes, in talking mode one model is designed. Both modes share the mutual gaze model and models of eye movements. Whereas the eyelid position is controlled by the eye blink models and the models of eye movements, the eye globe is only steered by the latter. Head motion and audio features are input parameters (*dashed arrows*)

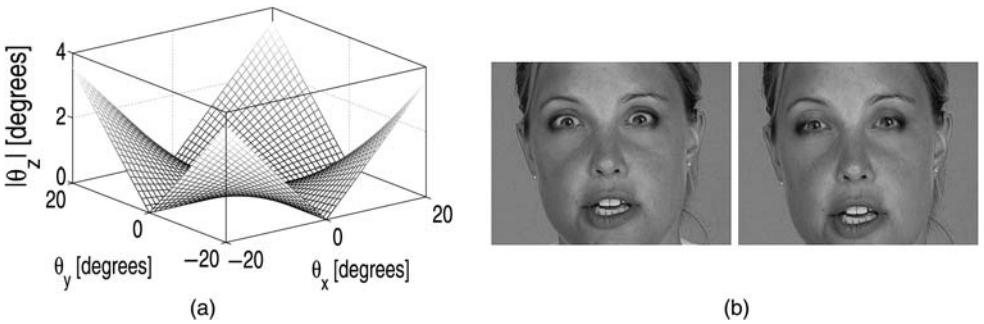


Fig. 8 (**a**) Torsional angle $|\theta_z|$ with respect to θ_x and θ_y. (**b**) While *on the left* the eyeball and lid movements are independently controlled, *the right* shows our proposed method where the eyelid follows the eye globe motion

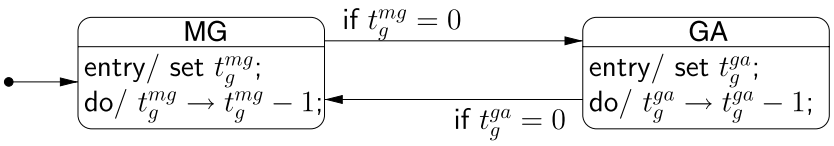


Fig. 9 FSM: Two states, mutual gaze (MG) and gaze away (GA), are used to generate gaze patterns. The duration t_g^{mg} and t_g^{ga} are determined by measured statistics

thus a rotation around the z-axis. We approximated the measured results of [40] by a cubic interpolation

$$\theta_z = p_1 \cdot \omega_z^3 + p_2 \cdot \omega_z^2 + p_3 \cdot \omega_z + p_4 \tag{7}$$

with ω_z equal to the head tilt and the parameters p_1 to p_4 equal to $3.7 \cdot 10^{-5}$, $-3.1 \cdot 10^{-6}$, $-2.1 \cdot 10^{-1}$, and $-2.3 \cdot 10^{-1}$, respectively. The sign in (7) depends on the direction of head tilt. The torsion component according to (6) and (7) may add up to 15°. Head tilts not only influence saccades but VOR as well.

In our system two types of eye movements are executed: saccades and vestibulo-ocular reflex (VOR), which compensates head motion. In order to fixate the retina onto an object during head rotation, the VOR executes eye movements compensating head motions. For perfect compensation, the rotation axis of the eye and head need to be parallel. It has been shown, however, that the axis of eye rotation during VOR neither meets the needs for perfect 3D gaze stabilization nor LL. It is a compromise of both constraints. Since the variations are negligible small, we assume a perfect compensation. Furthermore, the latency of VOR is less than 14 ms and therefore negligibly small in our system.

The generation of saccades is based on the work of [15]. We improve their model by including Listing's law and head tilts, which have been previously described, as well as eyelid movements. Direction and magnitude are generated by modeling the distributions in Fig. 6. Typically, the duration of the saccade is proportional to its magnitude, and the velocity is given by a measured velocity function. Vertical saccades and eyelid movements are coupled. For this, we define multiple saccadic magnitude thresholds in vertical direction (up and down). If the saccadic magnitude is larger than an empirically selected threshold, then the appropriate eyelid is selected from the database. Figure 8b shows two frames extracted from a synthesized video with and without considering this aspect.

5.2 Listening mode

In listening mode the animation system is waiting for input from the user, e.g., a mouse click. Hence, we only synthesize neutral head movements and small variations of the facial expression. Under this condition, we can design two independent control models, one for eye blinks and one for eye movements (Sect. 4.5).

A finite state machine (FSM) synthesizing new gaze patterns with two states is depicted as a statechart in Fig. 9. Initially, the duration t_g^{mg} or t_g^{ga} of remaining in the current state is determined by modeling the normalized lognormal distributions $\check{f}_{\ln}$. While the model remains in the same state, t_g^{mg} or t_g^{ga} decrease. If t_g^{mg} or t_g^{ga} is eventually equal to zero, a saccade is performed to switch the state. Note that the dependency between head motion and saccades is taken into account, too.

If the FSM is in mutual gaze, the eyes still perform small gaze shifts from one ROI to another. This observation is modeled by refining the state mutual gaze with a second FSM with four states each representing one ROI. All measured fixations are assigned to the corresponding ROI. The duration of remaining in one state is determined by modeling the exponential distribution (Fig. 5b). During this time the POR is fixed to the current ROI. Afterwards the state has to be changed. The transitions probabilities from each state model the experimental probability of being in one ROI (Fig. 5a). Note that the current state is immediately left if the duration of remaining in mutual gaze ends.

An FSM with two states, "no blink" and "blink", has the same structure as depicted in Fig. 9 and generates eye blink patterns. In the default state "no blink" the eyes are opened. The system remains in this mode for a certain duration, which is determined by modeling the normalized lognormal distribution in Fig. 4d. If the default state is left, an eye blink is executed, and the model returns to the default state.

5.3 Talking mode

In talking mode, two independent models controlling gaze shifts and blinks cannot be designed, since eye movements and blinks are coupled (Sect. 4.5). Hence, we propose an algorithm that iteratively determines an animation path, which contains information for eye movements and blinks for each frame of the animation (Fig. 10).

Firstly, each frame of the animation path is labeled with its corresponding observation o, which is extracted from the spoken output.

Secondly, for the entire animation, the gaze patterns mutual gaze and gaze away are determined and stored in the animation path. For this, to each frame a random number from a uniform probability distribution between zero and one is assigned. Now each frame has an observation o and random number. The animation system starts in the default state mutual gaze. A gaze shift (GS) is executed if the random number is smaller than $p(\text{GS}|o)$. Then the state is switched from

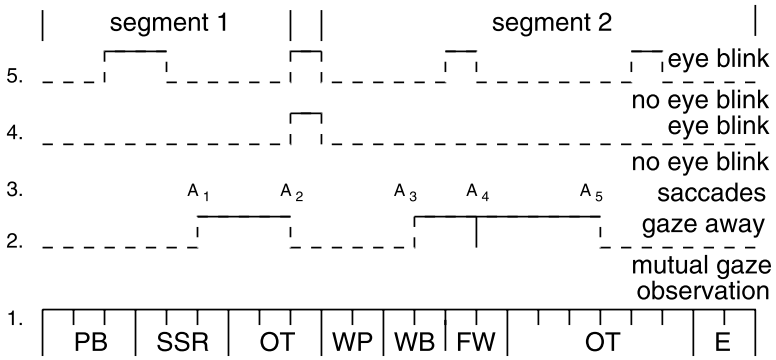


Fig. 10 An example illustrates the generation of an animation path in five steps ("word boundary or pause" (WB), "slow speech rate" (SSR), "word prominence" (WP), "filling word" (FW), "other" (OT), "end of sentence" (E))

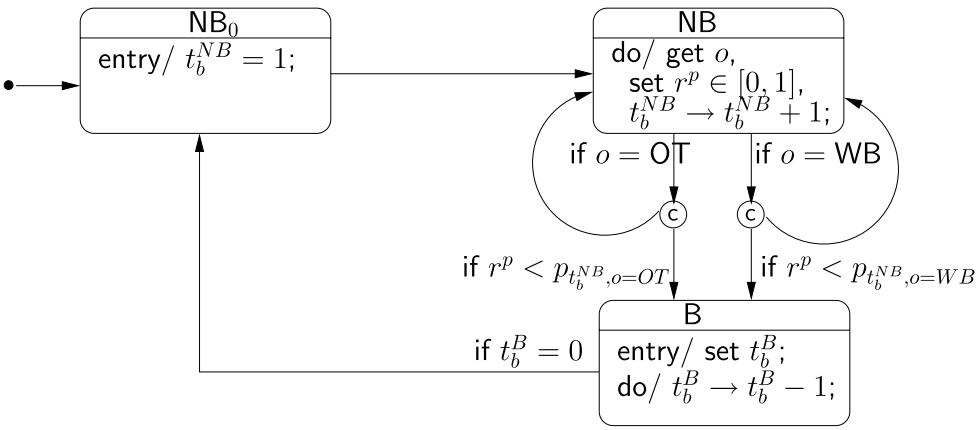


Fig. 11 Statechart: FSM with three states NB_0, NB, and B models eye blinks. The transition probability $p_{t_b^{NB},o}$ from the state NB to B depends on the duration t_b^{NB} and current observation o

mutual gaze to gaze away. Since the duration of remaining in gaze away is independent of the observation, the duration is determined by modeling the lognormal distribution in Fig. 4c. A second consecutive gaze shift is executed with a probability of 34%. After these one or two gaze shifts, the model returns to mutual gaze. At the end of the utterance, the talking-head is looking to the interlocutor (mutual gaze). Note that gaze shifts due to head motion are added (Sect. 4.4). While the system remains is mutual gaze, the POR is varied by an FSM as explained in Listening Mode.

Thirdly, saccades are generated. Preliminary studies indicate that there are no statistical dependencies between the saccadic magnitude of previously executed saccades, gaze away duration, and the current saccade. If necessary, the head motion in the background sequence is taken into account (Sect. 4.4).

Fourthly, eye blinks which are simultaneously executed with a gaze shift are added to the animation path. Since the magnitude A of large saccades and $p(\mathrm{B}|A)$ are known, the probability of executing a blink B can be calculated.

Finally, additional eye blinks are added to the animation path. While the model synthesizing new gaze patterns uses the conditional probability $p(\mathrm{GS}|o)$, eye blinks cannot be generated by only taking the observation o into account. The temporal dependency of blinks must be considered, since eye blinks fulfill the biological purpose to regularly wet the cornea and remove irritants from the surface of the cornea and therefore humans do regularly blink.

Hence, we determine the conditional probability $p(\mathrm{B}|t_b^{\mathrm{NB}}, o)$ of performing a blink B given observation o and time t_b^{NB} passed since the last blink. With simple algebraic manipulation we can easily derive

$$p\big(\mathrm{B}|t_b^{\mathrm{NB}}, o\big) = \frac{p(o|\mathrm{B}, t_b^{\mathrm{NB}}) \cdot p(\mathrm{B}|t_b^{\mathrm{NB}})}{p(o|t_b^{\mathrm{NB}})}. \tag{8}$$

Neglecting statistical dependencies between o and t_b^{NB}, we can rewrite (8) as

$$p\big(\mathrm{B}|t_b^{\mathrm{NB}}, o\big) = \frac{p(o|\mathrm{B}) \cdot p(\mathrm{B}|t_b^{\mathrm{NB}})}{p(o)}. \tag{9}$$

The conditional probability $p(o|\mathrm{B})$ is already measured (Table 3), $p(\mathrm{B}|t_b^{\mathrm{NB}})$ is modeled by the lognormal distribution (Fig. 4d), and $p(o)$ can easily be measured from the recorded corpus.

In order to generate eye blinks, we design an FSM with three states NB_0, NB, and B (Fig. 11). While in NB_0 and NB the eyes are open, in B an eye blink is executed. Initially and after the execution of an eye blink, the machine starts in the default state NB_0, which sets t_b^{NB} to one. After the initialization the state is changed from NB_0 to NB. Each time the current observation o is determined, a random

number r^p generated and the duration t_b^{NB} increased. The machine switches to another state if r^p is smaller than the transition probability $p_{t_b^{\mathrm{NB}},o}$. Since we do know the probability $p(\mathrm{B}_{t_b^{\mathrm{NB}},o})$ of switching to state B given the observation o and duration t_b^{NB}, we can relate the states NB and B with the transition probability $p_{t_b^{\mathrm{NB}},o}$ as

$$p(\mathrm{B}_{t_b^{\mathrm{NB}},o}) = p_{t_b^{\mathrm{NB}},o} \cdot p(\mathrm{NB}_{t_b^{\mathrm{NB}}-1}), \quad t_b^{\mathrm{NB}} > 1, \tag{10}$$

with

$$p(\mathrm{NB}_{t_b^{\mathrm{NB}}-1}) = 1 - \sum_{l=1}^{t_b^{\mathrm{NB}}-1} p(\mathrm{B}_l). \tag{11}$$

Since the probability $p(\mathrm{B}_{t_b^{\mathrm{NB}},o})$ is obviously equal to the conditional probability $p(\mathrm{B}|t_b^{\mathrm{NB}}, o)$, (9) and (10) can be combined resulting in

$$p_{t_b^{\mathrm{NB}},o} = \frac{p(o|B) \cdot p(\mathrm{B}|t_b^{\mathrm{NB}})}{p(o) \cdot p(\mathrm{NB}_{t_b^{\mathrm{NB}}-1})}, \tag{12}$$

giving the transition probability of the FSM of performing a blink given t_b^{NB} and o (Fig. 11). Initially, in state B the duration t_b^{B} of the eye blink is determined. After the blink, the FSM switches to the default state NB_0.

6 Rendering engine

Since modeling human eyes is a difficult task, a sophisticated rendering engine needs to be designed (Fig. 1). The iris contains specular reflections that need to be correctly modeled to achieve life-like looking eyes. In the image-based approach, however, the position of the specular reflections depends on the head's position in the recorded sequence. Hence, eye images cannot be normalized and rendered in a different position. Therefore, a rendering engine is developed which combines a 3D model and image-based rendering [18].

In order to animate a talking-head, the following data has to be initially prepared: The eye globe is modeled by a half sphere with eye texture, which consists of a high-resolution image of the human eye without specular reflections. Moreover, textures with specular lights need to be generated. The eye socket and eyelid is modeled by a number of images stored in a database in which the person executes a blink. The surface of the eye area including the eye socket is approximated by a 3D eye model, which is acquired by a 3D laser scan.

The eye animation is rendered in two steps: Firstly, the eye globes, which are synthesized by texturing half spheres, are rotated according to the eye control parameters. Afterwards specular lights are added at the appropriate positions on the eye globe by taking the eye pose and virtual spot light positions into account. The eye globes are combined with the eye socket image, which is retrieved from the database according to the eye control parameters. Different durations of eye blinks are generated by repeating or removing images from the recorded blink in the database. The rendered image is denoted as an eye sample. Secondly, image rendering overlays the eye sample over a background video sequence by warping the sample into the correct pose. In order to conceal illumination differences between an image of the background video and the eye sample, the samples are blended in the background sequence using alpha-blending.

7 Results

The quality of the synthesized animations of our models is evaluated by a subjective test. This test evaluates by measuring a participants' ability to tell the animated from the real video. Next, the setup of the subjective test is briefly explained.

Altogether 25 people whose age ranged from 17 to 59 years participated and who reported to have normal hearing and normal or corrected-to-normal vision. The number of professional participants was restricted to 20% in order to ensure better correlation of results with potential users. Participants are considered professionals if they have experience and knowledge in information technology or video processing. Participants with reluctant attitude towards technology were left out of the test in order to ensure to have participants interested in our type of research.

The general viewing conditions are set as recommended in [41, 42]. The videos with a resolution of 480 × 384 pel are MPEG-1 encoded with best image quality and displayed with the Windows Media Player. Videos of eye animations with our proposed system can be found on our web site.[1] Two types of test material are presented to participants: an English female and a German male speaker. Altogether 8 different utterances with duration between 2 s and 22 s are prepared. These clips are not used for previously training the models of the eye control unit. In the video clips, the speaker, on the one hand, utters typical sentences used by a virtual operator in a dialog system, and, on the other hand, the speaker describes his new apartment. Eye animations are generated by using the spoken output, and the speaker's head movements as input parameters to the eye animation system (Fig. 7). Only eye movements and blinks are varied, because we are only focusing on this part. We only test the talking mode on which we focus. Each test session begins with an introduction of the purpose and goals of the experiment, and instructions are given to the participants.

[1] http://www.tnt.uni-hannover.de/project/facialanimation/demo/index.html.

Pairs of real and synthetic image-sequences of the same utterance are presented as stimuli, one immediately after the other in randomized order. We compare the reference method of [15] (Type II), which is extended by the eye blink model for Listening mode for generating blinks and by our proposed method (Type III) with respect to the original video (Type I). The participants' task is to tell the order of the presented real and animated videos. If a participant cannot tell them apart, he has to guess. Hence, if a participant is never able to distinguish between real and animated video, then the percentage of correct answers is close to chance level (50%).

The average score of correctly identifying the order of the real and animated videos is presented in Table 4. Participants correctly identified the order of sequences of Type I and II with 78%. The average score of correctly identifying the order of real and our proposed method is only 54%, which is close to chance level. Hence, while participants were often able to discriminate between the original and reference method, they were usually not able to distinguish between the original and our proposed model.

For both pairs, we propose a hypothesis about the relation between real and synthetic sequences. Our null hypothesis states that the correctly identified orders are chance level, hence $H_0 : \mu = 0.5$. The alternative hypothesis is that it is not chance level $H_1 : \mu \neq 0.5$. Following we set the criterion of significance to 0.05, which equals a 95% confidence interval. Calculating the binomial distribution and computing the p value ($p < 10^{-6}$) between Type I and II indicates that p is located within the area of rejection. Therefore, the null hypothesis is rejected. The computed p value ($p \approx 0.32$) of the binomial distribution of Type I and III is not located within the area of rejection (Table 4). Thus, the null hypothesis is retained.

In Table 5 the correct answers and the duration of each video clip are presented. The number of correct answers of a clip does not increase by an increase of its duration. The variation of the correct answers is due to the small number of sample size. Figure 12 shows the binomial distribution of the subjective tests of Type I and III of the 25 participants. The probabilities of all samples are located inside the area of retaining H_0.

The subjective test was followed by an open interview on the participants' evaluation criteria during the test. In the following the differences in quality are analyzed in more detail.

Our designed model of generating saccades is based on the work of the reference method but improved by taking Listing's Law, head tilts, and eyelid movements into account. In comparison to our model, participants reported that some saccadic eye movements did look artificial in the reference method. Hence, our suggested extensions improve the saccadic eye movements.

The coupling between large saccades and eye blinks as integrated in our model does also improve the animation. Some participants reported that it appears unnatural to perform a large saccade without simultaneously executing a blink as done in the reference method.

Some participants disliked the observed jerky eye movements in the reference method. These movements occur if a large saccade with a short duration in gaze away is performed. We prevent these movements by refining the mutual gaze state. Thereby, we distinguish between small saccades performed in mutual gaze and saccades to shift the gaze from mutual gaze to gaze away and vice versa. These saccades have very different characteristics. For instance,

Table 4 Responses to pair presentations (sample mean $\bar{X}$, standard deviation S, and p-value). Type I: original, Type II: reference method [15], Type III: proposed method

	Type I versus Type II			Type I versus Type III		
	$\bar{X}$	S	p	$\bar{X}$	S	p
Correct answers	0.78	0.14	$< 10^{-6}$	0.54	0.19	≈ 0.32

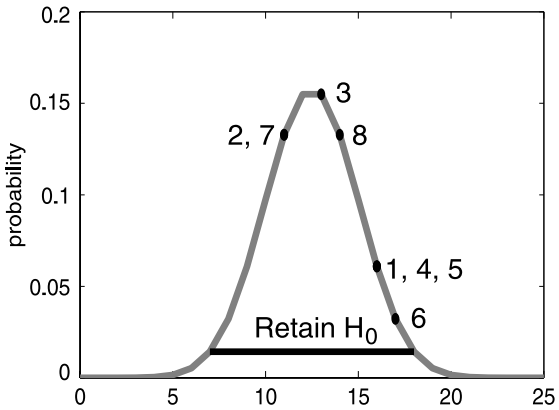


Fig. 12 The binomial distribution of the subjective test with 25 participants of Type I and III. The 8 video samples are marked as *black dots* on the distribution and are located within the area of retaining H_0

Table 5 Responses to each pair presentation. Type I: original, Type III: proposed method

	Type I versus Type III							
Sequence	1	2	3	4	5	6	7	8
Duration [s]	7	19	18	7	7	7	22	21
Correct answers	0.62	0.42	0.52	0.62	0.64	0.68	0.44	0.54

eye movements in mutual gaze have a shorter fixation duration than in gaze away. Participants, however, did not observe that the talking-head is varying his POR within mutual gaze. Hence, this aspect does not seem to improve the quality. On the other hand, if the image resolution is increased, this aspect may be realized.

In general, the timing of executing eye blinks and gaze shifts was preferred in our animations than in the reference method. For instance, some participants expressed that it seems inappropriate to perform a shift from mutual gaze to gaze away while simultaneously emphasizing a word.

The eye blink model was not mentioned by participants during the interview. Therefore, we showed 15 participants the same video sequences as before but with different eye blink patterns. One time these patterns are generated by a simple model (listening mode) and one time by our designed model in talking mode. After presenting both videos, participants were asked to select the preferred eye animation. 13 out of 15 participants (87%) selected the model used in talking mode. Hence, we conclude that the designed eye blink model in talking mode contributes to the overall animation quality, too.

In general, participants did not criticize the video samples created by our model during the interview.

8 Conclusions

In this work we developed a novel image-based eye animation system consisting of a ECU and a rendering engine. The designed ECU, which consists of different models, is based on eye movement physiology and the statistical analysis of recorded human subjects in a two-way conversation. The rendering engine synthesizes the eye animation with the provided control parameters from the eye control unit. The key idea of the designed rendering engine is the combination of 3D eye globes with the eye socket and eyelids, which are image-based modeled. In this way, the position of the pupil can be fully controlled, and specular reflections can be added to the iris in the rendering process.

We distinguish between listening and talking modes as already suggested in the literature. In listening mode two independent models control eye blink and gaze models, since gaze and blink patterns are statistically independent as presented in our work. We focused on talking mode and designed one integrated eye blink and gaze model, because we showed that eye blinks and gaze movements are coupled in talking mode.

Our analysis revealed statistical dependencies between eye blinks and gaze movements with spoken language. While eye blinks mainly occur at word boundaries, gaze shifts occur in thinking mode, e.g., indicated by filling word. On the other hand, if words are emphasized, the speaker usually looks to the interlocutor. This approach allows one to automatically generate appropriate eye animations to arbitrary spoken language.

While being in mutual gaze, humans vary the POR across the face. This observation is included in our work by extending the traditional simple model of mutual gaze and gaze away. In mutual gaze a second model generates small gaze shifts within mutual gaze. Our observation is also important, since the fixations after the execution of a gaze shift varies. After performing a gaze shift from mutual gaze to gaze away or vice versa, the eyes remain for a longer duration in the new position. If small gaze shifts within mutual gaze are executed, the eyes only remain for a short fixation in the new position. The distinction between these two types of gaze shifts allows one to better model human eye movements. That is because jerky eye movements may occur if large saccades with a short duration in mutual gaze are executed.

In our animation system two types of eye movements, saccades and VOR, are executed. Saccades are executed to perform a gaze shift, e.g., from mutual gaze to gaze away. VOR is performed in order to compensate head motion. We improve these eye movements by integrating Listing's law, head tilts, and eyelid movements in our model.

We conducted a subjective test using the original Ref. [15] and our proposed method. As null hypothesis, we stated that participants cannot distinguish between original and animated samples. Whereas this hypothesis was rejected for the reference method, which shows that most participants were able to distinguish between the reference and the original, it was retained for our proposed method. The main improvements are the integrated eye blink and gaze models in talking mode, which is fully controlled by audio features. Furthermore, we extended the mutual gaze state and improved the models of eye movements. Since the correctly identified video samples are close to chance level, the null hypothesis is retained, and participants did not criticize our video samples in the following interview, we conclude that the new eye animation system creates video-realistic eye animations for a talking-head, which has not been achieved before.

The ECU is not limited to image-based systems but may be also used in 3D facial animations. The computational effort of the models of the ECU are negligible, since the models are implemented as FSMs. The rendering engine is capable to render 25 frames per second, so that real-time animations as required in dialog systems can be generated. In the future our proposed system needs to be merged with a mouth animation, so that the system can be commercially used.

References

1. Ostermann, J., Weissenfeld, A.: Talking faces—technologies and applications. In: ICPR'04: Proceedings of the Pattern Recognition, vol. 3, pp. 826–833 (2004)

2. Pighin, F., Hecker, J., Lischinski, D., Szeliski, R., Salesin, D.H.: Synthesizing realistic facial expressions from photographs. Comput. Graph. **32**, 75–84 (1998)
3. Parke, F.I.: Computer generated animation of faces. In: ACM'72: Proceedings of the ACM Annual Conference, pp. 451–457 (1972)
4. Bregler, C., Covell, M., Slaney, M.: Video rewrite: driving visual speech with audio. In: Proc. ACM SIGGRAPH'97, in Computer Graphics Proceedings, Annual Conference Series (1997)
5. Ezzat, T., Geiger, G., Poggio, T.: Trainable videorealistic speech animation. In: Proc. ACM SIGGRAPH, pp. 388–397 (2002)
6. Cosatto, E., Graf, H.P.: Photo-realistic talking heads from image samples. IEEE Trans. Multimedia **2**(3), 152–163 (2000)
7. Cassell, J., Torres, O.: Turn taking vs. discourse structure: how best to model multimodal conversation. In: Wilks, Y. (ed.) Machine Conversations. Kluwer Academic, The Hague (1998)
8. Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., Stone, M.: Animated conversation: rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. Comput. Graph. **28**, 413–420 (1994)
9. Colburn, A., Cohen, M., Drucker, S.: The role of eye gaze in avatar mediated conversational interfaces MSR-TR-2000-81. Microsoft Research (2000)
10. Heylen, D., van Es, I., van Dijk, E.M.A.G., Nijholt, A.: Experimenting with the Gaze of a Conversational Agent, Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems. Kluwer Academic, Dordrecht (2005)
11. Poggi, I., Pelachaud, C., de Rosis, F.: Eye communication in a conversational 3D synthetic agent. AI Commun. **13**(3), 169–182 (2000)
12. Fukayama, A., Ohno, T., Mukawa, N., Sawaki, M., Hagita, N.: Messages embedded in gaze of interface agents—impression management with agent's gaze. In: CHI'02: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 41–48 (2002)
13. Garau, M., Slater, M., Bee, S., Sasse, M.A.: The impact of eye gaze on communication using humanoid avatars. In: CHI'01: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 309–316 (2001)
14. Deng, Zh., Lewis, J.P., Neumann, U.: Automated eye motion using texture synthesis. IEEE Comput. Graph. Appl. **25**, 24–30 (2005)
15. Park Lee, S., Badler, J.B., Badler, N.I.: Eyes alive. In: SIGGRAPH'02: Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques, pp. 637–644 (2002)
16. Freedman, E.G., Sparks, D.L.: Coordination of the eyes and head: movement kinematics. Exp. Brain Res. **131**, 22–32 (2000)
17. Maand, X., Deng, Z.: Natural eye motion synthesis by modeling gaze-head coupling. In: VR'09: Proceedings of the 2009 IEEE Virtual Reality Conference, pp. 143–150 (2009)
18. Cosatto, E.: Sample-based talking-head synthesis. PhD thesis, Signal Processing Lab, Swiss Federal Institute of Technology, Lausanne, Switzerland, 2002
19. Argyle, M., Cook, M.: Gaze and Mutual Gaze. Cambridge University Press, Cambridge (1976)
20. Masuko, S., Hoshino, J.: Generating head-eye movement for virtual actor. Syst. Comput. Jpn. **37**(12), 33–44 (2006)
21. von Cranach, M., Schmid, R., Vogel, M.W.: Über einige Bedingungen des Zusammenhanges von Lidschlag und Blickbewegung. Psychol. Forsch. **33**, 68–78 (1969)
22. Condon, W.S., Ogsten, W.D.: A segmentation of behavior. J. Psych. Res., 221–235 (1967)
23. Kuipers, J.: Quaternions and Rotation Sequences. Princeton University Press, Princeton (1998)
24. Young, S., Evermann, G., Gales, M., Hain, Th., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: HTK Book. Cambridge University Engineering Department, Cambridge (2005)
25. Huang, X., Hon, H.-W.: Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Prentice Hall PTR, New York (2001)
26. Zheng, J., Franco, H., Weng, F., Sankar, A., Bratt, H.: Word-level rate of speech modeling using rate-specific phones and pronunciations. Proc. ICASSP **3**, 1775–1778 (2000)
27. Terken, J.: Fundamental frequency and perceived prominence of accented syllables. J. Acoust. Soc. Am. **95**, 3662–3665 (1994)
28. Kennedy, L., Ellis, D.: Pitch-based emphasis detection for the characterization of meeting recordings. In: IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2003), pp. 243–248 (2003)
29. Arons, B.: Pitch-based emphasis detection for segmenting speech recordings. In: Proc. ICSLP'94, pp. 1931–1934 (1994)
30. LCTechnologies, Eyegaze systems. http://www.eyegaze.com (2007)
31. Kolmogorov, A.N.: Confidence limits for an unknown distribution function. Ann. Math. Stat. **12**, 461–483 (1941)
32. Aitchison, J., Brown, J.A.C.: The Lognormal Distribution. Cambridge University Press, Cambridge (1973)
33. Limpert, E., Stahel, W.A., Abbt, M.: Log-normal distributions across the sciences: keys and clues. BioScience **51**(5), 341–352 (2001)
34. Kendon, A.: Some functions of gaze-direction in social interaction. Acta Psychol. **26**, 22–63 (1967)
35. Barnes, G.R.: Vestibulo-ocular function during coordinated head and eye movements to acquire visual targets. J. Physiol. **287**, 127–147 (1979)
36. Stahl, J.S.: Amplitude of human head movements associated with horizontal saccades. Exp. Brain Res. **126**(1), 41–54 (1999)
37. Donders, F.C.: Beitrag zur Lehre von den Bewegungen des menschlichen Auges. Hollaendis. Beitr. Anat. Physiol. Wiss. **1**, 104–145 (1848)
38. Helmholtz, H.: On the normal movements of the human eye. Arch. Ophthalmol. **IX**, 153–214 (1863)
39. Haslwanter, T.: Mathematics of three-dimensional eye rotations. Vis. Res. **35**, 1727–1739 (1995)
40. Schworm, H.D., Ygge, J., Pansell, T., Lennerstrand, G.: Assessment of ocular counterroll during head tilt using binocular video oculography. Invest. Ophthalmol. Vis. Sci. **43**(3), 662–667 (2002)
41. ITU Telecom: Standardization Sector of ITU, Methodology for the Subjective Assessment of the Quality of Television Pictures. Recommendation ITU-R BT.500-11 (2002)
42. ITU International Telecom: Union, Telecom. sector, Subjective video quality assessment methods for multimedia applications. Recommendation ITU-T P.910 (1999)

Axel Weissenfeld received his Dipl.-Ing. degree in electrical engineering from the Leibniz University of Hanover in 2003. Since then he has been working toward the PhD degree at the Institut für Informationsverarbeitung of the Leibniz University of Hanover. His research interests are image processing, human–machine interfaces, multiple view geometry, and video coding.

Kang Liu was born in 1977. He studied Mechanical and Electrical Engineering at the Institute of Mechatronic Control Engineering, Zhejiang University, P.R. China. He received his Bachelor and Master Degrees from Zhejiang University in 2001 and 2004, respectively. Since March 2004, he has been working toward the PhD degree at the Institut für Informationsverarbeitung, Leibniz Universität Hannover, Germany. Currently, he is an active researcher in the 3DTV and facial animation projects. He has published several research papers. His current research interests are image processing, video coding, facial animation, and computer–human interfaces. He was awarded an ISCA Grant for participating in the ISCA sponsored conference, Interspeech 2008. His excellent research work on facial animation received the Golden Lips Award for Audiovisual Consistency in the first visual speech synthesis challenge: LIPS 2008, 22–26 September 2008, in Brisbane, Australia.

Jörn Ostermann studied Electrical Engineering and Communications Engineering at the University of Hannover and Imperial College London, respectively. He received Dipl.-Ing. and Dr.-Ing. from the University of Hannover in 1988 and 1994, respectively. From 1988 till 1994, he worked as a Research Assistant at the Institut für Theoretische Nachrichtentechnik conducting research in low bit-rate and object-based analysis–synthesis video coding. In 1994 and 1995 he worked in the Visual Communications Research Department at AT&T Bell Labs on video coding. He was a member of Image Processing and Technology Research within AT&T Labs—Research from 1996 to 2003. Since 2003 he is Full Professor and Head of the Institut für Informationsverarbeitung at the Leibniz Universität Hannover, Germany. In 2007, he became head of the Laboratory for Information Technology. From 1993 to 1994, he chaired the European COST 211 sim group coordinating research in low bitrate video coding. Within MPEG-4, he organized the evaluation of video tools to start defining the standard. He chaired the Adhoc Group on Coding of Arbitrarily-shaped Objects in MPEG-4 Video. Since 2008, he is the Chair of the Requirements Group of MPEG (ISO/IEC JTC1 SC29 WG11). Jörn was a scholar of the German National Foundation. In 1998, he received the AT&T Standards Recognition Award and the ISO award. He is a Fellow of the IEEE and member of the IEEE Technical Committee on Multimedia Signal Processing and past chair of the IEEE CAS Visual Signal Processing and Communications (VSPC) Technical Committee. Jörn served as a Distinguished Lecturer of the IEEE CAS Society. He published more than 100 research papers and book chapters. He is coauthor of a graduate level text book on video communications. He holds more than 30 patents. His current research interests are video coding and streaming, 3D modeling, face animation, and computer–human interfaces.