Video-Realistic Image-based Eye Animation System

A. Weissenfeld[1] K. Liu[1] and J. Ostermann[1]

[1]Institut fuer Informationsverarbeitung
Leibniz Universitaet Hannover, Germany

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. The designed eye control unit is based on the statistical analysis of recorded human subjects. We design a new model, which fully automatically couples eye blinks and movements with phonetic as well as prosodic information extracted from spoken language. Subjective tests showed that participants are not able to distinguish between real eye motions and our animations, which has not been achieved before.

1. Introduction

Talking-heads give a face to spoken output [OW04]. Dialog systems, as used in e-commerce, can integrate facial animations with synthesized speech 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.

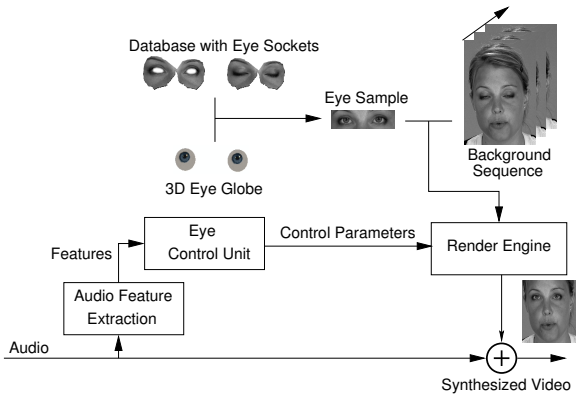

Figure 1: *Image-based eye animation system: Initially phonetic and prosodic information are extracted from the audio. The ECU generates eye blinks and movements and sends control parameters to the rendering engine.*

Eye animation systems (Fig. 1) consist of an eye control unit (ECU) and a rendering engine, which synthesizes eye animations by combining 3D and image-based models [Cos02]. Note that our ECU may steer arbitrary render engines. Optionally eye animation systems have a unit extracting audio features from the spoken output, which are sent to the ECU.

Image-based facial animation systems [BCS97] mainly concentrated on generating smooth mouth animations, however, facial expressions, head and eye movements are mainly neglected. This work focuses on replacing the eye area to generate video-realistic eye animations to spoken output. Most work on eye animations [HvEvDN05, GSBS01] concentrated on measuring the importance of the eyes as a major channel of non-verbal communication or designed very simple control models [DLN05, Cos02]. We regard the work of Lee et al. [LBB02] as a reference method, because they 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. To put it simply saccades are rapid eye movements with a direction and magnitude A repositioning the eye gaze to new locations in the visual environment. The gaze pattern consists of two states, looking at (MG) and away (GA) from the interlocutor. These states as well as the execution of saccades are mod-

eled by measured probability distributions. However, their method has still several shortcomings. Instead of manually adding the mode thinking as in [LBB02], 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. Our designed model will integrate possible statistical dependencies between eye movements and blinks. In order to model eye blinks, we will explore whether eye blinks can be controlled by spoken output.

2. Analysis of Eye Movements and Blinks

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, since we believe that the subjects acclimate after a while. For the analysis, only the frames of the 'talking segments' are labeled with their gaze and blink patterns as well as with audio information, which contains phonetic as well as prosodic features with the following labeling: 'pause' (WB), 'slow speech rate' (SSR), 'word prominence' (WP), 'filling word' (FW), and 'other' (OT).

3. 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.

3.1. Gaze Patterns, Eye Blinks and Spoken Language

In Tab. 1 the experimental conditional probability $p(\mathrm{GS}|o)$ is depicted, which illustrates the distribution of gaze shifts (GS) during an observation o. The three observations WB, SSR and FW, which may indicate that the speaker is in thinking mode, have a high probability $p(\mathrm{GS}|o)$. During WP the speaker usually looks to the interlocutor and therefore $p(\mathrm{GS}|o \neq \mathrm{WP}) \gg p(\mathrm{GS}|o = \mathrm{WP})$. We did not observe a dependency between observations and GA duration.

Condon and Ogston [CO67] 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' (V), 'consonant' (C) and 'word boundary' (WB). After, we calculate the experimental conditional probability $p(o|\mathrm{B})$ that the observation o occurs if a blink B is executed. A large number of blinks are performed at WB, $p(o = \mathrm{WB}|\mathrm{B}) = 0.61$ and 0.52 of subject 1 and 2. Since $p(o = \mathrm{V}|\mathrm{B}) \approx p(o = \mathrm{C}|\mathrm{B})$ the observations V and C are simply labeled as OT.

Table 1: *For two recorded human subjects, the experimental conditional probability $p(GS|o)$ is presented.*

WB	FW	SSR	WP	OT
7.6%	8.7%	4.4%	0.3%	1.3%
26.0%	20.0%	35.8%	4.5%	4.9%

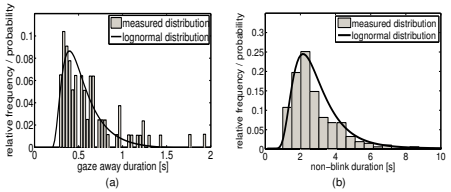

Figure 2: *Relative frequency and fitted lognormal distribution of the duration of (a) gaze away duration ($\bar{X} = 0.89s$, $S = 0.52s$), (b) duration between two consecutive eye blinks ($\bar{X} = 5.29s$, $S = 4.15s$).*

The relative frequency distributions of GA as well as the duration between two consecutive eye blinks are depicted in Fig. 2. These relative frequency distributions are modeled by lognormal distributions. For this, the lognormal distributions are 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.

3.2. Gaze Shifts and Eye Blinks

The analysis of the statistical dependency between gaze shifts and eye blinks indicates that we need to couple saccades and blinks in one control model, since $p(\mathrm{B}|\mathrm{GS})$ is equal to 23.2% and even 64.5% of subject 1 and 2. For this, we determine the experimental conditional probability $p(\mathrm{B}|A)$ of executing a blink B, given the saccadic magnitude A, which can be described by the following function

$$p(\mathrm{B}|A) = \begin{cases} 0.02 & ; \quad 5 \leq A^s < 7.5 \\ 0.09 & ; \quad 7.5 \leq A^s < 12.5 \\ 0.24 & ; \quad 12.5 \leq A^s < 17.5 \end{cases} \tag{1}$$

4. Eye Control Unit

The ECU consists of two models: Blink and gaze model and the models of eye movements. The latter is mainly based on the work of [LBB02], but improved by considering that

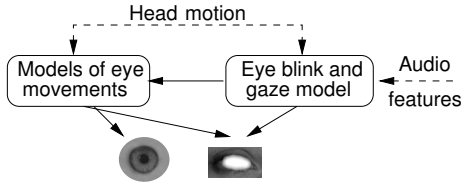

Figure 3: *Overview of the ECU.*

vertical saccades and eyelid movements are coupled. Head motion and audio features are input parameters. Our main focus is on the model controlling the eye movements and blinks (Fig. 3). We propose an algorithm that iteratively determines an animation path, which contains information for eye movements and blinks for each frame of the animation.

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 MG and GA are determined. A gaze shift is executed, if a random number is smaller than $p(\mathrm{GS}|o)$. Since the duration of remaining in GA is independent of o, the duration is determined by modeling the lognormal distribution in Fig. 2a. Thirdly, saccades are generated [LBB02]. Fourthly, eye blinks, which are simultaneously executed with a gaze shift, are added to the animation path by considering the magnitude A of the saccades and Eq. (1). 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. Hence, we determine the conditional probability $p(\mathrm{B}|t_b^{NB},o)$ 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

$$p(\mathrm{B}|t_b^{NB},o) = \frac{p(o|\mathrm{B},t_b^{NB})\cdot p(\mathrm{B}|t_b^{NB})}{p(o|t_b^{NB})}. \qquad (2)$$

Neglecting statistical dependencies between o and t_b^{NB} we can rewrite Eq. (2) as

$$p(\mathrm{B}|t_b^{NB},o) = \frac{p(o|\mathrm{B})\cdot p(\mathrm{B}|t_b^{NB})}{p(o)}. \qquad (3)$$

The conditional probability $p(o|\mathrm{B})$ is already calculated, $p(\mathrm{B}|t_b^{NB})$ is modeled by the lognormal distribution (Fig. 2b), and $p(o)$ can easily be measured from the recorded corpus. In order to generate eye blinks we design a FSM with three states NB_0, NB and B (Fig. 4). 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 $p_{t_b^{NB},o}$. Since we do know the probability $p(\mathrm{B}_{t_b^{NB},o})$ 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 $p_{t_b^{NB},o}$ as

$$p(\mathrm{B}_{t_b^{NB},o}) = p_{t_b^{NB},o}\cdot p(\mathrm{NB}_{t_b^{NB}-1}), \qquad \forall t_b^{NB} > 1 \qquad (4)$$

with

$$p(\mathrm{NB}_{t_b^{NB}-1}) = 1 - \sum_{l=1}^{t_b^{NB}-1} p(\mathrm{B}_l). \qquad (5)$$

Since the probability $p(\mathrm{B}_{t_b^{NB},o})$ is obviously equal to the conditional probability $p(\mathrm{B}|t_b^{NB},o)$, Eq. (3) and (4) can be combined resulting in

$$p_{t_b^{NB},o} = \frac{p(\mathrm{o}|B)\cdot p(\mathrm{B}|t_b^{NB})}{p(o)\cdot p(\mathrm{NB}_{t_b^{NB}-1})} \qquad (6)$$

giving the transition probability of the FSM of performing a blink given t_b^{NB} and o (Fig. 4). 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.

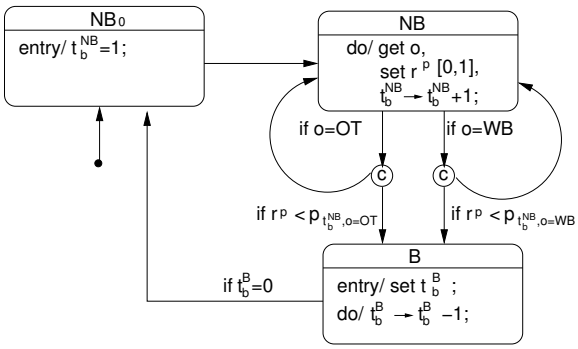

Figure 4: *Statechart: FSM with three states NB_0, NB, and B models eye blinks. The transition probability $p_{t_b^{NB},o}$ from the state NB to B depends on the duration t_b^{NB} and current observation o.*

5. RESULTS

The quality of the synthesized animations of our models is evaluated by a subjective test with 25 participants. The general viewing conditions are set as recommended in [ITU02]. The videos with a resolution of 480x384 are MPEG-1 encoded and displayed with the Windows Media Player. Videos of eye animations with our proposed system can be found on our web site: http://www.tnt.uni-hannover.de/project/facialanimation/demo/index.html. 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 and 22s are pre-

Table 2: *Correct answers to pair presentations (sample mean $\bar{X}$, standard deviation S and p-value). Type I: original, Type II: reference method [LBB02], Type III: proposed method.*

Type I versus Type II			Type I versus Type III		
$\bar{X}$	S	p	$\bar{X}$	S	p
0.78	0.14	$< 10^{-4}$	0.54	0.19	=0.33

pared. These clips are not used for previously training the models of the ECU. Eye animations are generated by using the spoken output and the speakers head movements as input parameters to the eye animation system. Each test session begins with an introduction of the purpose and goals of the experiment and instructions are given to the participants.

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 [LBB02] (Type II), as well as 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. Participants correctly identified the order of sequences of Type I and II with 78% (Tab. 2). Hence, most participants are able to distinguish between real and synthetic sequences. The average score of correctly identifying the order of real and our proposed method is only 54%, which is close to chance level. Note the number of correct answers of a clip does not increase by an increase of its duration.

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 $H_1 : \mu \neq 0.5$. While a t-test indicates that H_0 is rejected for of Type I and II, the null hypothesis is retained for Type I and III.

6. 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 the statistical analysis of recorded human subjects in a two-way conversation. We designed one integrated eye blink and gaze model, because we showed that eye blinks and gaze movements are coupled. Furthermore, 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 to automatically generate appropriate eye animations to arbitrary spoken language.

We conducted a subjective test using the original, reference [LBB02] and our proposed method. The test results show that most participants were able to distinguish between the reference and the original, whereas they were not able to distinguish between original and our proposed method. The new eye animation system creates video-realistic eye animations for a talking-head, which has not been achieved before.

References

[BCS97] BREGLER C., COVELL M., SLANEY M.: Video rewrite: Driving visual speech with audio. Proc. ACM SIGGRAPH 97, in Computer Graphics Proceedings, Annual Conference Series, 1997.

[CO67] CONDON W. S., OGSTEN W. D.: A segmentation of behavior. In *Journal of psychiatric research* (Great Britain, 1967), Pergamon Press Ltd, pp. 221–235.

[Cos02] COSATTO E.: *Sample-Based Talking-Head Synthesis*. Phd. thesis, Signal Processing Lab, Swiss Federal Institute of Techology, Lausanne, Switzerland, 2002.

[DLN05] DENG Z., LEWIS J. P., NEUMANN U.: Automated eye motion using texture synthesis. *IEEE Comput. Graph. Appl. 25*, 2 (2005), 24–30.

[GSBS01] 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* (New York, NY, USA, 2001), ACM Press, pp. 309–316.

[HvEvDN05] HEYLEN D., VAN ES I., VAN DIJK E., NIJHOLT A.: Experimenting with the gaze of a conversational agent. In *Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems*, van Kuppevelt J., Dybkjaer L., Bernsen N., (Eds.). Kluwer Academic Publishers, 2005.

[ITU02] ITU TELECOM. STANDARDIZATION SECTOR OF ITU: *Methodology for the Subjective Assessment of the Quality of Television Pictures, ITU-R BT.500-11*, August 2002.

[LBB02] LEE S. P., BADLER J. B., BADLER N. I.: Eyes alive. In *SIGGRAPH '02: Proceedings of the 29th annual conference on Computer graphics and interactive techniques* (New York, NY, USA, 2002), ACM Press, pp. 637–644.

[OW04] OSTERMANN J., WEISSENFELD A.: Talking faces - technologies and applications. In *ICPR '04: Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3* (2004), pp. 826–833.