Eyes Alive

Sooha Park Lee* Jeremy B. Badler† Norman I. Badler*

University of Pennsylvania * The Smith-Kettlewell Eye Research Institute †

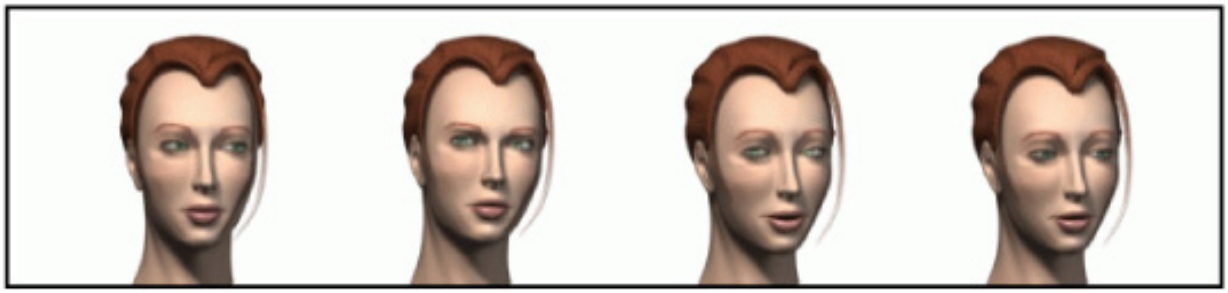


Figure 1: Sample images of an animated face with eye movements

Abstract

For an animated human face model to appear natural it should produce eye movements consistent with human ocular behavior. During face-to-face conversational interactions, eyes exhibit conversational turn-taking and agent thought processes through gaze direction, saccades, and scan patterns. We have implemented an eye movement model based on empirical models of saccades and statistical models of eye-tracking data. Face animations using stationary eyes, eyes with random saccades only, and eyes with statistically derived saccades are compared, to evaluate whether they appear natural and effective while communicating.

Keywords: Eye movement synthesis, facial animation, statistical modeling, saccades, HCI (Human-Computer Interface)

1 Introduction

In describing for artists the role of eyes, [Faigin 1990] illustrated that downcast eyes, upraised eyes, eyes looking sideways, and even out-of-focus eyes are all suggestive of states of mind. Given that eyes are a window into the mind, we propose a new approach for synthesizing the kinematic characteristics of the eye: the spatio-temporal trajectories of saccadic eye movement.

> *... Saccadic eye movements take their name from the French 'saccade', meaning 'jerk', and connoting a discontinuous, stepwise manner of movement as opposed to a fluent, continuous one. The name very appropriately describes the phenomenological aspect of eye movement [Becker 1989].*

We present a statistical eye movement model, which is based on both empirical studies of saccades and acquired eye movement data. There are three strong motivations for our work. First, for animations containing close-up views of the face natural-looking eye movements are desirable. Second, traditionally it is hard for an animator to obtain accurate human eye movement data. Third, the animation community appears to have had no proposals for saccadic eye movement models that are easily adopted for speaking or listening faces.

In recent years, there has been considerable interest in the construction and animation of human facial models. Applications include such diverse areas as advertising, film production, game design, teleconferencing, social agents and avatars, and virtual reality. To build a realistic face model, many factors including modeling of face geometry, simulation of facial muscle behavior, lip synchronization, and texture synthesis have been considered. Several early researchers [Parke 1974; Platt and Badler 1981; Waters 1987; Kalra et al. 1992] were among those who proposed various methods to simulate facial shape and muscle behavior. A number of investigators have recently emphasized building more realistic face models [Lee et al. 1995; Guenter et al. 1998; Pighin et al. 1998; Blanz and Vetter 1999; Brand 1999]. [DeCarlo et al. 1998] suggested automatic methods of building varied geometric models of human faces. [Petajan 1999] and [Essa and Pentland 1995] used motion capture methods to replay prerecorded facial skin motion or behaviors.

Research on faces has not focused on eye movement, although the eyes play an essential role as a major channel of non-verbal communicative behavior. Eyes help to regulate the flow of conversation, signal the search for feedback during an interaction (gazing at the other person to see how she follows), look for information, express emotion (looking downward in case of sadness, embarrassment, or shame), or influence another person's behavior (staring at a person to show power) [Duncan 1974; Pelachaud et al. 1996].

Recently, proper consideration of eye movement is getting more attention among researchers. Cassell and colleagues [Cassell et al. 1994; Cassell et al. 1999; Cassell et al. 2001] in particular have explored eye engagement during social interactions or discourse be-

*{sooha,badler}@graphics.cis.upenn.edu
†jbadler@ski.org

tween virtual agents. They discuss limited rules of eye engagement between animated participants in conversation. [Chopra-Khullar and Badler 1999] generated the appropriate attentional (eye gaze or looking) behavior for virtual characters existing or performing tasks in a changing environment (such as "walk to the lamp post", "monitor the traffic light", "reach for the box", etc. [Colburn et al. 2000] proposed behavioral models of eye gaze patterns for an avatar and investigated gaze behavior to see how observers reacted to whether an avatar was looking at or looking away from them. Vertegaal and colleagues [Vertegaal et al. 2000a; Vertegaal et al. 2000b; Vertegaal et al. 2001] presented experimental results which show that gaze directional cues of users could be used as a means of establishing who is talking to whom, and implemented probablistic eye gaze models for a multi-agent conversational system that uses eye input to determine whom each agent is listening or speaking to. Note that the above research focused on eye gaze patterns rather than how to generate detailed saccadic eye movements.

In this paper, we propose a new approach for synthesizing the trajectory kinematics and statistical distribution of saccadic eye movements. We present an eye movement model which is based on both empirical studies of saccades and statistical models of eye-tracking data.

The overview of our approach is as follows. First, we analyze a sequence of eye-tracking images in order to extract the spatio-temporal trajectory of the eye. Although the eye-tracking data can be directly replayed on a face model, its primary purpose is for deriving a statistical model of the saccades which occur. The eye-tracking video is further segmented and classified into two modes, a talking mode and a listening mode, so that we can construct a saccade model for each. The models reflect the dynamic[1] characteristics of natural eye movement, which include saccade magnitude, direction, duration, velocity, and inter-saccadic interval. Based on the model, we synthesize a face character with more natural looking and believable eye movements.

The remainder of this paper describes our approach in detail. In Section 2, we review pertinent research about saccadic eye movements and the role of gaze in communication. Section 3 presents an overview of our system architecture. Then, in Section 4, we introduce our statistical model based on the analysis of eye-tracking images. An eye saccade model is constructed for both talking and listening modes. Section 5 describes the architecture of our eye movement synthesis system. Subjective test results on the realism of our characters are presented in Section 6. Finally we give our conclusions and closing remarks.

2 Background

2.1 Saccades

Saccades are rapid movements of both eyes from one gaze position to another [Leigh and Zee 1991]. They are the only eye movement that can be readily, consciously, and voluntarily executed by human subjects. Saccades must balance the conflicting demands of speed and accuracy, in order to minimize both time spent in transit and time spent making corrective movements.

There are a few conventions used in the eye movement literature when describing saccades. The **magnitude** (also called the amplitude) of a saccade is the angle through which the eyeball rotates as it changes fixation from one position in the visual environment to another. Saccade **direction** defines the 2D axis of rotation, with zero degrees being to the right. This essentially describes the eye position in polar coordinates. For example, a saccade with magnitude 10 and direction 45 degrees is equivalent to the eyeball rotating 10 degrees in a rightward-upward direction. Saccade **duration** is the amount of time that the movement takes to execute, typically determined using a velocity threshold. The **inter-saccadic interval** is the amount of time which elapses between the termination of one saccade and the beginning of the next one.

The metrics (spatio-temporal characteristics) of saccades have been well studied (for a review, see [Becker 1989]). A normal saccadic movement begins with an extremely high initial acceleration (as much as $30,000\ deg/sec^2$) and terminates with almost as rapid a deceleration. Peak velocities for large saccades can be between 400 and 600 deg/sec. Saccades are accurate to within a few degrees. Saccadic reaction time is between 180 and 220 *msec* on average. Minimum intersaccadic interval ranges from 50 to 100 *msec*.

The duration and velocity of a saccade are functions of its magnitude. For saccades between 5 and 50 *deg*, the duration has a nearly constant rate of increase with magnitude and can be approximated by the linear function:

$$D = D_0 + d * A \tag{1}$$

where D and A are duration and amplitude of the eye movement, respectively. The slope, d, represents the increment in duration per degree. It ranges from $2-2.7\ msec/deg$. The intercept or catch-up time D_0 typically ranges from $20-30\ ms$ [Becker 1989].

Saccadic eye movements are often accompanied by a head rotation in the same direction (gaze saccades). Large gaze shifts always include a head rotation under natural conditions; in fact, naturally occurring saccades rarely have a magnitude greater than 15 degrees [Bahill et al. 1975]. Head and eye movements are synchronous [Bizzi 1972; Warabi 1977].

2.2 Gaze in social interaction

According to psychological studies [Kendon 1967; Duncan 1974; Argyle and Cook 1976], there are three functions of gaze: (1) **sending social signals**—speakers use glances to emphasize words, phrases, or entire utterances while listeners use glances to signal continued attention or interest in a particular point of the speaker, or in the case of an averted gaze, lack of interest or disapproval; (2) **open a channel to receive information**—a speaker will look up at the listener during pauses in speech to judge how their words are being received, and whether the listener wishes them to continue while listeners continually monitor the facial expressions and direction of gaze of the speaker; and (3) **regulate the flow of conversation**—the speaker stops talking and looks at the listener, indicating that the speaker is finished and conversational participants can look at a listener to suggest that the listener be the next to speak.

Aversion of gaze can signal that a person is thinking. For example, someone might look away when asked a question as they compose their response. Gaze is lowered during discussion of cognitively difficult topics. Gaze aversion is also more common while speaking as opposed to listening, especially at the beginning of utterances and when speech is hesitant.

[Kendon 1967] found additional changes in gaze direction, such as the speaker looking away from the listener at the beginning of an utterance and towards the listener at the end. He also compared gaze during two kinds of pauses during speech: phrase boundaries, the pause between two grammatical phrases of speech, and hesitation pauses, delays that occur when the speaker is unsure of what to say next. The level of gaze rises at the beginning of a phrase boundary pause, similar to what occurs at the end of an utterance in order to collect feedback from the listener. Gaze level falls at a hesitation pause, which requires more thinking.

[1] In this paper, 'dynamic' refers to spatio-temporal.

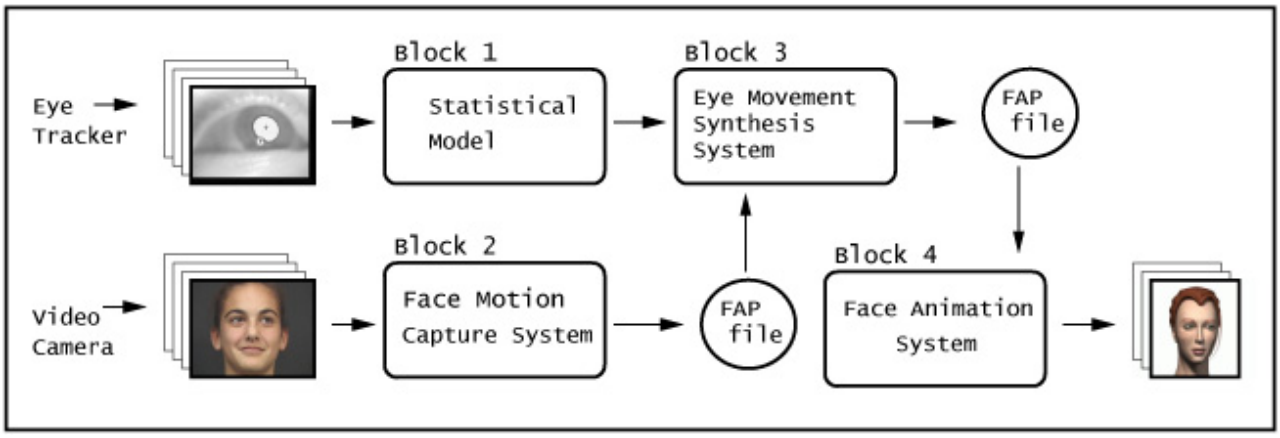


Figure 2: Overall system architecture

3 Overview of system architecture

Figure 2 depicts the overall system architecture and animation procedure. First, the eye-tracking images are analyzed and a statistically based eye movement model is generated using MATLAB™(The MathWorks, Inc.) (Block 1). Meanwhile, for lip movements, eye blinks, and head rotations, we use the alterEGO face motion analysis system (Block 2), which was developed at face2face™, Inc[2]. The alterEGO system analyzes a series of images from a consumer digital video camera and generates a MPEG-4 Face Animation Parameter (FAP) file [Petajan 1999; N3055 1999; N3056 1999]. The FAP file contains the values of lip movements, eye blinking, and head rotation [Petajan 1999]. Our principal contribution , the Eye Movement Synthesis System (EMSS) (Block 3) takes the FAP file from the alterEGO system and adds values for eye movement parameters based on the statistical model. As an output, the EMSS produces a new FAP file that contains eyeball movement as well as the lip and head movement information. We constructed the Facial Animation System (Block 4) by adding eyeball movement capability to face2face's Animator plug-in for 3D Studio Max™(Autodesk, Inc.)

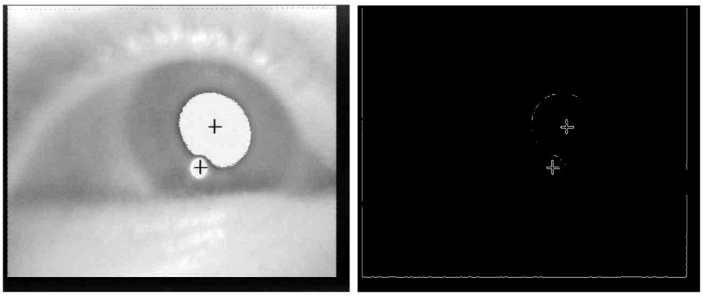
In the next section, we will explain the analysis of the eye-tracking images and the building of the statistical eye model (Block 1). More detail about the EMSS (Block 3) will be presented in Section 5.

4 Analysis of eye tracking data

4.1 Images from the eye tracker

We analyzed sequences of eye-tracking images in order to extract the dynamic characteristics of the eye movements. Eye movements were recorded using a light-weight eye-tracking visor (ISCAN Inc.). The visor is worn like a baseball cap, and consists of a monocle and two miniature cameras. One camera views the visual environment from the perspective of the participant's left eye and the other views a close-up image of the left eye. Only the eye image was recorded to a digital video tape using a DSR-30 digital VCR (Sony Inc.). The ISCAN eye-tracking device measures the eye movement by comparing the corneal reflection of the light source (typically infra-red) relative to the location of the pupil center. The position of the pupil center changes during rotation of the eye, while the corneal reflection acts as a static reference point.

[2]http://www.f2f-inc.com



(a) (b)

Figure 3: (a) Original eye image from the eyetracker (left), (b) Output of Canny Enhancer (right) distribution

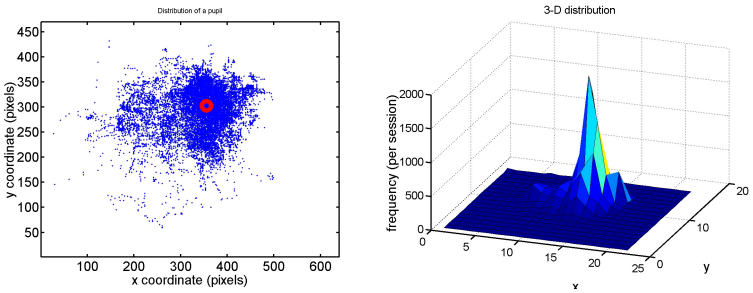
The sample video we used is 9 minutes long and contains informal conversation between two people. The speaker was allowed to move her head freely while the video was taken. It was recorded at the rate of 30 frames per second. From the video clip, each image was extracted using Adobe Premiere™(Adobe Inc.). Figure 3 (a) is an example frame showing two crosses, one for the pupil center and one for the corneal reflection.

We obtained the image (x,y) coordinates of the pupil center by using a pattern matching method. First, the features of each image are extracted by using the Canny operator [Canny 1986] with the default threshold grey level. Figure 3(b) is a strength image output by the Canny enhancer. Second, to determine a pupil center the position histograms along the x and y axes are calculated. Then, the coordinates of the two center points with maximum correlation values are chosen. Finally, the sequences of (x,y) coordinates are smoothed by a median filter.

4.2 Saccade statistics

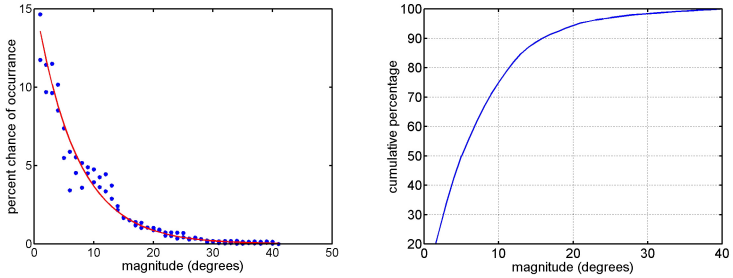
Figure 4(a) shows the distributions of the eye position in image coordinates. The red circle is the primary position (PP), where the speaker's eye is fixated upon the listener. Figure 4(b) is the same distribution plotted in 3 dimensions, with the z-axis representing the frequency of occurrence at that position. The peak in the 3-D plot corresponds to the primary position.

The saccade magnitude is the rotation angle between its starting position S(x_s,y_s) and ending position E(x_e,y_e), which can be com-



(a) (b)

Figure 4: (a) Distribution of pupil centers, (b) 3-D view of same distribution



(a) (b)

Figure 5: (a) Frequency of occurrence of saccade magnitudes, (b) Cumulative percentage of magnitudes

puted by

$$\theta \approx arctan(d/r) = arctan(\frac{\sqrt{(x_e - x_s)^2 + (y_e - y_s)^2}}{r}), \quad (2)$$

where d is the Euclidean distance traversed by the pupil center and r is the radius of the eyeball. The radius r is assumed to be one half of x_{max}, the width of the eye-tracker image (640 pixels).

The frequency of occurrence of a given saccade magnitude during the entire recording session is shown in Figure 5(a). Using a least mean squares criterion the distribution was fitted to the exponential function

$$P = 15.7e^{-\frac{A}{6.9}}, \quad (3)$$

where P is the percent chance to occur and A is the saccade magnitude in degrees. The fitted function is used for choosing a saccade magnitude during synthesis.

Figure 5 (b) shows the cumulative percentage of saccade magnitudes, i.e. the probability that a given saccade will be smaller than magnitude x. Note that 90% of the time the saccade angles are less than 15 degrees, which is consistent with a previous study [Bahill et al. 1975].

Saccade directions are also obtained from the video. For simplicity, the directions are quantized into 8 evenly spaced bins with centers 45 degrees apart. The distribution of saccade directions is shown in Table 1. One interesting observation is that up-down and left-right movements happened more than twice as often as diagonal movements. Also, Up-down movements happened equally as often as left-right movements.

Saccade duration was measured using a velocity threshold of 40 deg/sec (1.33 $deg/frame$). The durations were then used to derive an instantaneous velocity curve for every saccade in the eye-track record. Sample curves are shown in Figure 6 (black dotted lines). The duration of each eye movement is normalized to

Direction	0 deg	45 deg	90 deg	135 deg
Percent(%)	15.54	6.46	17.69	7.44
Direction	180 deg	225 deg	270 deg	315 deg
Percent(%)	16.80	7.89	20.38	7.79

Table 1: Distribution of saccade directions

6 frames. The normalized curves are used to fit a 6-dimensional polynomial (red solid line),

$$Y = 0.13X^6 - 3.16X^5 + 31.5X^4 - 155.87X^3 + 394X^2 - 465.95X + 200.36, \quad (4)$$

where X is frame 1 to 6 and Y is instantaneous velocity ($degrees/frame$).

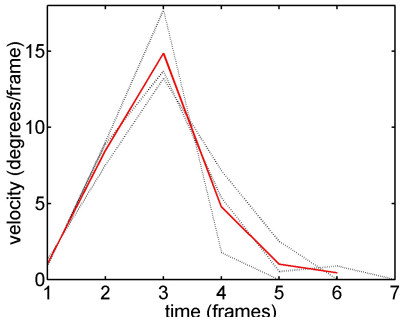


Figure 6: Instantaneous velocity functions of saccades

The inter-saccadic interval is incorporated by defining two classes of gaze, *mutual* and *away*. Mutual gaze indicates that the subject's eye is in the primary position, while gaze away indicates that it is not. The duration that the subject remains in one of these two gaze states is analogous to the inter-saccadic interval. Figures 7(a) and 7(b) plot duration distributions for the two types of gaze while the subject was talking. They show the percent chance of remaining in a particular gaze mode (i.e., not making a saccade) as a function of elapsed time. The polynomial fitting function for mutual gaze duration is

$$Y = 0.0003X^2 - 0.18X + 32, \quad (5)$$

and for gaze away duration is

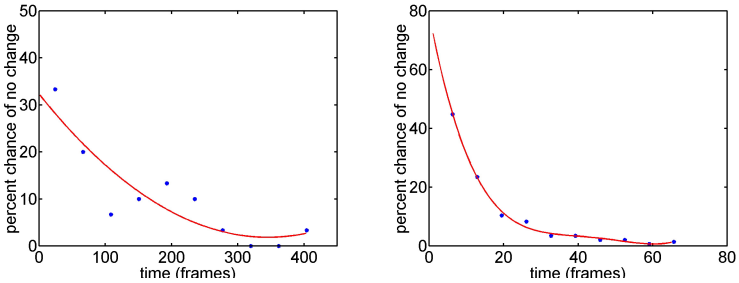
$$Y = -0.0034X^3 + 0.23X^2 - 6.7X + 79 \quad (6)$$

Note that the inter-saccadic interval tends to be much shorter when the eyes are not in the primary position.

4.3 Talking mode vs. Listening mode

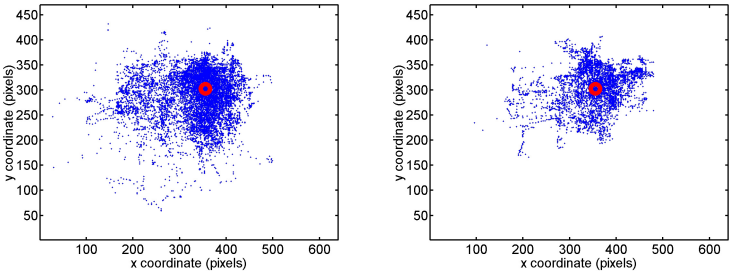
It can be observed that the characteristics of gaze differ depending on whether a subject is talking or listening [Argyle and Cook 1976]. In order to model the statistical properties of saccades in talking and listening modes, the modes are used as a basis to further segment and classify the eye movement data. The segmentation and classification were performed by a human operator inspecting the original eye-tracking video.

Figures 8 (a) and (b) show the eye position distributions for talking mode and listening mode, respectively. In talking mode, 92%



(a) (b)

Figure 7: (a) Frequency of mutual gaze duration while talking, (b) Frequency of gaze away duration while talking



(a) (b)

Figure 8: Distribution of saccadic eye movements (a) in talking mode, (b) in listening mode

of the time saccade magnitude is 25 degrees or less. In listening mode, over 98% of the time the magnitude is less than 25 degrees. The average magnitude is 15.64 ± 11.86 degrees ($mean \pm stdev$) for talking mode and 13.83 ± 8.88 degrees for listening mode. In general the magnitude distribution of listening mode is much narrower than that of talking mode, indicating that when the subject is speaking eye movements are more dynamic and active. This is also apparent while watching the eye-tracking video.

Inter-saccadic intervals also differ between talking and listening modes. In talking mode, the average mutual gaze and gaze away durations are 93.9 ± 94.9 frames and 27.8 ± 24.0 frames, respectively. The complete distributions are shown in figures 7(a) and 7(b). In listening mode, the average durations are 237.5 ± 47.1 frames for mutual gaze and 13.0 ± 7.1 frames for gaze away. These distributions were far more symmetric and could be suitably described with Gaussian functions. The longer mutual gaze times for listening mode are consistent with earlier empirical results [Argyle and Cook 1976] in which the speaker was looking at the listener 41% of the time, while the listener was looking at the speaker 75% of the time.

5 Synthesis of natural eye movement

A detailed block diagram of the statistical eye movement synthesis model is illustrated in Figure 9. The key components of the model are (1) **Attention Monitor (AttMon)**, (2) **Parameter Generator (ParGen)**, and (3) **Saccade Synthesizer (SacSyn)**.

AttMon monitors the system state and other necessary information, such as whether it is in talking or listening mode, whether the direction of the head rotation has changed, or whether the current frame has reached the mutual gaze duration or gaze away duration. By default, the synthesis state starts from the *mutual gaze state*.

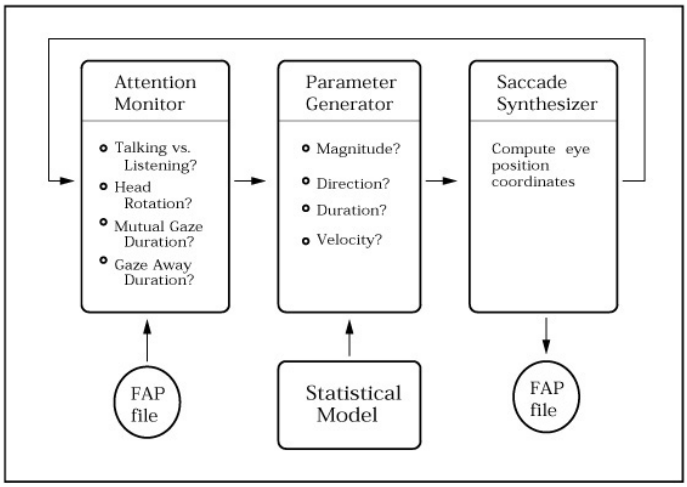


Figure 9: Block diagram of the statistical eye movement model

The agent mode (talking or listening mode) can be provided by a human operator using linguistic information. The head rotation is monitored by the following procedure:

```
1: Initialize start and duration index for head rotation
2: for each frame
3:     Determine direction and amplitude of head rotation
         for current frame by comparing with head rotation
         FAP values of current frame and previous frame
4:     if direction has been changed in this frame
5:       Calculate head rotation duration by searching
           backwards until reaching starting index value
6:       Set starting index to the current frame number
7:       Set duration index to 0
       else
8        Increment duration index
9: end
```

If the direction of head rotation has changed and its amplitude is bigger than an empirically chosen threshold then it invokes **ParGen** to initiate eye movement. Also, if the timer for either mutual gaze or gaze away duration is expired, it invokes **ParGen**.

ParGen determines saccade magnitude, direction, duration and instantaneous velocity. It also decides the gaze away duration or mutual gaze duration depending on the current state. Then, it invokes the **SacSyn**, where appropriate saccade movement is synthesized and coded into the FAP values.

Saccade magnitude is determined using the inverse of the exponential fitting function shown in Figure 5(a). First, a random number between 0 and 15 is generated. The random number corresponds to the y-axis (percentage of frequency) in Figure 5(a). Then, the magnitude can be obtained from the inverse function of Equation 3,

$$A = -6.9 * log(P/15.7) \quad (7)$$

where A is saccade magnitude in degrees and P is the random number generated, i.e., the percentage of occurrence. This inverse mapping using a random number guarantees the saccade magnitude has the same probability distribution as shown in Figure 5(a). Based on the analysis result in section 4.3, the maximum saccade magnitude is limited to 27.5 degrees for talking mode and 22.7 degrees

for listening mode.[3]

Saccade direction is determined by two criteria. If the head rotation is larger than a threshold, the saccade direction follows the head rotation. Otherwise, the direction is determined based on the distribution shown in Table 1. A uniformly distributed random number between 0 to 100 is generated and 8 non-uniform intervals are assigned to the respective directions. That is, a random number between 0 to 15.54 is assigned to the direction 0 deg (right), a number between 15.54 to 22.00 to the direction 45 deg (up-right), and so on. Thus 15.54% of the time a pure rightward saccade will occur, and 6.46% of the time a up-rightward saccade will be generated.

Given a saccade magnitude A, the duration is calculated using Equation 1 with values $d = 2.4\ msec/deg$ and $D_0 = 25\ msec$. The velocity of the saccade is then determined using the fitted instantaneous velocity curve (Equation 4.) Given the saccade duration D in frames, the instantaneous velocity model is resampled at D times the original sample rate (1/6). The resulting velocity follows the shape of the fitted curve with the desired duration D.

In talking mode, the mutual gaze duration and gaze away duration are determined similarly to the other parameters, using inverses of the polynomial fitting functions (equations 5 and 6). Using the random numbers generated for the percentage range, corresponding durations are calculated by root solving the fitting functions. The resulting durations have the same probability distributions. In listening mode, inter-saccadic intervals are obtained using Gaussian random numbers with the duration values given in section 4.3: 237.5 ± 47.1 frames for mutual gaze and 13.0 ± 7.1 frames for gaze away.

The **SacSyn** collects all synthesis parameters obtained above and calculates the sequence of the coordinates of the eye centers. The coordinate values for eye movements are then translated into the FAP values for the MPEG4 standard [N3055 1999; N3056 1999]. For facial animation, we merge the eye movement FAP values with the parameters for lip movement, head movement, and eye blinking provided by the alterEGO system. Each frame is rendered in the 3D StudioMax environment. After synthesizing a saccade movement, the **SacSyn** sets the synthesis state to either *gaze away state* or *mutual gaze state*. Again, the **AttMon** checks the head movement, internal mode of the agent, and the timer for gaze away duration. When a new eye movement has to be synthesized, the **ParGen** is invoked in order to determine the next target position. Depending on the next target position, the state either stays at the *gaze away state* or goes back to the *mutual gaze state*.

We generate facial animation using the face2face Animation Plug-In by applying FAP values to the face model in 3D StudioMax. We added the eye animation capability to the Plug-In. In addition to applying the saccade data from the FAP file, our modified plug-in incorporates the vestibulo-ocular reflex (VOR). The VOR stabilizes gaze during head movements (as long as they are not gaze saccades) by causing the eyes to counter-roll in the opposite direction [Leigh and Zee 1991].

6 Results

In order to compare the proposed saccade model to simpler techniques, we synthesized eye movements on our face model using three different methods. In the first (Type I), the subject does not have any saccadic movements. The eyeballs remain fixated on the camera. In the second (Type II), the eye movement is random. The saccade magnitude, direction and inter-saccadic interval are chosen by random number generators. In the third (Type III), the eye movements are sampled from our estimated distributions. The statistical model reflects the dynamic characteristics of natural eye movements. Also, the model eye movements are synchronized with head movements and speech acts. Figure 1 shows several samples of the output images.

[3]The maximum magnitude thresholds are determined by the average magnitude plus one standard deviation for each mode.

Questions	p-values	
	Type I vs. Type III	Type II vs. Type III
Overall	0.0000	0.0000
Q1	0.1321	0.0588
Q2	0.1127	0.0006
Q3	0.0037	0.0029
Q4	0.0000	0.1310

Table 2: Results of Newman-Keuls test

We conducted a subjective test to evaluate the three types of eye movement. The three characters (Type I, II, III) were presented in random order to 12 subjects. The subjects were asked the following questions:

- Q1: Did the character on the screen appear interested in (5) or indifferent (1) to you?
- Q2: Did the character appear engaged (5) or distracted (1) during the conversation?
- Q3: Did the personality of the character look friendly (5) or not (1)?
- Q4: Did the face of the character look lively (5) or deadpan (1)?
- Q5: In general, how would you describe the character?

Note that higher scores correspond to more positive attributes in a conversational partner. Most of the subjects were naive in the sense that they were not familiar with computer graphics or neurobiology, and none of the subjects were authors of the study. For questions 1 to 4, the score was graded on a scale of 5 to 1.

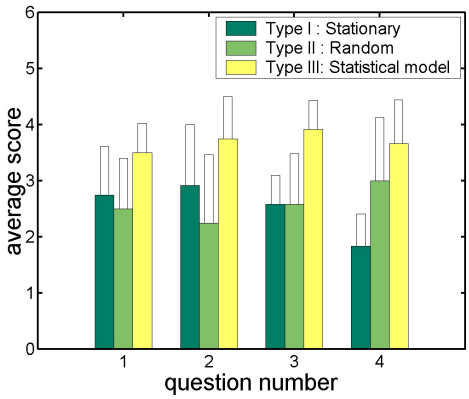


Figure 10: Results of subjective evaluations. Average score and standard deviation

Figure 10 summarizes the average score and standard deviation for the first four questions. The scores were analyzed using the STATISTICA^{TM} software package (StatSoft, Inc.). A Kruskal-Wallis ANOVA indicated that the three character types had significantly different scores (p = 0.0000). To further quantify the differences between the characters, a standard 2-way ANOVA and

Newman-Keuls post-hoc test were performed (Table 2). The interactions between the three models and four questions were tested while the subjects were pooled. Overall, the scores for type III characters were significantly higher than either type I or type II characters, while type I and II characters scored the same (not shown in table; p = 0.7178). The results for individual questions agree well with intuition. Type I characters (staring eyes) were not rated as significantly less interested in (Q1) or engaged with (Q2) the subjects than type III characters (normal eyes). Type II characters (erratic eyes) were not significantly less lively (Q4) than type III characters. They were also not significanly less interested than type III characters, though only marginally. In all other cases type III characters scored significantly higher than the others.

According to the general remarks in Q5, the subjects tended to believe the following:

1. Type I looked interested in the viewers, but it seemed to have a cautious, demanding, sleepy-looking (not lively) and cold personality.

2. Type II's eye movement was unnatural, jittery and distracted, but more lively and friendly. No head-eye synchronization was jarring. Resulting in a character who looked unstable and schizophrenic.

3. Type III had better eye movement, which was purposeful, natural and realistic. The character looked more friendly and outgoing.

[Argyle and Dean 1965] found that very high amounts of eye contact (i.e. direct gaze)may be perceived as too intimate for that particular encounter and hence may be less favorably rated. Our findings using characters with no saccadic eye movement are consistent with those conclusions. In summary, 10 out of 12 subjects chose Type III as the most natural, while two subjects had no preference.

7 Conclusions

In this paper, we presented eye saccade models based on the statistical analysis of eye-tracking video. The eye-tracking video is segmented and classified into two modes: talking mode and listening mode. A saccade model is constructed for each of the two modes. The models reflect the dynamic characteristics of natural eye movement, which include saccade magnitude, duration, velocity, and inter-saccadic interval.

We synthesized a face character using 3 different types of eye movements: stationary, random, and model-based. We conducted a subjective test using these movements. The test results show that the model generated eyeball movement that made the face character look more natural, friendly and outgoing. No eye movement gave the character a lifeless quality, while random eye movement gave the character an unstable quality.

Another way to generate eye movements on a face model is to replay the eye tracking data previously recorded from a subject. Preliminary tests using this method indicated that the replayed eye movements looked natural by themselves, but were often not synchronized with speech or head movement. An additional drawback to this method is that it requires new data to be collected every time a novel eye-track record is desired. Once the distributions for the statistical model are derived, any number of unique eye movement sequences can be animated.

The eye movement video used to construct the saccade statistics was limited to a frame rate of 30 *Hz*, which can lead to aliasing. In practice this is not a significant problem, best illustrated by an example. Consider a small saccade of 2 *degrees*, which will have a duration of around 30 *msec* (Equation 1). To completely lose all information on the dynamics of this saccade, it must begin within three *msec* of the first frame capture, so that it is completely finished by the second frame capture 33 *msec* later. This can be expected to happen around 10 % of the time (3 / 33). From Figure 5 (b), it can be seen that saccades this small comprise about 20 % of all saccades in the record, so only around 2 % of all saccades should be severely aliased. This small percentage has little effect on the instantaneous velocity function of Figure 6. Since saccade starting and ending positions are still recoverable from the video, magnitude and direction are much less susceptible to aliasing problems.

A more important consideration is the handling of the VOR during the eye movement recording. A change in eye position that is due to a saccade (e.g., up and to the left) must be distinguishable from a change that is due to head rotation (e.g., down and to the right). One solution is to include a sensor which monitors head position. When head position is added to eye position, the resultant gaze position is without the effects of the VOR. However, this introduces the new problem that eye and head movements are no longer independent. An alternate approach is to differentiate the eye position data, and threshold the resultant eye velocity (e.g, at 80 deg/sec) to screen out non-saccadic movements. Although this can be performed post-hoc, it is not robust at low sampling rates. For example, revisiting the above example, a 2 *degree* position change that occurred between two frames may have taken 33 *msec* (velocity = 60 deg/sec) or 3 *msec* (velocity = 670 deg/sec). In this study, head movements in subjects occurred infrequently enough that they were unlikely to severely contaminate the saccade data. However, in future work they can be better accounted for, using improved equipment, more elaborate analysis routines, or a combination of the two.

There are a number of enhancements to our system which could be implemented in the future. During the analysis of eye-tracking images, we noticed a high correlation between the eyes and the eyelid movement which could be incorporated. Only the cognitive states of talking and listening were considered. The number of states could be expanded to model gaze patterns during other phases of speech, such as the tendency to look away at the beginning of an utterance, look toward the listener at the end, or to look up when thinking of what to say next. A scan-path model could be added, using not only the tracking of close-up eye images but also the visual environment images taken from the perspective of the participant's eye. Additional subjects could be added to the pool of saccade data, reducing the likelihood of idiosyncracies in the statistical model. Other modeling procedures themselves could be investigated, such as neural networks or Markov models. Improvements such as these will further increase the realism of a conversational agent.

8 Acknowledgment

We would like to thank Eric Petajan, Doug DeCarlo, and Ed Roney for their valuable comments. We thank face2face,inc for providing the face tracking software. A special thanks goes to Minkyu Lee for his endless discussion and support in making this work possible. We greatly acknowledge John Trueswell for the eye tracking data and Andrew Weidenhammer for the face model and the subject data. Finally, gratitude is given to everybody in the University of Pennsylvania Graphics Lab, especially Jan Allbeck, Karen Carter, and Koji Ashida. This research is partially supported by Office of Naval Research K-5-55043/3916-1552793, NSF IIS99-00297, and NSF-STC Cooperative Agreement number SBR-89-20230.

References

ARGYLE, M., AND COOK, M. 1976. *Gaze and Mutual Gaze*. Cambridge University Press, London.

ARGYLE, M., AND DEAN, J. 1965. Eye-contact, distance and affiliation. *Sociometry*, 28, 289–304.

BAHILL, A., ANDLER, D., AND STARK, L. 1975. Most naturally occuring human saccades have magnitudes of 15 deg or less. In *Investigative Ophthalmol.*, 468–469.

BECKER, W. 1989. Metrics, chapter 2. In *The Neurobiology of Saccadic Eye Movements, R H Wurtz and M E Goldberg (eds)*, 13–67.

BEELER, G. W. 1965. *Stochastic processes in the human eye movement control system*. PhD thesis, California Institute of Technology, Pasadena, CA.

BIZZI, E. 1972. Central programming and peripheral feedback during eye-head coordination in monkeys. In *Bibl. Ophthal. 82*, 220–232.

BLANZ, V., AND VETTER, T. 1999. A morphable model for the synthesis of 3D faces. In *Computer Graphics (SIGGRAPH '99 Proceedings)*, 75–84.

BRAND, M. 1999. Voice puppetry. In *Computer Graphics (SIGGRAPH '99 Proceedings*, 21–28.

CANNY, J. 1986. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-8*, 679–698.

CASSELL, J., PELACHAUD, C., BADLER, N., STEEDMAN, M., ACHORN, B., BECHET, T., DOUVILLE, B., PREVOST, S., AND STONE, M. 1994. Animated conversation: Rule-based generation of facial expression gesture and spoken intonation for multiple converstaional agents. In *Computer Graphics (SIGGRAPH '94 Proceedings)*, 413–420.

CASSELL, J., TORRES, O., AND PREVOST, S. 1999. Turn taking vs. discourse structure: How best to model multimodal conversation. In *In Machine Conversations, Y. Wilks (eds)*, 143–154.

CASSELL, J., VILHJALMSSON, H., AND BICKMORE, T. 2001. BEAT:The Behavior Expression Animation Toolkit. In *Computer Graphics (SIGGRAPH '01 Proceedings)*, 477–486.

CHOPRA-KHULLAR, S., AND BADLER, N. 1999. Where to look? automating visual attending behaviors of virtual human characters. In *Autonomous Agents Conf.*

COLBURN, R., COHEN, M., AND DRUCKER, S. 2000. Avatar mediated conversational interfaces. In *Microsoft Technical Report*.

DECARLO, D., METAXAS, D., AND STONE, M. 1998. An anthropometric face model using variational techniques. In *Computer Graphics (SIGGRAPH '98 Proceedings)*, 67–74.

DUNCAN, S. 1974. Some signals and rules for taking speaking turns in conversations. Oxford University Press, New York.

ESSA, I., AND PENTLAND, A. 1995. Facial expression recognition using a dynamic model and motion energy. In *ICCV95*, 360–367.

FAIGIN, G. 1990. The artist's complete guide to facial expression. Watson-Guptill Publications, New York.

GUENTER, B., GRIMM, C., AND WOOD, D. 1998. Making faces. In *Computer Graphics (SIGGRAPH '98 Proceedings)*, 55–66.

ISO/IEC JTC 1/SC 29/WG11 N3055. Text for CD 14496-1 Systems MPEG-4 Manual. 1999.

ISO/IEC JTC 1/SC 29/WG11 N3056. Text for CD 14496-2 Systems MPEG-4 Manual. 1999.

KALRA, P., MANGILI, A., MAGNENAT-THALMANN, N., AND THALMANN, D. 1992. Simulation of muscle actions using rational free form deformations. In *Proceedings Eurographics '92 Computer Graphics Forum*, Vol. 2, No. 3, 59–69.

KENDON, A. 1967. Some functions of gaze direction in social interaction. *Acta Psychologica 32*, 1–25.

LEE, Y., WATERS, K., AND TERZOPOULOS, D. 1995. Realistic modeling for facial animation. In *Computer Graphics (SIGGRAPH '95 Proceedings)*, 55–62.

LEIGH, R., AND ZEE, D. 1991. *The Neurology of Eye Movements*, 2 ed. FA Davis.

PARKE, F. 1974. *Parametrized Models for Human Faces*. PhD thesis, University of Utah.

PELACHAUD, C., BADLER, N., AND STEEDMAN, M. 1996. Generating facial expressions for speech. *Cognitive Science 20*, 1, 1–46.

PETAJAN, E. 1999. Very low bitrate face animation coding in MPEG-4. In *Encyclopedia of Telecommunications, Volume 17*, 209–231.

PIGHIN, F., HECKER, J., LISCHINSKI, D., SZELISKI, R., AND SALESIN, D. 1998. Synthesizing realistic facial expressions from photographs. In *Computer Graphics (SIGGRAPH '98 Proceedings)*, 75–84.

PLATT, S., AND BADLER, N. 1981. Animating facial expressions. In *Computer Graphics (SIGGRAPH '81 Proceedings)*, 279–288.

VERTEGAAL, R., DER VEER, G. V., AND VONS, H. 2000. Effects of gaze on multiparty mediated communication. In *Proceedings of Graphics Interface 2000*, Morgan Kaufmann Publishers, Montreal,Canada: Canadian Human-Computer Communications Society, 95–102.

VERTEGAAL, R., SLAGTER, R., DER VEER, G. V., AND NIJHOLT, A. 2000. Why conversational agents should catch the eye. In *Summary of ACM CHI 2000 Conference on Human Factors in Computing Systems*.

VERTEGAAL, R., SLAGTER, R., DER VEER, G. V., AND NIJHOLT, A. 2001. Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. In *ACM CHI 2001 Conference on Human Factors in Computing Systems*, 301–308.

WARABI, T. 1977. The reaction time of eye-head coordination in man. In *Neurosci. Lett. 6*, 47–51.

WATERS, K. 1987. A muscle model for animating three-dimensional facial expression. In *Computer Graphics (SIGGRAPH '87 Proceedings)*, 17–24.