Natural Eye Motion Synthesis by Modeling Gaze-Head Coupling

Xiaohan Ma* Zhigang Deng†

Computer Graphics and Interactive Media Lab, Department of Computer Science
University of Houston, Houston, TX

ABSTRACT

Due to the intrinsic subtlety and dynamics of eye movements, automated generation of natural and engaging eye motion has been a challenging task for decades. In this paper we present an effective technique to synthesize natural eye gazes given a head motion sequence as input, by statistically modeling the innate coupling between gazes and head movements. We first simultaneously recorded head motions and eye gazes of human subjects, using a novel hybrid data acquisition solution consisting of an optical motion capture system and off-the-shelf video cameras. Then, we statistically learn gaze-head coupling patterns using a dynamic coupled component analysis model. Finally, given a head motion sequence as input, we can synthesize its corresponding natural eye gazes based on the constructed gaze-head coupling model. Through comparative user studies and evaluations, we found that comparing with the state of the art algorithms in eye motion synthesis, our approach is more effective to generate natural gazes correlated with given head motions. We also showed the effectiveness of our approach for gaze simulation in two-party conversations.

Keywords: Gaze-Head Coupling, Eye Motion, Facial Animation, Digital Avatars

Index Terms: I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Animation; H.5.2 [Information Interfaces and Presentation]: User Interfaces—Graphical user interfaces (GUI)

1 INTRODUCTION

Engaging and natural eye gazes play an indispensable role to the realism of computer-generated avatars centered to a broad variety of computer graphics, animation and virtual reality applications. Although various efforts have been attempted to produce realistic eye movements in recent years [21, 12, 24], the automated synthesis of natural eye gazes accompanying head movements is still a challenging task in computer animation and virtual reality research.

In this paper, we take *head motion compensated eye gazes* into consideration and perform in-depth statistical analysis and learning to model the association between gazes and head movements. Based on the statistical modeling, we present an effective approach to synthesize natural eye gazes for novel head motion sequence input. This approach proceeds as follows. First, eye movements and head motions of real humans are recorded using a novel hybrid data acquisition solution that consists of an optical motion capture system and off-the-shelf video cameras. In this hybrid system, the motion capture system is used to capture 3D head movements, and the video cameras are used for recording eye motion video. Then, the Dynamic Coupled Component Analysis (DCCA) model [30] is adapted for processing the joint saccade-head movements. Based on the constructed eye-head DCCA model, we can generate natural eye gazes to comply with novel head motion input. Through comparative user studies, we validated that our approach is more effective to generate natural gazes correlated with given head motions than the state of the art eye motion synthesis algorithms [21, 12]. We also demonstrated the effectiveness of our approach for gaze simulation in two-party conversations. Figure 1 shows the schematic illustration of our approach.

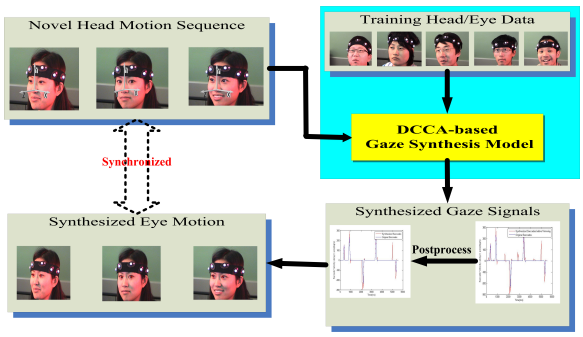

Figure 1: Schematic illustration of our eye motion synthesis approach. First, given the collected eye-head motion data (training set), we build a DCCA-based statistical model. After that, given a novel head motion input, we generate its corresponding natural eye motions.

The major contributions of this work include: (1) our statistical model encodes the intrinsic temporal dynamics and correlation between head movements and eye motions, which is distinguished from the state of the art algorithms in eye motion synthesis [21, 12, 24]. (2) Different from previous approaches that heavily rely on a large number of empirical rules or parameters (*e.g.*, thresholds) [21, 12, 24], most of parameters in our DCCA-based gaze synthesis model are learned automatically and no user-specified threshold is required for our gaze synthesis algorithm; and (3) besides synthesizing the kinematic details of the saccades of a single talking avatar, our approach can be effectively applied for gaze simulation in two-party conversations, which was evaluated and validated through our comparative user studies.

The remainder of this paper is organized as follows. Section 2 briefly reviews recent related work. Section 3 describes how we recorded and processed joint gaze and head movement data. Section 4 briefly describes the Dynamic Coupled Component Analysis (DCCA) model. Section 5 describes how to synthesize natural eye gazes based on the constructed gaze-head coupling model, followed by various results synthesized by our approach and our comparative user studies (Section 6).

2 RELATED WORK

Since an extensive survey on facial animation research is beyond the scope of this work, readers can refer to recent facial animation book [13]. In this section we only briefly review recent work that is closely related to the synthesis of eye motions and head movements.

A significant amount of research interests have been drawn to produce realistic eye motions [18, 31, 32, 8, 28, 21, 11, 33, 19,

*e-mail: xiaohan@cs.uh.edu
†e-mail: zdeng@cs.uh.edu

IEEE Virtual Reality 2009
14-18 March, Lafayette, Louisiana, USA
978-1-4244-3943-0/09/$25.00 ©2009 IEEE

12, 24] and head movements [7, 9, 16, 10, 29, 6, 5]. For example, Chopra-Khullar *et al.* [18] compute eye gazes and head motions of avatars in virtual environments based on high-level scripts provided by users. Vertegaal *et al.* [31, 32] conducted user studies to look into the role of gaze clues in order to answer cognitive questions including "who is talking to whom" in multi-party conversation scenarios. The "Eyes Alive" model proposed by Lee *et al.* [21] analyzes the *textural* aspects of gazes using the first order statistics. Through comparative user studies, they demonstrated that the gazes synthesized by their approach achieve noticeable visual realisms. However, in their approach, only the first-order statistics are employed, and a significant number of empirical parameters are involved. Deng *et al.* [11, 12] present an efficient texture synthesis based approach to simultaneously synthesize realistic eye gaze and blink motion, accounting for any possible correlation between the two. But the success of this data-driven technique largely depends on the quality and quantity of the used eye motion data. Masuko and Hoshino [24] came up a set of empirical rules for the generation of coordinated eye and head motions, and demonstrated its selected applications on talking avatars. Nevertheless, the generality of these proposed empirical rules and how to generate dynamic details of saccades are not addressed in their work.

Saccades are defined as the rapid movements of the eyes from one position to another [22, 21]. It was reported that large gaze shifts typically accompany a head rotation under natural conditions [1, 2, 35, 3]. Specifically, saccadic eye movements are often accompanied by a head rotation in the same direction; and the amplitude of a head movement significantly influences its saccadic peak velocity. However, as pointed out by Freeman *et al.* [14], saccade kinematics, even the subtle eye-head interaction Vestibulo-Ocular Reflex(VOR), are predictable if head movements are taken into account. Hence, an open research question is: *can we automatically synthesize engaging and natural eye movements accompanying with any given head motion?*

To tackle the above research question, we need to address the following two fundamental issues:

1. When will saccades happen when the head is moving?

2. What are its kinematics details including amplitude, velocity variation, and duration, when a saccade happens?

In this work, we pose these two issues as a form of learning the dynamic behaviors of saccade motions and their dependence with head movements. However, choosing sound statistical models for eye motion synthesis is non-trivial. Hidden Markov Models (HMMs) [4], Linear Dynamic Systems (LDS) [23], and Gaussian Process [34] are natural considerations due to their successes in dynamic human motion synthesis. By conducting experiments to test these models for the purpose of gaze synthesis (Fig. 2), we found that it is impossible to directly apply these models for novel gaze synthesis without considerable efforts, due to the following reasons: (1) HMMs provide a powerful framework for modeling temporal dynamic signals, as evident from its successful application for automatic speech recognition, but determining proper configurations of HMMs for eye gaze synthesis is demanding due to the intrinsic characteristics of eye movements. Furthermore, it is difficult to model the accurate dynamical behaviors of saccade motions, *e.g.*, velocity variation, using discrete hidden state sequences of HMMs. (2) Linear Dynamic Systems (LDS) and its variations [25, 23] have been extensively used for describing temporal variations of motion signals, but using LDS for modeling sparse stepwise motion sequences (*i.e.*, saccade signals in this case) has not been reported and validated, to the best of our knowledge. (3) Theoretically, we may be able to learn the mappings of head motions and saccades using various regression/mapping models including Neural Networks and Gaussian Process model. However, it is often impractical to train a well-behaviored mapping model for head motions and saccades due to its demanding requirement on the volume of training data. Figure 2 shows experiment results using the above three statistical models (HMMs, SLDS, and Gaussian Process) for gaze synthesis. As shown in this figure, gaze signals by these three models do not capture the intrinsic characteristics of natural human gazes.

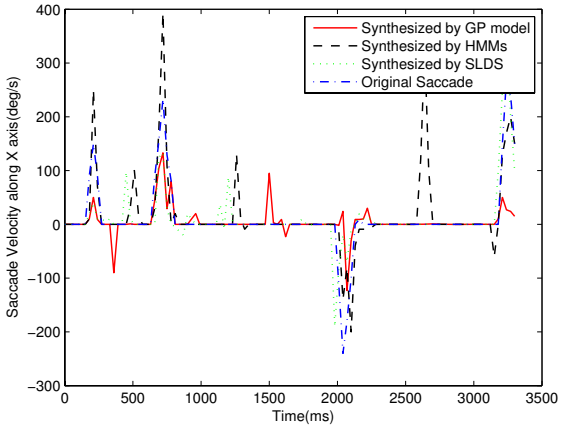

Figure 2: Experiment results by using HMMs, Switched Linear Dynamic Systems (SLDS), and Gaussian Process (GP) for gaze synthesis.

3 Data Acquisition and Processing

We introduce a novel hybrid data acquisition solution consisting of an optical motion capture system and off-the-shelf video cameras to capture simultaneous head motions and eye movements. In the capture setup, a human subject with six markers on his/her head sits within the volume of the optical motion capture system, and the video cameras are specifically set up to record the close-view of his/her eye movements. When the captured subject is involved with natural conversations, the movements of his/her eyes and head are recorded simultaneously using this hybrid scheme. During the capture, an auxiliary person (not for capture) was asked to naturally communicate with the captured subject. Figure 3 shows the snapshot of our hybrid data acquisition solution.

A total of six human subjects participated in our data acquisition. The duration of each subject's data capture was average about two and half minutes. The total amount of gaze and head motion data that were eventually used in this work is 960 seconds. In this work, we used 75% (720 seconds) for training statistical models while retaining the remaining 25% (240 seconds) for the test and validation purpose. It should be noted that the sampling frequency of the optical motion capture system is 120 Hz, therefore, we down-sampled the 3D head marker data (120 frames/second) to 30 frames/second to match with the frame rate of the eye video recorded by the off-the-shelf cameras.

After 3D motions of six head markers were recorded, we extracted head rotation matrices as follows: construct a local orthogonal 3D coordinate system for each motion capture frame based on four chosen head markers (out of the six head markers), and then calculate rotation matrices between these 3D coordinate systems. After that, we converted head rotation matrices to the Euler angle representation, *i.e.*, each head motion frame is represented as three Euler angles (H_x, H_y, H_z) along X, Y and Z directions, respectively.

To extract gaze signals from the recorded eye video, we manually picked the pupil center and then calculated its relative position (x, y), frame by frame. The range of x and y is between 0 and 1, where (0, 0) represents the left-bottom corner of the eyes and (1,1)

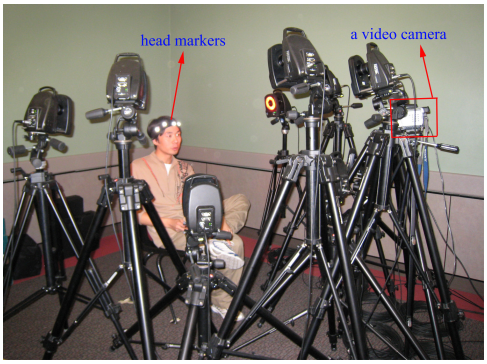

Figure 3: A snapshot of our hybrid data acquisition system used in this work. It consists of an optical motion capture system and off-the-shelf video cameras.

represents the right-top corner of the eyes. It is noteworthy that compared with the data acquired by a special-purpose eye-tracking device [21], the manually estimated gazes are not completely accurate, but it qualitatively captures the characteristics and cadence of gaze movements of real humans, and the obtained gaze durations are frame-accurate. Furthermore, manually estimated gazes had been successfully used by other researchers [12]. For various techniques for identifying gazes from video, please refer to the recent survey by Salvucci and Goldberg [26].

Gaze and Head Velocity Feature Computation: We further transformed the extracted 2D gaze signals and three Euler angles (of head motion) to a velocity-based representation that is used in the follow-up statistical modeling of gaze-head movements. Intrigued by the "Eyes Alive" model [21], we also assume the maximum saccade magnitude is 30 degree from left to right (X direction) and 15 degree from top to bottom (Y direction). As such, the velocity of saccadic motion (VE_x, VE_y) can be computed as follows:

$$VE_x = f \times \Delta x \times \frac{30}{eye\_width} \tag{1}$$

$$VE_y = f \times \Delta y \times \frac{15}{eye\_height} \tag{2}$$

Here f (= 30) is the frame rate per second in the eye video. $eye\_width = 1$ and $eye\_height = 1$ according to the range of x and y. Analogously, we applied a differential operator to three Euler angles of head motion to compute its velocity features, $(\Delta H_x, \Delta H_y, \Delta H_z)$.

4 Dynamic Coupled Component Analysis

Torre and Black [30] proposed a Dynamic Coupled Component Analysis (DCCA) model that learns dependencies between two different data sets by coupling them into a hidden parameter space. Motivated by this work, we develop a DCCA-based, head-dependent gaze synthesis model (detailed in Section 5). Here we briefly describe the concept of the DCCA model.

Assuming $\mathbf{D} \in \mathbf{R}^{d_1 \times n}$ and $\hat{\mathbf{D}} \in \mathbf{R}^{d_2 \times n}$ are two multi-dimensional observations (*e.g.*, head and eye motion sequences in this work), the DCCA model first finds the linear transformation $\mathbf{B} \in \mathbf{R}^{d_1 \times d_2}$ that bridges $\mathbf{D}$ with $\hat{\mathbf{D}}$ in a latent parameter space, by optimizing the following energy function where $\mathbf{d}_i$ and $\hat{\mathbf{d}}_i$ are ith column vectors in matrix $\mathbf{D}$ and $\hat{\mathbf{D}}$ respectively.

$$E_{couple}(\mathbf{B}) = \sum_{i=1}^{n} ||\hat{\mathbf{d}}_i - \mathbf{B}^T \mathbf{d}_i||_2^2 \tag{3}$$

To give a fast convergence and allow a generalization to the dynamic extension, a hidden coefficient matrix $\mathbf{C}$ is introduced to extend Eq. 3 as follows.

$$E_{couple}(\mathbf{B}, \hat{\mathbf{B}}, \mathbf{C}) = \sum_{i=1}^{n} ||\hat{\mathbf{d}}_i - \hat{\mathbf{B}}\, \mathbf{c}_i||_2^2 + \lambda \sum_{i=1}^{n} ||\mathbf{c}_i - \mathbf{B}^T \mathbf{d}_i||_2^2 \tag{4}$$

Here $\mathbf{c}_i$ indicates the ith column vector in matrix $\mathbf{C}$, $\hat{\mathbf{B}} \in \mathbf{R}^{d_2 \times d_2}$ gives the transformation from $\mathbf{C}$ to $\hat{\mathbf{D}}$, and the constant, λ, weights the importances of different terms. Eq. 4 is eventually derived to the following DCCA form [30]:

$$E_{dynamic} = E_{couple} + \lambda_2 \sum_{i=1}^{n-1} ||\mathbf{c}_i - \mathbf{A}\, \mathbf{c}_{i-1}||_{\overline{W}_i} \tag{5}$$

Here $\mathbf{A}$ is the dynamic transition matrix to be estimated, and $||\mathbf{x}||_{\overline{\mathbf{W}}}$ denotes the weighted L_2 norm of a vector $\mathbf{x}$, $\mathbf{x}^T\, \overline{\mathbf{W}}\, \mathbf{x}$. Given the training data $\hat{\mathbf{D}}$ and $\mathbf{D}$, the DCCA model [30] can iteratively estimate $\hat{\mathbf{B}}$, $\mathbf{B}$, and $\mathbf{A}$.

5 Head-Dependent Gaze Synthesis Model

The main idea of this work is to learn a dynamic DCCA model from the training head and eye motion data, and then synthesize novel eye gazes based on the constructed model. As described in Section 3, we obtained saccade feature vectors (VE_x, VE_y) (Eq. 1-2). Meanwhile, we formed a head feature vector by concatenating the $(\Delta H_x, \Delta H_y, \Delta H_z)$ of two continuous frames. We will explain why a head motion feature vector is formed by concatenating two continuous frames in this work shortly.

Given the training set of head feature vectors $\mathbf{d} \in \mathbf{R}^{6 \times 1}$, and the training set of saccade feature vector $\hat{\mathbf{d}} \in \mathbf{R}^{2 \times 1}$, we compute a DCCA model by minimizing Eq. 5 in Section 4, and estimate $\mathbf{B}$, $\hat{\mathbf{B}}$, and $\mathbf{A}$. In our implementation, both $\mathbf{B}$ and $\hat{\mathbf{B}}$ are initialized as an identity matrix, $\lambda = 0.5$, and $\lambda_2 = 0.8$. The weight matrix $\overline{\mathbf{W}}$ is initialized as a matrix where all the entries on diagonal places are 1, *i.e.*, each feature element has the same weight.

Since $\mathbf{B}$, $\hat{\mathbf{B}}$, and $\mathbf{A}$ are obtained from the constructed DCCA model, now we describe how we synthesize corresponding saccades $\mathbf{S}$ given a new head motion sequence $\mathbf{H}$. We first compute the hidden variable $\mathbf{C}$ as follows:

$$\mathbf{C} = \mathbf{B}^T\, \mathbf{H} \tag{6}$$

Then, the transition matrix $\mathbf{A}$ is used to further smooth the hidden variable $\mathbf{C}$ using the following Eq. 7.

$$c_i = \kappa_1\, (\mathbf{B}^T\, \mathbf{h}_i) + \kappa_2\, (\mathbf{A}\, \mathbf{c}_{i-1}) \tag{7}$$

In the above Eq. 7, we empirically set two weighting coefficients κ_1 and κ_2: $\kappa_1 = 0.7$, and $\kappa_2 = 0.3$. Finally, new saccade *feature vectors* $\mathbf{S}$ are reconstructed using Eq. 8.

$$\mathbf{S} = \hat{\mathbf{B}}\, \mathbf{C} \tag{8}$$

Figure 4 and 5 show synthesized eye gaze signals when the retained test head motion sequences are inputted to our DCCA-based gaze synthesis algorithm. It should be noted that in these two figures, we also show synthesized gaze signals (green curves) if a head feature vector is just the velocity feature, $(\Delta H_x, \Delta H_y, \Delta H_z)$, of a single head motion frame. As shown in Figs 4 and 5, when a head feature vector is formed by concatenating the velocity features of two continuous head motion frames, the synthesized gaze signals are closer to original gaze signals. A legitimate explanation is that the original DCCA model [30] focuses on the modeling of dynamic behaviors in dataset $\hat{\mathbf{D}}$ (*i.e*, eye motion in this work), while concatenating the velocities of two continuous frames of $\mathbf{D}$ (*i.e*, head

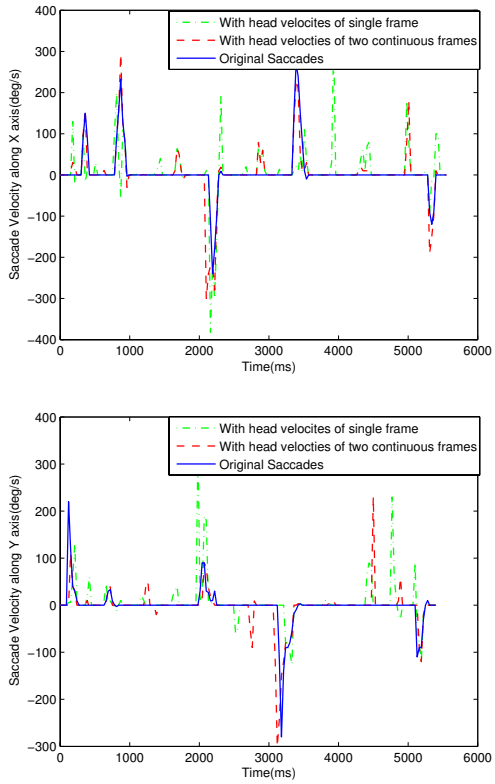

Figure 4: A comparison between the original gaze signals and the gaze signals synthesized by our approach (before a trimming operation is applied). The top panel shows the comparison of gaze velocities in X direction, and the bottom panel shows the comparison of gaze velocities in Y direction.

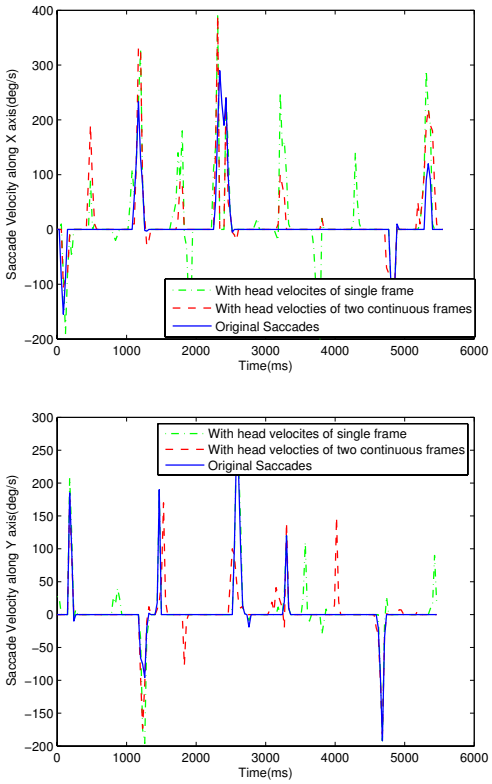

Figure 5: The second comparison (example) between the original gaze signals and the gaze signals synthesized by our approach (before a trimming operation is applied). The top panel shows the comparison of gaze velocities in X direction, and the bottom panel shows the comparison of gaze velocities in Y direction.

motion in this work) together essentially encloses its dynamic information into the constructed DCCA-based gaze synthesis model as well.

After we compute the saccade *feature vectors* for given head motion input, we can generate corresponding gaze motion trajectory straightforwardly. However, the DCCA-based gaze synthesis model may lead to some small saccadic movements whose peak velocities are less than 100 deg/sec in X direction and 80 deg/sec in Y direction (refer to red curves in Figs 4 and 5). These small gaze variations seldom happen in the eye movements of real humans, as evident from the frequency distribution of the peak velocities of saccades in our data set (Figure 8). Hence, we postprocess the synthesized gaze velocities by trimming these small gaze variations as noises. Figures 6 and 7 show the trimmed results for Figs 4 and 5, respectively. As clearly shown in Figs 6 and 7, the trimmed gaze signals are measurably close to the groundtruths (the retained test gaze signals).

6 Results and Evaluation

We generated eye and head animation clips using four different methods (the original recorded motion data, the "Eyes Alive" model [21], the texture synthesis-based method [12], and our approach), and then conducted a comparative user study on these animation clips. The first method (Method I) is to simply transfer the original recorded eye and head motions onto the chosen 3D face models. The second method (Method II) is to combine the "Eyes Alive" model [21] with given head motion sequences. The third method (Method III) is to combine the texture synthesis-based eye motion synthesis approach [12] with the given head motion sequences. The fourth method (Method IV) is to combine eye gazes synthesized by our approach with given head motion sequences. The same realistic 3D face models were used for producing all the animation clips used in this user study. It should be noted that the eye/head motion generation approach proposed by Masuko and Hoshino [24] was not chosen to this comparative study due to the difficulty of performing sound comparisons, because it is a descriptive and rule-based approach (not data-driven). Figure 9 shows some frames of the synthesized eye and head animation clips by the four different methods.

A total of 40 animation clips were produced and used for this user study: 24 of them contain a single avatar (*i.e.*, based on 6 chosen head motion sequences as input, each method generated corresponding 6 eye and head animation clips. In this paper, we also call these inputted head motion sequences as *samples*), and the other 16 clips contain two-party conversations (each method generated 4 animation clips). For the two-party conversation animation clips, we directly applied our statistical gaze synthesis model to the two involved avatars separately. The duration of each animation clip is about 10-20 seconds. Fig. 9 shows some animation frames generated by the four different methods.

We employed the "paired comparison" evaluation scheme proposed by Ledda *et al.* [20] for our user study. Its basic idea is that instead of explicitly rating visual stimuli, participants are asked to select the perceptually better one between two visual stimuli (a pair). As such, the participants can avoid to make forced, inaccurate perception decisions, *e.g.*, assign a subjective and quantitative

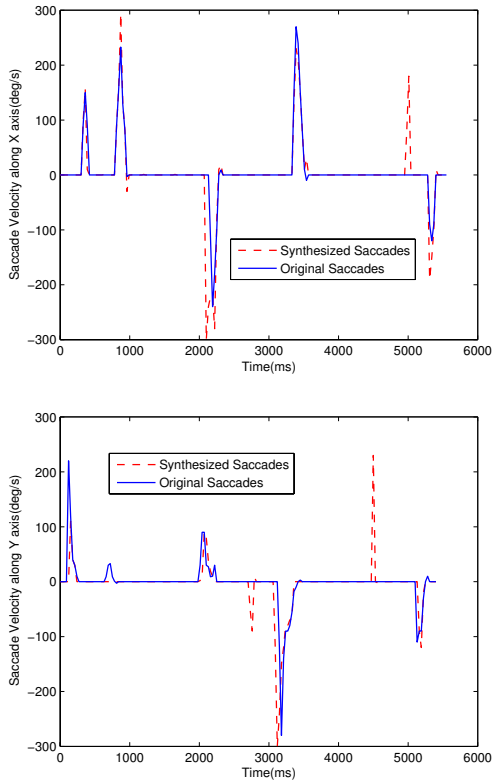

Figure 6: Comparisons between the original gaze signals and the synthesized gaze signals after a trimming operation is applied. The top panel shows the comparison of saccade velocities in X direction, and the bottom panel shows the comparison of saccade velocities in Y direction.

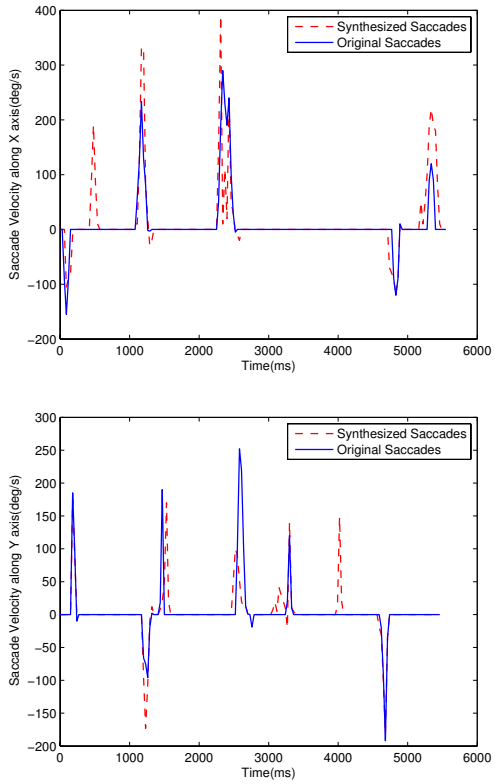

Figure 7: The second comparison between the original gaze signals and the synthesized gaze signals after a trimming operation is applied. The top panel shows the comparison of saccade velocities in X direction, and the bottom panel shows the comparison of saccade velocities in Y direction.

rating to each stimulus. On the contrary, selecting "the perceptually better one" between two visual stimuli (a pair) is qualitative and easier for the participants, which increases the accuracy and robustness of the experiment outcomes.

In this study, based on the above 40 animation clips, we produced a total of 60 (=$6 \times C_2^4 + 4 \times C_2^4$) side-by-side comparison pairs, and each comparison pair consists of two animation clips generated by two different methods with the same head motion input. A total of 22 subjects in a university participated in this user study. Most of these participants are graduate students majoring in a variety of fields, *e.g.*, Computer Science, Electrical Engineering, Linguistics, etc. After the participants viewed each comparison pair for a maximum of four times, they were asked to select "the better one" from the two animation clips in the pair, by answering the following questions (Q1 for single-avatar pairs and Q2 for two-party conversation pairs).

- **Q1-1**. Overall behavior (abbreviated as **Question O**): Which avatar is more perceptually believable in terms of the overall coordination of eye and head motions?
- **Q1-2**. Eye motion reality (abbreviated as **Question E**): Which avatar is more realistic based on its overall eye movements?
- **Q1-3**. Participant involvement (abbreviated as **Question P**): Which avatar appears to let you feel engaged if you're chatting with him/her?
- **Q2-1**. Overall behavior (abbreviated as **Question O**): Which conversation is more perceptually believable in terms of the overall coordination of eye and head motions?
- **Q2-2**. Eye motion reality (abbreviated as **Question E**): Which conversation is more realistic based on its eye activities?
- **Q2-3**. Participant involvement (abbreviated as **Question P**): Which conversation seems more natural and engaged?

Based on the participants' votings, for each sample (a given head motion input), we computed a *preference matrix* based on the voting results of the twenty-two participants. Table 1 shows the preference matrix for Sample #10 (single avatar), as an example. The number in each cell denotes the frequency count of a specific synthesis method chosen as the perceptually better one in terms of one of the asked research questions. For instance, "18" in the 4th cell of the 2nd row indicates that the method I (transferring original motion data) was voted a total of 18 times better than the method II (the "Eyes Alive" synthesis method) in terms of the Question E (*i.e.*, Q1-1 - overall behavior).

Prior to performing vote comparisons for the four different methods, we carried out two statistical tests: (1) *consistency test* to see for each participant whether there was any intransitive vote [20], and (2) *agreement test* to see whether the participants voted all the pairs in a similar way. In this work, we used the consistency test method proposed by Kendall and Smith [17] to compute the Coefficient Of Consistency (COC) for each participant. For the sake of a clear explanation, we take #10 sample (shown in Table 1) as an example, its computed COC for Q1-1 (overall behavior) is $\zeta = 0.643$,

	Question	ζ	u	χ^2	p value, 6 $d.f.$	Method I	Method II	Method III	Method IV
Sample 1	**O**	0.611	0.231	17.553	<0.01	46	14	28	44
	E	0.724	0.019	4.197	<0.01	46	29	31	26
	P	0.649	0.176	14.088	<0.01	39	25	32	36
Sample 2	**O**	0.473	0.579	39.477	<0.05	47	31	11	43
	E	0.581	0.353	25.239	<0.05	43	23	20	46
	P	0.757	0.099	9.237	<0.01	50	16	20	46
Sample 3	**O**	0.576	0.138	11.694	<0.05	41	19	29	43
	E	0.408	0.288	17.364	<0.05	51	30	26	25
	P	0.512	0.197	15.411	<0.05	49	28	18	37
Sample 5	**O**	0.713	0.572	39.036	<0.01	42	26	24	40
	E	0.486	0.273	20.199	<0.05	48	23	25	36
	P	0.877	0.199	15.537	<0.001	39	41	17	35
Sample 6	**O**	0.734	0.353	25.239	<0.01	41	26	22	43
	E	0.819	0.050	6.15	<0.001	46	28	18	40
	P	0.641	0.016	4.008	<0.01	66	24	13	29
Sample 10	**O**	0.643	0.241	18.183	<0.01	50	30	13	39
	E	0.468	0.150	12.45	<0.05	49	25	24	34
	P	0.796	0.051	6.213	<0.01	43	24	29	36

Table 2: Comparisons of consistency and agreement test statistics for single avatar samples (total six). The numbers shown in the right part of the table denote the total votes that a specific method won.

	Question	ζ	u	χ^2	p value, 6 $d.f.$	Method I	Method II	Method III	Method IV
Sample 4	**O**	0.665	0.375	26.625	<0.01	44	19	28	41
	E	0.462	0.190	14.97	<0.05	39	30	20	43
	P	0.562	0.020	4.26	<0.05	46	20	26	40
Sample 7	**O**	0.709	0.304	22.152	<0.01	46	22	21	43
	E	0.695	0.059	6.717	<0.01	49	21	26	36
	P	0.713	0.159	13.017	<0.01	38	33	25	36
Sample 8	**O**	0.720	0.378	26.814	<0.01	49	39	18	26
	E	0.622	0.287	21.081	<0.01	41	41	24	26
	P	0.538	0.510	35.13	<0.05	35	33	28	36
Sample 9	**O**	0.559	0.431	30.153	<0.05	49	25	22	36
	E	0.615	0.011	3.693	<0.01	38	22	32	40
	P	0.826	0.309	22.467	<0.01	52	19	25	36

Table 3: Comparisons of consistency and agreement test statistics for two-party avatar conversation samples (total four). The numbers shown in the right part of the table denote the total votes that a specific method won.

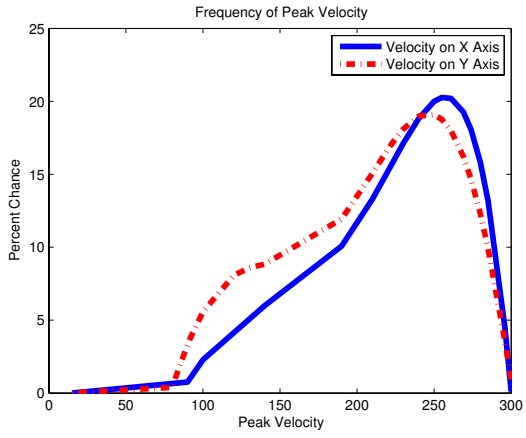

Figure 8: The frequency distribution of the peak velocities of saccades in our training data set. As shown in this figure, the peak velocities of saccades are usually larger than 100 deg/sec in X direction and 80 deg/sec in Y direction.

	Question	I	II	III	IV	Total
I	**O**	-	**18**	20	12	50
	E	-	17	16	16	49
	P	-	15	17	11	43
II	**O**	4	-	15	11	30
	E	5	-	11	9	25
	P	7	-	10	7	24
III	**O**	2	7	-	4	13
	E	6	11	-	7	24
	P	5	12	-	12	29
IV	**O**	10	11	18	-	39
	E	6	13	15	-	34
	P	11	15	10	-	36

Table 1: The computed preference matrix for sample #10. Here I, II, III and IV denote four different synthesis methods: transferring original motion data (Method I), "Eyes alive" model (Method II), texture synthesis-based approach (Method III), and our approach (Method IV).

Figure 9: Some frames of the synthesized eye and head animation clips with a single avatar. From top to bottom: Method I, Method II, Method III, and Method IV.

which means its corresponding p-value is 0.05 given DOF = 6; as such, this consistency test result is considered statistically significant. The consistency test results of all the samples and questions are shown in Tables 2 (single avatar case) and 3 (two-party conversation).

After the voting consistency of each participant was computed and ensured, we checked the overall agreement among all the participants by calculating the Coefficient Of Agreement (COA) [17], and then further used the Chi-Square test statistics (χ^2) to compute the statistical significance of the COA [27]. Here we take #10 sample as an example again, its COA for Q1-1 (overall behavior) is $u = 0.241$, its χ^2 is 18.183, and its corresponding p-value is 0.001 (1% χ^2 is 16.27) given DOF = 6. For the other questions and samples, we obtained similar statistics results (refer to Tables 2 and 3). Thus, our Chi-Square test statistics results indicate that for the used research questions and samples there is a statistically strong agreement among the participants.

Using the above consistency and agreement tests, we validated that the voting results of our paired comparisons were statistically valid and significant. Then, we compared the mean values obtained by the four different methods. Figures 10 and 11 compare and plot the mean values and standard deviations of the four different methods. As shown in these two figures, the mean values of our method is measurably higher than those of the "Eyes Alive" approach [21] and texture synthesis-based approach [12].

We also used a logistic regression analysis technique [15] to check and ensure the significance of the above mean value comparison. Tables 4 (single avatar cases) and 5 (two-party conversations) show the increase of the deviances of Chi-Square test statistics if one variable (synthesis methods, the used face models, or the duration of the animation clips) is not considered. The duration (of the used animation clips) variable only has two discrete values (categories): > 15 seconds and ≤ 15 seconds. As clearly shown in these two tables, the "synthesis method" variable is the most significant one among the three variables to account for the statistical difference of experiment outcomes.

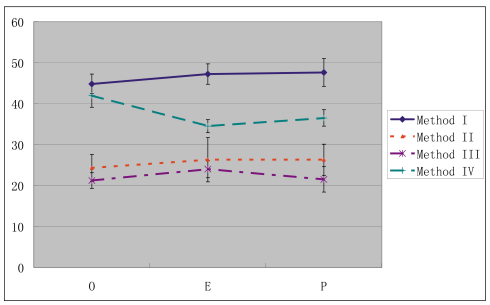

Figure 10: Plotting of the mean value and standard deviation comparisons of four different methods (for the single avatar case). The X axis represents different research questions, and the Y axis represents the average (mean) values of the four different methods. Vertical bars represent the standard deviations.

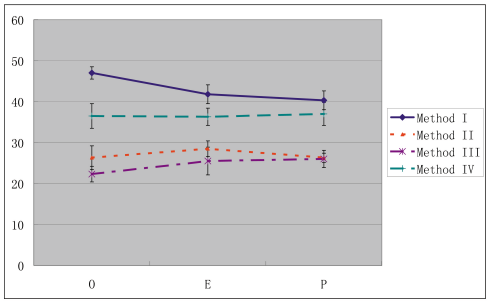

Figure 11: Plotting of the mean value and standard deviation comparisons of four different methods (for two-party conversations). The X axis represents different research questions, and the Y axis represents the average (mean) values of the four different methods. Vertical bars represent the standard deviations.

Variable	Question	χ^2	d.f.	5% χ^2
Method	**O**	30.0	6	12.59
	E	12.8	6	12.59
	P	29.3	6	12.59
Model	**O**	3.4	1	3.84
	E	1.6	1	3.84
	P	0.4	1	3.84
Duration	**O**	0.0	1	3.84
	E	1.5	1	3.84
	P	2.3	1	3.84

Table 4: Logistic regression analysis for single avatar samples.

Variable	Question	χ^2	d.f.	5% χ^2
Method	**O**	33.4	6	12.59
	E	26.4	6	12.59
	P	22.3	6	12.59
Model	**O**	1.4	1	3.84
	E	1.7	1	3.84
	P	2.3	1	3.84
Duration	**O**	0.4	1	3.84
	E	0.0	1	3.84
	P	0.3	1	3.84

Table 5: Logistic regression analysis for two-party avatar conversation samples.

7 Discussion and Conclusions

In this paper, we present a simple while effective technique for synthesizing natural and engaging eye gazes given a head motion sequence as input. A novel hybrid data acquisition solution, consisting of an optical motion capture system and off-the-shelf video cameras, is specially introduced for recording simultaneous saccades and head movements. Based on the recorded joint gaze-head motion data, we statistically learn a DCCA-based gaze head interaction model that sufficiently captures the dynamics and kinematic characteristics of natural saccades and their intrinsic associations with accompanying head movements. Furthermore, by conducting in-depth comparative user studies, we validated that gazes synthesized by our approach are measurably outperform the state of the art "Eyes Alive" model [21] and texture synthesis-based approach [12].

Two limitations exist in our current approach. First, we recorded our gaze and head motion data by asking the captured subjects to sit face-to-face. Hence, an implicit assumption of our approach is that the Point of Interest (POI) of an avatar is looking at his/her front area. If a given head motion sequence does not satisfy this assumption, our model may fail to generate its corresponding natural saccade movements. For instance, in a three-party conversation scenario, one subject may focus his/her POI on one of the other two parties. In this case, our statistical model may not be able to synthesize engaging gazes for all the involved three parties. Second, speech content is not taken into consideration in current gaze synthesis algorithm. People may have slightly different gaze patterns depending on whether they are talking or listening, which was also observed and reported by Lee *et al.* [21]. In the future, we also plan to further improve the intelligence and usability of our approach by integrating speech content into our gaze synthesis algorithm.

Acknowledgments

This work is funded by the Texas Norman Hackerman Advanced Research Program (project number: 003652-0058-2007). We would like to thank Wei Song for implementing the eyes alive model and insightful discussion, Chang Yun, Susu Liao and Benjamin Soibam for data capture and processing.

References

[1] A. Bahill, D. Andler, and L. Stark. Most naturally occurring human saccades have magnitudes of 15 deg or less. *Investigative Ophthalmol*, 14(6):468–469, 1975.

[2] E. Bizzi. Central programming and peripheral feedback during eye-head coordination in monkeys. *Bibl. Ophthal.*, (82):220–232, 1972.

[3] C. Blakemore and M. Donaghy. Co-ordination of head and eyes in the gaze changing behaviour of cats. *J Physiol.*, (300):317–335, 1980.

[4] M. Brand and A. Hertzmann. Style machines. In *Proc. of Siggraph 2000*, pages 183–192, 2000.

[5] C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan. Rigid head motion in expressive speech animation: Analysis and synthesis. *IEEE Transaction on Audio, Speech and Language Processing*, 15(3):1075–1086, March 2007.

[6] C. Busso, Z. Deng, U. Neumann, and S. Narayanan. Natural head motion synthesis driven by acoustic prosody features. *Computer Animation and Virtual Worlds*, 16(3-4):283–290, July 2005.

[7] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone. Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In *Proc. of ACM SIGGRAPH'94*, pages 413–420, 1994.

[8] J. Cassell, H. Vilhjalmsson, and T. Bickmore. Beat: The behavior expression animation toolkit. In *Computer Graphics (Proc. of ACM SIGGRAPH'01)*, pages 477–486, Los Angeles, 2001.

[9] M. Costa, T. Chen, and F. Lavagetto. Visual prosody analysis for realistic motion synthesis of 3d head models. In *Proc. of Int'l Conf. on Augmented, Virtual Environments and Three-Dimensional Imaging*, Ornos, Mykonos, Greece, 2001.

[10] Z. Deng, C. Busso, S. S. Narayanan, and U. Neumann. Audio-based head motion synthesis for avatar-based telepresence systems. In *Proc. of ACM SIGMM 2004 Workshop on Effective Telepresence*, pages 24–30, New York, NY, Oct. 2004.

[11] Z. Deng, J. P. Lewis, and U. Neumann. Practical eye movement model using texture synthesis. In *Proc. of ACM SIGGRAPH 2003 Sketches and Applications*, San Diego, 2003.

[12] Z. Deng, J. P. Lewis, and U. Neumann. Automated eye motion synthesis using texture synthesis. *IEEE Computer Graphics and Applications*, pages 24–30, March/April 2005.

[13] Z. Deng and U. Neumann. *Data-Driven 3D Facial Animation*. Springer-Verlag Press, 2007.

[14] E. G. Freedman and D. L. Sparks. Coordination of the eyes and head: movement kinematics. *Experimental Brain Research*, 131(1):22–32, 3 2000.

[15] M. Garau, M. Slater, S. Bee, and M. A. Sasse. The impact of eye gaze on communication using humanoid avatars. In *Proc. of ACM CHI'01*, pages 309–316, 2001.

[16] H. P. Graf, E. Cosatto, V. Strom, and F. J. Huang. Visual prosody: Facial movements accompanying speech. In *Proc. of IEEE Int'l Conf. on Automatic Face and Gesture Recognition(FG'02)*, May 2002.

[17] M. G. Kendall and B. Babington-Smith. On the method of paired comparisons. *Biometrika*, 31:324–345, 1940.

[18] S. C. Khullar and N. Badler. Where to look? automating visual attending behaviors of virtual human characters. In *Proc. of the Third ACM Conf. on Autonomous Agents*, pages 16–23, 1999.

[19] N. D. L. Itti and F. Pighin. Realistic avatar eye and head animation using a neurobiological model of visual attention. In *Proc. SPIE*, volume 5200, pages 64–78, 2004.

[20] P. Ledda, A. Chalmers, T. Troscianko, and H. Seetzen. Evaluation of tone mapping operators using a high dynamic range display. In *ACM SIGGRAPH 2005 Papers*, pages 640–648, 2005.

[21] S. P. Lee, J. B. Badler, and N. I. Badler. Eyes alive. *ACM Trans. Graph.*, 21(3):637–644, 2002.

[22] R. J. Leigh and D. S. Zee. *The Neurology of Eye Movements*. Oxford University Press, 1999.

[23] Y. Li, T. Wang, and H.-Y. Shum. Motion texture: a two-level statistical model for character motion synthesis. *ACM Trans. Graph.*, 21(3):465–472, 2002.

[24] S. Masuko and J. Hoshino. Head-eye animation corresponding to a conversation for cg characters. *Computer Graphics Forum (Proc. of Eurographics 07)*, 26(3):303–312, 2007.

[25] V. Pavlovic, J. M. Rehg, and J. MacCormick. Learning switching linear models of human motion. In *NIPS*, pages 981–987, 2000.

[26] D. D. Salvucci and J. H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proc. of ETRA 2000*, pages 71–78, 2000.

[27] S. Siegel. *Nonparametric statistics for the behavioral sciences*. McGRAW-HILL BOOK COMPANY, INC., 1956.

[28] W. Steptoe and A. Steed. High-fidelity avatar eye-representation. In *Proc. of IEEE Virtual Reality'08*, pages 111–114, 2008.

[29] M. Stone, D. DeCarlo, I. Oh, C. Rodriguez, A. Stere, A. Lees, and C. Bregler. Speaking with hands: creating animated conversational characters from recordings of human performance. In *ACM SIGGRAPH 2004 Papers*, pages 506–513, 2004.

[30] D. L. Torre and M. J. Black. Dynamic coupled component analysis. In *Proc. of CVPR'01*, pages 643–650, 2001.

[31] R. Vertegaal, G. V. Derveer, and H. Vons. Effects of gaze on multiparty mediated communication. In *Proc. of GI'00*, pages 95–102, 2000.

[32] R. Vertegaal, R. Slagter, G. V. Derveer, and A. Nijholt. Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. In *Proc. of ACM CHI'01*, pages 301–308, 2001.

[33] V. Vinayagamoorthy, M. Garau, A. Steed, and M. Slater. An Eye Gaze Model for Dyadic Interaction in an Immersive Virtual Environment: Practice and Experience. *Computer Graphics Forum*, 23(1):1–11, 2004.

[34] J. M. Wang, D. J. Fleet, and A. Hertzmann. Gaussian process dynamical models for human motion. *IEEE Trans. PAMI*, pages 283–298, Feb 2008.

[35] T. Warabi. The reaction time of eye-head coordination in man. *Neurosci. Lett.*, (6):47–51, 1977.