Eurographics/ ACM SIGGRAPH Symposium on Computer Animation (2008)
M. Gross and D. James (Editors)

Perceptually Guided Expressive Facial Animation

Zhigang Deng[†] and Xiaohan Ma[‡]

Computer Graphics and Interactive Media Lab, Department of Computer Science
University of Houston, Houston, TX, USA
http://graphics.cs.uh.edu/

Abstract

Most of current facial animation approaches largely focus on the accuracy or efficiency of their algorithms, or how to optimally utilize pre-collected facial motion data. However, human perception, the ultimate measuring stick of the visual fidelity of synthetic facial animations, was not effectively exploited in these approaches. In this paper, we present a novel perceptually guided computational framework for expressive facial animation, by bridging objective facial motion patterns with subjective perceptual outcomes. First, we construct a facial perceptual metric (FacePEM) using a hybrid of region-based facial motion analysis and statistical learning techniques. The constructed FacePEM model can automatically measure the emotional expressiveness of a facial motion sequence. We showed how the constructed FacePEM model can be effectively incorporated into various facial animation algorithms. For the sake of clear demonstrations, we choose data-driven expressive speech animation generation and expressive facial motion editing as two concrete application examples. Through a comparative user study, we showed that comparing with the traditional facial animation algorithms, the introduced perceptually guided expressive facial animation algorithms can significantly increase the emotional expressiveness and perceptual believability of synthesized facial animations.

Categories and Subject Descriptors (according to ACM CCS): I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism-Animation H.1.2 [Models and Principles]: User/Machine Systems-Human Factors

1. Introduction

Facial expression is arguably one of the subtlest parts in computer-generated virtual faces and characters. In the past several decades, various research efforts have been attempted to produce realistic facial animations with human-like emotions; nevertheless, how to efficiently measure and synthesize highly-believable and expressive facial animations is still a challenging research topic.

The ultimate measuring stick for the visual fidelity of expressive facial animations is *human perception*. Currently, in order to measure the visual fidelity and emotional expressiveness of an expressive facial animation, the most popular means is to conduct subjective user studies where participants first view the animation and then complete evaluation forms [POM99]. This process is not automated (tedious human involvements), inefficient (time-consuming experiment setup and user studies), and costly (participant cost). Therefore, the above subjective evaluation process is typically limited to an offline, post-production evaluation tool.

In this work, we inject human perception insights into facial animation algorithms by introducing a novel computational facial perceptual metric that models the association between high-dimensional, dynamic facial motion patterns and ultimate perceptual outcomes. First, based on a pre-recorded, high-fidelity expressive facial motion dataset, we conducted subjective evaluation experiments (*i.e.*, asking participants to evaluate the expressiveness of facial animation clips as a nonforced-choice task) and performed region-based facial motion analysis and modeling. Then, we learned a statistical perceptual prediction model (termed as the *FacePEM* in this work) that is able to measure and predict the perceptual outcomes of new facial motion sequences. Finally, we showed how to effectively incorpo-

[†] zdeng@cs.uh.edu
[‡] xiaohan@cs.uh.edu

rate the FacePEM into various facial animation algorithms. For the sake of clear demonstrations, we choose data-driven speech animation synthesis and expressive facial motion editing as two concrete application examples.

The major contributions of this work include: (1) it introduces *a novel computational perceptual metric* (FacePEM) for measuring and predicting the emotional expressiveness of facial motion sequences. It naturally bridges high-dimensional, dynamic expressive facial motions and ultimate human perceptual outcomes; and (2) on top of the constructed FacePEM, it introduces *perceptually guided expressive facial animation algorithms*, and our user study reveals that the introduced perceptually guided algorithms are able to significantly increase the algorithmic intelligence and the perceptual believability of the synthesized expressive facial animations.

The remainder of this paper is organized as follows. Section 2 briefly reviews recent research efforts most related to this work. Section 3 describes how we collected and pre-processed expressive facial motion data for this work. Section 4 details how we construct our facial perceptual metric (FacePEM). Section 5 describes how we incorporate the constructed FacePEM into data-driven expressive speech animation synthesis (Section 5.1), expressive facial motion editing applications (Section 5.2), and user studies (Section 5.3). Finally, discussion and conclusions are presented in Section 6.

2. Related Work

Significant research efforts have been attempted to generate realistic facial and character animations. In this section, we briefly review recent research efforts that are most related to this work.

2.1. Facial Animation and Expression

Various techniques were developed to model and animate computerized faces [DN07], including geometric deformations [SF98, PHL*98, NN01, SP04, LMDN05], statistical face models [BV99], physically-based approaches [LTW95, SNF05], performance-driven facial animations [Wil90, ZSCS04], and facial expression synthesis and editing [ZLGS03,JTDP03]. Recently data-driven approaches for facial animation have achieved noticeable successes [Bra99, CDB02, VBPP05, WSZP07]. The work of [BCS97, KT03, CFKP04,DN06,DN08] essentially optimally recombines facial motion frames from a pre-recorded facial motion dataset to meet novel inputs, *e.g.*, spoken or typed input. The above data-driven approaches often focus on the mathematical accuracy or efficiency of their algorithms, while little attention has been paid to incorporate perceptual insights to their algorithms. For example, these techniques do not provide any mechanism for automatically measuring the realism or expressiveness of facial animations being synthesized. In addition, automatic analysis of facial expressions (*e.g.*, recognize the movements of facial action units [EF78]) from images or video streams has been a hot topic in computer vision community [PR00,TKC01,VP06]. While these approaches work on the analysis aspect of facial expressions, our work focuses on the synthesis side of facial expressions and animations by inventing and exploiting a computational perceptual metric.

2.2. Perceptual Approaches for Animation

Due to the importance of human perception and cognition [Ado02, SWCCG06], perceptual approaches for graphics and animations have attracted increasing interdisciplinary interests in recent years [OHM*04].

A number of studies have been conducted to measure the association between human perception and different factors of character animations [HOT98,OD01,ODGK03,WFM01, RP03, WB04, MDCO06, MNO07]. Watson *et al.* [WFM01] studied the visual perceptions of static 3D models (animals and man-made artifacts) that are simplified by different algorithms, in terms of the following three measurements: naming times, rating, and preferences. Researchers also looked into perceptual sensitivity to errors in ballistic motion [RP03] and dynamic abnormalities in physically-based simulation [ODGK03]. It was experimentally reported that humans are more sensitive to horizontal velocity than vertical velocity, and more sensitive to added accelerations than added decelerations [RP03]. Perceptual studies were also conducted to measure how different factors contribute to human perception of character motion in various settings including collision perception [OD01], different Level Of Details (LODs) representation of clothing [MDCO06], and pose update rate [MNO07]. However, measures used in character animation perceptual studies, *e.g.*, horizontal/vertical velocity [RP03] and angular/momentum [ODGK03], cannot be used for measuring expressive facial animations due to the significant difference between facial animations and character animations. As such, the findings from the above perceptual studies and approaches cannot be directly applied to expressive facial animations without considerable efforts.

Subjective evaluation and psychophysical experiments have been also conducted to gain human perceptual insights on facial expressions [CBK*03,CKBW04,WBCB05, WBCB08] or lip-sync [GEP03, CPM*05]. Geiger *et al.* [GEP03] use two different types of perceptual discrimination experiments (an explicit visual detection task and an implicit lip-reading task) to evaluate the realism of synthesized visual speech. Cosker *et al.* [CPM*05] exploit and adopt the “McGurk Effect" for the purpose of evaluating the realism of lip-sync. Psychophysical experiments have been conducted to study which and how animation parameters affect human perception [WBCB05,WBCB08] or study which/how different facial regions affect the perception of facial expressions [CKBW04]. Most of these efforts are still centered at the *qualitative* side of the perceptual realism of synthetic faces. In this work, our aim is to construct a com-

putational perceptual metric that *quantitatively* models the association between dynamic 3D facial motions and its perceptual outcomes, and further exploit this metric to build perceptually guided expressive facial animation algorithms.

3. Data Acquisition and Preprocessing

We used a VICON optical motion capture system with ten cameras to record high-fidelity, expressive facial motions of human subjects at a 120 Hz sampling frequency (Fig. 1). Four college students majoring in theatre/performing art in a university were selected for this motion data acquisition. Attached with a total of 103 markers (95 face markers, 4 head markers, and 4 neck markers), they were directed to speak a number of pre-designed sentences three times. Each time the captured subjects spoke with a different emotion (happy, angry, or sad). Therefore, we are aware of the intended emotion label of each recorded facial motion sequence. A total of about seventy sentences (counting all the captured subjects) were recorded. The duration of each sentence recording is from six seconds to twenty seconds.

Figure 1: *Snapshots of the used motion capture system. The left two panels show the system, and the right panel shows the used facial marker configuration.*

After the data acquisition, we removed head motion from the data as follows: construct a local orthogonal coordinate system for each motion capture frame based on the four head markers, and then calculate rotation matrices between these coordinate systems [NN01]. Due to the difference of the 3D face geometries of the captured subjects, we picked one of them as *the reference face*, and then transformed and aligned the facial motions of other captured subjects with the reference face using the close-form solution proposed by Horn *et al.* [HHN88]. Following their approach, we computed the translational offset between two subjects as the difference of their coordinate centroids. The scaling ratio was set to the ratio of the root-mean-square deviations from their respective coordinate centroids. In this way, all the recorded facial motion data are normalized to the same 3D coordinate system (of the reference face).

4. Construction of the Computational Perceptual Metric for Expressive Facial Animation

In this section, we describe how we construct a statistical facial perceptual metric. In this work, we call this perceptual metric *FacePEM*. Figure 2 shows the schematic view of the construction of the FacePEM metric. It consists of the following steps: (1) high-fidelity expressive facial motion data of human subjects are recorded, (2) objective facial motion patterns are extracted by applying region-based facial motion analysis and modeling algorithms, (3) subjective perceptual studies are conducted on facial animation clips generated by transferring pre-recorded 3D facial motion data to a photorealistic 3D face model, (4) a statistical perceptual prediction model (FacePEM) that directly maps facial motion patterns to perceptual outcomes is trained and constructed, and (5) finally, given a new facial motion sequence, the constructed FacePEM is able to automatically compute and measure its corresponding perceptual outcomes. The above paradigm (Fig. 2) is built on the following key observation: *3D visual facial motions are qualitatively correlated with the perception of facial emotion in a consistent manner* [DBLN06].

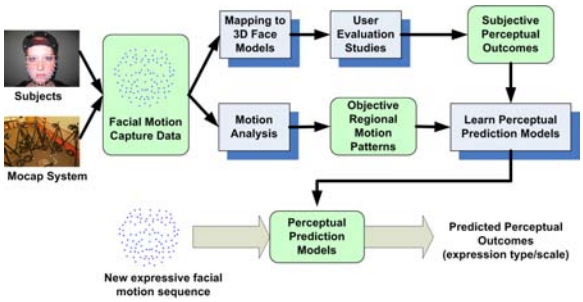

Figure 2: *Schematic view of the construction of the computational perceptual metric (FacePEM) for measuring expressive facial animations.*

4.1. Subjective Evaluations

We used 68 of the recorded facial motion sequences in our subjective evaluation experiment. We first transferred these facial motion sequences to a photorealistic 3D face model using a feature point-based deformation technique [KGT00]. The resulting facial animation clips (the left of Fig. 3) enclose three different emotions (happy, angry and sad). Then, we conducted a subjective evaluation experiment at a university classroom by showing these facial animation clips in a random order to 30 participants. Most of the participants are computer science undergraduate or graduate students. After viewing each facial animation clip for a maximum of three times, the participants were required to identify their perceived emotion as a nonforced-choice task (four options: happy, angry, sad, and N/A) and corresponding emotional expressiveness scale (1 to 10, 1 represents “very little emotion", and 10 represent “full/maximum emotion"). The addition of a “N/A" category could have alleviated discrimination in the experiment [FS01]. Figure 3 shows a used facial animation clip (the left panel) and a snapshot of our subjective experiment scenario (the right panel).

After subjective evaluation results were collected, for

Figure 3: *Subjective evaluation experiment on the facial animation clips. The left shows an example of the used expressive facial animation clips, and the right shows a snapshot of our subjective evaluation experiment scenario.*

each facial animation clip A_i, we computed its **Perceptual Outcome Vector (POV)**, $POV_i = (S_i^{happy}, S_i^{angry}, S_i^{sad}, S_i^{n/a})$. S_i^{θ} (θ is *happy*, *angry*, *sad*, or *n/a*) is calculated using the following equation:

$$S_i^{\theta} = (\sum_{j=1}^{N} \delta_{i,j}^{\theta} * E_{i,j}^{\theta})/N \quad (1)$$

Where N is the number of the participants; $\delta_{i,j}^{\theta}$ is a Kronecker Delta function which returns 1 when the j^{th} participant perceived emotion type θ from the facial animation clip A_i, otherwise 0; $E_{i,j}^{\theta}$ indicates the perceived expressiveness scale of emotion type θ on A_i by the j^{th} participant. Figure 4 shows three examples of expressive facial animation clips and their computed POVs.

Figure 4: *Examples of expressive facial animation clips and their computed Perceptual Outcome Vectors (POVs). (1) POV: (happy=8.6, angry=0.7, sad=0.7, n/a=0), (2) POV: (happy=0.0, angry=0.3, sad=9.7, n/a=0), and (3) POV: (happy=1.0, angry=7.0, sad=2.0, n/a=0). The picked frame of these clips is #110.*

4.2. Facial Motion Analysis and Modeling

In the above subjective evaluation experiment, we obtained a POV for each facial motion sequence. In this part, we employ statistical learning techniques to analyze and model the recorded 3D expressive facial motions.

4.2.1. Face Segmentation

As described in Section 3, 95 facial markers were captured. If concatenating 3D positions of these markers forms a vector, its dimension is high (95*3=285). If a single Principal Component Analysis (PCA) space is constructed for these motion vectors, and PCA is essentially a global transformation/reduction, there is no explicit and intuitive correspondence between global PCA eigen-vectors and localized facial movements. In this work we adopt a divide-and-conquer strategy to partition the whole face into different facial regions: first apply a feature point based deformation technique [KGT00] to deform a static 3D face model based on the 95 facial markers, and then use a physically-motivated segmentation scheme proposed by Joshi *et al.* [JTDP03] to divide the face into meaningful regions. The left panel of Figure 5 shows the used 3D face model, and its right panel shows its segmentation result. In this work, the segmentation threshold used in the work of [JTDP03] is set to 0.3.

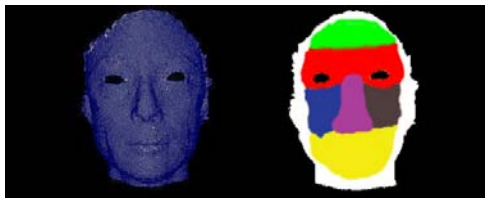

Figure 5: *Illustration of face segmentation in this work. The left panel shows the used 3D face model, and the right panel shows the face segmentation result. Distinct colors represent different regions.*

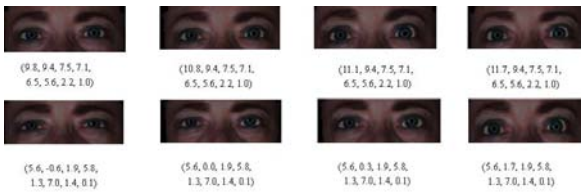

Figure 6: *Illustration of how the first and second most dominant PCA eigen-vectors affect the movements of the eye region. The 8-tuples are corresponding PCA coefficients. When the PCA coefficient of the first or second most dominant eigen-vector of the eye region is increased, the eyes (eyebrows) tend to be more open (raised).*

4.2.2. Region-Based Motion Reduction

Based on the above face segmentation, we obtain the following six facial regions: *forehead, eye, the left cheek, the right cheek, mouth*, and *nose*. For each facial region, we apply PCA to reduce its dimensionality while retaining more than 95% of its motion variation, and construct a truncated PCA space for each region. In this work, to retain more than 95% of the motion variation, the retained dimensionality is 4 for the forehead region, 8 for the eye region, 3 for the left cheek region, 3 for the right cheek region, 4 for the mouth region, and 5 for the nose region. In addition, we found that region-based PCA eigen-vectors typically intuitively correspond to meaningful, localized facial movements in the specific facial

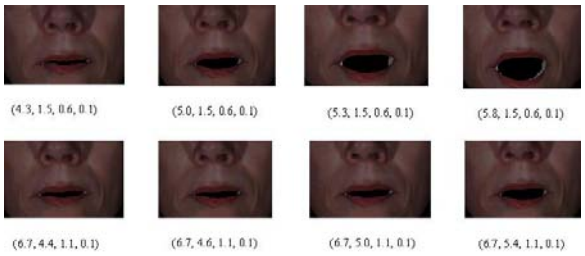

Figure 7: *Illustration of how the first and second most dominant PCA eigen-vectors affect the movements of the mouth region. The 4-tuples are corresponding PCA coefficients. When the PCA coefficient of the first or second most dominant eigen-vector of the mouth region is increased, the mouth tends to be more open.*

region [LD07]. Figures 6 and 7 show two examples of how the first and second most dominant PCA eigen-vectors affect regional facial movements (the eye region and the mouth region).

In this way, we can project any facial motion capture frame FRM_i into a reduced representation, termed as **the Locally Reduced** PCA **Coefficients** (**LRC**). In this paper, the LRC of FRM_i is represented as κ_i. Specifically, the κ_i consists of the following six components: κ_i^h (the forehead region), κ_i^e (the eye region), κ_i^{lc} (the left cheek region), κ_i^{rc} (the right cheek region), κ_i^m (the mouth region), and κ_i^n (the nose region).

4.2.3. Modeling Expressive Facial Motion Patterns

For a facial motion sequence S_k (assume a total of n facial motion frames), we obtain its corresponding LRC sequence, $S\kappa_k = \kappa_{k,1}, \kappa_{k,2}, \cdots$, and $\kappa_{k,n}$, by using the above region-based motion reduction. The dynamic behaviors of facial motion sequences are controlled by a continuous control state sequence, hence we model expressive facial motion patterns using the M-order Linear Dynamical Systems (LDS) [PRM00, CH07] given the generated LRC sequences. We used 54 out of the 68 expressive facial motion sequences described in Section 4.1 for this modeling, and retained the remaining 14 for test and validation.

For each emotion and each region, a separate M-order LDS is constructed. In other words, a total of 18 LDSs (6 regions $\times$ 3 emotions) are constructed. A LDS can be described in the following equation (Eq. 2):

$$x_n = \sum_{i=1}^{m} (A_i\, x_{n-i}) + v_i \tag{2}$$

Here m is the order of the LDS, x_i is the system state at time step i, v_i is an independent Gaussian noise, and A_i is a coefficient matrix. In this work, m (the order of the LDS) is experimentally set to 2, and x_i is $\kappa_{k,i}^{\theta}$ where $\theta \subseteq \{h, e, lc, rc, m, n\}$ (representing all the six facial regions).

As mentioned in the data acquisition (Section 3), the intended emotion label of every pre-recorded facial motion sequence is known. Thus, the above equation (Eq. 2) is fitted with the LRC sequences with a specific emotion and further solved using the least square method. For example, if S_{A_1}, S_{A_2},..., and S_{A_m} are the pre-recorded facial motion sequences with the angry emotion, then we use $S\kappa_{A_1}^e$,$S\kappa_{A_2}^e$,..., and $S\kappa_{A_m}^e$ to fit the above Eq. 2 to obtain the *Angry-EyeRegion LDS*, $LDS_{A,e}$. In this paper, we represent the constructed 18 LDSs as $LDS_{emo,reg}$ where $emo \subseteq \{H, A, S\}$ (representing *Happy*, *Angry*, and *Sad* respectively) and $reg \subseteq \{h, e, lc, rc, m, n\}$. It should be noted that in this step we did not construct LDSs for *N/A* (*i.e.*, $LDS_{n/a,reg}$) due to the lack of proper and sufficient training data.

Based on the fitted $LDS_{emo,reg}$, we further define a closeness function $P_{emo,reg}(S\kappa_i)$ that describes the closeness (match) of the facial motion sequence S_i at a specific facial region *reg* representing a specific emotion *emo*. If this value is larger, it means S_i is better matched with the dynamical motion patterns of the emotion *emo* at the facial region *reg*, and vice versa. Similar to the work of [CH07], we create this closeness function using the following equations (Eq. 3 and 4). We create a total of 18 closeness functions $P_{emo,reg}$ (all possible combinations between three emotions and six facial regions). Therefore, for S_i, we obtain its 18 closeness values $\{P_{emo,reg}(S\kappa_i)\}$. If these closeness values are concatenated together in a certain order as a vector, we term this vector as the **Objective Matchness Vector (OMV)** of S_i, represented as OMV_i in this paper.

$$\begin{aligned} P_{emo,reg}(S\kappa_i) &= G(LDS_{emo,reg}, S\kappa_i^{reg}) \\ &= e^{-E} \end{aligned} \tag{3}$$

$$\begin{aligned} E &= -\ln F(S\kappa_i^{reg}) = -\ln F(x_{1:T}) \\ &= -\ln \Pi_{t=m+1}^{T} F(x_t | x_{t-m:t-1}) \\ &\approx C * \sum_{t=m+1}^{T} ||x_t - \sum_{j=1}^{m} (A_j\, x_{t-j}) - v_j||^2 \end{aligned} \tag{4}$$

Here the function G calculates the closeness value given $S\kappa_i^{reg}$ and its corresponding fitted $LDS_{emo,reg}$, the function F computes the error (deviation) when $S\kappa_i^{reg}$ is fitted to the constructed LDS, $LDS_{emo,reg}$, and C in Eq. 4 is a user-defined constant. In this work, it is experimentally set to 1.

4.3. Learning Facial Perceptual Prediction

For the training dataset (54 selected facial motion sequences), represented as $TrS_{i=1}^{54}$, we have their corresponding POVs (represented as $TrPOV_{i=1}^{54}$) and OMVs (represented as $TrOMV_{i=1}^{54}$). As described in Section 4.1, the dimensionality of the original POV_i is 4. In this step, we discard its fourth component N/A and retain the other three components (angry, happy and sad). In other words, the dimensionality of POV_i is changed to 3. Note that the dimen-

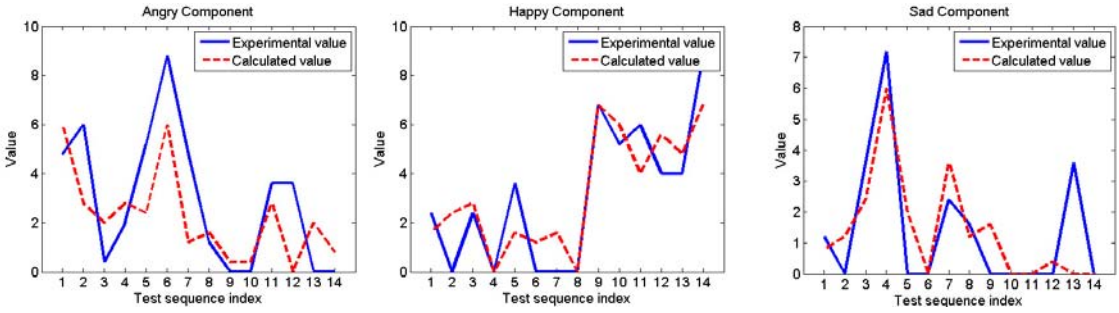

Figure 8: *Cross-validation results by applying the trained SVMs to our test/validation dataset (14 facial motion sequences). The left is for the angry component of their POVs, the middle is for the happy component of their POVs, and the right is for the sad component of their POVs. Here red dot lines represent ground-truth perceptual outcomes obtained from the subjective evaluation experiment, and blue solid lines represent the computed perceptual outcomes.*

sionality of OMV_i is still 18. Essentially, now given 54 validated mappings (correspondences) between perceptual outcomes $TrPOV_{i=1}^{54}$ and objective facial motion pattern descriptions $TrOMV_{i=1}^{54}$, we need to train a statistical regression model (*i.e.*, a facial perceptual prediction model) that can predict POV (as output) for any OMV (as input). Least-square based linear fitting, the Radial Basis Functions Network (RBFs), and Support Vector Machines (SVMs) were chosen and trained respectively. To compare the performance of these three different approaches, we used the 14 retained test/validation facial motion sequences (not used for training) and define the following error metric (Eq. 5).

$$err = \sqrt{\frac{\sum_{i=1}^{TNum}(TePOV_i - \widehat{TePOV_i})^2}{TNum}} \qquad (5)$$

Here $TNum$ is the number of test/validation sequences (=14 here), $TePOV_i$ is the POV of the i^{th} test motion sequence, and $\widehat{TePOV_i}$ is the computed POV of the i^{th} test motion sequence using our trained statistical model.

In this work, the Matlab RBFs implementation and the LIBSVM with a RBF kernel [CL01] were used. For the above three methods we obtained the following errors: SVMs (1.337), RBFs (1.5749), and Linear (2.9402). SVMs achieved the smallest error on our validation dataset. Therefore, we chose the SVMs as the statistical model for this regression step. Figure 8 shows cross-validation results by applying the trained SVMs to our validation dataset. Note that because a POV encloses three components (angry, happy, and sad), a separate panel of Fig. 8 is used to show the comparison of each component respectively.

Figure 9 shows the inside view of the constructed facial perceptual prediction model (FacePEM). Given a new facial motion sequence, this constructed model automatically outputs its corresponding POV. Then, we determine the emotion type and its scale based on the element with maximum value.

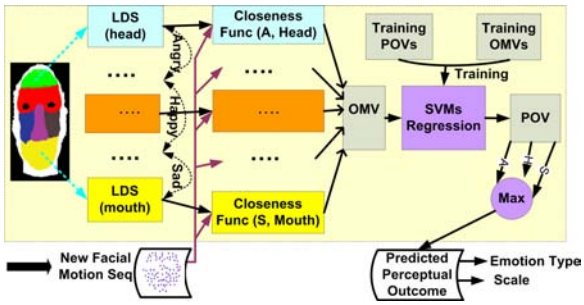

Figure 9: *Inside view of the constructed facial perceptual prediction model (FacePEM).*

5. FacePEM-Guided Facial Animation Algorithms

In this section, we describe how to effectively incorporate the FacePEM into various facial animation algorithms and applications. Specifically, for the sake of clear demonstrations, we chose data-driven expressive speech animation generation (Section 5.1) and expressive facial motion editing (Section 5.2) as two concrete application examples. For the two chosen application examples, we performed comparative user studies to evaluate the new FacePEM-guided facial animation algorithms (Section 5.3).

5.1. Perceptually Guided Speech Animation Synthesis

To meet new spoken or typed input, data-driven speech animation synthesis approaches either synthesize novel speech animations by sampling from the learned statistical models [Bra99, EGP02, VBPP05, WSZP07], or optimally recombine motion frames from a pre-recorded facial motion dataset [BCS97, KT03, CFKP04, DN06, DN08]. The key part of the latter [BCS97, KT03, CFKP04, DN06, DN08], a search cost function can be generalized to the following formula (Eq. 6):

$$Cost = PhoMtchCost + ConstrCost + SmoCost \qquad (6)$$

Here *SmoCost* describes the smoothness of the facial motion sequence being synthesized, *ConstrCost* describes the match between the facial motion sequence being synthesized and specified constraints (*e.g.*, emotion constraints), and *PhoMtchCost* describes the match between the phonemes of the inputted novel speech and the phonemes of the facial motion sequence being synthesized.

In order to test and validate the usability of the FacePEM, we chose and implemented the speech animation synthesis part of [DN06, DN08] due to the reason that an emotion match cost is included in its search cost function. In their system, its emotion match cost *EC* is defined and incorporated into the search cost function (Eq. 6) in the following heuristic way:

$$EC(s, Emo) = C * (1 - Same(Emo, EmoLabel(s))) \quad (7)$$

Here s represents a facial motion subsequence of the pre-recorded facial motion dataset, *Emo* represents the target emotion specified by users, C is a constant penalty value, $EmoLabel(s)$ represents the intended emotion label (pre-planned when s was recorded, and it can only take one of the following three discrete values - *angry*, *happy*, and *sad*). The *Same* function returns 1 if two input emotion types are the same, otherwise 0. However, in the real world, generally human subjects cannot always display facial emotions in the 0/1 mode (1 for "full specific emotion", *e.g.*, angry, and 0 for "no emotion"), and they show emotions in the middle of the two extreme cases. As such, the 0/1 emotion judgment (Eq. 7) used in their work [DN06, DN08] is non-optimal.

Based on the FacePEM, we are able to reliably and automatically measure and predict the emotion type and expressiveness scale of any facial motion sequence in the runtime of the synthesis algorithm. We modified the traditional way of computing *EC* (Eq. 7) to the following perceptually guided way (Eq. 8).

$$EC(s, Emo) = C * (1 - CalcEmo(s)[Emo]) \quad (8)$$

Here $CalcEmo(s)$ is the emotion vector (normalized to $0 \sim 1$) computed from the FacePEM, and $CalcEmo(s)[Emo]$ represents its expressiveness scale of emotion type *Emo*. Given the same inputs, we generated expressive speech animation clips using both the traditional algorithm [DN06, DN08] and the new FacePEM-guided algorithm. Figure 10 shows some speech animation frames synthesized by the two approaches. We also conducted subjective user studies to evaluate the expressiveness of these synthetic speech animation clips. Detailed user study results are described in Section 5.3.

5.2. Expressive Facial Motion Editing Enhanced with Expression Cues

A number of data-driven, expressive facial motion editing techniques [CDB02, CFP03, JTDP03, VBPP05, LD07] had been proposed to edit facial motion sequences, *e.g.*, increasing the expressiveness or changing their affective states. However, all these approaches do not provide any feedback or expressiveness cue to users when the users are performing editing operations, which imposes great difficulty and inconvenience for the users.

To test and validate the usability of the FacePEM for facial motion editing applications, on top of an existing expressive facial motion editing system [LD07], we incorporated the FacePEM into the editing system in the following way: when users modify one or several facial motion frames, or the whole motion sequence, our FacePEM model will measure and display its updated emotion type and expressiveness scale to the users. This timely emotion and expressiveness cue adds a new kind of intelligence into the facial motion editing system and greatly improves the efficiency of editing operations, *e.g.*, alleviate the pains of the back-and-forth tuning/editing.

Three users were asked to use both the traditional editing system (without the emotional expressiveness cue) and the new editing system enhanced with the FacePEM. A number of edited expressive facial animation clips were used for a subjective user study. Results of the subjective user study are detailed in follow-up Section 5.3.

5.3. Results and Evaluations

In order to evaluate and quantify the effects of the above FacePEM-guided expressive facial animation algorithms (Section 5.1 and Section 5.2), we generated a total of 30 facial animation clips from both the traditional algorithms (without perceptual metrics) and the new FacePEM-guided algorithms, and then conducted subjective evaluations on these clips. Half of the clips are from the traditional/new algorithms. 20 out of the total 30 clips are from the data-driven expressive speech animation synthesis (Section 5.1), and the other 10 are from the expressive facial motion editing application (Section 5.2). These 30 clips were mixed in a random order. Similar to the evaluation procedure in Section 4.1, we conducted a comparative user study experiment to evaluate the emotion fidelity and expressiveness of these facial animation clips. A total of twenty participants were asked to identify the perceived emotion type and expressiveness scale of these clips.

We performed the One-Way ANOVA analysis on the collected experiment results. As shown in Fig. 11, facial animation clips generated by the new FacePEM-guided algorithms achieved significantly higher average ratings. The only exception is the angry clips from the facial motion editing application: although the FacePEM-guided algorithm still achieved a slightly higher average rating than the traditional

Figure 10: *Side-by-side frame comparisons of expressive speech animations synthesized by a speech animation synthesis system with/without FacePEM.*

one, their scores were quite close. Our subjective evaluation results reveal that the FacePEM metric can be effectively incorporated into data-driven expressive speech animation synthesis and facial motion editing, and it measurably increases the perceptual believability of synthesized expressive facial animations.

6. Discussion and Conclusions

In this paper, we present a novel computational framework for constructing a perceptual metric (called the *FacePEM*) to measure and predict the emotional fidelity of expressive facial animations, by bridging human perceptual insights with objective facial motion patterns. The constructed FacePEM enables the automated computation of the emotion and expressiveness scale of facial animation sequences.

We further demonstrated how the FacePEM can be effectively incorporated into various expressive facial animation algorithms and applications. In this work, we choose expressive visual speech animation synthesis and expressive facial motion editing as two concrete application examples. Through comparative user studies, we found that in most cases the FacePEM-guided algorithms are able to significantly improve the intelligence and efficiencies of facial animation algorithms and measurably increase the perceptual believability of synthesized expressive facial animations.

We employ statistical learning algorithms to construct this computational framework including the region-based Principal Component Analysis (PCA) for facial motion analysis, the *M*-order Linear Dynamical Systems (LDS) for facial motion pattern modeling, and the Support Vector Machines (SVMs) for learning the mapping between objective facial patterns and subjective perceptual outcomes.

One common limitation of statistical learning approaches is that it is hard to in advance know or predict how much data would be enough to train well-behaviored statistical models. The similar limitation exists in our current approach. As a future work, we plan to look into meta learning algorithms to alleviate this issue. Another limitation of current approach is that we did not consider the effects of eye gaze/motion. In our subjective evaluation experiments, we simplified the eye motion by setting it to a fixed position. As “the windows to the soul", eye movements are generally believed to provide important cues to the mental and emotional state of human beings. As a next step, we plan to incorporate statistical eye motion models [DLN05] into our perceptual metric and investigate the emotion perception effect of the eye movements.

In current work, we only considered three basic emotions: angry, happy and sad. However, as pointed out by Ekman and Friesen [EF78], there exist six universal facial emotions: angry, happy, sad, fear, surprise, and disgust. In addition, cultures might play an indispensable role in emotion perception and understanding. Currently the captured subjects and the majority of experiment participants are Americans.

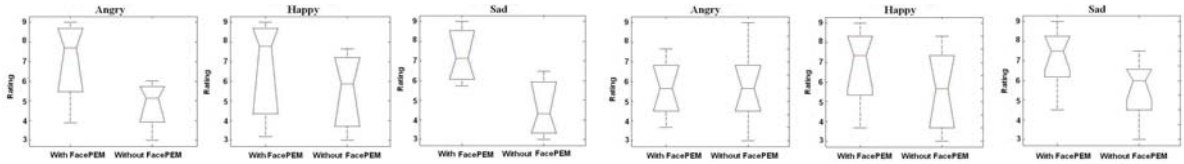

Figure 11: *One-Way ANOVA results of comparative subjective evaluation experiments. The left three panels are for the expressive speech animation synthesis, and the right three panels are for the expressive facial motion editing. The P-values are 0.026, 0.138, 0.011, 0.850, 0.2495 and 0.017 from left to right.*

We plan to extend our framework to enclose more emotion types and model the culture-dependent issue of the computational facial perceptual metrics. In the future we also plan to remove the idiosyncratic behaviors of recorded data and extract pure facial emotional signals in the data preprocessing step [JL08].

Acknowledgments

This work is funded by the University of Houston new faculty startup fund and the Texas Norman Hackerman Advanced Research Program (project number: 003652-0058-2007). We would like to thank Qing Li for sharing her facial motion editing codes, Jose Baez-Franceschi for his 3D model cleaning, Tanasai Sucontphunt for his facial deformation implementation, and other member of UH CGIM Lab for their insightful suggestions. We also thank Dr. Ioannis Kakadiaris and his UH Computational Biomedicine Lab for the help in 3D face modeling.

References

[Ado02] ADOLPHS R.: Recognizing emotion from facial expressions: Psychological and neurological mechanisms. *Behavioral and Cognitive Neuroscience Reviews 1*, 1 (2002), 21–62.

[BCS97] BREGLER C., COVELL M., SLANEY M.: Video rewrite: Driving visual speech with audio. *Proc. of ACM SIGGRAPH'97 31* (1997), 353–360.

[Bra99] BRAND M.: Voice puppetry. In *Proc. of ACM SIGGRAPH'99* (1999), pp. 21–28.

[BV99] BLANZ V., VETTER T.: A morphable model for the synthesis of 3D faces. In *Proc. of ACM SIGGRAPH'99* (1999), pp. 187–194.

[CBK*03] CUNNINGHAM D. W., BREIDT M., KLEINER M., WALLRAVEN C., BÜLTHOFF H. H.: How believable are real faces? towards a perceptual basis for conversational animation. In *Proc. of IEEE CASA'03* (2003).

[CDB02] CHUANG E., DESHPANDE H., BREGLER C.: Facial expression space learning. In *Proc. of PG'02* (2002).

[CFKP04] CAO Y., FALOUTSOS P., KOHLER E., PIGHIN F.: Real-time speech motion synthesis from recorded motions. In *Proc. of SCA'04* (2004), pp. 345–353.

[CFP03] CAO Y., FALOUTSOS P., PIGHIN F.: Unsupervised learning for speech motion editing. In *Proc. of SCA'03* (2003).

[CH07] CHAI J., HODGINS J. K.: Constraint-based motion optimization using a statistical dynamic model. *ACM Trans. Graph. 26*, 3 (2007).

[CKBW04] CUNNINGHAM D. W., KLEINER M., BÜLTHOFF H. H., WALLRAVEN C.: The components of conversational facial expressions. In *Proc. of APGV '04* (2004), pp. 143–150.

[CL01] CHANG C.-C., LIN C.-J.: LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

[CPM*05] COSKER D., PADDOCK S., MARSHALL D., ROSIN P. L., RUSHTON S.: Toward perceptually realistic talking heads: Models, methods, and mcgurk. *ACM Trans. on Appl. Percep. 2*, 3 (2005), 270–285.

[DBLN06] DENG Z., BAILENSON J., LEWIS J. P., NEUMANN U.: Perceiving visual emotions with speech. In *Proc. of the 6th Int'l Conf. on Intelligent Virtual Agents (IVA) 2006* (August 2006), pp. 107–120.

[DLN05] DENG Z., LEWIS J. P., NEUMANN U.: Automated eye motion synthesis using texture synthesis. *IEEE CG&A* (March/April 2005), 24–30.

[DN06] DENG Z., NEUMANN U.: eFASE: Expressive facial animation synthesis and editing with phoneme-isomap controls. In *Proc. of SCA'06* (2006), pp. 251–259.

[DN07] DENG Z., NOH J. Y.: Computer facial animation: A survey. In *Data-Driven 3D Facial Animation, Z. Deng and U. Neumann (Eds.)* (2007), Springer-Verlag Press, pp. 1–28.

[DN08] DENG Z., NEUMANN U.: Expressive speech animation synthesis with phoneme-level control. *Computer Graphics Forum 27*, 6 (2008), in press.

[EF78] EKMAN P., FRIESEN W. V.: *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, 1978.

[EGP02] EZZAT T., GEIGER G., POGGIO T.: Trainable videorealistic speech animation. In *ACM Trans. on Graph.* (2002), vol. 21, pp. 388–398.

[FS01] FRANK M., STENNETT J.: The forced-choice paradigm and the perception of facial expressions of emotion. *Journal of Personality and Social Psychology 80* (2001), 75–85.

[GEP03] GEIGER G., EZZAT T., POGGIO T.: Perceptual evaluation of video-realistic speech. *MIT-AI-Memo 2003-003* (Feb. 2003).

[HHN88] HORN B. K. P., HILDEN H. M., NEGAHDARIPOUR S.: Closed-form solution of absolute orientation using orthonormal matrices. *J. Opt. Soc. Am. A 5*, 7 (1988), 1127.

[HOT98] HODGINS J. K., O'BRIEN J. F., TUMBLIN J.: Perception of human motion with different geometric models. *IEEE Trans. Vis. Comp. Graph. 4*, 4 (1998), 307–316.

[JL08] JU E., LEE J.: Expressive facial gestures from motion capture data. *Computer Graphics Forum 27*, 2 (2008).

[JTDP03] JOSHI P., TIEN W., DESBRUN M., PIGHIN F.: Learning controls for blend shape based realistic facial animation. In *Proc. of SCA'03* (2003), pp. 35–42.

[KGT00] KSHIRSAGAR S., GARCHERY S., THALMANN N. M.: Feature point based mesh deformation applied to mpeg-4 facial animation. In *Proc. Deform'2000* (November 2000), pp. 23–34.

[KT03] KSHIRSAGAR S., THALMANN N. M.: Visyllable based speech animation. *Computer Graphics Forum 22*, 3 (2003).

[LD07] LI Q., DENG Z.: Facial motion capture editing by automated orthogonal blendshape construction and weight propagation. *IEEE Computer Graphics and Applications* (accepted for publication 2007).

[LMDN05] LEWIS J. P., MOOSER J., DENG Z., NEUMANN U.: Reducing blendshape interference by selected motion attenuation. In *Proc. of I3DG'05* (2005), pp. 25–29.

[LTW95] LEE Y., TERZOPOULOS D., WATERS K.: Realistic modeling for facial animation. In *Proc. of ACM SIGGRAPH'95* (1995), pp. 55–62.

[MDCO06] MCDONNELL R., DOBBYN S., COLLINS S., O'SULLIVAN C.: Perceptual evaluation of lod clothing for virtual humans. In *Proc. of SCA'06* (2006), pp. 117–126.

[MNO07] MCDONNELL R., NEWELL F., O'SULLIVAN C.: Smooth movers: perceptually guided human motion simulation. In *Proc. of SCA'07* (2007), pp. 259–269.

[NN01] NOH J.-Y., NEUMANN U.: Expression cloning. In *Proc. of ACM SIGGRAPH'01* (2001), pp. 277–288.

[OD01] O'SULLIVAN C., DINGLIANA J.: Collisions and perception. *ACM Trans. on Graph. 20*, 3 (2001), 151–168.

[ODGK03] O'SULLIVAN C., DINGLIANA J., GIANG T., KAISER M. K.: Evaluating the visual fidelity of physically based animations. In *Proc. of ACM SIGGRAPH'03* (2003), pp. 527–536.

[OHM*04] O'SULLIVAN C., HOWLETT S., MORVAN Y., MCDONNELL R., O'CONOR K.: Perceptually Adaptive Graphics. In *STAR - Proc. of Eurographics 2004* (2004), pp. 141–164.

[PHL*98] PIGHIN F., HECKER J., LISCHINSKI D., SZELISKI R., SALESIN D. H.: Synthesizing realistic facial expressions from photographs. *Proc. of ACM SIGGRAPH'98* (1998), 75–84.

[POM99] PANDZIC I., OSTERMANN J., MILLEN D.: Users evaluations: synthetic talking faces for interactive services. *The Visual Computer 15* (1999), 330–340.

[PR00] PANTIC M., ROTHKRANTZ L.: Automatic analysis of facial expressions: The state of the art. *IEEE Trans. on PAMI 22*, 12 (2000), 1424–1445.

[PRM00] PAVLOVIC V., REHG J. M., MACCORMICK J.: Learning switching linear models of human motion. In *NIPS* (2000), pp. 981–987.

[RP03] REITSMA P. S. A., POLLARD N. S.: Perceptual metrics for character animation: sensitivity to errors in ballistic motion. In *ACM Trans. on Graph.* (2003), vol. 22, pp. 537–542.

[SF98] SINGH K., FIUME E.: Wires: A geometric deformation technique. In *Proc. of ACM SIGGRAPH'98* (1998), pp. 405–414.

[SNF05] SIFAKIS E., NEVEROV I., FEDKIW R.: Automatic determination of facial muscle activations from sparse motion capture marker data. *ACM Trans. on Graph. 24*, 3 (2005), 417–425.

[SP04] SUMNER R. W., POPOVIĆ J.: Deformation transfer for triangle meshes. *ACM Trans. Graph. 23*, 3 (2004), 399–405.

[SWCCG06] SCHWANINGER A., WALLRAVEN C., CUNNINGHAM D. W., CHILLER-GLAUS S. D.: Processing of facial identity and expression a psychophysical, physiological and computational perspective. *Progress in Brain Research 156* (2006).

[TKC01] TIAN Y., KANADE T., COHN J.: Recognizing action units for facial expression analysis. *IEEE Trans. on PAMI 23*, 2 (February 2001), 97 – 115.

[VBPP05] VLASIC D., BRAND M., PFISTER H., POPOVIĆ J.: Face transfer with multilinear models. *ACM Trans. Graph. 24*, 3 (2005), 426–433.

[VP06] VALSTAR M., PANTIC M.: Fully automatic facial action unit detection and temporal analysis. In *CVPRW '06: Proc. of the 2006 Conf. on Computer Vision and Pattern Recognition Workshop* (2006), IEEE Computer Society.

[WB04] WANG J., BODENHEIMER B.: Computing the duration of motion transitions: an empirical approach. In *Proc. of SCA'04* (2004), pp. 335–344.

[WBCB05] WALLRAVEN C., BREIDT M., CUNNINGHAM D., BULTHOFF H. H.: Psychophysical evaluation of animated facial expressions. In *Proc. of APGV'05* (August 2005), pp. 17–24.

[WBCB08] WALLRAVEN C., BREIDT M., CUNNINGHAM D. W., BULTHOFF H. H.: Evaluating the perceptual realism of animated facial expressions. *ACM Trans. on Appl. Percep.* (January 2008), 1–20.

[WFM01] WATSON B., FRIEDMAN A., MCGAFFEY A.: Measuring and predicting visual fidelity. In *Proc. of ACM SIGGRAPH '01* (2001), pp. 213–220.

[Wil90] WILLIAMS L.: Performance-driven facial animation. In *Proc. of ACM SIGGRAPH'90* (1990), pp. 235–242.

[WSZP07] WAMPLER K., SASAKI D., ZHANG L., POPOVIĆ Z.: Dynamic, expressive speech animation from a single mesh. In *Proc. of SCA'07* (2007), pp. 53–62.

[ZLGS03] ZHANG Q., LIU Z., GUO B., SHUM H.: Geometry-driven photorealistic facial expression synthesis. In *Proc. of SCA'03* (2003), pp. 177–186.

[ZSCS04] ZHANG L., SNAVELY N., CURLESS B., SEITZ S. M.: Spacetime faces: High-resolution capture for modeling and animation. *ACM Trans. on Graph. 23*, 3 (2004), 548–558.