Facial animation: overview and some recent papers

Benjamin Schroeder

January 25, 2008

Outline

I'll start with an overview of facial animation and its history; along the way I'll discuss some common approaches. After that I'll talk about some notable recent papers and finally offer a few thoughts about the future.

1 Defining the problem

2 Historical highlights

3 Some recent papers

4 Thoughts on the future

Defining the problem

We'll take facial animation to be the process of turning a character's *speech* and *emotional state* into *facial poses* and *motion*. (This might extend to motion of the whole head.) Included here is the problem of modeling the *form* and *articulation* of an expressive head.

There are some related topics that we won't consider today. Autonomous characters require *behavioral models* to determine what they might feel or say. *Hand* and *body gestures* are often used alongside facial animation. Faithful animation of *hair* and rendering of *skin* can greatly enhance the animation of a face.

Defining the problem

This is a tightly defined problem, but solving it is difficult.

We are intimately aware of how human faces should look, and sensitive to subtleties in the form and motion.

Lip and mouth shapes don’t correspond to individual sounds, but are context-dependent. These are further affected by the emotions of the speaker and by the language being spoken.

Many different parts of the face and head work together to convey meaning.

Facial anatomy is both structurally and physically complex: there are many layers of different kinds of material (skin, fat, muscle, bones).

Defining the problem

Here are some sub-questions to consider.

How should the motion be produced? That is, how is the face model defined and what are its capabilities?

How do the constituent parts of speech correspond to facial motion?

What non-verbal expressions are produced during speech, and why?

How does a character’s emotional state affect his or her face?

Parke

Fred Parke created the first 3D parametric model of a human face. The model is discussed in his 1974 dissertation.

> A parametric model for human faces
> Frederic Parke, 1974 Ph.D. dissertation, Utah

Parke

The facial geometry is broken into parts and controlled by parameters - for example, the rotation of the jaw or the direction of an eye's gaze.

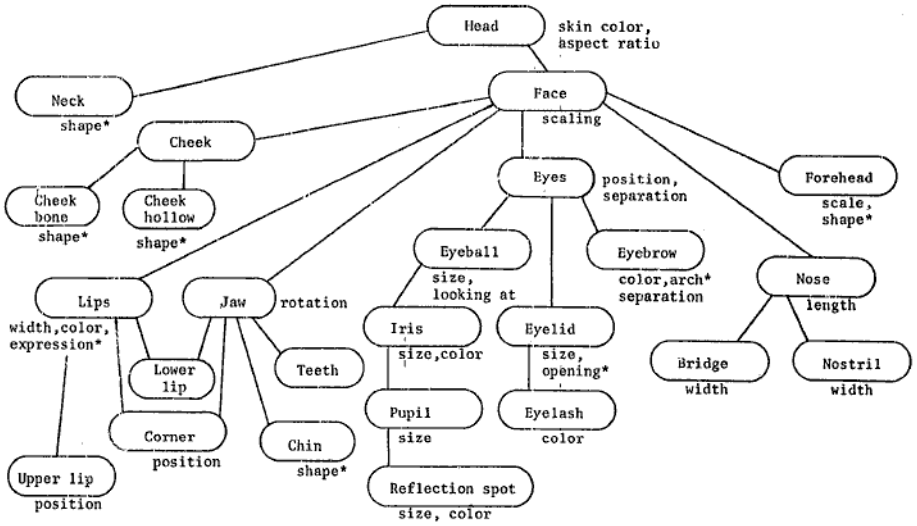


Figure 3.1 - The structure of the parametric model. The parameters affecting the various nodes are shown. An * indicates the use of interpolation to implement the parameter.

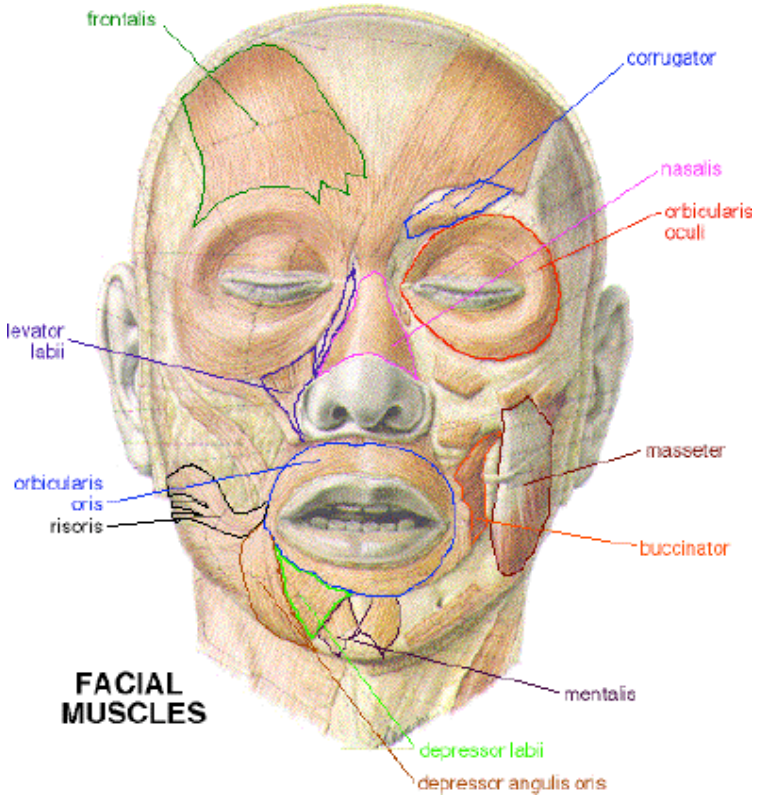
Platt and Badler

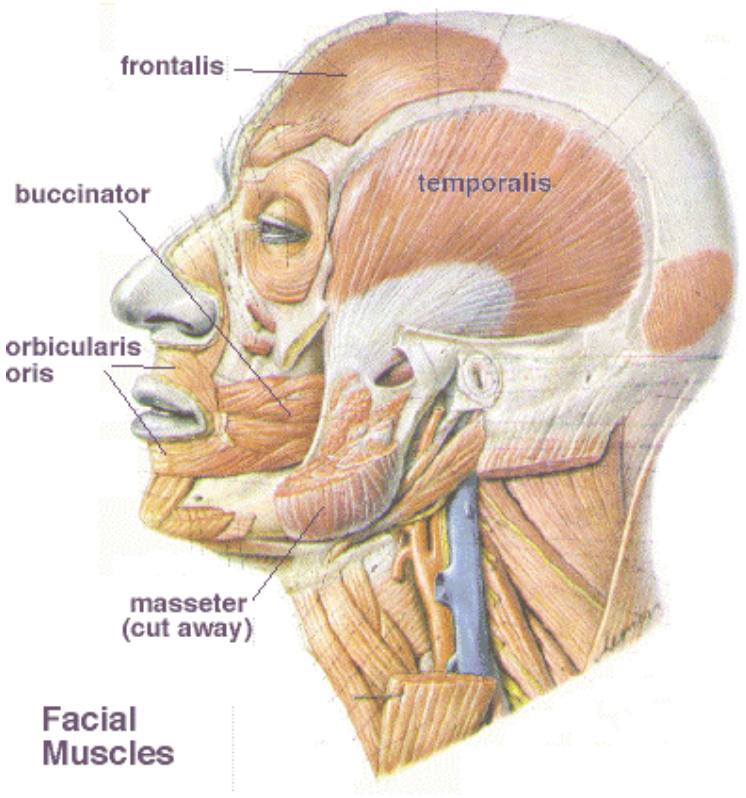
In their 1981 SIGGRAPH paper, Platt and Badler describe how to construct expressions using a muscle-based facial model.

> Animating Facial Expressions
> Platt and Badler, SIGGRAPH 1981

This work used the Facial Action Coding System (FACS), a model from psychology, to determine which muscles to activate in the underlying model.

Sidebar: Muscles of the Face





(Images from http://www.yorku.ca/earmstro/journey/facial.html)

Sidebar: FACS

FACS describes the face in terms of “Action Units”. These may be combined to describe any facial expression.

For example, AU 23 is “Lip Tightener”; AU 19 is “Tongue Out”.

Some of these correspond directly to actions of facial muscles; others involve things like the movement of the tongue or air filling the cheeks.

Facial Action Coding System
Ekman and Friesen, 1978
(The system has subsequently been revised several times.)

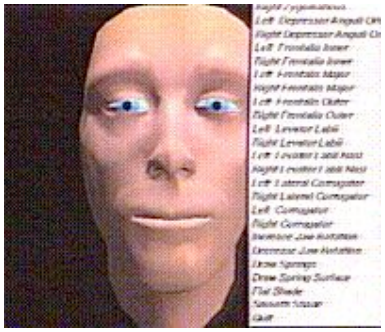
Tony de Peltrie

“Tony de Peltrie” (1985) marked the first time computer facial animation played an important role in telling a story.

The facial expressions were produced by photographing an actor with a control grid on his face, and then matching points to those on a 3D computer face (itself obtained by digitizing a clay model).

Waters

Keith Waters described a more developed, more general muscle model in his 1987 SIGGRAPH paper.



He also used FACS to relate expressions to muscle activation.

A Muscle Model for Animating Three-Dimensional Facial Expression
Waters, SIGGRAPH 1987

Waters

Figure 16
Neutral face with the muscles relaxed

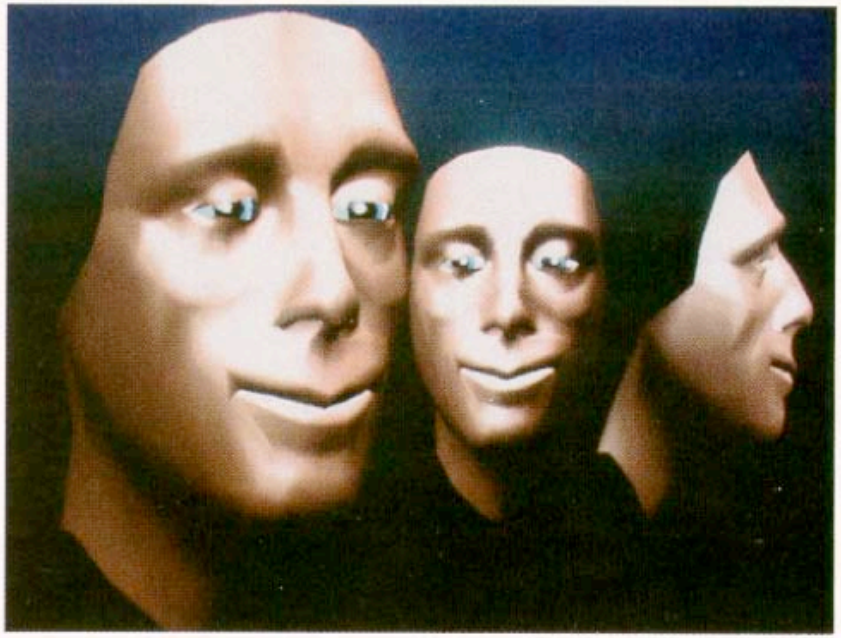


Figure 17
Happiness the corners of the lips are drawn back and raised obliquely by the zygomatic major muscle.

Waters

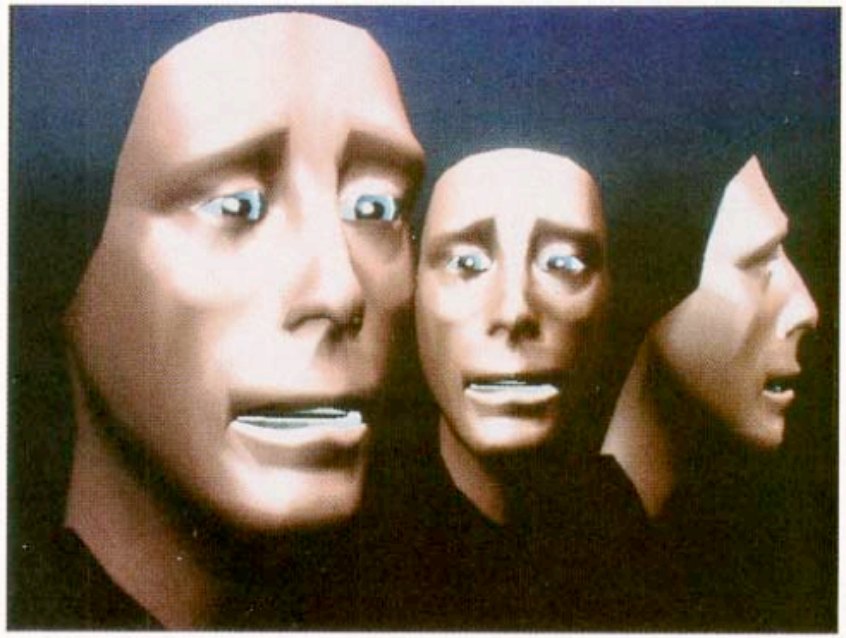


Figure 19
Fear the inner brows are raised by the inner frontalis muscle, the eyes are wide with pupils dilated. The jaw is rotated and the lips drawn back.

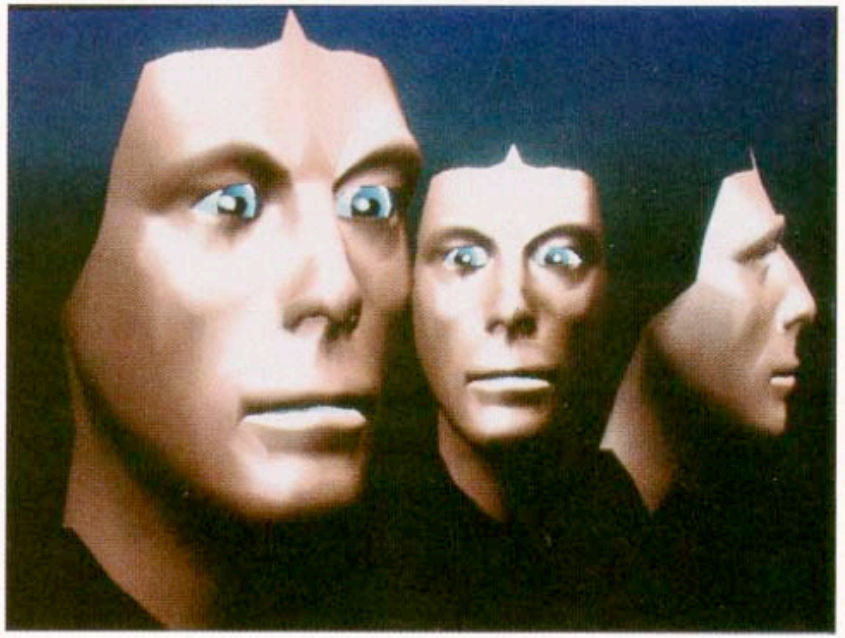


Figure 21
Anger the brows are lowered and the inner part drawn together. The jaw is not rotated and the lips are tight.

Waters

Figure 20
Disgust the alaeque nasi muscle raises the upper lip pulling the skin around the nose and causing the nostrils to dilate.

Figure 22
Surprise the brows are curved and high, the eyelids wide and the pupils dilated.

Tin Toy

The Pixar short “Tin Toy” (1988) was the first computer-animated film to win an Oscar.

The child’s face is animated using a Waters-style model.

It’s interesting to note how effective the toy’s simple, geometric expressions are as well.

J.P. Lewis

Lewis and Parke describe their lip sync system in a 1987 CHI paper. It breaks recorded speech into phonemes and then changes the mouth shape of a parametric model.

Earlier systems had specified speech using text. One advantage of using recorded speech is that it is easy to obtain a natural speech rhythm.

> Automated Lip-Synch and Speech Synthesis
> for Character Animation
> Lewis and Parke, CHI 1987

Sidebar: Phonemes

Phonemes are logical parts of words. For example, the first phoneme in the word “rip” is /r/; the first phoneme in “fun” is /f/, which is also the first is the first phoneme in “physics”.

Note that the same phoneme might have several slightly different sounds (or *phones*) due to context.

Phonemes are language-specific.

Cohen and Massaro

Cohen and Massaro (in 1990) also produced a lip-sync system using a parametric model and studied it in the context of speech perception. They later extended the model to include a tongue and to model coarticulation effects.

Synthesis of Visible Speech
Cohen and Massaro, 1990

Perception of synthesized audible and visible speech
Cohen and Massaro, 1990

Modeling coarticulation in synthetic visual speech
Cohen and Massaro, 1993

Sidebar: Coarticulation

Coarticulation refers to the way visual speech changes based on surrounding segments.

Cohen and Massaro (1993) give the examples of the articulation of the final consonant in “boot” and “beet” - backward coarticulation - and the way the lips round at the beginning of “stew” in anticipation of the “t”.

Pelachaud and Badler

When a person is speaking, their emotional state and the intonation used affects the expression on their face and the mouth shapes used to produce speech.

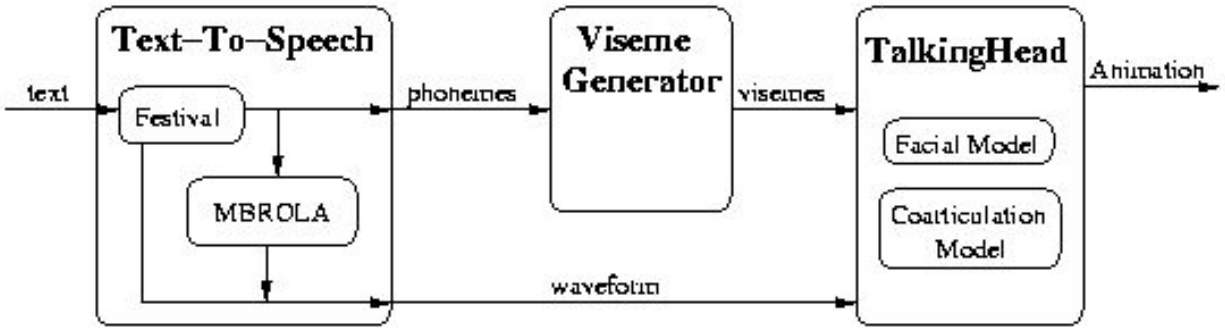
Pelachaud, Badler, and Steedman considered these effects, as well as coarticulation, producing a Platt-style system that incorporates such things as blinks, head movements, and modified mouth motion.

Linguistic Issues in Facial Animation
Pelachaud, Badler, and Steedman, Computer Animation 1991

Generating Facial Expressions for Speech
Pelachaud, Badler, and Steedman, 1994

King

King and Parent developed highly deformable models of the lips and tongue to support realistic speech animation. They consider coarticulation effects, following Cohen and Massaro, but notably convert phonemes to curves instead of single keyframe targets, more accurately modeling the action of the lips and tongue.



A Facial Model and Animation Techniques for Animated Speech
Scott A. King, Ph.D. thesis, 2001, OSU

Somasundaram

Somasundaram and Parent produced a system which synthesizes emotion-laden animation and audio from neutral input audio. The system includes a coarticulation model that takes emotion into account. It makes use of a muscle model and motion capture data in the synthesis of emotion.

A Facial Animation Model for Expressive Audio-Visual Speech
Arun Somasundaram, Ph.D. thesis, 2006, OSU

Image 2: *Expressive visemes created by activating expression muscles in neutral viseme poses.*

BEAT

BEAT, the Behavior Expression Animation Toolkit, produces expressive facial and full-body animation given input text. It discerns intonation and gesture from the content of the text, and is user-extensible.

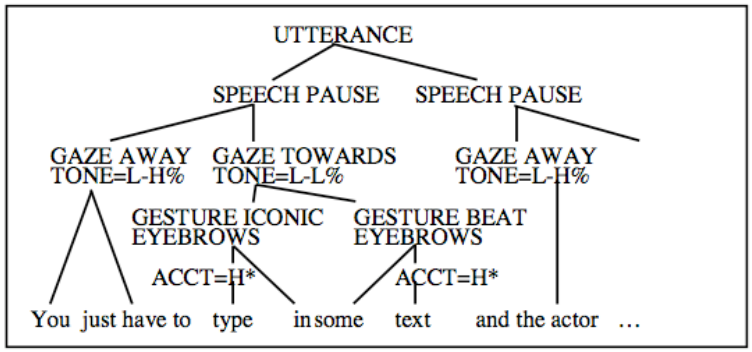


Figure 9. Part of the output XML tree for first example

Figure 10. "You just have to type in some text…"

BEAT: the Behavior Expression Animation Toolkit

Cassell, Vilhjálmsson, and Bickmore, SIGGRAPH 2001

FacEMOTE

The EMOTE system described a way to specify gestures using Laban Movement Analysis, a system for describing motion that evolved from dance. EMOTE maps the components of LMA into movement of a parametric model.

FacEMOTE is an extension of EMOTE to handle facial animation.

> The EMOTE model for effort and shape
> Chi, Costa, Zhao, and Badler, SIGGRAPH 2000
>
> FacEMOTE: Qualitative Parametric Modifiers for Facial Animation
> Byun and Badler, SIGGRAPH 2002

FacEMOTE

Here are some sample elements of the Effort component of LMA as it might be applied to facial animation.

Space	Indirect	Scanning the party floor. Rolling the eyes with no particular focus.
	Direct	Focusing on a ball player at the ball field. Squinting at the object an artist is drawing. Blowing out a candle.
Weight	Light	Whispering to a child to sleep. Lightly tickled into giggling. Whining in a muffled sound. Licking ice cream.
	Strong	Spelling out a word at a spelling bee. Snarling at an offender. Putting on a stern face when scolding a child.
Time	Sustained	Relaxed expression while daydreaming. Taking a deep breath. Yawning.
	Quick	Nervous fidgeting. Coughing. Clearing the throat. Sobbing of a child after a screaming fit.
Flow	Free	Crying of a baby when it is hungry. Bursting into uncontrollable laughter. Shouting in raging fury.
	Bound	Holding back tears. Chuckling instead of laughing loudly. Grimacing when touching a repulsive object.

Table 1. Effort elements

RUTH

RUTH, the Rutgers University Talking Head, is a system for producing conversational sound and animation given input text tagged with information about gesture and intonation. This might come from manual input, a system like BEAT, or from natural-language generation.
RUTH uses a parametric model with King-style coarticulation.

((far ((register "HL") (accent "L+H*") (jog "TR")))
(greater ((accent "!H*") (tone "H-") (blink) (jog)))

Making discourse visible: Coding and animating conversational facial displays

DeCarlo, Revilla, Stone, and Vendetti, Computer Animation 2002

Finer-grained Emotion Modeling

Albrecht et al. present a model for producing finer-grained emotions than those previously considered. They create new emotions by combining basic emotions on a polar-coordinate (“wheel”) model.

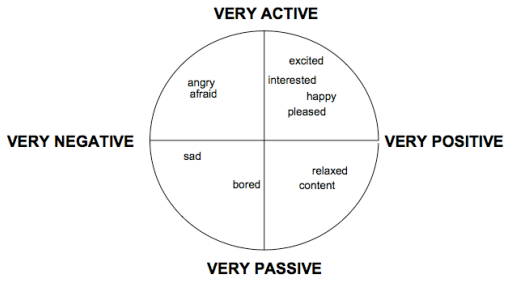


Fig. 1 The two-dimensional, disk-shaped activation-evaluation space proposed by Cowie et al.

Mixed feelings: Expression of non-basic emotions
in a muscle-based talking head
Albrecht, Schröder, Haber, Seidel, Virtual Reality 2005

Finer-grained Emotion Modeling

The synthesized emotions are modified according to additional parameters. From Schröder's website:

I believe that representing emotions in terms of "basic" emotion categories, such as "anger", "fear", "joy" etc., is not the most useful way to obtain a flexible speech synthesis system capable of expressing emotions. Instead, I argue in favour of **emotion dimensions** as a simple means for capturing basic properties of the emotional state in a gradual way. The emotion dimensions generally agreed upon as being most basic are "**activation**" (or "arousal", i.e. the readiness to act in some way) and "**evaluation**" (or "valence", "pleasure", in terms of positive/negative, liking/disliking). In social interaction settings, a third dimension "**power**" (or "control", "dominance", the social status) has shown to be useful.

Finer-grained Emotion Modeling

anxiety	**fear**	**panic fear**
$a = 8$	$a = 14.8$	$a = 20$
$e = -24.1$	$e = -44.4$	$e = -60.1$
$r = 25.3$	$r = 46.8$	$r = 63.4$
$\omega = 288.4$	$\omega = 288.4$	$\omega = 288.4$

Fig. 3 *Anxiety* and *panic* belong to the same fundamental class as *fear*, but differ in intensity. Therefore their facial expressions can be generated from fear by scaling with the ratio of the radii. The angle on the emotion disc is kept fixed for both new expressions, while the radii are varied, thereby yielding new values for activation and evaluation.

Finer-grained Emotion Modeling

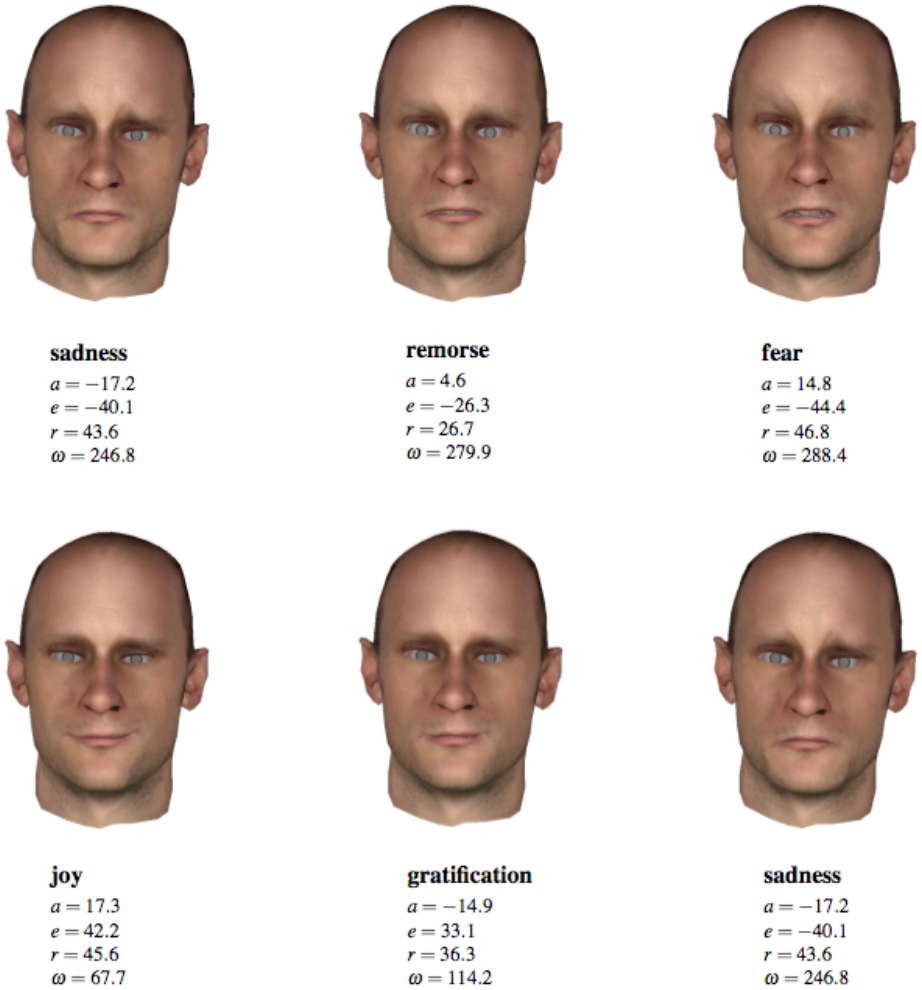


Fig. 4 The emotional expression in the middle has been obtained from those at the left and right using the blending algorithm. The radius r and the angle on the emotion disc ω determine the influence of each generating expression and hence the degree of similarity to the new one. The coordinates in emotion space have been obtained from the NECA data.

Finer-grained Emotion Modeling

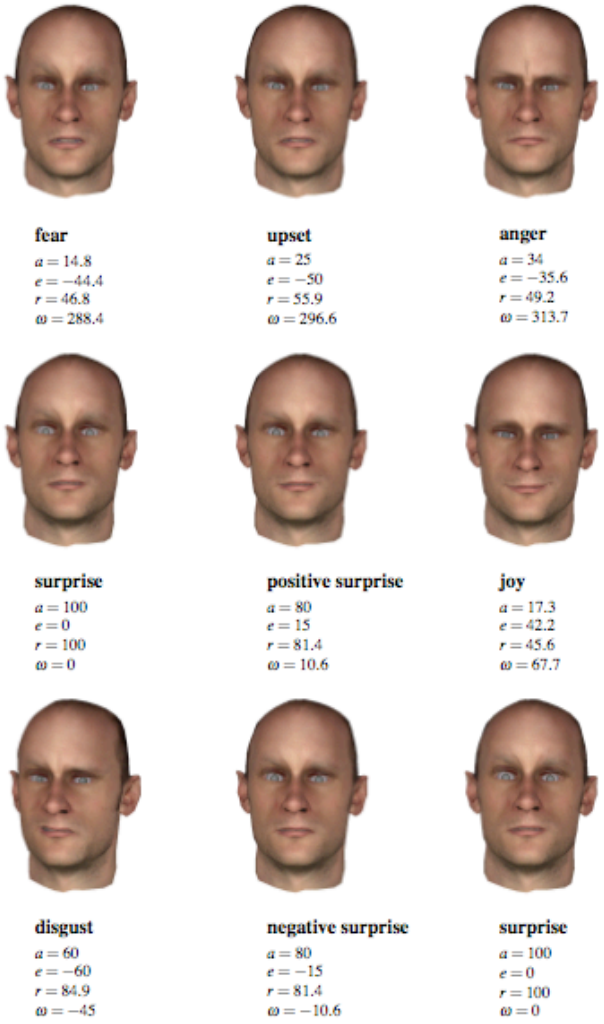
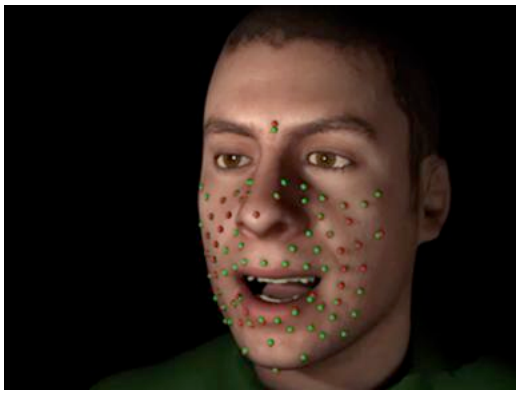
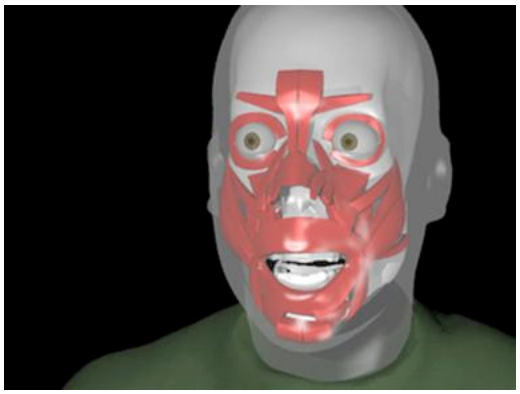


Fig. 5 The emotional expression in the middle has been obtained from those at the left and right using the blending algorithm. The radius r and the angle on the emotion disc ω determine the influence of each generating expression and hence the degree of similarity to the new one. The first example is a not too active, but rather negative emotion, while the second one could be pleasant surprise, and the last one unpleasant surprise.

Physics-based Muscles and Mocap

Sifakis, Neverov, and Fedkiw present a muscle-model based head with a physically-simulated flesh model. They determine muscle activations using motion capture data. The combination of the muscle model and simulated flesh allows for retargeting to novel physical situations.





Automatic Determination of Muscle Activations
from Sparse Motion Capture Data
Sifakis, Neverov, and Fedkiw, SIGGRAPH 2005

Physics-based Muscles and Mocap

This model was later extended for use in lip-sync animation. Mocap data is segmented to determine muscle activations for individual phonemes.

Simulating Speech with a Physics-Based Facial Muscle Model

Sifakis, Selle, Robinson-Mosher, and Fedkiw, SCA 2006

Reducing Blendshape Interference

Blendshapes are popular and (of course) easy to adapt to many models, but suffer from poor orthogonality of slider controls. Lewis et al propose allowing animators to hold a set of points more-or-less fixed while adjusting one slider, and solve for values for the other sliders so that those values will be retained.

Fig. 1: (a) We attempt to mimic the "Jack Nicholson" expression of partially closed eyes with an arched eyebrow. First the eyelids are partially closed.

(b) The model has three controls over eyebrow shape. The desired arched eyebrow is easily obtained, but the eyelid is changed as a side effect.

(c) The model is capable of approximating the desired expression however, by readjusting the eyelid control (or, by using our technique).

Reducing Blendshape Interference by Selected Motion Attenuation

Lewis, Mooser, Deng, and Neumann, I3D 2005

Learning Coarticulation

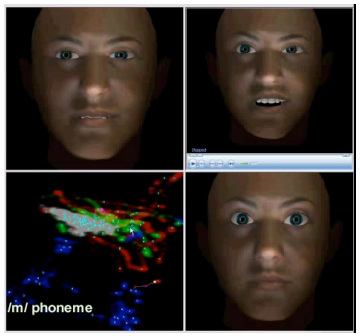
Deng et al. present a system that uses controlled mocap data to learn both a coarticulation model and a phoneme-independent model of non-verbal facial expressions. The models may then be used to synthesize novel speech.

Expressive Facial Animation by Learning Coarticulation and Expression Spaces

Deng, Neumann, Lewis, Kim, Bulut, and Narayanan, TVCG 2006

Controls for Use of Mocap

Segmented motion capture data may be used to generate speech animation from novel input. Deng and Neumann present a technique and system to give animators more control over what motion is used for each phoneme. “Hard constraints” allow direct specification of motion, and “soft constraints” specify that a certain emotion should be used if possible.



eFASE: Expressive Facial Animation Synthesis and Editing with Phoneme-Isomap Controls

Deng and Neumann, SCA 2006

Reconstructing Facial Models

Kähler, Haber, and Seidel present a use of their muscle model in forensics: for reconstructing facial models using scanned skull data.

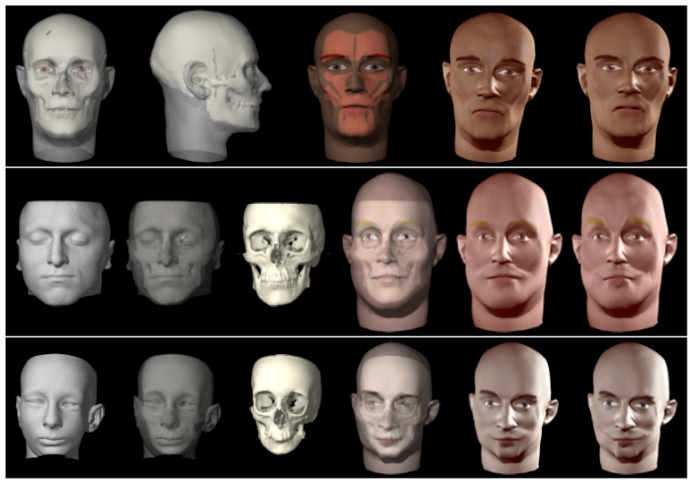


Figure 9: Examples of facial reconstructions created with our system. Top: model created from a scanned real skull, showing fit of skin to skull, transferred muscles, and two facial expressions. Middle: Reconstruction from a volume scan of a male, showing the actual face as contained in the data, superimpositions of the actual and the reconstructed face with the skull, and the reconstruction with neutral and "worried" expression. Bottom: Reconstruction from volume scan of a female with strong skull deformations. The CT data sets don't contain the top and bottom of the heads, thus the source skull and face models are cut off. The actual head height had to be guessed in these cases.

Reanimating the Dead: Reconstruction of Expressive Faces from Skull Data

Kähler, Haber, Seidel, SIGGRAPH 2003

Wrinkles from Mocap and Video

Bickel et al. present a method for blending mocap and video data with a high-resolution face scan to add wrinkles to a face. The resulting model is not “drivable” to produce synthetic expressions or speech, but I include this here to show the compelling expressions resulting from the high-resolution scan and wrinkle model.

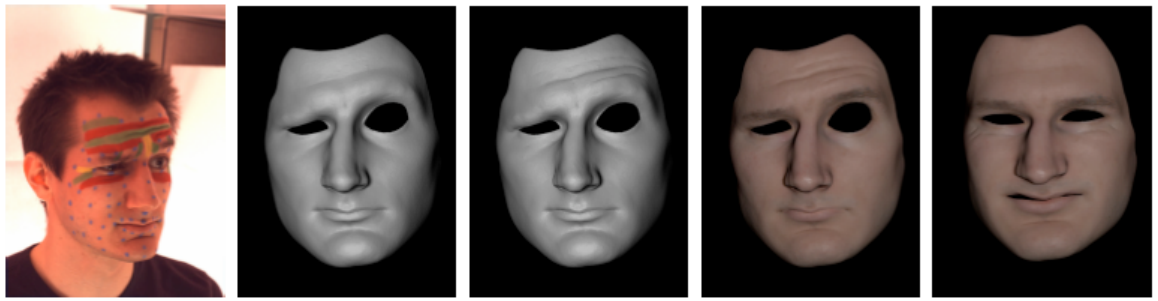


Figure 1: *Animation of a high-resolution face scan using marker-based motion capture and a video-driven wrinkle model. From left to right: video frame, large-scale animation without wrinkles, synthesis of medium-scale wrinkles, realistic skin-rendering, different expression.*

Multi-Scale Capture of Facial Geometry and Motion
Bickel, Botsch, Angst, Matusik, Otaduy,
Pfister, and Gross, SIGGRAPH 2007

The Future?

In the near term, I think we can expect to see more capable faces along several mostly incremental lines, some which continue research we've seen. For example:

better models of non-verbal expressions;

use of captured data for direct playback or model control;

models for multiple languages;

speech changes in relation to simple environmental changes (a la Sifakis's lollipop example);

control systems for practicing animators.

The Future?

Further out, many problems remain to be explored. Here are a few thoughts on some that we could see in the medium-term:

deeper models of auxiliary actions: swallowing, tics, illness;

more integration with the body and environment - say, scratching the face or batting at flies;

more modeling of realistic tissue and skin, especially things like fat, wrinkles, and facial hair;

relatedly, modeling of faces with different ages, such as children or the elderly.

The Future?

Consider the wide range of expression seen in cartoons. This expression is often at the same time simplified and exaggerated.

What can we do to facilitate this kind of animation?

The Future?

Of course, we don't need to go as far as cartoons to see expressiveness. The realistic human face offers plenty of examples, from the subtle to the absurd.

Some Web References

Cohen and Massaro's lab has a good (if outdated) general reference page.

http://mambo.ucsc.edu/psl/fan.html

Wikipedia currently has a good overview of the area with several jumping-off points.

http://en.wikipedia.org/wiki/Computer_facial_animation