

Facial Image Synthesis<sup>1</sup>

Barry-John Theobald and Jeffrey F. Cohn

**1 Introduction**

Facial expression <cross-ref to facial expression of emotion > has been central to the study of emotion < cross-ref to emotion > for over a hundred years (Darwin, 1872/1998). Much of what we have learned was made possible by technological breakthroughs: photography in the nineteenth century and film and later video in the twentieth. Today, two new technologies are just beginning to make their potential felt. These are automated facial image analysis and facial image synthesis. A recent review of the former can be found in (Cohn & Kanade, in press). Here we summarize major approaches to facial image synthesis of identity and static and dynamic changes in facial expression. Synthesis of dynamic sequences is referred to as animation.

The anatomical structure of the face consists of layers of bone, muscle, subcutaneous fat, and skin. Facial expression results from complex, non-linear interactions among these layers. For the animator attempting to simulate facial appearance or expression, the question is how best to represent these layers and model interactions between them.

These questions can be approached from three perspectives. One utilizes computer graphics, a second image processing, and a third integrates both. Graphics-based approaches offer a high degree of flexibility, are computationally inexpensive, and can be integrated easily into animation systems for full-bodied avatars. Most commercial systems use this approach. Image-based approaches are computationally more expensive, more

limited in the range of facial expression, and cannot animate full-body avatars. They can, however, produce stunning realism. Hybrid approaches that combine computer graphics and image processing are just emerging.

**2 Computer graphics-based synthesis**

Graphics-based approaches represent the surface of the face as vertices in a three-dimensional (3D) space. The vertices are connected to form a triangulated mesh that approximates the skin. By manipulating the position of the vertices, changes in face appearance and expression result. Early systems adopted a *key-frame* approach, in which the animator painstakingly specified the movement of each vertex at set points in time. The vertex positions then were interpolated to generate the *in-between* frames (a technique known as *tweening*). A breakthrough occurred when animators learned to control groups of vertices using a single control parameter. All graphics-based approaches synthesize images by acting on the facial mesh either directly or indirectly.

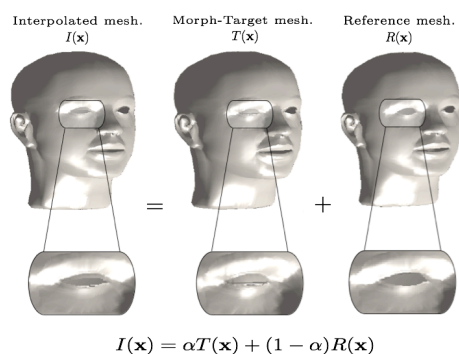
**2.1 Directly parameterized models**

Directly parameterized models make no attempt to represent the detailed anatomical structure of the face. Instead, a collection of meshes, known as morph-targets, is defined that specify changes in a mesh relative to a default pose (Figure 1). Interpolating the morph-targets and combining the resultant facial gestures produces facial expressions. Each parameter of a directly parameterized model controls the amount of interpolation between the reference mesh and a morph-target. An advantage of this approach is that the parameters have a direct, physical, intuitive meaning (e.g., lip-rounding or eye-blinking), which makes them attractive for non-experts in facial kinematics.

**2.2 Indirectly parameterized models**

Whereas the directly parameterized approach models only the changes visible on the face surface, the indirectly-

<sup>1</sup> Acknowledgements: Preparation of this manuscript was supported in part by grants NIMH R01-501435, NSF HSD-0(49)527444, and EPSRC EP/D0490751. Barry-John Theobald is with the University of East Anglia, Jeffrey F. Cohn with the University of Pittsburgh. This chapter is to appear in D. Sander & K. R. Scherer (Eds.). *Oxford companion to affective sciences: An encyclopedic dictionary for the affective sciences*. NY: Oxford University Press.



**Figure 1.** Illustration of directly parameterized facial animation. Each image consists of 5,828 vertices in a 3D space, connected to form 11,370 triangles. The image to the right is a reference face model in a default pose. The center image illustrates a morph-target, where the right (from the point of view of the model) eye-lid has closed. The image to the left is a weighted average of the vertices of the reference and morph-target meshes. A complete facial animation system would contain a sufficient number of morph-targets to define all facial actions of interest. The greater the number of morph-targets, the more subtle the animation produced. Adapted from (Pighin, Hecker, Lischinski, Szeliski, & Salesin, 1998).

parameterized approach models interactions among all layers of the face. For this reason, they often are referred to as physically-based models. They have two advantages over directly parameterized models:

- Because the parameters do not act directly on the mesh, the model is not tied to a particular mesh topology. The mesh can be modified without the cost of updating parameters, and any number of meshes can be animated using a single parameterization.
- Face models can be defined in terms of the Facial Action Coding System (FACS) (Ekman, Friesen, & Hager, 2002) < cross ref to Facial Action Coding System >. Parameters can be designed to have a one-to-one mapping to FACS action units < cross-ref to Action Units >.

Disadvantages include: One, because knowledge of facial kinematics is assumed, non-experts may require trial-and-error to generate realistic expressions. Two, while muscle vectors are

not tied to a particular mesh topology, they must be *married* to a mesh before they can be used. The location of each muscle within the model must be defined, and errors can produce unexpected results. Three, physically-based animation is more computationally expensive than directly parameterized animation, although with advances in hardware this difference fades.

### 2.3 Summary of graphics-based synthesis

For both directly parameterized and physically-based animation, long sequences of realistic and complex expression are difficult to generate. Key-framing even for groups of vertices can quickly become tedious as each parameter must be manually specified in all key-frames. A more efficient but less flexible alternative is to capture motion data from the face of an actor and then use a predefined correspondence between points on the actor's face and the vertices of the mesh model.

Computer graphics-based models are usually adopted when video-realism is not required or when an application must be computationally efficient. They are particularly suited to web-based applications as they require relatively low bandwidth. When video-realism is required, image-based methods are usually preferable.

### 3 Image-based synthesis

In contrast to graphics based approaches, image-based synthesis uses images of real faces rather than a geometric model. Image-based animation uses a number of images to capture subtle variation in face shape and appearance. The key is how to create a realistic transition from one expressive image to another. Two main approaches are *image morphing* and *image concatenation*.

#### 3.1 Image morphing

Image morphing is similar to traditional graphics-based key-frame animation using morph-targets. However, rather than representing key-frames as mesh poses, static images of pre-specified facial expressions are selected. The pixel-



**Figure 2.** Example of image synthesis by 3dMM (Blanz, 2006) (© 2006 IEEE).

values within key-frames are interpolated to create the in-between frames.

The pixels in the key-frames are interpolated using *optical flow*. This approach involves a dense pixel-wise morph between the images. Alternatively, the approach can be area based, where a coarse mesh is divided into triangles, and the pixels in one triangle in one image are mapped to the corresponding triangle in a second image.

Image morphing has been extended to 3D with great effect. One example (Pighin, Hecker, Lischinski, Szeliski, & Salesin, 1998) successfully synthesized facial expression by capturing eight expressions in five different poses, recovering the 3D geometry, then interpolating the static expressions. The image quality was greatly improved by blending the original image from the different viewpoints.

### 3.2 Image concatenation

The idea underpinning image concatenation is to mimic *flip-book* animation, where each page of a book has a picture of an object in a slightly different position. Flipping the pages quickly enough gives the illusion that the object is moving seamlessly around the pages. Animation based on image concatenation is similar. Images from a video sequence are re-ordered to generate new sequences. The challenge is to select images that match the desired expression at each point in time while maintaining a smooth transition from one to the next.

### 3.3 Summary of image-based synthesis

The main advantage of image-based synthesis is photorealism - images of real faces are used. However, image-based synthesis has several limitations:

- The range of expressions is limited to those available in a

database or video sequence. While the expression space can be extended by capturing new images, acquisition conditions (e.g. the lighting) must be identical to those of the original recording.

- To animate a new person's face requires the capture of complete training data for that person.
- Animation is generally limited to full-frontal views.

## 4 Hybrid graphics and image synthesis

With suitable hardware for capturing the shape and reflectance properties of the face, stunning animated sequences can be generated by coupling image-based and graphics-based animation. An essentially image-based face model can be viewed in different poses and under a number of lighting conditions.

Two such hybrid approaches are 3D morphable models (3dMMs) (Blanz, 2006) and active appearance models (AAMs) (Xiao, Baker, Matthews, & Kanade, 2004). The idea behind both is to model both the shape (graphics-based synthesis) and appearance (image-based synthesis) of the face.

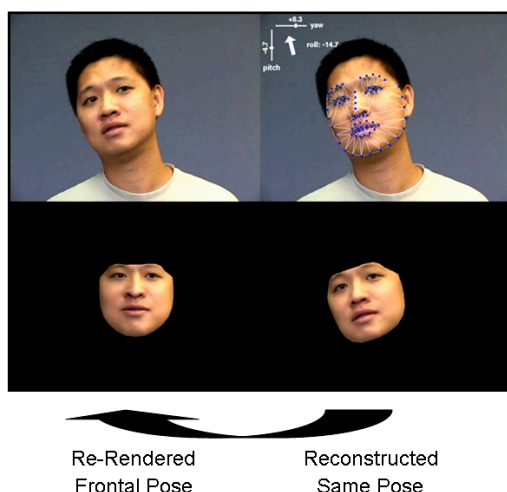
Morphable models are learned from Cyberware scans of human faces. The scan provides both an image and the vertices of a dense mesh sampled at tens of thousands of points on the face surface. Faces are represented as a linear combination of training faces; new faces are synthesized by applying a weighted combination of faces

from the training faces. 3dMMs can synthesize photo-realistic images of faces that have altered lighting, pose, identity, and expression. Figure 2 shows examples of variation in face characteristics.

AAMs lack the photorealism of 3dMMs but have two important advantages. They use normal video rather than special-purpose scanning. And AAMs process video at frame rate, which makes real-time applications possible. Recent uses include synthesis of the visual aspects of speech production (Theobald, Bangham, Matthews, & Cawley, 2004) and pose normalization and expression synthesis (Figure 3).

## 5 Conclusion

Approaches to facial animation include graphics-based, image-based, and hybrid. Modern graphics-based systems are attractive as they are able to animate a wide range of facial identity and expression, are computationally efficient, and can animate the face of a full-bodied virtual character. Their major limitation is lack of photorealism. Image-based techniques can produce sequences that approach real video but lack the efficiency and flexibility of graphics-based approaches. Commercial packages tend to be graphics-based. Hybrid techniques may offer improved realism, efficiency, and flexibility.



**Figure 3.** Example of image synthesis by AAM. Upper left is original image. Upper right is 3D shape overlaid onto original image. Bottom row shows reconstructed (i.e. synthesized) views. From (Xiao, Baker, Matthews, & Kanade, 2004). (©2004 IEEE).

With recent developments in facial animation, exciting possibilities emerge for research on the facial expression of emotion. One, it becomes possible for the first time to separate stable characteristics of a facial image, such as those associated with sex, race, or attractiveness, from the dynamics of facial expression. Faces with different appearance, such as those of men and women, may all be animated using the same dynamics. Stable and dynamic characteristics no longer need be confounded. Two, also for the first time, hypotheses about rapid facial actions can be experimentally tested. Because facial action parameters can be manipulated on the fly, in real time, expressive behavior can be scaled to exaggerate or attenuate actual behavior. One could, for instance, experimentally manipulate the occurrence, amplitude, and timing of the Duchenne marker, micro-expressions, or incongruent facial actions to discover their influence on social dynamics. Until now, such questions could be addressed only by using quasi-experimental approaches or static images. These are only some of the topics that may prove fruitful to investigate with these new tools.

## Further reading

(Blanz, 2006; Cohn & Kanade, in press; Massaro, 1998; Parke & Waters, 1996)

## 8 References

- Blanz, V. (2006). Computing human faces for human viewers: Automated animation in photographs and paintings *Proceedings of the IEEE International Conference on Multimodal User Interfaces*, Banff, Canada, 249-256.
- Cohn, J. F., & Kanade, T. (in press). Use of automated facial image analysis for measurement of emotion expression. In J. A. Coan & J. B. Allen (Eds.), *The handbook of emotion elicitation and assessment*. New York, NY: Oxford.
- Darwin, C. (1872/1998). *The expression of the emotions in man and animals (3rd Edition)*. New York, New York: Oxford University.
- Ekman, P., Friesen, W. V., & Hager, J. C. (Eds.). (2002). *Facial action coding system: Research Nexus, Network Research Information*, Salt Lake City, UT.

- Massaro, D. W. (1998). *Perceiving talking faces*. Cambridge: MIT Press.
- Parke, F. I., & Waters, K. (1996). *Computer facial animation*. Wellesley, MA: A.K. Peters.
- Pighin, F., Hecker, J., Lischinski, D., Szeliski, R., & Salesin, D. (1998). Synthesizing realistic facial expressions from photographs. *Proceedings of the ACM SIGGRAPH*, Orlando, 75 – 84.
- Theobald, B.-J., Bangham, J. A., Matthews, I., & Cawley, G. C. (2004). Near-videorealistic synthetic talking faces: Implementation and evaluation. *Speech Communication*, 44, 127-140.
- Xiao, J., Baker, S., Matthews, I., & Kanade, T. (2004). Real-time combined 2D+3D active appearance models. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Washington, D.C., 535-542.