Siggraph 2005 course notes - Digital Face Cloning

Introduction

Frédéric Pighin J.P. Lewis

University of Southern California

Overview

By far the most challenging aspect of a photoreal actor is the creation of a digital face that can stand up to close scrutiny. The human face is an extremely complex biomechanical system that is very difficult to model. Human skin has unique reflectance properties that are challenging to simulate accurately. Moreover the face can convey subtle emotions through minute motions. We do not know the control mechanism of these motions. All these issues combine to make the human face one of the most challenging object to model using computer graphics.

In recent years, a novel approach to synthesizing realistic faces has appeared. Driven by the need to replace actors in difficult stunts, recent efforts have focused on capturing digital replicas of real actors. This approach, called *Digital Face Cloning*, attempts to create a digital face by replicating in parts the face of a real-performer. Digital face cloning describes the process of capturing an actor's performance and optionally their likeness in a digital model. With this technique, the human face is no longer considered as a biomechanical system but as a real object that can be digitized to produce a synthetic replica. This replication process has been made possible through a set of recently developed technologies that allow the recording of a performer's face. For instance, a face scanner can be used to to recover the geometry and some of the reflectance properties of the face. The motion of a face can be recorded using a motion capture system. Hundreds of points on the face can be tracked and mapped onto a virtual character whose facial motions will mimic those of the performer.

Advances in software technology have also been instrumental for cloning digital faces.

One of the most promising technology for replicating the motion of the human face is dense motion capture technology. Traditionally motion capture systems are limited to at most a few hundred markers on a human face. At this resolution, many of the small scale details (e.g. wrinkles) of the face might be missed. Dense motion capture allows the tracking of tens of thousands of facial locations. Generally the output of dense motion capture is a sequence of detailed face meshes. The meshes can then be registered so that facial features are in topological correspondence throughout the sequence. The end result is high-fidelity three-dimensional representation of the actor's performance.

Equally important is the ability to render faces realistically. Human skin is particularly difficult to render. This difficulty has two origins. First, human skin is somewhat transparent and exhibit multiple scattering effects. Simulating this phenomena requires considering skin as a volumetric object. Second, the skin's reflectance varies in space but also in time. Through the contraction and dilatation of blood vessels the reflectance of the skin varies as the face changes facial expression. To make things worse, the appearance of the face is also determined by wrinkles, pores, follicles, and other surface details whose size is often less than a millimeter. Given these challenges recent research has focussed on digitally capturing these properties with

measuring devices. The purpose of these devices is to measure a particular performer's skin BRDF.

In the rest of this course, we will describe in more details the state of the art in facial geometry, motion, and BRDF acquisition. A critical issue in face cloning is the ability to remap motion captured on one performer into a digital character. This issue, called cross-mapping will be discussed in the course.

Finally and most importantly, we will present a few practical examples of successful digital face cloning systems that have been used in production.

In the rest of this introduction, we provide a brief historical overview of the field and we discuss some major issues related to digital face cloning.

History

The idea of creating artificial humans greatly predates computer graphics. It has haunted popular imagination well before computers were invented. For instance, in Jewish folklore the Golem is an artificially created human supernaturally endowed with life. More recently Mary Shelley's Frankenstein is a classic of gothic literature.

From a computer graphics point of view, we can look at the progression of digital face cloning from two perspectives. First, the film industry has long tried to include synthetic actors in movies. This has been motivated primarily by the need to replace actors in difficult stunts, but other purposes are becoming more common, including the reproduction of persons that are no longer living. Second, we will survey university research on this problem and consider cloning techniques that may be adopted in the future.

Film industry

The emergence of believable computer graphics (CG) actors remains a significant challenge in our field, and a successful demonstration of photoreal actors would be considered a milestone.

In fact considerable albeit gradual progress towards this goal has already occurred, and a singular "breakthrough" appearance of virtual actors is perhaps unlikely. CG humans have appeared in movies since *Futureworld* (1976), and CG stunt doubles have made distant and brief appearances in films in the 1990s including *Terminator 2* (1991), *Jurrasic Park* (1993, body only) and *Titanic* (1997). In recent years these stunt doubles have had somewhat larger and longer (though still brief) appearances, and have taken the form of recognizable actors in films such as *Space Cowboys* (2000) and *Enemy at the Gates* (2001). The current state of the art is represented by digital stuntpeople in *Spiderman II*, *The Matrix: Revolutions* (with a more-than-10-second full-screen shot of two virtual clones), and *Lemony Snicket* (with a number of CG shots of a baby intercut with the real baby). The corresponding technical developments, including implementations of lighting capture, subsurface scattering, and dense motion capture, are outlined in Siggraph sketches [19, 9, 2].

Although the progress in the last decade has been substantial, there is even further to go. Current technology can produce only brief cloned shots at considerable cost, and although the results are remarkable examples of computer graphics, they do not actually deceive a majority of observers for more than a few seconds. Moreover, several large scale film industry attempts to produce a CG "lead" character have been attempted and subsequently abandoned. These include a proposed remake of *The Incredible Mr. Limpet* in 1998, and Disney's Gemini Man attempt in 2000.

Academia

Facial animation has long been a very active research area. We do not pretend to cover all relevant research in this section but provide a broad historical overview of the field. More details studies can be found in other sections of the course notes.

We can tentatively date the beginnings of research on digital face cloning with the use of photographs for face modeling. One of the earliest efforts in this field was the pioneering system developed by Fred

Parke [14, 15] who made use of two orthogonal photographs and patterns painted on a performer's face to recover 3D facial geometry. Later Pighin et al. [16] extended this technique to an arbitrary number of images and texture extraction. With the advent of range scanning researchers have explored the use of facial scan to automatically model faces. The resulting range and color data can be fitted with a structured face mesh, augmented with a physically based model of skin and muscles [11, 10, 20, 22]. Haber et al. [7] propose a similar system using an improved facial model.

The idea of recovering facial geometry from images led naturally to motion estimation from video. An early system by Lance Williams [23] tracks the 2D face motion of a performer from a single video stream. Later, Brian Guenter et al. [6] extend this approach to 3D recovery from multiple video streams. Other researchers have explored marker-less tracking techniques, often relying on more sophisticated models such as blend shape (linear) models [17, 1], bi-linear models [21], or physically-based models [20]. In this perspective, a promising new research direction is the recovery of dense geometry from video. The UCAP system [3] used five hi-def cameras to recover the facial geometry and texture at every frame (requiring manual assistance, but no structured light), whereas the system developed by Li Zhang et al. [24] uses six video cameras and two structured light projectors to recover the facial geometry completely automatically.

To enable rendering from a different point of view and with a different lightning environment, researchers have explored how to capture the reflectance property of a face. For instance Marschner et al. [13] measured the human skin BRDF using a small number of images under different lighting conditions and re-used these measurements to realistically render a performance driven face [12]. In a more data-driven approach, Debevec et al. [4] used an image-based technique to render human faces in arbitrary lighting environments by recombining a large set of basis images from different lighting directions. Finally, Tim Hawkins et al. [8] extended this work by capturing the performer in a variety of expressions to build a blend shape model that includes reflectance information.

Main issues

The process of digital face cloning can be broken into two main steps. First, during the *recording step*, the face of the performer is recorded. This recording can take several aspects depending on how the recorded data will be used. To build an accurate replica, the geometry, the reflectance, and the motion of the performer all need to be recorded.

Second, during the *synthesis step*, the recorded data is reused to create a digital sequence. In the simplest case, only a specific performance needs to be captured and synthesized. In such a case there is usually no much need to modify the recorded data. In more complex cases, the data might need to be extensively modified. For instance, this may be because the data is remapped onto a different character, or because the character is placed in a different lighting environment, or maybe the motion has to be modified.

In an abstract sense, the data collected represent a set of samples from a space of facial properties. For instance, a face scan might be considered as a sample from a space of facial expressions, a recorded motion would belong to a space of motions. During the second step, these samples are then used to reconstruct portions of that space. In this perspective, the synthesis step can be seen as a resampling problem where the original samples are altered to meet the needs of a production. How this resampling is done is key to the process of digital face cloning. In general, this raises the issue of figuring out how the recorded properties of the face behave between the samples.

For example let us consider the space of facial expressions for a performer: we might have scanned his face in a neutral expression and in a angry expression. To generate an intermediate expression between these two sampled expressions we have to make some assumptions about how the face behave between the samples. These assumptions form what we will call a *representation* for the data. For instance, a BRDF representation of the skin might be used to resample reflectance data, or frequency decomposition might be used to modify facial motions. Another interesting example is a blend shape model. In this case resampling is done using correspondences between facial features (i.e., domain knowledge) and linear interpolation.

Many representations or resampling techniques can be used to modify the data. Some of them are very simple such as nearest neighbor, where the closest (according to some appropriate distance metrics) is selected, others can be quite complex and for instance involve physically-based modeling. In general, there seems to be a wide spectrum of representations starting with the *weak representations* (point sampling, linear interpolation) up to the *strong representations* (biomechanics, physics-based models). Choosing the best representation depends on many factors. This decision mainly reflects a tradeoff between how many samples need to be recorded versus how complex the representation is. If the reconstructed space is densely sampled then a weak representation is more appropriate since the behavior between the samples can usually be modeled using some simple approximation (e.g. linear approximation). If on the other hand there are few samples to work with then a stronger (or richer) representation might be needed in order to model complex behavior between the samples. Often weak representations are limited to interpolating the samples so that the reconstructed space lies within the convex hull of the samples – whereas with a strong representation, using the same samples, it might be possible to extrapolate from the samples and cover a much wider portion of the space.

As we have discussed the use of weak models often relies on a dense set of samples. Depending on the properties of the face that is modeled obtaining sufficient sample data might be more or less tractable. For instance, it is relatively easy to take photographs of a face using different camera positions (e.g. for view interpolation) but it is significantly more difficult to densely sample the set of potential face motions. It also depends on the extend of the portion of the space that needs to be reconstructed. Obviously the smaller that extent is the easier it is to sample that portion of the space. Finally, the properties of the face that vary little or in smooth ways usually require much fewer samples to be modeled compared to other properties that are less smooth.

The representation issue is not only relevant to the synthesis step but also to the data gathering step. Data gathering sometimes necessarily makes use of a representation. For instance in model-based tracking, a representation of the face is used to track facial motion from a video stream. The representation itself might be based on some initial data (or prior in a Bayesian framework). The representation used for data gathering might not be appropriate for synthesis, in this case a different representation is needed. This raises the question of whether the same representation can be used for analysis and synthesis. One of the interesting issues is the parameterization of the data. Often during synthesis we are interested in choosing parameters that have an intuitive meaning so that they can be manipulated easily by an animator. On the other hand, for analysis, a good set of parameters is sometime selected according to different criteria such as orthogonality. This issue arises for instance in the parameterization of a blend shape model.

The synthesis step dictates what kind of samples need to be collected during the capture step. However, given a particular sample space, it is often unclear which specific samples need to be recorded. If the sampling process is expensive (e.g. motion capture) the issue of how many samples need to be recorded and which to record becomes critical. Unfortunately, often there are no rules to answer these questions. More likely, it is a trial and error process that can be guided by some study of the human face. For instance, even though work in psychology [5] stresses the role of six basic expressions, most blend shape systems use a significant larger number of expressions (e.g. 946 for Gollum in the Lord of the Rings [18]).

References

[1] V. Blanz, C. Basso, T. Poggio, and T. Vetter. Reanimating faces in images and video. In *Proc. of Eurographics*, 2003.

[2] George Borshukov and J. P. Lewis. Realistic human face rendering for “the matrix reloaded”. In *Proceedings of SIGGRAPH conference Sketches & applications*. ACM Press, 2003.

[3] George Borshukov, Dan Piponi, Oystein Larsen, J. P. Lewis, and Christina Tempelaar-Lietz. Universal capture: image-based facial animation for "the matrix reloaded". In *Proceedings of SIGGRAPH conference on Sketches & applications*. ACM Press, 2003.

[4] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *SIGGRAPH 2000 Conference Proceedings*, pages 35–42. ACM SIGGRAPH, July 2000.

[5] P. Ekman and W.V. Friesen. *Unmasking the face. A guide to recognizing emotions fron facial clues*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1975.

[6] B. Guenter, C. Grimm, D. Wood, H. Malvar, and F. Pighin. Making faces. In *SIGGRAPH 98 Conference Proceedings*, pages 55–66. ACM SIGGRAPH, July 1998.

[7] J. Haber, K. Khler, I. Albrecht, H. Yamauchi, and H.-P. Seidel. Face to face: From real humans to realistic facial animation. In *Proceedings Israel-Korea Binational Conference on Geometrical Modeling and Computer Graphics*, pages 37–46, 2001.

[8] Tim Hawkins, Andreas Wenger, Chris Tchou, Andrew Gardner, Fredrik Göransson, and Paul Debevec. Animatable facial reflectance fields. In *Rendering Techniques 2004: 15th Eurographics Workshop on Rendering*, pages 309–320, June 2004.

[9] Christophe Hery. Implementing a skin bssrdf (or several). SIGGRAPH course notes: Renderman, Theory and Practice, 2003.

[10] Y. Lee, D. Terzopoulos, and K. Waters. Realistic modeling for facial animation. In *SIGGRAPH 95 Conference Proceedings*, pages 55–62. ACM SIGGRAPH, August 1995.

[11] Y.C. Lee, D. Terzopoulos, and K. Waters. Constructing physics-based facial models of individuals. In *Proceedings of Graphics Interface 93*, pages 1–8, May 1993.

[12] Stephen Marschner, Brian Guenter, and Sashi Raghupathy. Modeling and rendering for realistic facial animation. In *Rendering Techniques 2000: 11th Eurographics Workshop on Rendering*, pages 231–242, June 2000.

[13] Stephen R. Marschner, Stephen H. Westin, Eric P. F. Lafortune, Kenneth E. Torrance, and Donald P. Greenberg. Image-based brdf measurement including human skin. In *Eurographics Rendering Workshop 1999*, June 1999.

[14] F.I. Parke. Computer generated animation of faces. *Proceedings ACM annual conference.*, August 1972.

[15] F.I. Parke. *A parametric model for human faces*. PhD thesis, University of Utah, Salt Lake City, Utah, December 1974. UTEC-CSc-75-047.

[16] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D.H. Salesin. Synthesizing realistic facial expressions from photographs. In *SIGGRAPH 98 Conference Proceedings*, pages 75–84. ACM SIGGRAPH, July 1998.

[17] F. Pighin, R. Szeliski, and D.H. Salesin. Resynthesizing facial animation through 3d model-based tracking. In *Proceedings, International Conference on Computer Vision*, 1999.

[18] B. Raitt. The making of Gollum. Presentation at U. Southern California Institute for Creative Technologies's *Frontiers of Facial Animation* Workshop, August 2004.

[19] Mark Sagar. Reflectance field rendering of human faces for "spider-man 2". In *Proceedings of SIGGRAPH conference Sketches & applications*. ACM Press, 2004.

[20] D. Terzopoulos and K. Waters. Techniques for realistic facial modeling and animation. In Nadia Magnenat Thalmann and Daniel Thalmann, editors, *Computer Animation 91*, pages 59–74. Springer-Verlag, Tokyo, 1991.

[21] D. Vlasic, M. Brand, H. Pfister, and J. Popovic. Face transfer with multilinear models. In *Proceedings of ACM SIGGRAPH 2005*. ACM Press/Addison-Wesley Publishing Co., 2005.

[22] K. Waters. A muscle model for animating three-dimensional facial expression. In *SIGGRAPH 87 Conference Proceedings)*, volume 21, pages 17–24. ACM SIGGRAPH, July 1987.

[23] L. Williams. Performance-driven facial animation. In *SIGGRAPH 90 Conference Proceedings*, volume 24, pages 235–242, August 1990.

[24] Li Zhang, Noah Snavely, Brian Curless, and Steven M. Seitz. Spacetime faces: high resolution capture for modeling and animation. *ACM Trans. Graph.*, 23(3):548–558, 2004.