Facial Capture and Animation in Visual Effects

Darren Cosker, Peter Eisert, Volker Helzle

20.1 Introduction

In recent years, there has been increasing interest in facial animation research from both academia and the entertainment industry. Visual effects and video game companies both want to deliver new audience experiences – whether that is a hyper-realistic human character [Duncan 09] or a fantasy creature driven by a human performer [Duncan 10]. Having more efficient ways of delivering high quality animation, as well as increasing the visual realism of performances, has motivated a surge of innovative academic and industrial developments.

Central to many of these developments are key technical advances in computer vision and graphics. Of particular note are advances in multi-view stereo reconstruction, facial tracking and motion capture, dense non-rigid registration of meshes, measurement of skin rendering attributes (e.g. BRDFs for skin and skin subsurface scattering models) and sensing technology.

This chapter builds on concepts already described earlier in this book – such as 3D capture, rigging, and non-rigid registration – and takes a more practical look at how they might typically be applied in visual effects. First, methods and applications for facial static capture and rendering are considered, before dynamic capture is addressed. Finally, a case study is examined called *The Gathering* involving the creation of an animated face – from animation to final composite.

20.2 Static Facial Realism and Capture

In today's world, static facial realism has reached a level where humans cannot distinguish 3D facial models from real photographs anymore. Technology such as the Light-Stage [Ma et al. 07] allow the capture of highly detailed facial surface information. Coupled with sub-surface reflectance data [Donner et al. 08] facial models now display photo-realistic likenesses to real faces. Such technology is now widely used in modern motion pic-

tures, e.g. *Spider Man 2* [Fordham 04] being one recent high-profile use of the Light-Stage. In these circumstances, the high-detail static scans are composited onto either a stunt-actor's body, or a digital double. This is where the actor is to be placed in situations that may not be practical or safe (such as explosions). However, it is still the case that the actor's expression is typically static in these situations, and close examination of such shots reveal a dead-like facial quality. An early high-profile example of facial replacement was in the Matrix sequels [Borshukov et al. 05], where passive facial scanning was used to obtain 3D faces with high detail facial texture. An important aspect of the use of facial scans for movies and video games is that faces must be renderable in a wide range of environments so that the face can be convincingly composited into the overall scene. Therefore, the UV map (texture data) is typically diffuse albedo. Skin detail is enhanced through high-resolution normal maps [Ma et al. 07] or geometry [Beeler et al. 10], and rendering is enhanced through sophisticated BRDFs/BSSRDFs modeling material properties [Donner and Jensen 06,Jensen et al. 01]. Acquisition of such reflection properties has advanced widely over recent years, resulting in highly detailed rendering [Donner et al. 08].

An important aspect to the realism of synthetic humans is the realistic rendering of hair, which has made a significant progress in the last years. Single hair fibers have been modeled [Marschner et al. 03] as semi-transparent elliptical cylinders. By defining surface reflection as well as scattering inside the hair, the complex lighting characteristics of real hair with its view dependency, highlights, and color changes can be accurately reproduced with moderate rendering complexity [Ren et al. 10]. The possibilities to model several fibers up to a complete hairstyle range from NURBS surfaces via thin shell volumes to strain-based modeling by parameterized clusters, fluid flow, or vector and motion fields [Ward et al. 07]. In order to simulate the complex lighting interaction between strands of hair, Lokovich et al. [Lokovic and Veach 00] propose a deep shadow map which relates visibility to depth for each pixel, yielding realistic but computationally expensive self-shadowing. Approximation algorithms for making the simulation of multiple scattering among hair fibers tractable have been proposed using methods like photon mapping [Moon and Marschner 06b] or spherical harmonics [Moon and Marschner 06a].

Whereas laser scanning technology was initially the most accurate way to derive static facial detail, passive scanning technology using consumer hardware is now popular [Beeler et al. 10,Blumenthal-Barby and Eisert 14]. Multiple consumer-level SLR cameras are used to acquire high-detail images which provide strong features for stereo matching algorithms, Chapter 8, and can result in captures with skin pore (mesoscopic-level) facial details [Beeler et al. 10]. One aspect to consider when using such data is practicality, as the meshes can contain millions of vertices. This is a differ-

ent approach to those methods currently considered in, for example, movies where low polygonal meshes are used along with high-detail normal maps to display facial meso-structure. There is therefore still a great deal of work to be done on practically using such technology for video games and modern VFX.

Many state-of-the-art stereo and multiview approaches are local in the sense that they reconstruct the 3D location, and sometimes orientation, of isolated image patches [Furukawa and Ponce 10]. While this strategy is beneficial for parallelization, it requires a post-processing stage to generate a mesh. The reconstruction yields a point cloud with outliers which has to be filtered and meshed with appropriate algorithms, Chapter 10, such as Poisson meshing [Kazhdan et al. 06]. Smoothness priors are often only considered at the meshing stage. Local reconstruction is difficult to combine with efficient interactive tools. As each patch is unaware of its neighbors, the correction of a single mismatched patch by the user will not affect its neighbors, although they are likely to be erroneous as well. Therefore, [Blumenthal-Barby and Eisert 14] follows a similar approach as [Beeler et al. 10] but uses mesh-based deformable image alignment for the reconstruction of high-detail face geometry (including hair) from two or more SLR cameras, Fig. 20.1. Instead of iteratively matching small image patches along the epipolar line, an entire view is warped to target views in an uncalibrated framework incorporating a mesh-based deformation model. The additional connectivity information enables the incorporation of surface-dependent smoothness priors and optional user guidance for robust and interactive geometry estimation [Schneider and Eisert 12].

Figure 20.1: Static reconstruction of the head including hair from two images [Blumenthal-Barby and Eisert 14].

Most digital face replacement in movies involves static face replacement, with the actor having little or no movement in facial expression. Although this might be satisfactory for a few frames, as soon as the face moves, or the shot continues for more than a few seconds, this illusion becomes hard to maintain. In the next section, the movement of faces is considered,

especially with respect to maintaining an illusion of realism.

20.3 Dynamic Facial Capture and Animation

The holy-grail of facial animation research is the portrayal of characters indistinguishable from real humans. This is extremely difficult since humans are experts in detecting the slightest flaws in faces. Even minor defects can break the illusion of realism. In the previous section, static faces are considered where realism has reached a point where it is impossible to distinguish computer graphics from real photographs. However, in order to display a synthetic human that is truly life-like, the movement of the face remains a major challenge.

Arguably, it is easier to convey dynamic realism in the play-back of actual recorded performances than to author of new animation. In order to highlight this, the acquisition of dynamic 3D facial sequences (termed here as 4D for brevity) is first considered.

There are now many commercial companies that market 4D facial capture systems, i.e. those that can obtain 3D mesh data at video recording rate (e.g. Dimensional Imaging [1], 3DMD [2]). However, the focus here is primarily on academic research in this area. One of the first compelling uses of dynamic facial capture in movies was in the Matrix sequels [Borshukov et al. 05]. A passive stereo capture system was constructed where 3D mesh data can be acquired from a face at video rate along with high-resolution texture. The recorded sequences were then composited onto the actors in key action sequences. An extension of this system called *Universal Capture* was later used in several Electronic Arts (EA) promotions and video games (e.g. Tiger Woods Golf [Borshukov et al. 03]). Here, the system was made more robust by adding markers to the actor's face. This could be used to stabilize and track a canonical mesh (i.e. mesh with a known topology) through the captured sequence. Bickel et al. adopt a similar approach with the addition of extra facial paint to appropriately capture wrinkles on the face [Bickel et al. 07].

The use of markers has overcome previous issues related to tracking a face mesh using optical flow. Such methods are notorious to drift, caused primarily by fast facial changes, for example during speech. Early approaches to avoid the drift in markerless tracking over longer sequences are the incorporation of additional constraints from the silhouette [DeCarlo and Metaxas 00] or the use of an analysis-by-synthesis estimation as in [Eisert 03]. Both methods ensure that estimates are referred to a global

[1] http://www.di3d.com
[2] http://www.3dmd.com

reference and avoid error accumulation over time. This approach is also followed by Bradley et al. who propose a multi-view stereo capture system comprising of 14 HD cameras mosaiced together [Bradley et al. 10]. This results in a highly detailed set of images upon which to apply optical flow for mesh tracking. Referencing the initial frames of the sequence results in improved mesh stabilization over time. Expanding further on this work, Beeler et al. introduced the concept of anchor frames for stabilizing 4D passive facial capture [Beeler et al. 11]. In this work, neutral frames in the sequences are searched for and then used to essentially reinitialize mesh tracking where possible. This also has the added benefit of offering robustness to certain facial self-occlusions (e.g. as caused by the lips). Although having a lower geometric resolution than previous, passive static capture work [Beeler et al. 10] – which includes approximated skin pore geometry – the extension to 4D including the impressive temporal mesh coherence is a high current benchmark in contemporary facial capture research and development.

One highly successful recent demonstration of the use of 4D capture in industry is from the video game LA Noire [3]. Hundreds of hours of actor footage were recorded in a controlled lighting environment. Key 3D character scenes were then composited with the volumetric facial performances resulting in highly detailed and realistic results. Another high profile use of 3D technology for industrial use was by Alexander et al. [Alexander et al. 10]. The Digital Emily Project was a collaboration between Image-metrics and USC using Light Stage technology to capture high detail normal map and surface reflectance properties from an actor's face [Hawkins et al. 07]. A facial blendshape rig was constructed from captured 3D data and then matched to the performance of the actor using proprietary Image-metrics markerless facial capture technology. Blendshapes are facial poses of different expressions – from stereotypical (happy, sad) to extremely subtle (narrow eyes). The term *rig* is used to describe the complete facial model with all its control parameters. The degrees of freedom of the facial rig are a function of the number and complexity of the blendshapes, and new facial poses are created by combining blendshapes with different weights, Chapter 13. More recently, the Digital Ira project [von der Pahlen et al. 14] demonstrated how high levels of static and dynamic realism could be animated and rendered in real-time. Thirty high resolution facial scans were captured using the new Light Stage X system [Ghosh et al. 11], providing data for a facial rig. Video performance was then captured of an actor and used to animate the facial performance, combined with sophisticated real-time rendering of multiple effects [Jimenez et al. 12].

While the LA Noire production is a high-profile use of 4D capture, it is

[3] http://www.rockstargames.com/lanoire

essentially play-back of the recorded data. On the other hand, the Digital Emily and Ira projects demonstrate a degree of performance-driven animation, or retargeting . This type of animation is highly popular in academia and industry, where a performer animates a *puppet* via motion capture or speech (audio only or phonemes). In the case of the both the Digital Emily and Ira projects, the rigs are tracked and animated directly from the actor reference video footage. However, in other cases it is often necessary to retarget between two different rigs – one created as a likeness to the actor's face (which is tracked to the input performance) and a second (often a creature or non-human character) animated from output controls of the first rig. In such cases, a rule-based or example-based mapping must be learned between the two rigs – this is a current active area of academic and industry research [Bhat et al. 13]. While it is not the intention of this chapter to give a detailed review of retargeting methods, the excellent course material in Ref. [Havaldar et al. 06] encompasses many of the ideas in this area still used today. The aim here is rather to make the distinction between direct playback of captured volumetric animation and the creation of realistic character animation given some reference (e.g. actor performance). However, one important point to make is that even given the best tracking or analysis of a human performer, *automatic* animation of a rig to a level satisfactory for visual effects is still an open problem. Typically, after automatically animating a face in this way, an artist is still required to spend considerable time matching and adding secondary rig movements to the reference performance. In video games, this process can be a hindrance – where hundreds of hours of generated performance may be required for delivery under a short time constraint. In this scenario, a lower level of quality than VFX may therefore be acceptable, as generating VFX quality for current video game productions would add unrealistic burden on third party facial animation production or in-house game studios.

The movie *The Curious Case of Benjamin Button* [Duncan 09] contains another successful example of human realistic performance driven animation and retargeting. MOVA[4] performance capture technology was used to collect 3D scans of Brad Pitt's face and used for blendshape rig construction. Animation was then carried out with the aid of markerless performance mapping from reference footage of the actor. The movie *Avatar* [Duncan 10] also pushed forward the technology of facial performance capture and retargeting. Although the characters were not human, the movie demonstrated that modern techniques involving motion capture and artistic input could be used for producing large volumes of high-quality performances. The production involved the use of head-mounted cameras, targeted at the actor's face for recording the movements of painted markers.

[4]http://www.mova.com

These movements were transferred into a combination of blendshapes per-frame, and the resulting animation used to *block out* an initial animation as a first pass for artists – who later edited and enhanced the performance with the aid of additional video reference (akin to the method previously described earlier in this chapter).

While marker-based motion capture techniques are widely popular e.g. using commercial optical capture systems or painted markers (e.g. Vicon's CARA system [5]), markerless methods provide the potential to capture areas of the face where marker placement is too obtrusive. In addition, it raises the possibility of obtaining a dense capture field for the face, for example based on skin pores. Where the facial rig is based on blendshapes [Havaldar et al. 06], the aim is to optimize a set of weights that approximate the positions of the markers. In marker-less systems, such markers might be located using image-based deformable tracking techniques such as Active Appearance Models [Cootes et al. 98]. Another alternative is to fit the blendshapes to 4D surface data [Weise et al. 09,Valgaerts et al. 12]. This latter method has also been shown to work with consumer 2.5D capture devices such as the Kinect [Weise et al. 11]. However, whether this technology alone can provide the fidelity required for VFX to move beyond optical or marker-based methods remains unclear. It may therefore be sensible in the future to consider a combined approach: markers, high quality RGB, and depth sensors.

In the examples so far, facial dynamics have been captured and replayed – often with considerable artistic manual intervention [Alexander et al. 10]. However, the concept of using such data to author entirely novel performances without reference footage remains a difficult challenge. The success of such methods still largely depends on artistic talent. Advances in interactive facial models, and new methods to create efficient rigs, are promising avenues for improvement. In the last part of this section, some recent advances in blendshape rig construction are briefly considered that could help animators use performance capture data more efficiently and provide better artist tools.

One challenge is how to create effective blendshapes. A standard approach in modern VFX is based on Action Units (AUs) from the Facial Action Coding System (FACS) [Ekman and Friesen 78], e.g. in *Monsters House* [Havaldar et al. 06] and *Watchmen* [Fordham 09]. Having a FACS basis can potentially provide a mapping between different facial rigs. This can be especially useful if one blendshape model is based on actor facial scans and fitted to an actor, and then the weights are transferred onto a puppet model, perhaps of a creature. More recently, Li et al. considered creating blendshape rigs given only a few example expressions and a

[5] http://www.vicon.com/System/Cara

generic blendshape rig with a wider number of expressions [Li et al. 10]. Such systems can potentially reduce artist time when manually sculpting blendshapes for rigs, and also for reusing existing blendshape models when new rig creating is required. Facial rigs in movies can potentially become very large, with hundreds of blendshapes for *hero rigs*, i.e. rigs required to deliver close-up expressive performances [Fordham 03]. Any technique for increasing efficiency is therefore of high value to industry.

Facial animation bases – or Blendshape bases – are also not restricted to artistically sculpted facial expressions or captured 3D scans. Principle Component Analysis (PCA) also offers a basis for animating faces. However, although this basis is orthogonal – meaning that each expression has a unique solution with respect to the basis - these are often not intuitive enough for artistic animation. In order to address this, Tena et al. [Tena et al. 11] recently proposed a region-based PCA modeling approach that allows more intuitive direct manipulation of local facial regions. Their method also highlights how solving for expression weights locally can provide better approximation of motion capture data. Ultimately however, what an artist will desire of the facial model is a set of controls that are both intuitive and also orthogonal such that altering one expression does that interfere too much with others. To counter this, blendshape rigs become highly complex, with additional shapes (corrective blend shapes) included to counter interference cases. In an ideal world such correctives would not be necessary given the extra work burden they impose, and future work is still required to address this core problem.

Given the discussion of static and dynamic facial capture, the next section considers a case study where facial models based on the likeness of real people were animated and composited onto real footage. This builds on many of the ideas expressed previously in this chapter, and also highlights many of the practical and real world constraints such a project imposes on animators and technical directors.

20.4 Case Study: The Gathering

The Gathering is a final year short film of Filmakademie Baden-Wuerttemberg[6]. The protagonists in this short film were created digitally based on photo references. The process relied completely on artistic skills since no real reference for 3D scanning or reflectance measurement could be employed. The budget and time constraints of the project demanded to create all assets digitally.

Once the digital models were completed, their facial animation rigs were

[6]http://www.svendreesbach.com/the-gathering/

created by applying the *Adaptable Setup for Performance Driven Facial Animation* [7] [Helzle et al. 04]. This extension for Autodesk Maya[8] enables a rigging artist to apply a generalized library of muscle group movements conforming to the FACS [Ekman and Friesen 78] system to any humanoid geometry. The deformations are driven by a dense data model which includes the non-linear characteristic of facial actions. Compared to static modeling and interpolation of blendshapes, this approach allows for fast and flexible control over the individual facial deformations. The toolset allows complete control in how this data is applied and adjusts to the physiognomies of the geometry. The rigging artist has to manually apply facial landmarks that drive the deformation. The approach has its limitations mainly with respect to the amount and influence of the 69 deformation objects. One way to overcome this limitation was with the use of a limited number of corrective blendshapes, i.e. new blendshapes that trigger other key blendshape combinations to alleviate unwanted or unnatural movements.

Custom extensions to the toolset allow controlling the stickiness of the lips as they part when speaking. This effect is due to moisture on the lips, causing them to open from the inside to the outside as the lips part. Furthermore, the effect of the eyes cornea bulging the upper or lower lids as the gaze changes was realized using a complex constellation of additional deformation rig objects. A fast animation rig allowed quick iterations when animating the sequences and kept the animation artists motivated. All facial animation was realized by rotoscoping the movements recorded from the real actors on set. The head movement could be extracted by rigid body tracking of the markers applied to the actors' heads as shown in the top left of Fig. 20.2.

The animated models were rendered using Newteks Lightwave 3D[9] software package. The top right of Fig. 20.2 shows the raw rendering which included additional information like motion vectors, reflection and diffuse values embedded inside Open-EXR files, which were provided for compositing. The final compositing was accomplished in The Foundry's Nuke[10] software, integrating the CGI elements into the plates (lower left of Fig. 20.2) before final coloring and additional effects like cigarette smoke were added (lower right of Fig. 20.2).

The Gathering shows that it is possible to create convincing digital faces by relying mostly on artistic skills and powerful tools for facial animation. However, it also highlights that the complexities of creating a face and its movements digitally demand a wide set of skills.

[7] http://fat.research.animationsinstitut.de
[8] http://www.autodesk.com/maya
[9] http://www.lightwave3d.com/
[10] http://www.thefoundry.co.uk/nuke/

Figure 20.2: The Gathering: Case study on facial animation. ©Filmakademie Baden-Wuerttemberg, The Gathering, 2011.

20.5 Summary

Capturing real faces with modern technologies like Light Stages provides 3D face models with high geometric detail and sophisticated material properties that enable face synthesis that is almost indistinguishable from a real picture. While static face models can be used for replacing an actor's face for a few frames, often dynamic face capturing is also desired – which is still a challenging task due to the sensitivity of a human observer for subtle inconsistencies in facial motion. Facial dynamics are usually modeled by a blendshape rig, either from scans, multiview images, or manual work of an artist, while animation data is often derived from marker-based or, more recently, markerless motion capture systems. As shown in the presented case study, convincing digital faces can be animated with such techniques. However, creating realistic facial animations and models still requires significant manual work by artists, leaving room for novel algorithms and toolsets to simplify and automate the process. In terms of research challenges, there is still a large scope for further work in this area. Creating rigs is a time consuming process, and methods to automate this at production quality are highly desirable. Another core area for future work is in the retargeting of actor faces to new creatures. One challenge in achieving this aim, however, lies perhaps in the contrast between the academic and industrial worlds: academia often has the time to focus on algorithms that could solve this problem while lacking the complex rigs required to test their idea. Conversely, industry has the expertise to produce such rigs but

given practical movie constraints often does not have the time to focus on the algorithms. It is therefore unsurprising that the best advances in this area have been from academic and industrial collaboration.