CREATING A PHOTOREAL DIGITAL ACTOR: THE DIGITAL EMILY PROJECT

Oleg Alexander[1] **Mike Rogers**[1] **William Lambeth**[1] **Matt Chiang** [2] **Paul Debevec**[2]

[1] Image Metrics
1918 Main St, 2nd Floor, Santa Monica, CA 90405, USA
e-mail: {oleg.alexander,mike.rogers,william.lambeth}@image-metrics.com

[2] University of Southern California Institute for Creative Technologies
13274 Fiji Way, Marina del Rey, CA 90292, USA
e-mail: {chiang,debevec}@ict.usc.edu

Abstract

The Digital Emily Project is a collaboration between facial animation company Image Metrics and the Graphics Laboratory at the University of Southern California's Institute for Creative Technologies to achieve one of the world's first photorealistic digital facial performances. The project leverages latest-generation techniques in high-resolution face scanning, character rigging, video-based facial animation, and compositing. An actress was first filmed on a studio set speaking emotive lines of dialog in high definition. The lighting on the set was captured as a high dynamic range light probe image. The actress' face was then three-dimensionally scanned in thirty-three facial expressions showing different emotions and mouth and eye movements using a high-resolution facial scanning process accurate to the level of skin pores and fine wrinkles. Lighting-independent diffuse and specular reflectance maps were also acquired as part of the scanning process. Correspondences between the 3D expression scans were formed using a semi-automatic process, allowing a blendshape facial animation rig to be constructed whose expressions closely mirrored the shapes observed in the rich set of facial scans; animated eyes and teeth were also added to the model. Skin texture detail showing dynamic wrinkling was converted into multiresolution displacement maps also driven by the blend shapes. A semi-automatic video-based facial animation system was then used to animate the 3D face rig to match the performance seen in the original video, and this performance was tracked onto the facial motion in the studio video. The final face was illuminated by the captured studio illumination and shading using the acquired reflectance maps with a skin translucency shading algorithm. Using this process, the project was able to render a synthetic facial performance which was generally accepted as being a real face.

Keywords: Digital Actors; Facial Animation; 3D Scanning

1 Introduction

Creating a photoreal digital actor with computer graphics has been a central goal of the field for at least thirty years [Parke 1972]. The Digital Emily project undertaken by Image Metrics and the University of Southern California's Institute for Creative Technologies (USC ICT) attempted to achieve an animated, photoreal digital face by bringing together latest-generation results in 3D facial capture, modeling, animation, and rendering. The project aimed to cross the "uncanny valley" [20], producing a computer-generated face which appeared to be a real, relatable, animated person. Some of the key technologies employed included a fast high-resolution digital face scanning process using the light stage at USC ICT, and the Image Metrics video-based facial animation system. The result of the project was by several accounts the first public demonstration of a photoreal computer-generated face able to convincingly speak and emote in a medium closeup.

2 Previous Efforts at Photoreal Digital Humans

A variety of laudable efforts have been made to create realistic digital actors over the last decade, each leveraging numerous advances in computer graphics technology and artistry. In this section, we overview some of these key efforts in order to compare and contrast them with the Digital Emily project.

The SIGGRAPH 1999 Electronic Theater featured "The Jester" [15], a short animation by Life/FX of a woman reading poetry in jester's cap. The actor's face was three-dimensionally laser scanned and textured using photographic textures stitched together with artistic effort to minimize the original shading and specular reflectance effects, and her performance was recorded with a traditional arrangement of motion capture markers. The motion capture dots were used to drive a volumetric finite element model which allowed a high-resolution facial mesh to produce simulated buckling and wrinkling from anisotropic stress at a scale significantly more detailed than the original motion capture data was able to record. While the skin shading lacked realistic specular

reflectance and pore detail, the face conveyed realistic motion and emotion in significant part due to the skin dynamics model. The process was later extended with significant additional artistic effort to create an aged version of the actor in a follow-up animation "Young at Heart".

Disney's "Human Face Project" developed technology to show an older actor encountering a younger version of himself [27, 12]. A facial mold of the actor's face was taken and the resulting cast was scanned using a high-resolution face scanning process, and twenty-two Cyberware scans in various expressions were also acquired. A medium-resolution animated facial rig was sculpted to mimic the expression scans used as reference. A multi-camera facial capture setup was employed to film the face under relatively flat cross-polarized lighting. Polarization difference images isolating specular reflections of the face [5] were acquired to reveal high-resolution texture detail, and analysis of this texture was used to add skin pores and fine wrinkle detail as displacement maps to the scans. Optical flow performed on video of the actor's performance was used to drive the digital face, achieving remarkably close matches to the original performance. A younger version of the actor was artistically modeled based on photographs of him in his twenties. HDR lighting information was captured on set so that the digital actor could be rendered with image-based lighting [4] to match the on-set illumination. Final renderings achieved convincing facial motion and lighting in a two-shot but less convincing facial reflectance; a significant problem was that no simulation of the skin's translucency [14] had been performed.

The Matrix Sequels (2003) used digital actors for many scenes where Keanu Reeves acrobatically fights many copies of Hugo Weaving. High-quality facial casts of Reeves and Weaving were acquired and scanned with a high-resolution laser scanning system (the XYZRGB system based on technology from the Canadian National Research Council) to provide 3D geometry accurate to the level of skin pores and fine wrinkles. A six-camera high-definition facial capture rig was used to film facial performance clips of the actors under relatively flat illumination. The six views in the video were used to animate a facial rig of the actor and, as in [10], to provide time-varying texture maps for the dynamic facial appearance. Still renderings using image-based lighting [4] and a texture-space approximation to subsurface scattering [2] showed notably realistic faces which greatly benefitted from the high-resolution geometric detail texture found in the XYZRGB scans. However, the animated facial performances in the film were shown in relatively wide shots and exhibited less realistic skin reflectance, perhaps having lost some appearance of geometric detail during normal map filtering. The animated texture maps, however, provided a convincing degree of dynamic shading and albedo changes to the facial performances. A tradeoff to the use of video textures is that the facial model could not easily generate novel performances unless they too were captured in the complete performance capture setup.

Spider Man 2 (2004) built digital stunt doubles for villain Doc Ock (Alfred Molina) and hero Spider-Man (Tobey Maguire) using facial reflectance scans in USC ICT's Light Stage 2 device [24]. Each actor was filmed with four synchronized 35mm film cameras in several facial expressions from 480 lighting directions. Colorspace techniques as in [5] were used to separate diffuse and specular reflections. The relightable texture information was projected onto a 3D facial rig based on geometry from a traditional laser scan, and illuminated variously by HDRI image-based lighting [4] and traditional CG light sources using a custom shading algorithm for approximately 40 digital double shots. Using additional cameras, the technique was also used to construct digital actors for Superman Returns (2006), Spider Man 3 (2007), and Hancock (2008). Due to the extensive reflectance information collected, the technique yielded realistic facial reflectance for the digital characters, including close-up shots with mild degrees of facial animation, especially in *Superman Returns*. However, results of the process did not demonstrate emotive facial performances in closeup; significant facial animation was shown only in wide shots.

Beowulf (2007) used a multitude of digital characters for a fully computer-rendered film. The performance capture based film following the approach of the 2001 film Final Fantasy of constructing as detailed characters as possible and then driving them with motion capture, keyframe animation, and simulation. Beowulf substantially advanced the state of the art in this area by leveraging greater motion capture fidelity and performance volume, employing more complex lighting simulation, and using better skin shading techniques. While some characters were based closely in appearance their voice and motion capture actors (e.g. Angelina Jolie and Anthony Hopkins), other characters bore little resemblance (e.g. Ray Winstone), requiring additional artistic effort to model them. Static renderings of the faces achieved impressive levels of realism, although the single-layer subsurface scattering model produced a somewhat "waxy" appearance. Also, the artist-driven character creation process did not leverage high-resolution 3D scans of each actor in a multitude of facial expressions. According to *Variety* [3], the "digitized figures in 'Beowulf' look eerily close to storefront mannequins ... suspended somewhere between live-action and animation, fairy tale and videogame."

The Curious Case of Benjamin Button (2008) was the first feature film to feature a photoreal human virtual character. The aged version of Brad Pitt seen in the film's first 52 minutes was created by visual effects studio Digital Domain leveraging "recently developed offerings from Mova, Image Metrics, and the [USC] Institute for Creative Technologies" [23]. Silcone maquettes of Brad Pitt as an old man were constructed by Kazuhiro Tsuji at Rick Baker's makeup studio and used as the basis for the digital character. Detailed facial reflectance capture was performed for the age 70 maquette in USC ICT's Light Stage 5 device [26] from eight angles and 156 lighting directions. Medium-resolution meshes of Brad Pitt's face in a multitude of facial expressions were captured using

Mova's Contour facial capture system [22], using a mottled pattern of glow-in-the-dark makeup to create facial geometry from multi-view stereo. Digital Domain used special software tools and manual effort to create a rigged digital character whose articulations were based closely on the facial shapes captured in the Mova system. Image Metrics' video-based facial animation system was used to provide animation curves for the Benjamin facial rigs by analyzing frontal, flat-lit video of Brad Pitt performing each scene of the film in a studio; these animation curves were frequently refined by animators at Digital Domain to create the final performances seen in the film. Extensive HDRI documentation was taken on each film set so that advanced Image-Based Lighting techniques based on [4] could be used to render the character with matching illumination to the on-set lighting. Image-based relighting techniques as in [5] were used to simulate the reflectance of the scanned maquette in each illumination environment as cross-validation of the digital character's lighting and skin shaders. The final renderings were accurately tracked onto the heads of slight, aged actors playing the body of Benjamin in each scene. The quality of the character was universally lauded as a breakthrough in computer graphics, winning the film an Academy Award for Best Visual Effects. An estimate of more than two hundred person-years of effort was reportedly spent [23] creating the character in this groundbreaking work.

3 Acquiring high-resolution scans of the actor in various facial expressions

Emily O'Brien, an Emmy-nominated actress from the American daytime drama "The Young and the Restless", was cast for the project. After being filmed seated describing aspects of the Image Metrics facial animation process on an informal studio set, Emily came to USC ICT to be scanned in its Light Stage 5 device on the afternoon of March 24, 2008. A set of 40 small facial dots were applied to Emily's face with a dark makeup pencil to assist with facial modeling. Emily then entered Light Stage 5 for approximately 90 minutes during which data for thirty-seven high-resolution facial scans were acquired. Fig. 1 shows Emily in the light stage during a scan, with all 156 of its white LED lights turned on.

The light stage scanning process used for Digital Emily was described in [18]. In contrast to earlier light stage processes (e.g. [5, 11]), which photograph the face under hundreds of illumination directions, this newer capture process requires only fifteen photographs of the face under different lighting conditions as seen in Fig. 2 to capture geometry and reflectance information for a face. The photos are taken with a stereo pair of Canon EOS 1D Mark III digital still cameras, and the images are sufficiently few so that they can be captured in the cameras' "burst mode" in under three seconds, before any data needs to be written to the compact flash cards.

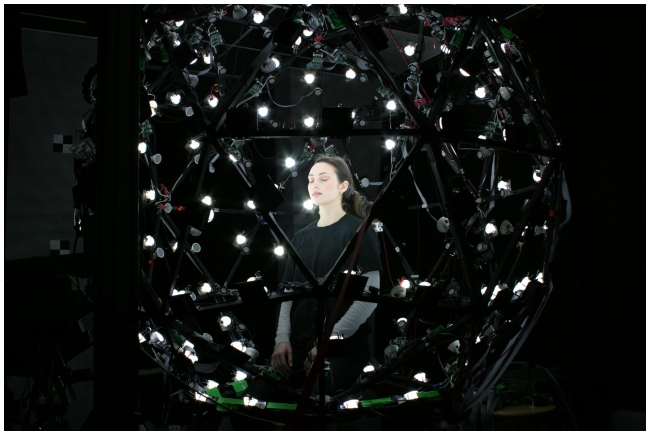


Figure 1: Actress Emily O'Brien being scanned in Light Stage 5 at USC ICT for the Digital Emily project.

3.1 Estimating Subsurface and Specular Albedo and Normals

Most of the images are shot with essentially every light in the light stage turned on, but with different gradations of brightness. All of the light stage lights have linear polarizer film placed on them, affixed in a special pattern of orientations, which permits the measurement of the specular and subsurface reflectance components of the face independently by changing the orientation of a polarizer on the camera.

The top two rows of Fig. 2 show Emily's face under four spherical gradient illumination conditions and then a point-light condition, and all of the images in this row are cross-polarized to eliminate the shine from the surface of her skin – her specular component. What remains is the skin-colored "subsurface" reflection, often referred to as the "diffuse" component. This is light which scatters within the skin enough to become depolarized before re-emerging. Since this light is depolarized, approximately half of this light can pass through the horizontal polarizer on the camera. The top right image is lit by a frontal flash, also cross-polarizing out the specular reflection.

The middle row of Fig. 2 shows parallel-polarized images of the face, where the polarizer on the camera is rotated vertically so that the specular reflection returns, in double strength compared to the attenuated subsurface reflection. We can then reveal the specular reflection on its own by subtracting the first row of images from the second row, yielding the specular-only images shown in Fig. 3.

Fig. 4(a) is a closeup of the "diffuse-all" image of Emily. Every light in the light stage is turned on to equal intensity, and the polarizer on the camera is oriented to block the specular reflection from every single one of the polarized LED light sources. Even the highlights of the lights in Emily's eyes are eliminated.

This is about as flat-lit an image of a person's face as can conceivably be photographed, and thus it is nearly a perfect

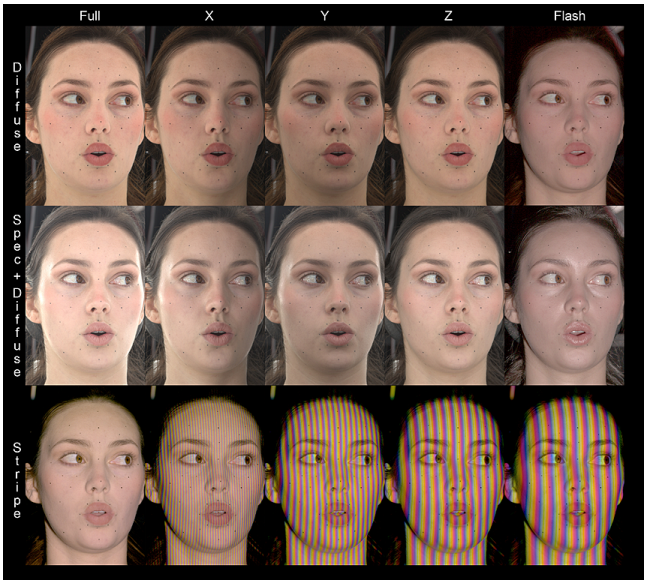


Figure 2: The fifteen photographs taken by a frontal stereo pair of cameras for each Emily scan. The photo sets, taken in under three seconds, record Emily's diffuse albedo and normals (top row), specular albedo and normals (second row), and base 3D geometry (bottom row).

Figure 3: Emily's specular component under the four gradient lighting conditions (all, left, top, front) and single frontal flash condition, obtained by subtracting the cross-polarized images from the parallel-polarized images.

image to use as the diffuse texture map for the face in building a digital actor. The one issue is that it is affected to some extent by self-shadowing and interreflections, making the concavities around the eyes, under the nose, and between the lips appear somewhat darker and more color-saturated than it inherently is. Depending on the rendering technique chosen, having these effects of occlusion and interreflection "baked in" to the texture map is either a problem or an advantage. For real-time rendering, the effects can add realism (as if the environmental light were entirely uniform) given that simulating them accurately might be computationally prohibitive. If new lighting is being simulated on the face using a more accurate global illumination technique, then it is problematic to calculate self-shadowing of a surface whose texture map already has self-shadowing present; likewise for interreflections. In this case, one could perform inverse rendering by using the actor's 3D geometry and approximate reflectance to predict the effects of self-shadowing and/or interreflections, and then divide these effects out of the texture

(a) (b) (c)

Figure 4: Closeups of Emily's (a) cross-polarized subsurface component, (b) parallel-polarized subsurface plus specular component, and (c) isolated specular component formed by subtracting (a) from (b). The black makeup dots on her face are easily removed digitally and help with aligning and corresponding her scans.

image as in [6].

Fig. 4(a) also shows the makeup dots we put on Emily's face which help us to align the images in the event there is any drift in her position or expression over the fifteen images; they are relatively easy to remove digitally. Emily was extremely good at staying still for the three-second scans and many of her datasets required no motion compensation at all. (Faster capture times are already possible: 24fps capture of such data using high-speed video cameras is described in [19]).

The shinier image in Fig. 4(b) is also lit by all of the light stage lights, but the orientation of the polarizer has been turned 90 degrees which allows the specular reflections to return. Her skin exhibits a specular sheen, and the reflections of the lights are now evident in her eyes. In fact, the specular reflection is seen at double the strength of the subsurface (or diffuse) reflection, since the polarizer on the camera blocks about half of the unpolarized subsurface reflection.

Fig. 4(b) shows the combined effect of specular reflection and subsurface reflection. For modeling facial reflectance, we would ideally observe the specular reflection independently. As a useful alternative, we can simply subtract the diffuse-only image Fig. 4(a) from this one. Taking the difference between the diffuse-only image and the diffuse-plus-specular image yields an image of primarily the specular reflection of the face as in 4(c). A polarization difference process was used previously for facial reflectance analysis in [5], but only for a single point light source and not for the entire sphere of illumination. The image is mostly colorless since this light has reflected specularly off the surface of the skin, rather than entering the skin and having its blue and green colors significantly absorbed by skin pigments and blood before reflecting back out.

This image provides a useful starting point for building a digital character's specular intensity map, or "spec map"; it shows for each pixel the intensity of the specular reflection at that pixel. However, the specular reflection becomes amplified near grazing angles, such as at the sides of the face according to the denominator of Fresnel's equations. We generally model and compensate for this effect using Fresnel's equations but also discount regions of the face at extreme grazing angles.

The image also includes some of the effects of "reflection occlusion" [16]. The sides of the nose and innermost contour of the lips appear to have no specular reflection since self-shadowing prevents the lights from reflecting in these angles. This effect can be an asset for real-time rendering, but should be manually painted out for offline rendering.

Recent work [9] reports that this sort of polarization difference image also contains the effects of single scattering, wherein light refracts into the skin but scatters exactly once before refracting back toward the camera. Such light can pick up the color of the skin's melanocytes, adding some color to the specular image. However, the image is dominated by the specular component's first-surface reflection, which allows us to reconstruct high-resolution facial geometry.

The four difference images of the face's specular reflection under the gradient illumination patterns (Fig. 3) let us derive a high-resolution normal map for the face: a map of its local surface orientation vector at each pixel. If we examine the intensity of one pixel across this four-image sequence, its brightness in the X, Y, and Z images divided by its brightness in the fully-illuminated image uniquely encodes the direction of the light stage reflected in that pixel. From the simple formula in [18] involving image ratios, we can derive the reflection vector at each pixel, and from the camera orientations (calibrated with the technique of [28]) we also know the view vector.

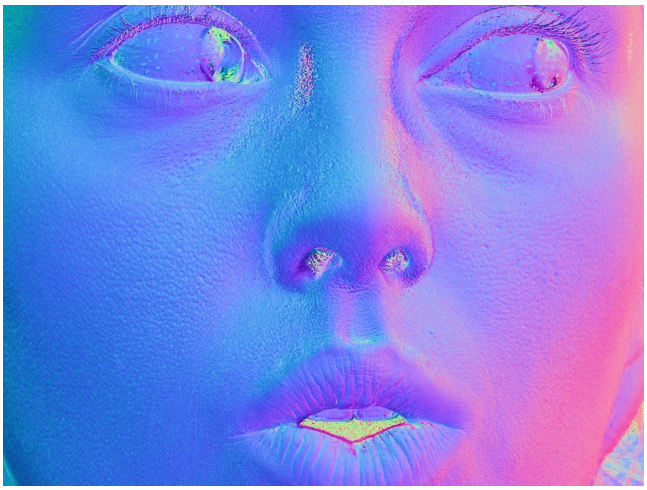


Figure 5: The specular normal map for one of Emily's expressions derived from the four specular reflection images under the gradient illumination patterns.

Computing the vector halfway between the reflection vector and the view vector yields a surface normal estimate for the face based on the specular reflection. Fig. 5 shows the face's normal map visualized using the common color map where red, green and blue indicate the X, Y, and Z components of the surface normal. The normal map contains detail at the level of skin pores and fine wrinkles. The point-lit polarization difference image (Fig. 3, far right) provides a visual indication of the BRDF's specular lobe shape on the nose, forehead, cheeks, and lips. This image provides visual reference for choosing specular roughness parameters for the skin shaders. The image can also be used to drive a data-driven specular reflectance model as in [9].

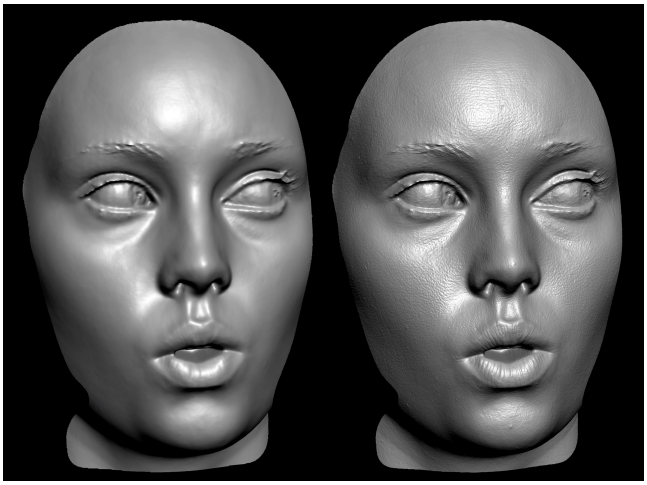
3.2 Deriving High-Resolution 3D Geometry

The last set of images in the scanning process (Fig. 2, bottom row) are a set of colored stripe patterns from a video projector which allow a stereo correspondence algorithm to robustly compute pixel correspondences between the left and right viewpoints of the face. The projector is also cross-polarized so that the stereo pair of images consist of only subsurface reflection and lack specular highlights; such highlights would shift in position between the two viewpoints and thus complicate the stereo correspondence process. The patterns form a series of color-ramp stripes of different frequencies so that a given facial pixel receives a unique set of RGB irradiance values over the course of the sequence [17]. The first projected image, showing full-on illumination, is used to divide out the cross-polarized facial BRDF from the remaining stripe images; this also helps ensure that pixels in one camera have very nearly the same pixel values as in the other camera, facilitating correspondence. From these correspondences and the camera calibration [28], we can triangulate a three-dimensional mesh of the face with vertices at each pixel, applying bilateral mesh denoising [8] to produce a smooth mesh without adding blur to geometric features. However, the surface resolution observable from the diffuse reflection of skin is limited by the scattering of the incident light beneath the skin. As a result, the geometry appears relatively smooth and lacks the skin texture detail that we wish to capture in our scans.

We add in the skin texture detail by embossing the specular normal map onto the 3D mesh. By doing this, a high-resolution version of the mesh is created and the vertices of each triangle are allowed to move forward and back until they best exhibit the same surface normals as the normal map. The ICT Graphics Lab first described this process – using diffuse normals estimated in Light Stage 2 – in [13] 2001; more recent work in the area includes Nehab et al. [21].) This creates a notably high-resolution 3D scan, showing different skin texture detail clearly observable in different areas of the face (Fig. 6).

3.3 Scanning a Multitude of Expressions

Emily was captured in thirty-three different facial expressions based loosely on Paul Ekman's Facial Action Coding System (FACS) [7] as seen in Fig. 7; fourteen of these individual scans are shown in Fig. 8. By design, there is a great deal of variety in the shape of her skin and the pose of her lips, eyes, and jaw across the scans. Emily was fortunately very good at staying still for all of the expressions. Two of the expressions, one with eyes closed and one with eyes open, were also scanned from the sides with Emily's face rotated to the left and right as seen in the inset figures of the neutral-mouth-closed and neutral-mouth-open scans. This allowed us to merge together a 3D model of the face with geometry stopping just short of her ears to create the complete "master mesh" (Fig. 12(a)) and



(a) (b)

Figure 6: (a) Facial geometry obtained from the diffuse (or subsurface) component of the reflection. (b) Far more detailed facial geometry obtained by embossing the specular normal map onto the diffuse geometry.

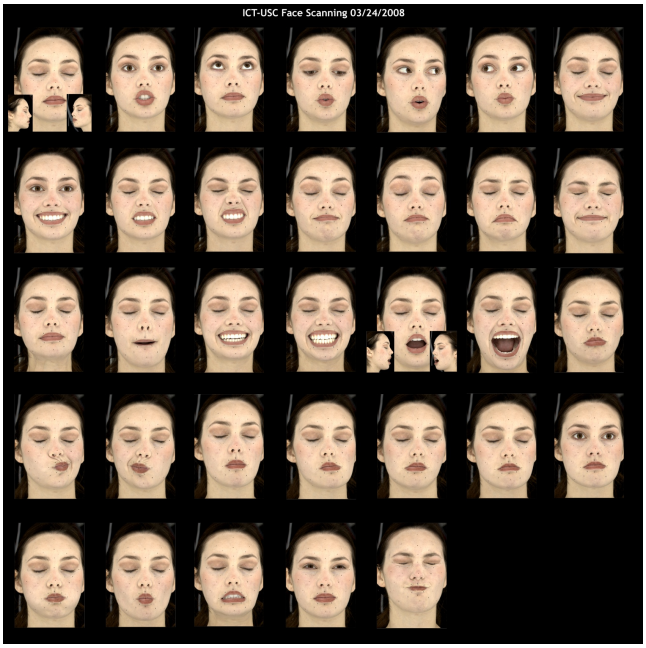


Figure 7: The thirty-three facial expressions scanned for creating the Digital Emily character.

to extrapolate full ear-to-ear geometry for each partial (frontal-only) facial scan as in Fig. 15(c).

We note that building a digital actor from 3D scans of multiple facial expressions is a commonly practiced technique; for example, the Animatable Facial Reflectance Fields project [11] followed this approach in scanning actress Jessica Vallot in approximately forty facial expressions. Going further back, visual effects company Industrial Light + Magic acquired several Cyberware 3D scans of actress Mary Elizabeth Mastrantonio in different expressions to animate the face of the water creature in 1989's The Abyss.

3.4 Phenomena Observed in the Facial Scans

The fourteen faces in Fig. 8 show a sampling of the high-resolution scans taken of Emily in different facial expressions, and offer an opportunity to observe three-dimensional facial dynamics in more detail than has been previously easy to do. A great deal of dynamic behavior can be observed as a face moves, exemplified in the detailed images in Fig. 9. For example, the skin pore detail on the cheek in Fig. 9(a) changes dramatically when Emily pulls her mouth the side in Fig. 9(b): the pores significantly elongate and become shallower. When Emily stretches her face vertically, small veins pop out from her eyelid Fig. 9(c).

When Emily raises her eyebrows, the relatively isotropic skin pores of her forehead transform into rows of fine wrinkles in Fig. 9(d) – her skin is too elastic to develop deep furrows. When she scowls in Fig. 9(e), indentations above her eyebrows appear where her muscles attach beneath the skin. And when she winces in Fig. 9(f), the muscles in her forehead bulge out. Examining the progression through Figs. 9(d,e,f), we see how the bridge of Emily's nose significantly shrinks and expands as the rest of the face pushes and pulls tissue into and out of the area. While we may not consciously notice these kinds of phenomena when interacting with others, they are present and visible in all faces, and failing to reproduce them accurately imperils the realism of a digital character.

3.5 Scanning Emily's Teeth

Finally, we also scanned a plaster cast of Emily's teeth, which required adapting the 3D scanning system to work with greater accuracy in a smaller scanning volume. Fig. 10(a) shows the cast and Fig 10(b) shows a rendering of Emily's digital teeth model; the upper and lower teeth each were the result of merging eight sinusoid-pattern structured light scans to form meshes with approximately 600,000 polygons.

4 Building the digital character from the scans

4.1 Constructing the Animatable Base Mesh

The stitched ear-to-ear neutral-expression scan of Emily was remeshed to create a 4,000 polygon animatable mesh as seen in Figure 11(a). This drastic polygon reduction from the several million polygons of the original scan was done to make animation of the mesh tractable and to ease corresponding the geometry across scans; the geometric skin texture detail would be added back using displacement maps calculated from the high-resolutions scans. This was done principally using the commercial product ZBrush to create the facial topology and then Autodesk's Maya package to create the interior of the mouth and the eye sockets. Then, UV texture coordinates for the neutral animatable mesh were mapped out in Maya, yielding the complete master mesh.

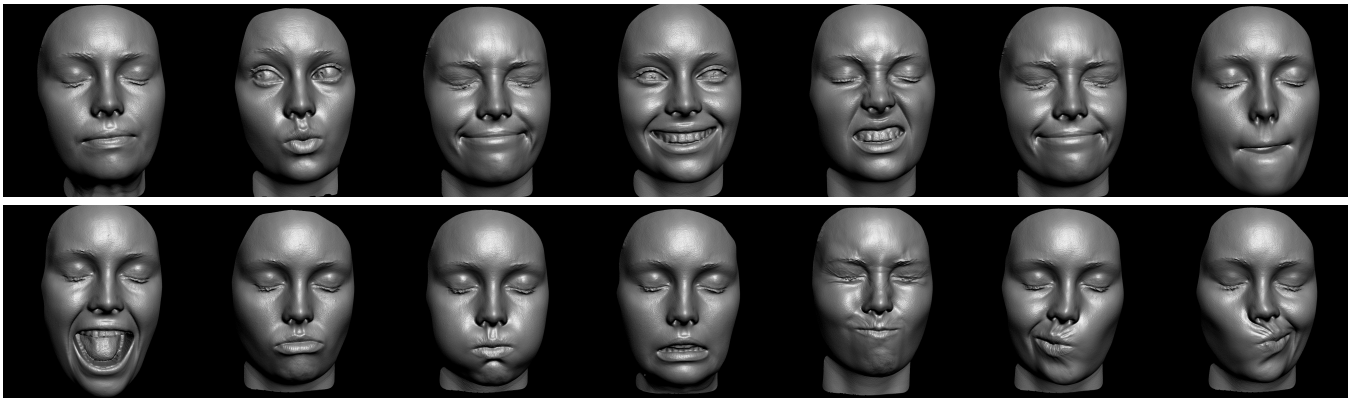


Figure 8: High-resolution 3D geometry from fourteen of the thirty-three facial scans. Each mesh is accurate to 0.1mm resolution and contains approximately three million polygons.

4.2 Building the Blendshapes

Image Metrics originally planned to use the scans captured in the Light Stage as artistic reference for building the blendshapes. However, the scans proved to be much more useful than just reference. The tiny stabilization dots drawn on Emilys face during the scanning session were visible in every texture map of every scan. Rather than sculpt the blendshape meshes artistically and then project (or "snap") the vertices to the corresponding scans, we used the dots directly to warp the neutral animatable mesh into the different expressions. This achieved not only an accurate shape for each blendshape, but also accurate skin movement between blendshapes.

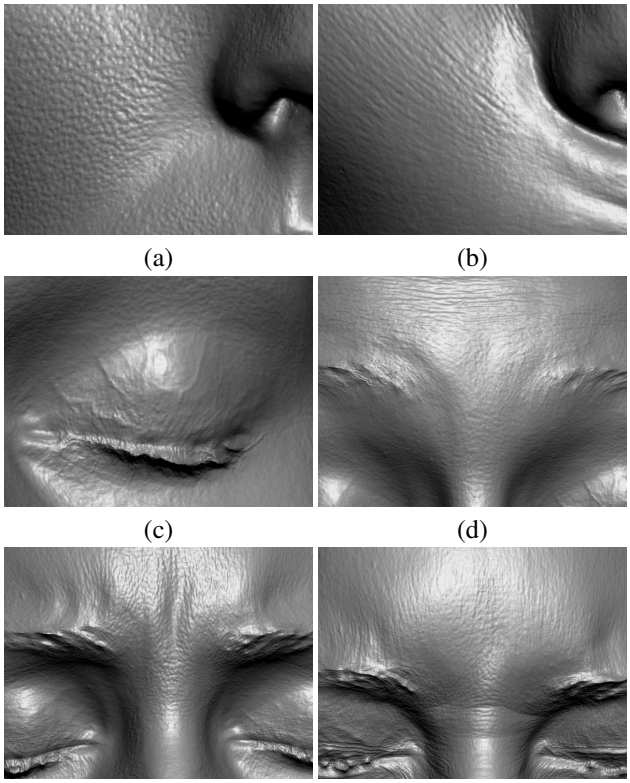
The challenge in constructing blendshapes from these data was to use the sparse set of stabilization dots to find a dense correspondence between the animatable master mesh and each expression scan. Figure 12 shows an example of three such partial frontal scans alongside the animatable master mesh.

Although highly detailed and accurate, the expression scans required some pre-processing before appropriate blendshapes could be constructed. For example, many meshes have irregular edges with poor triangulation and mesh artifacts around the teeth and eye regions. Also, many of the scans contained surface regions not represented in the master mesh, such as the surface of the eyes and teeth. These regions should not be corresponded with master mesh vertices. Finally, the scans, captured from a frontal stereo camera pair, did not cover the full facial region of the animatable master mesh. Figures 13 and 14 show examples of each of these issues. These aspects of the data meant that a fully automatic correspondence algorithm would be unlikely to provide adequate results.

The correspondence method we used required some simple manual data cleaning and annotation to produce data more amenable to automatic processing. For each expression scan, we removed regions from both the master mesh and the expression mesh to achieve rough edge consistency. For example, we removed the neck region from the expression scans and face sides from the master to produce consistent facial coverage. We also removed uncorrespondable regions from the meshes by deleting vertices; these were commonly the teeth and eyeballs regions of the expression scans, as such regions are not present in the master mesh. This process also removed most of the mesh artifacts associated with discontinuities in the expression scans. The 3D locations of the stabilization dots were annotated manually and served as a sparse set of known points of correspondence between master mesh and each expression scan. For each expression, this sparse correspondence information was used to initialize and stabilize a proprietary automatic method of determining a dense correspondence. The method used a 3D spatial location and mesh normal agreement measure within a 2D conformal mapping frame (such as in [25]) to obtain the required dense correspondence. This process resulted in a mapping of each master mesh vertex to a position on the surface of the appropriate expression mesh.

With the correspondence between our partial, cleaned-up meshes it was possible to calculate the shape change required to transform the neutral master mesh to each expression. However, a rigid translation/rotation between the master and expression scans was also generally present in the data. This rigid transformation must be calculated to ensure only motion due to the changing facial expression and not residual head motion is represented in each blendshape. To do this, a 3D rigid transformation for each expression was calculated using a manually selected subset of texture markers. The texture markers selected were chosen independently for each expression so as to be the least affected by the skin and jaw motion for that expression, and thus provide a good basis for calculating a rigid transformation between scans. Once the rigid transformations were calculated and was factored out of each blendshape, blendshape deltas for each vertex in the partial master mesh were calculated. This produced partial blendshapes as shown in Figure 15(a,b).

Any jaw movement had to be subtracted from each blendshape; this step was a very important because it eliminated redundancy in the rig. One example was the `lipSuck` shape where both top and bottom lips are sucked inside the mouth. In order to capture the maximum amount of data for the lips,

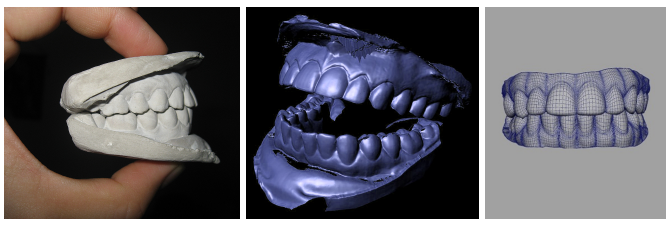


(a) (b) (c) (d) (e) (f)

Figure 9: Details of the 3D face scans showing various dynamic facial deformation phenomena.

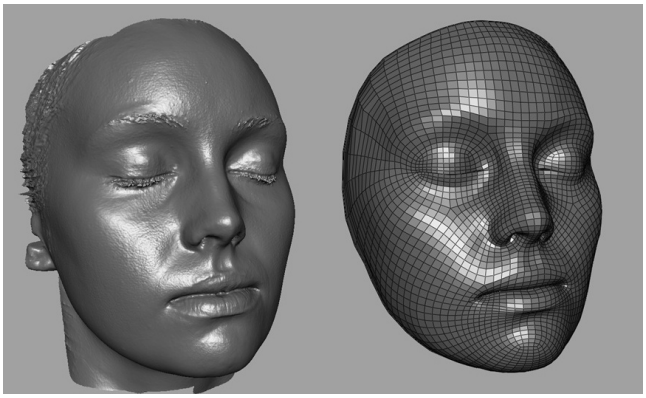
this expression was scanned and modeled with the jaw slightly open. After the blendshape was modeled the jaw movement was subtracted by going negative with the `jawOpen` shape.

A custom algorithm was used to map the partial blendshapes onto the full master mesh. The missing parts of the full master mesh that were not observed in the expression scan were interpolated using the assumption that the extreme edges of the face remained static, which we found to be a reasonable assumption for most expressions. Where this assumption was not appropriate, the results were artistically corrected to give the desired shape. The unknown internal parts of the master mesh, such as the inner mouth and lips, were interpolated



(a) (b) (c)

Figure 10: (a) A plaster cast of Emily's upper and lower teeth. (b) Resulting merged 3D model before remeshing. (c) Remeshed model with 10,000 polygons each for the upper and lower teeth.



(a) (b)

Figure 11: (a) Stitched ear-to-ear neutral-expression mesh comprising several millions polygons made from merging left, front, and right scans (b) Remeshed neutral mesh with 4,000 polygons and animation-friendly topology.

(a) (b) (c) (d)

Figure 12: (a) Ear-to-ear master mesh with (b,c,d) three partial expression scans to be used as blend shapes.

to move appropriately with the known vertices. Figure 15(c) illustrates the result of this interpolation process.

The final results of the automatic blendshape creation process were then artistically cleaned up by a small team of rigging artists to provide a full set of quality-assured expression blend shapes.

After the blendshape modeling and cleanup process, Image Metrics ended up with approximately 30 "pre-split" blendshapes, one blendshape corresponding to each scan. However, most of the scans were captured with "doubled up" facial expressions. For example, the `browRaise` and `chinRaise` were captured in the same scan. This was done to get through the scanning session quicker and to save processing time. But now the blendshapes had to be split up into localized shapes – the "post-split" shape set. To do this, we used the *Paint Blendshape Weights* feature in Maya, creating a set of normalized "split maps" for each facial region. For example, this is how the `browRaise` blendshape was split up. Three normalized split maps were painted for the `browRaise` blendshape: `browRaise_L`, `browRaise_C`, and `browRaise_R`, roughly following the eyebrow raisers described in [7]. A custom MEL script was then run which applied each split map to the `browRaise`

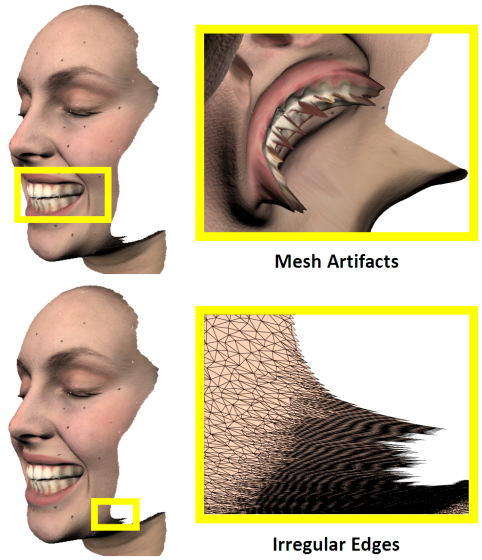


Figure 13: Mesh artifacts and irregular edges in the raw Emily scans

Figure 14: Inconsistent coverage at the periphery of the face and uncorrespondable regions (the eyeball and eyelid, also the teeth) in the raw Emily scans

shape in the *Paint Blendshape Weights* tool. The left, center, and right `browRaise` blendshapes were then duplicated out. Splitting up the `browRaise` shapes in this way allowed the animator to control every region of the brows independently. However, because the split maps were normalized, turning on all three regions together summed to the original pre-split `browRaise` shape. All the pre-split blendshapes underwent this process resulting in a post-split shape set of roughly 75 blendshapes. This shape set gave the Image Metrics animators an unprecedented amount of control over each region of the face.

4.3 The Facial Rig's User Interface

The rig user interface (Fig. 16) was inspired by the facial muscles and their "direction of pull". Many of the controls were represented as arrow-shaped NURBS curves, with each arrow representing a certain facial muscle. Pulling on the

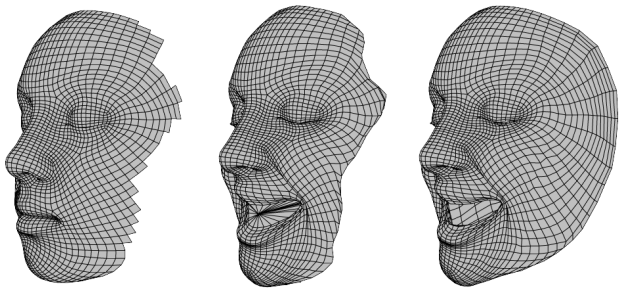


Figure 15: (a) A partial master mesh with (b) a partial blendshape expression mesh. (c) A complete expression scan, created by finding correspondences from (b) to (a) to the complete master mesh.

arrows toward its point was equivalent to contracting those muscles. Additional animation controls were placed in the Maya channel box allowing numerical input.

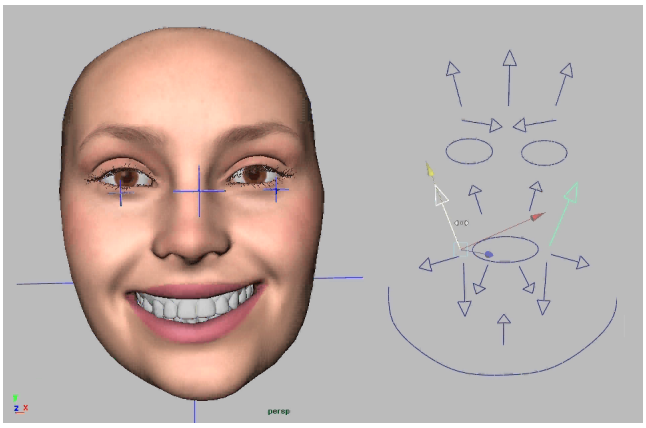


Figure 16: The user interface (right) for the Digital Emily facial rig (left).

4.4 Soft Eyes and Sticky Lips

The Emily facial rig included two notable special effects: *soft eyes* and *sticky lips*. Soft eyes is the effect of the rotation of the cornea pushing and tugging the eyelids and skin around the eye. The soft eyes setup created for Emily was relatively simple in that each eye had a separate blendshape node with four shapes: `lookUp`, `lookDown`, `lookLeft`, and `lookRight`. The envelope of this blendshape node was negated by the `blink` shape. In other words, whenever the eye blinked, the soft eyes effect was turned off. This was done to prevent a conflict between the `blink` and `lookDown` blendshapes.

Sticky lips is the subtle effect of the lips peeling apart during speech, and has been a part of several high-end facial rigs including the Gollum character built at WETA Digital for the Lord of the Rings sequels. The Emily rig had a relatively elaborate sticky lips setup involving a series of Maya deformers which provided animators a "sticky" control for both corners of the lips.

4.5 Adding Blendshape Displacement Maps

In order to preserve as much of the scan data as possible, the Emily render had 30 animated displacement maps. The displacement maps were extracted using Pixologic's ZBrush software by placing each pre-split blendshape on top of its corresponding scan and calculating the difference between the two. Then each displacement map was cleaned up in Photoshop and divided according to the same normalized split maps used to split the blendshapes. This yielded a displacement map for each blendshape in the rig (although only a subset of these were used in the final renderings). It is important to note that the highest frequency, pore-level detail displacement map came only from the neutral scan. All the other displacement maps had a median filter applied to them to remove any high frequency detail, while still keeping the wrinkles. This was done because adding two maps with high frequency detail would result in the pores "doubling up" in the render, since the high-resolutions scans had not been aligned to each other at the level of skin pores and fine wrinkles. Therefore, only the neutral displacement map had the pore detail; all the other maps had only wrinkles without pores. The displacement map animation was driven directly by the corresponding blendshape animation.

4.6 Adding the Teeth

Plaster casts of Emily's upper and lower teeth were made using standard dental casting techniques. The two casts were scanned with a structured light scanning system to produce two 600,000 polygon meshes from sixteen merged scans acquired using structured lighting patterns based on Gray codes. These teeth scans were manually remeshed to have far fewer triangles and smoother topology. The top and bottom teeth meshes were remeshed to 10,000 polygons each as in Fig. 10(c). We then carefully placed the teeth geometry inside the neutral mesh using the high-resolution smile scans as reference for the teeth positions.

The teeth scanning was done relatively late in the facial model construction process. Until the teeth were scanned, a generic teeth model was used in the animated character. We found this to be significantly less believable than using the model of Emily's actual teeth.

5 Video-Based Facial Animation

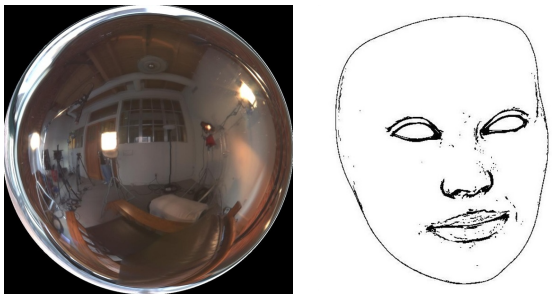
Digital Emily's facial animation was created using Image Metrics' proprietary video analysis and animation system which allows animators to associate character poses with a small subset of performance frames. The analysis of the performance requires video from a single standard video camera and is designed to capture all the characteristics of the actor. The animation technology then uses the example poses provided by an animator to generate predictions of the required character pose for each frame of the performance. The animator can then iteratively refine this prediction by adding more example poses until the desired animation is achieved. As the process is example driven, it presents a natural framework to allow artistic and stylistic interpretation of an actor's performance, for example when using a human to drive a cartoon or animal character. However, in this case the technology was used for the purpose of producing a faithful one-to-one reproduction of Emily's performance.

The Image Metrics facial animation process has several advantages over traditional performance-driven animation techniques which significantly added to the realism of the digital character. First, the process is based on video of the actor performing, which provides a great deal of information regarding the motion of the actor's eyes and mouth; these are problematic areas to capture faithfully with traditional motion capture markers, but are the most important part of the face for communicating an actor's performance. Second, the process leverages an appropriate division of labor between the animator and the automated algorithms. Because an artist is part of the process, they can ensure that each of the key animation poses reads in an emotionally faithful way to the appearance of the actor in the corresponding frame of video. This process would be difficult to automate, since it involves reading and comparing the emotional content of faces with significant detail. Conversely, the trajectories and timing of the facial motion between key poses are successfully derived from the automated video analysis; achieving realistic timing in photoreal facial animation requires a great deal of skill and time for an animator to achieve manually. As a result, the process combines what an animator can do quickly and well with what an automatic process can accomplish, yielding facial animation with higher quality and greater efficiency than either fully manual or fully automatic techniques currently allow.

6 Tracking, Lighting, Rendering, and Compositing

Emily's performance was shot from two angles using high-definition cameras: a front-on closeup for use in the Image Metrics analysis pipeline, and a medium shot in a three-quarter view to provide the background plate for the final video. The final renderings needed to be match-moved to Emily's head position in the three-quarter shot. Complicating the process (but representative of many practical production scenarios) the cameras were uncalibrated, and there were no markers on Emily's face to assist in tracking. Match moving requires sub-pixel accuracy and temporal consistency of pose, otherwise unwanted "floating" effects become easily visible. To achieve this, we developed a manually guided match-moving process. We manually set the 3D pose of the animated character on several example frames so that the projected location of the model closely matched that of Emily in the video frame. Using these frames as templates, we applied a model-based optical flow algorithm [1] to calculate the required character pose in the intermediate frames. To ensure a smooth temporal path within the pose space, we developed a novel weighted combination of template frames that smoothly favors the temporally closest set of templates. Any errors were

manually corrected and added as a new template to derive a new pose-space tracking result until the results no longer exhibited artifacts.



(a) (b)

Figure 17: (a) The light probe image acquired to simulate the live-action lighting conditions on Digital Emily's face. (b) A frame from a facial contour pass used in the compositing process.

Since Emily's face would be composited onto live-action background plate including real video of her hair, ears, neck, and body, it was imperative for the rendered version of Emily's face to look completely convincing. Refining the rendering parameters took a rendering artist approximately three months to perfect. The rendering was done in Mental Ray using the Fast Subsurface Scattering skin shader. The lighting was based on an high dynamic range light probe image captured on the day of the shoot [4] as seen in Figure 17(a). Many different passes were rendered, including diffuse, specular, matte, and contour passes. In particular, the contour pass (Fig. 17(b)) was very important for visually integrating (or "marrying") the different components of facial geometry using different amounts of blur in the composite. Otherwise, for example, the line where the eyelid meets the eyeball would appear too sharp and thus unrealistic. Compositing was done in the Eyeon Fusion package. Emily's fingers were rotoscoped whenever she moved them in front of her face so they could also obscure her digital face. Small paint fixes were done to the final renderings around the eye highlights. Two frames from a final Emily animation are shown in Fig. 18, and animated results may be seen at the web site **http://gl.ict.usc.edu/Research/DigitalEmily/**.

7 Discussion

Some timeframes and personnel for the Digital Emily Project were:

- Scanning: 1.5 hours, 3 seconds per scan, 3 technicians
- Scan Processing: 10 days, 37 processed scans, 1 artist
- Rig Construction: 3 months, 75 blendshapes, 1 artist
- Animation: 2 weeks, 90 seconds of animation, 2 animators
- Rendering/Compositing: 3 months, 1 artist
- Output: 24fps, 1920x1080 pixel resolution

A great deal of information was learned and many tools were developed and improved in the context of this project, so to apply the process again would likely require significantly fewer resources to achieve results of equal or even better quality.

***Figure 18: A final rendering from the Digital Emily animation. The face is completely computer-rendered, including its eyes, teeth, cheeks, lips, and forehead. Emily's ears, hair, neck, arms, hands, body, and clothing come from the original background plate. Animated results may be seen at:* http://gl.ict.usc.edu/Research/DigitalEmily/**

Based on the experiences of the Digital Emily project, five sufficient (and perhaps necessary) steps for achieving a photoreal digital actor are:

1. Sufficient facial scanning resolution accurate to the level of skin pores and fine wrinkles, achieved with the scanning process of [18]
2. 3D geometry and appearance data from a wide range of facial expressions, also achieved through [18]
3. Realistic facial animation, as based on a real actor's performance, including detailed motion of the eyes and mouth, which was achieved through the semi-automatic Image Metrics video-based facial animation process
4. Realistic skin reflectance including translucency, leveraging a subsurface scattering technique based on [14]
5. Accurate lighting integration with the actor's environment, done using HDRI capture and image-based lighting as described in [4]

7.1 Conclusion: Lessons Learned

The experience of creating Digital Emily taught us numerous lessons which will be useful in the creation of future digital characters. These lessons included:

- Having consistent surface (u,v) coordinates across the scans accurate to skin pore detail would be of significant use in the process; a great deal of the effort involved was forming correspondences between the scans.
- Although the Image Metrics facial animation system requires no facial markers for animation, including at

least a few facial markers as the actor performs would make the head tracking process much easier.

- Giving the digital character their own teeth, accurately scanned and placed within the head and jaw, is important for the believability of the character and their resemblance to the original person.

References

[1] Simon Baker and Iain Matthews. Lucas-kanade 20 years on: A unifying framework. *Int. J. Comput. Vision*, 56(3):221–255, 2004.

[2] G. Borshukov and J. P. Lewis. Realistic human face rendering for 'The Matrix Reloaded'. In *ACM SIGGRAPH 2003 Sketches & Applications*, 2003.

[3] Chang. Beowulf. *Variety*, Nov 2007.

[4] Paul Debevec. Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *Proceedings of SIGGRAPH 98*, Computer Graphics Proceedings, Annual Conference Series, pages 189–198, July 1998.

[5] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *Proceedings of ACM SIGGRAPH 2000*, Computer Graphics Proceedings, Annual Conference Series, pages 145–156, July 2000.

[6] Paul Debevec, Chris Tchou, Andrew Gardner, Tim Hawkins, Charis Poullis, Jessi Stumpfel, Andrew Jones, Nathaniel Yun, Per Einarsson, Therese Lundgren, Marcos Fajardo, and Philippe Martinez. Estimating surface reflectance properties of a complex scene under captured natural illumination. Technical Report ICT-TR-06.2004, USC ICT, Marina del Rey, CA, USA, Jun 2004. http://gl.ict.usc.edu/Research/reflectance/Parth-ICT-TR-06.2004.pdf.

[7] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, 1978.

[8] Shachar Fleishman, Iddo Drori, and Daniel Cohen-Or. Bilateral mesh denoising. *ACM Transactions on Graphics*, 22(3):950–953, July 2003.

[9] Abhijeet Ghosh, Tim Hawkins, Pieter Peers, Sune Frederiksen, and Paul Debevec. Practical modeling and acquisition of layered facial reflectance. *ACM Transactions on Graphics*, 27(5):139:1–139:10, December 2008.

[10] Brian Guenter, Cindy Grimm, Daniel Wood, Henrique Malvar, and Frédéric Pighin. Making faces. In *Proceedings of SIGGRAPH 98*, Computer Graphics Proceedings, Annual Conference Series, pages 55–66, July 1998.

[11] 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.

[12] Walter Hyneman, Hiroki Itokazu, Lance Williams, and Xinmin Zhao. Human face project. In *ACM SIGGRAPH 2005 Course #9: Digital Face Cloning*, New York, NY, USA, July 2005. ACM.

[13] ICT-Graphics-Laboratory. Realistic human face scanning and rendering. Web site, 2001. http://gl.ict.usc.edu/Research/facescan/.

[14] Henrik Wann Jensen, Stephen R. Marschner, Marc Levoy, and Pat Hanrahan. A practical model for subsurface light transport. In *Proceedings of ACM SIGGRAPH 2001*, Computer Graphics Proceedings, Annual Conference Series, pages 511–518, August 2001.

[15] Debra Kaufman. Photo genesis. *WIRED*, 7(7), July 1999. http://www.wired.com/wired/archive/7.07/jester.html.

[16] Hayden Landis. Production-ready global illumination. In *Notes for ACM SIGGRAPH 2005 Course #16: RenderMan in Production*, New York, NY, USA, July 2002. ACM.

[17] Wan-Chun Ma. *A Framework for Capture and Synthesis of High Resolution Facial Geometry and Performance*. PhD thesis, National Taiwan University, 2008.

[18] Wan-Chun Ma, Tim Hawkins, Pieter Peers, Charles-Felix Chabert, Malte Weiss, and Paul Debevec. Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In *Rendering Techniques*, pages 183–194, 2007.

[19] Wan-Chun Ma, Andrew Jones, Jen-Yuan Chiang, Tim Hawkins, Sune Frederiksen, Pieter Peers, Marko Vukovic, Ming Ouhyoung, and Paul Debevec. Facial performance synthesis using deformation-driven polynomial displacement maps. *ACM Transactions on Graphics*, 27(5):121:1–121:10, December 2008.

[20] Masahiro Mori. Bukimi no tani (the uncanny valley). *Energy*, 7(4):33–35, 1970.

[21] Diego Nehab, Szymon Rusinkiewicz, James Davis, and Ravi Ramamoorthi. Efficiently combining positions and normals for precise 3d geometry. *ACM Transactions on Graphics*, 24(3):536–543, August 2005.

[22] Steve Perlman. Volumetric cinematography: The world no longer flat. *Mova White Paper*, Oct 2006.

[23] Barbara Robertson. What's old is new again. *Computer Graphics World*, 32(1), Jan 2009.

[24] Mark Sagar, John Monos, John Schmidt, Dan Ziegler, Sing-Choong Foo, Remington Scott, Jeff Stern, Chris Waegner, Peter Nofz, Tim Hawkins, and Paul Debevec. Reflectance field rendering of human faces for s̈pider-man 2.̈ In *SIGGRAPH '04: ACM SIGGRAPH 2004 Technical Sketches*, New York, NY, USA, 2004. ACM.

[25] Yang Wang, Mohit Gupta, Song Zhang, Sen Wang, Xianfeng Gu, Dimitris Samaras, and Peisen Huang. High resolution tracking of non-rigid motion of densely sampled 3d data using harmonic maps. *Int. J. Comput. Vision*, 76(3):283–300, 2008.

[26] Andreas Wenger, Andrew Gardner, Chris Tchou, Jonas Unger, Tim Hawkins, and Paul Debevec. Performance relighting and reflectance transformation with time-multiplexed illumination. *ACM Transactions on Graphics*, 24(3):756–764, August 2005.

[27] Ellen Wolff. Creating virtual performers: Disney's human face project. *Millimeter magazine*, April 2003.

[28] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(11):1330–1334, 2000.