THE BIRTH OF A DIGITAL ACTOR

THABO BEELER, DEREK BRADLEY

Digital humans have become of central relevance in current film production. Out of the five top grossing films released to date (2021), four make use of digital humans of some sort. The outlier, Titanic, was shot long before digital humans were practical, back in 1997. Since then, technology has leaped forward, making photoreal digital humans a reality. Still, the creation of such digital characters poses a formidable challenge and is oftentimes considered to be the Holy Grail of visual effects. The quality bar is tremendously high, since we humans have been conditioned by evolution to scrutinize human faces and even the slightest imperfection can destroy the illusion and trigger rejection of the character—the infamous uncanny valley phenomenon as introduced by Masahiro Mori (2012 [1970]). A digital actor is typically born by first scanning his or her likeness, then building up a digital model, which can then be animated or retargeted to a separate digital character.

SCANNING THE LIKENESS

In order to cross this uncanny valley, the industry relies on reproducing reality as faithfully as humanly possible. To create effects such as explosions, water or destruction, physically based simulation has been invented. To emulate the way light interacts with materials we rely on physically based rendering. To create digital characters, studios rely on capturing real humans at different stages. While this holds in general for the entire human, we will focus on the most challenging aspect in the remainder of this article—the human face.

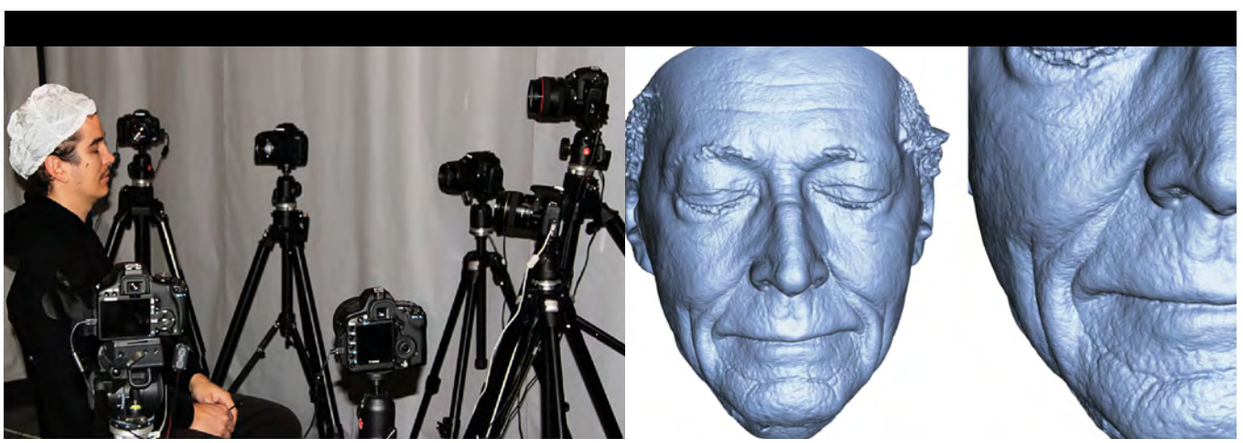

FIG. 1
MULTIVIEW RECONSTRUCTION SETUP AND EXAMPLE 3D GEOMETRY.

First, the shape and appearance of a human face will be acquired using scanning technologies. Shape refers to the 3D geometric structure of the face, where appearance denotes its color and how it interacts with light. Interestingly, the delineation between shape and appearance is scale-dependent, and small scale structures such as skin detail may sometimes be considered geometry or modeled as part of skin reflectance. To acquire the shape, an actor is typically captured in a photogrammetry or videogrammetry setup (Beeler/Bickel/Sumner/Beardsley/Gross 2010). Such setups consist of multiple cameras, ranging anywhere from two to two-hundred, and are accompanied by reconstruction software that converts the multi-view face images to a 3D face model. At their core, these algorithms all function similarly as they rely on the fact that points at different distances from the cameras will project to different locations inside the captured images. Hence, identifying corresponding features across views allows to triangulate them in space to recover their 3D position. While there are other methods that employ more advanced hardware for 3D scanning, such as structured light (Weise/Li/Van Gool/Pauly 2009) or depth sensors (Li/Yu/Ye/Bregler 2013), nowadays passive photogrammetry has become the method of choice since it is highly accurate and requires relatively inexpensive hardware.

As indicated, shape alone is not sufficient to render realistic images since it lacks the appearance information. Skin exhibits very intricate appearance properties due to its physical structure, which makes both acquisition and reproduction very challenging. A fraction of the incident illumination is reflected off the top layer of the skin, the so-called stratum corneum, causing the highlights called specular reflections. How much light is reflected depends on the incoming light direction relative to the local orientation of the skin surface. These reflections preserve the color of the light source, since the rays do not interact with the lower skin layers at all. The rest of the incident light traverses the stratum corneum entering lower skin layers, specifically epidermis and dermis. While it travels through this volume, bouncing from molecule to molecule it gradually changes its color as skin absorbs certain parts of the light spectrum more than others. At some point it will exit the skin again, now appearing flesh colored. The exact color depends on the parts of the skin it travelled through and the distance it covers. This property is called subsurface scattering and is the reason why our skin appears soft and translucent. To measure these reflectance properties, typically studios rely on a lightstage, which is a large device that allows to illuminate the face from a large number of light directions (up to several hundred) in a controlled way (Debevec/Hawkins/Tchou/Duiker/Sarokin/Sagar 2000). By observing the varying appearance at a specific point on the face lit from multiple of these light sources, it is possible to estimate the reflectance properties of the face, such as albedo (color without shading) or specular attenuation. The need to be captured in two different setups—one for shape and model building (described next), and one for reflectance—is obviously suboptimal and recent research has suggested a way to add appearance acquisition to the well-established videogrammetry setups, yielding a one-stop-shop to digitize the likeness of the human face (Gotardo/Riviere/Bradley/Ghosh/Beeler, 2018; Riviere/Gotardo/Bradley/Ghosh/Beeler 2020).

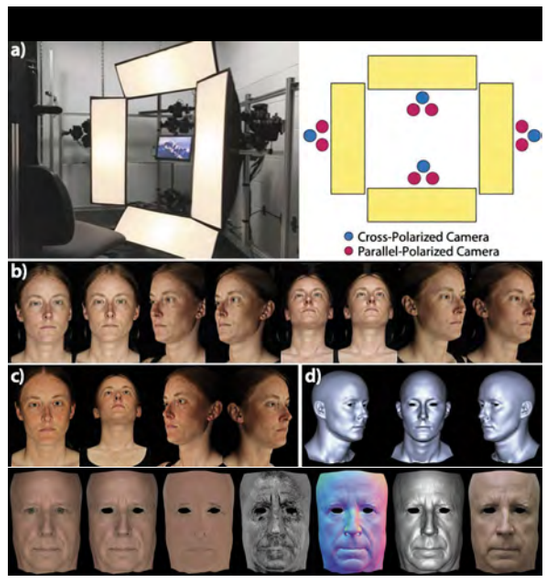

FIG. 2
COMBINED FACIAL GEOMETRY AND APPEARANCE CAPTURE FROM RIVIERE, GOTARDO, BRADLEY, GHOSH, BEELER 2020. A) CAPTURE SETUP, B) EIGHT PARALLEL-POLARIZED VIEWS, C) FOUR CROSS-POLARIZED VIEWS, D) RECONSTRUCTED GEOMETRY. BOTTOM: EXAMPLE RECONSTRUCTED APPEARANCE, INCLUDING (FROM LEFT TO RIGHT) ORIGINAL IMAGE, RECONSTRUCTED RENDER, DIFFUSE ALBEDO, SPECULAR INTENSITY, NORMALS, HIGH QUALITY GEOMETRY, ANOTHER RENDER UNDER DIFFERENT LIGHTING. © DISNEY.

MODEL BUILDING

The scanning technologies introduced in the previous section are employed to capture a human actor performing a number of expressions. How many and which expressions depend on the requirement of the model building stage (also known as *rigging)*. The typical approach is to model the human face holistically using a *linear blend-shape model*—a strategy that represents all facial expressions as a linear combination of a set of base expressions (Lewis/Anjyo/Rhee/Pighin/Deng 2014). After scanning the base expressions and converting them into a common mathematical representation where vertices correspond between shapes, new expressions or entire performances can be created and animated by manipulating the weights of the base shapes in the linear combination. These models are extremely robust and intuitive

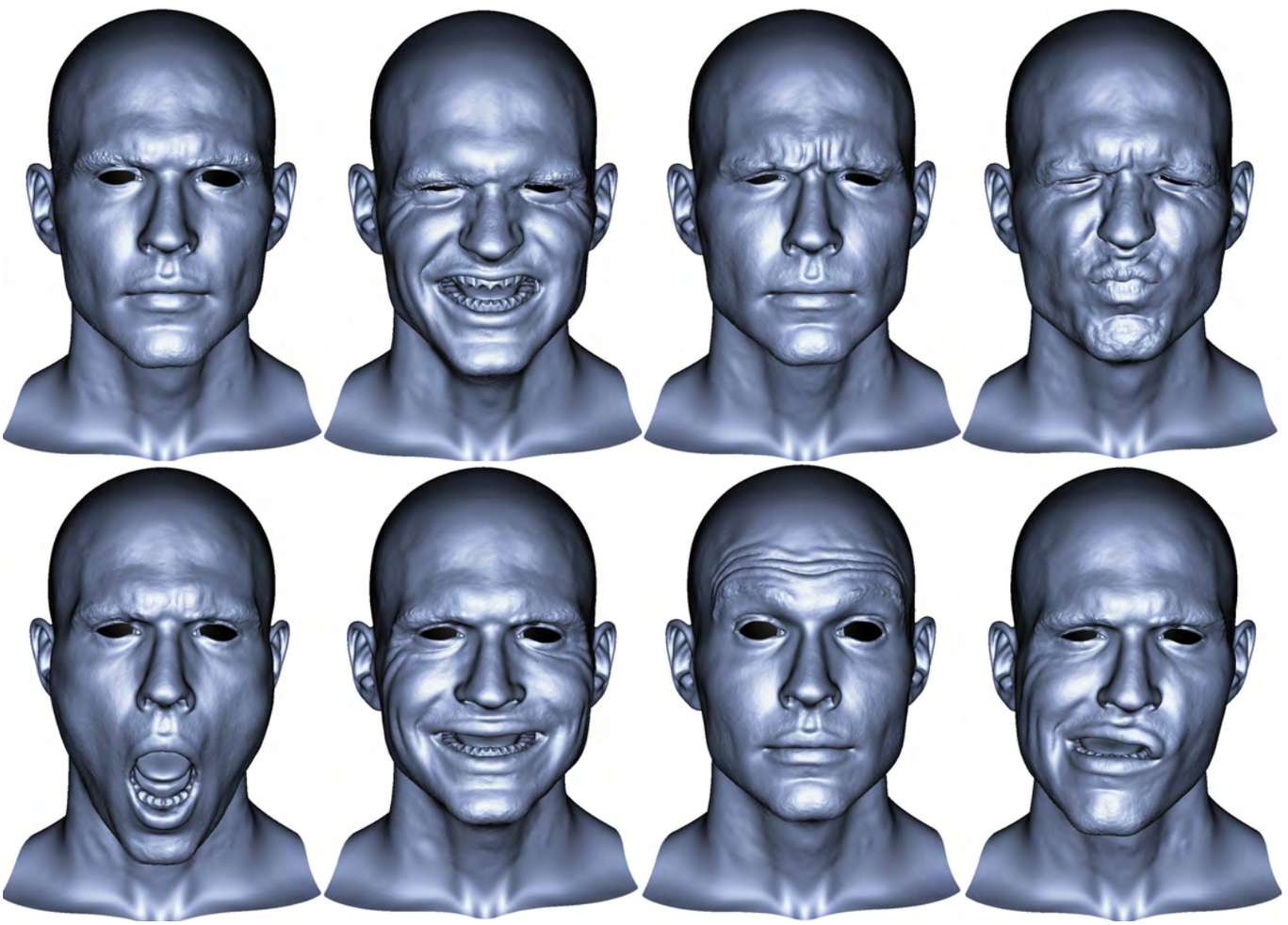

FIG. 3
EXAMPLE FACIAL EXPRESSIONS THAT ARE SCANNED USING MEDUSA TO FORM AN ACTOR MODEL. © DISNEY.

for artists, yet they require to capture a large number of base shapes in order to provide sufficient flexibility and expressiveness. For example, a model consisting of two shapes, say neutral and smile, will be able to express these two shapes and any shape in-between, such as a half smile, but it will not be able to express, for example, a frown. Hence people rely on sampling the human expression space as systematically as possible, oftentimes following the *Facial Action Coding System* or FACS system as introduced by Ekman and Friesen (1978). This system isolates facial motion into atomic "Action Units" (AU) as a function of the underlying muscle structure. For example a smile may be produced as a combination of

AU 6 (cheek raiser), AU 14 (dimpler), AU 15 (lip corner depressor), AU 17 (chin raiser), and AU 24 (lip pressor), among others (Schmidt/Cohn 2001). As a result, global blend-shape facial rigs typically require the acquisition of hundreds of base shapes. The challenge of this system in practice is, however, that humans are typically not capable of consciously activating individual muscle groups. Furthermore, since most scanning systems are seated, effects such as secondary dynamics due to motion are not captured and hence cannot be modelled. Nevertheless, building an actor model from a fixed number of scanned base shapes is still the method of choice for most productions today.

To overcome some of the limitations of global blend-shape models, recent research has proposed more flexible *local* face models, which model the deformation of faces in smaller local regions (Tena/de la Torre/Matthews 2011; Wu/Bradley/Gross/Beeler 2016). This allows to obtain more variation in the combined facial expressions from fewer scanned base shapes, but this approach comes at the cost of robustness, as inconsistent behavior across different local regions can result in uncanny shapes that no longer resemble faces. Such a problem, however, can be alleviated by considering spatial regularization and the global anatomical structure of the face, in particular modeling the expression-specific skin thickness between the skin surface and the bones (e.g. skull and mandible) (Wu/Bradley/Gross/Beeler 2016). Such an anatomical local face model has become a new option for digital face modeling in recent film productions.

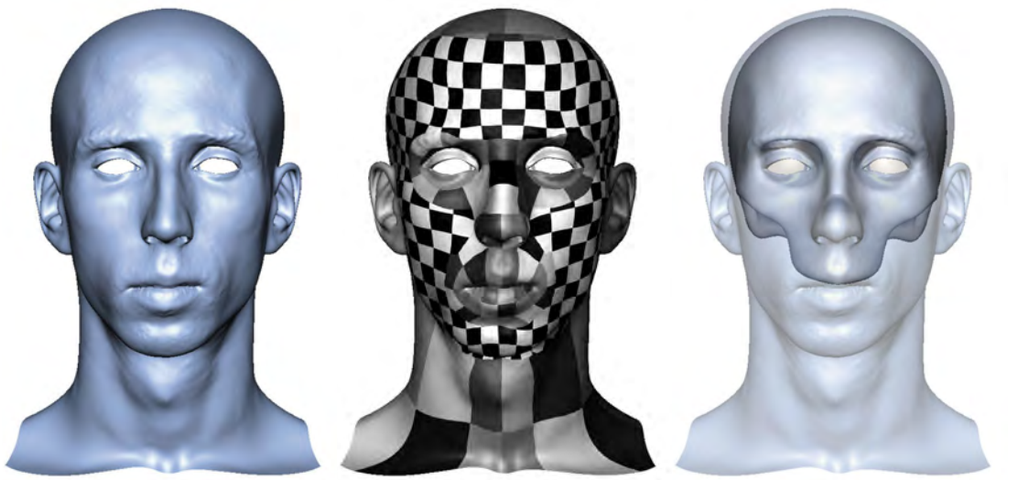

FIG. 4

LOCAL ANATOMICAL FACE MODEL FROM WU, BRADLEY, GROSS, BEELER 2016, SHOWING THE FACE GEOMETRY, THE DISTRIBUTION OF SKIN PATCHES, AND THE UNDERLYING BONE STRUCTURES. © DISNEY.

PERFORMANCE ANIMATION

Once an actor's face model is built, it can be animated by artists using the rig to create novel performances for the digital avatar. However, obtaining precisely accurate facial motion is extremely difficult by hand, and oftentimes digital characters that look realistic in still life, suddenly look uncanny when they move or speak. For this reason, it is customary to also capture performance animation from the real actors, and map this onto the digital character.

Performance capture has a long history, and by now most people are familiar with the idea of motion capture suits for body tracking, where actors wear tight fitted clothing covered in retro-reflective balls that can be tracked using infrared cameras. This notion of "sparse" mo-cap is then translated into full body character animation. For the case of facial performances, similar sparse motion capture has been achieved by painting small black or colored dots on the actor's face and tracking them in video. This provides a low-resolution repre-

sentation of the facial performance, which can be used to approximately drive the higher-resolution actor model. A major drawback of marker-based performance capture is the need to place the markers on the actor's face, which not only requires time, but also requires the makeup artist to place the markers in exactly the same locations from day to day over the course of a production. Furthermore, the result is only a rough approximation of the facial motion, since accurate information can only be obtained at the marker locations. For this reason, follow up research focused on markerless "dense" performance capture of faces (Beeler/Hahn/Badley/Bickel/Beardsley/Gotsman/Sumner/Gross 2011), which uses high resolution synchronized video cameras and tracks the face at the skin pore level, yielding up to a million accurate point locations tracked for each frame of a performance. Such a system, called Medusa, was developed by DisneyResearch|Studios and was awarded an Academy Science and Technology Oscar in 2019.[1]

FIG. 5

SEVERAL FRAMES FROM AN ANYMA PERFORMANCE CAPTURE, SHOWING THE INPUT WITH WIREFRAME RESULT OVERLAID (TOP) AND STABILIZED GEOMETRY WITH RIGID SKULL MOTION REMOVED (BOTTOM). © DISNEY.

Markerless facial performance capture systems like Medusa typically require a well-controlled studio setup, consisting of many cameras, bright uniform illumination, and the actor is usually required to stay seated with their head in a fixed position—all in an effort to make the complex problem tractable. In order to allow a more "unencumbered" performance, more recent research has investigated model-based performance capture, where an actor-specific motion prior is used during the reconstruction process (eg. Tewari, Zollöfer, Kim, Garrido, Bernard, Perez, Christian 2017). This prior is often just the actor model (or rig) built from the facial scans as described above. Optimizing for the model parameters directly to match the video (so-called 'analysis-by-synthesis') is a convenient way to recover the performance, since the output is already in the required format (e.g. mapped onto the actor rig). Using the prior also constrains the set of possible facial shapes that can be reconstructed, and this allows to lessen the physical burden, allowing performance capture using less cameras (even just 1), and in less constrained environments like outdoors. Along these lines is the Anyma performance capture system,[2] developed by DisneyResearch|Studios, which uses the anatomical local face model mentioned earlier (Wu/Bradley/Gross/Beeler 2016) as the prior. This approach provides among the highest quality 3D facial performances and has been used in several recent film productions.

1 https://studios.disneyresearch.com/medusa/ (02.02.2022)

2 https://studios.disneyresearch.com/anyma/ (02.02.2022)

RETARGETING

There are several practical scenarios for creating digital actors. One very common one is when there is a need for a digital character that does not exist in the real world, but can still be "performed" by a real actor (e.g. Marvel's Hulk performed by Mark Ruffalo). In such scenarios, a digital version of the actor's performance is generated, and then transferred to the ultimate digital character seen on screen—a process known as performance *retargeting* (Ribera/Zell/Lewis/Noh/Botsch 2017).

When retargeting a human actor performance to a separate digital character, it is common practice to undergo most of the process of scanning and model building for the live actor, even though they may never appear on screen. This helps to reduce the problem to a purely

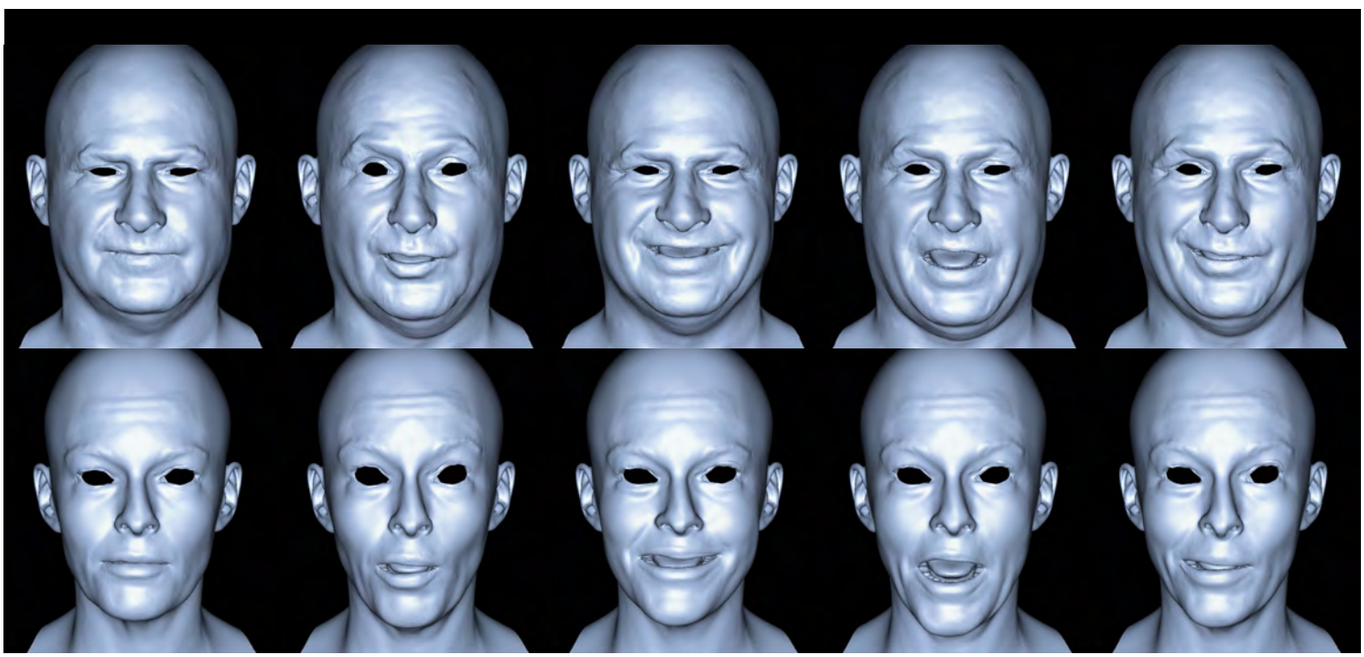

FIG. 6
SEVERAL FRAMES OF A PERFORMANCE RETARGETED FROM THE SOURCE CHARACTER (TOP) TO A TARGET CHARACTER (BOTTOM). © DISNEY.

3D geometric one—how to map the deformation of one 3D face (the source) onto another (the target). The problem is challenging because oftentimes the source and target faces have different proportions, different dynamics and could even have very different bone and muscle structures (for example, a human mapped to a dragon). Several different methods for performance retargeting have been developed over the years. The most straightforward approach is to build complete, identical blend-shape models for both the source and target characters, and once the captured performance of the source actor is modeled by a time series of blend weights, the same weights can simply be used on the target blend-shape model to obtain the retargeted performance. The obvious downside of this approach is the time and effort required to build two complete and expressive blend-shape rigs, carefully crafted such that the expressions artistically correspond in the desired way. But once the models are constructed, large numbers of performance sequences can be readily converted from the source character to the target in no time, and thus this approach is popular in production environments.

When constructing multiple blend-shape models is impractical or otherwise unattractive, an alternative approach is to retarget the surface motion directly. If the source and target characters do not differ greatly in proportions, the time-varying per-vertex

displacements of the source performance can be directly added to the target neutral face. This approach, often referred to as *delta transfer,* is particularly handy when the target character should retain the fine scale details (e.g. expression wrinkles) of the source actor. When the source and target geometry are not compatible enough for simple delta transfer, a common, more elaborate approach is *deformation transfer* (Sumner/Popovic 2004), where the set of transformations induced by the triangles in the source mesh are transferred to the triangles of the target mesh, and the target vertex positions are then solved as an optimization problem. While offering flexibility in the performance transfer, this method is sometimes prone to geometric artifacts such as pinching and folding of the surface when the deformation is extreme. Compared to the blend-shape weight transfer method described first, these direct transfer operations do not require elaborate character rigs but do require a spatial correspondence between the source and target faces, and the resulting retargeted performance can be difficult to edit as no rig exists. In practice, many people use some form of hybrid approach, where large scale deformation is retargeted using a *reduced* blend-shape model of limited size and expressivity, and then finer scale details are retarged with a direct method, in some sense retaining the best of both worlds.

CONCLUSION

In this article we discuss the birth of a digital actor, involving several important steps like scanning the likeness, model building, performance animation and retargeting. Each of these steps has seen tremendous growth in technology and research over the last decades. With every new innovation, we get closer and closer to being able to cross the uncanny valley, and perhaps some digital avatars already have.

Despite the progress, researchers continue to push the limits of "actor to avatar" technology. One current drawback is that most methods for high quality performance capture require a lot of processing time and are thus executed offline after the performance video has been recorded. In this era of real-time virtual production where film-makers are eager to direct digital characters in digital environments in real-time,[3] the next big challenge will be production-quality face capture at real-time frame rates, from a single camera directly on a film set. While some of the latest capture methods do achieve real-time performance (Feng/Feng/Black/Bolkart 2021), they do not yet reach the quality of offline approaches. Still, with the speed that technology is developing in this field, we likely won't have to wait much longer.

3 https://www.fxguide.com/fxfeatured/how-virtual-production-worked-on-set-of-the-lion-king/ (02.02.2022)

REFERENCES

Beeler, Thabo/Hahn, F./Badley, Derek/Bickel, B./Beardsley, P./Gotsman, C./Sumner, R. W./Gross, M. (2011), High-quality Passive Facial Performance Capture Using Anchor Frames. ACM Trans. Graphics (Proc. SIGGRAPH) 30, 75:1–75:10.

Beeler, Thabo/Bickel, B/Sumner, R./Beardsley, P./Gross, M. (2010), High-quality single-shot capture of facial geometry. ACM Trans. Graphics (Proc. SIGGRAPH).

Debevec, Paul/Hawkins, Tim/Tchou, Chris/Duiker, Haarm-Pieter/Sarokin, Westley and Sagar, Mark (2000), Acquiring the Reflectance Field of a Human Face, in: Proceedings of the 27th annual conference on Computer graphics and interactive techniques. ACM Press/Addison-Wesley Publishing Co., ACM, p. 145–156.

Ekman, Paul/Friesen, W (1978), Facial Action Coding System: A Technique for the Measurement of Facial Movement. Palo Alto.

Feng, Yao/Feng, Haiwen/Black, Michael J./Bolkart, Timo (2021), Learning an Animatable Detailed 3D Face Model From In-the-wild images, ACM Transactions on Graphics. Volume 40, Issue 4, Article No.: 88, pp. 1–13.

Gotardo, Paulo/Riviere, Jérémy/Bradley, Derek/Ghosh, Abhijeet/Beeler, Thabo (2018), Practical Dynamic Facial Appearance Modeling and Acquisition. ACM Transactions on Graphics (TOG) 37, 6, 232:1–232:13.

Lewis J. P./Anjyo, K./Rhee T./Zhang M./Pighin F./Deng Z. (2014), Practice and Theory of Blendshape Facial Models, in Eurographics—State of the Art Reports, Lefebvre S., Spagnuolo M., (Eds.), The Eurographics Association, p. 199–218.

Li, H., Yu, J., Ye, Y., Bregler, C. (2013), Realtime Facial Animation With on-the-fly Correctives, ACM Trans. Graphics (Proc. SIGGRAPH) 32, 4, 42:1–42:10.

Mori, Masahiro (2012 [1970]), The Uncanny Valley. IEEE Robotics & Automation Magazine. 19 (2): p. 98–100.

Ribera, R./ Zell, E./Lewis, J. P./Noh, J./Botsch, M. (2017), Facial Retargeting with Automatic Range of Motion Alignment, ACM Trans. Graph. 36, 4, 154:1–154:12.

Riviere, Jérémy/Gotardo, Paulo/Bradley, Derek/Ghosh, Abhijeet and Beeler, Thabo (2020), Single-shot High-quality Facial Geometry and Skin Appearance Capture. ACM Transactions on Graphics (TOG) 39, 4, 81:1–81:12.

Schmidt, K.L./Cohn, J.F. (2001), Dynamics of Facial Expression: Normative Characteristics and Individual Differences. IEEE International Conference on Multimedia and Expo, 2001. ICME 2001.

Sumner, R. W./Popovic, J. (2004), Deformation Transfer for Triangle Meshes, ACM Trans. Graph. 23, 3), p. 399–405.

Tena J. R./De la Torre, F./Matthews, I. (2011), Interactive Region-based Linear 3D Face Models. ACM Trans. Graph. 30, 4, 76:1–76:10.

Tewari, A./Zollöfer, M./Kim, H./Garrido, P./Bernard, F./Perez, P./Christian, T. (2017), MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction, in: Proceedings IEEE International Conference on Computer Vision (ICCV).

Weise, T., Li, H., Van Gool, L., Pauly, M. (2009), Face/off: Live Facial Puppetry, in: Proc. SCA, p. 7–16.

Wu C./Bradley, D./Gross M/Beeler, T. (2016), An Anatomically-constrained Local Deformation Model for Monocular Face Capture, ACM Trans. Graph. 35, 4, 115:1–115:12.