Automatic Generation and Stylization of 3D Facial Rigs

Fabien Danieau* — Technicolor, France
Ilja Gubins — Utrecht University, Netherlands
Nicolas Olivier — ESIR, France
Olivier Dumas — Technicolor, France
Bernard Denis — Technicolor, France
Thomas Lopez — Technicolor, France
Nicolas Mollet — Technicolor, France
Brian Frager — Technicolor Experience Center, USA
Quentin Avril — Technicolor, France

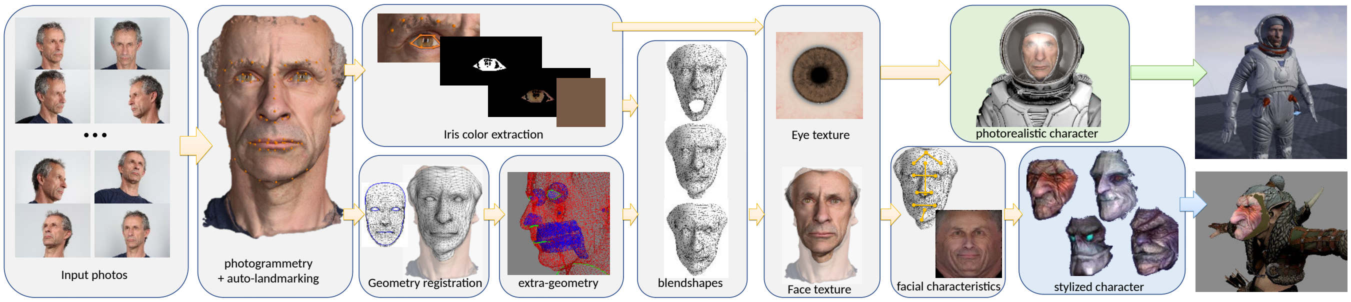

Figure 1: Overview of the automatic pipeline for generating high quality characters. Input is a set of photos of one's face and the output is a fully rigged character. The face is first reconstructed using photogrammetry and automatic landmarking. A generic face is then automatically registered on top while the color of the iris is extracted. Extra-geometry such as jaws, teeth, or nostrils are transferred. Blendshapes are transferred from the generic face. Facial and eye texture are applied to the registered mesh. The face is eventually merged to the generic body. Facial characteristics may also be extracted to apply the unique facial morphology to a non-human character.

Abstract

In this paper, we present a fully automatic pipeline for generating and stylizing high geometric and textural quality facial rigs. They are automatically rigged with facial blendshapes for animation, and can be used across platforms for applications including virtual reality, augmented reality, remote collaboration, gaming and more. From a set of input facial photos, our approach is to be able to create a photorealistic, fully rigged character in less than seven minutes. The facial mesh reconstruction is based on state-of-the art photogrammetry approaches. Automatic landmarking coupled with ICP registration with regularization provide direct correspondence and registration from a given generic mesh to the acquired facial mesh. Then, using deformation transfer, existing blendshapes are transferred from the generic to the reconstructed facial mesh. The reconstructed face is then fit to the full body generic mesh. Extra geometry such as jaws, teeth and nostrils are retargeted and transferred to the character. An automatic iris color extraction algorithm is performed to colorize a separate eye texture, animated with dynamic UVs. Finally, an extra step applies a style to the photorealistic face to enable blending of personalized facial features into any other character. The user's face can then be adapted to any human or non-human generic mesh. A pilot user study was performed to evaluate the utility of our approach. Up to 65% of the participants were successfully able to discern the presence of one's unique facial features when the style was not too far from a humanoid shape.

Keywords: character, animation, pipeline, virtual reality

Index Terms: I.2.10 [artificial intelligence]: Vision and Scene Understanding—Intensity, color, photometry, and thresholding; I.3.7 [computer graphics]: Three-Dimensional Graphics and Realism—Animation

1 Introduction

Digital humans are key aspects of the rapidly evolving areas of virtual reality, augmented reality, virtual production and gaming. Even outside of the entertainment world, they are becoming more and more commonplace in retail, sports, social media, education, health and many other fields. In the context of virtual reality, the digital personalized representation of the user highly increases immersion, presence and emotional response [38]. However, the fast creation of photorealistic characters is still challenging. Setting up a facial rig remains a long, manual and tedious artistic task. This is because people are extremely sensitive to subtle variations in facial morphology. The well-known concept of the uncanny valley encapsulates the central challenge in creating digital humans in general, and especially digital doubles of real people [32]. Many current solutions avoid this problem by skewing towards a very stylized or abstracted character. Nonetheless, our digital lives are increasingly intertwined with our identities. Setting up a quick, automated and photoreal facial rig pipeline for real-time usage encompasses many important scientific and technical challenges. The geometry, the texture, the material of the face and all of the extra geometry elements (eyes, jaws, teeth, etc.) must be properly captured and modeled.

Capturing the 3D static mesh of a face in high resolution with high-frequency details remains a key-issue. It has been studied for decades and still suffers from expensive and bulky hardware to set up, a long capture protocol to capture all the deformations of the face, and significant computation time to reconstruct meshes and textures. Photogrammetry has become increasingly popular in visual effects pipelines for almost every aspect of production, starting from the capture of a film set for previsualization and reference for artists [44], to the creation of digital doubles for starring actors [9]. This makes photogrammetry a favorable choice for creating photorealistic virtual humans [1].

In addition to the steps of capture and modeling, the ideal

*e-mail:fabien.danieau@technicolor.com

pipeline would also allow the blending of the constructed facial morphology with any other style of character. Blending personalized facial features into other characters extends the use cases beyond photoreal facsimiles of people, which are useful but limited in context. One can imagine many entertainment and gaming applications for embodying characters from favorite science fiction or fantasy worlds and infusing those creatures with one's own facial morphology.

In this context, this paper presents two contributions:

1. A complete automatic pipeline for the creation of high quality facial rigs. It relies on state-of-the-art photogrammetry, facial landmarking, mesh registration and deformation transfer algorithms. To the best of our knowledge, this is the first combination of these algorithms into an automatic system.

2. A novel style transfer method for facial meshes. Geometry and texture are modified and adapted to match a specific content. We have conducted a preliminary pilot study to identify the possibilities of such an approach with both humanoid and non-humanoid faces.

2 Related work

We first review techniques for acquiring the geometry and appearance of a face. In a second section, we detail existing approaches for facial animation. Then we survey the existing pipelines for creating real-time characters. Finally, research results of the recent field of style transfer are detailed.

2.1 Facial acquisition

The problem of facial acquisition can be split into 3D facial geometry acquisition, and facial appearance acquisition.

The methods for 3D facial geometry capture developed in the last two decades can be divided into active and passive systems. Active capture systems require special-purpose hardware, and extra constraints in setup. Such systems are usually based on laser, structured light, gradient-based illumination [24], or even requiring spatial multiplexing [41]. While the results they provide are often very robust, passive systems are much more versatile and adaptive, allowing different arrangements of setup, numbers of camera, and virtually no constraint on camera position [3]. Passive techniques have the advantage of non-intrusiveness and capture what is observed. Beeler et al. presented a passive stereo vision system that computes the accurate 3D geometry of the face with a laser scanner [3]. This work makes assumption of constant omni-directional illumination. This constraint can be released by estimating the environment map [42].

Facial appearance acquisition is the way to record the complex interaction of the light with the skin. Two general categories of such methods are distinguished: image-based methods and parametric methods. Image-based methods exhaustively capture the exact face appearance under various lighting and viewing conditions, and then solve the rendering problem through weighted image combinations [17]. Whereas the parametric methods aim at modeling the structure of the skin with suitable approximations. Such representation is more flexible but at the cost of a potentially inexact reproduction [11, 13].

Photogrammetry is an image-based passive system [3]. Thus, with a simple setup, a precise model and a basic skin texture can be captured.

2.2 Facial animation by deformation transfer

Facial animation can be achieved by a large variety of different methods: skeletons and joints [25], physically-based muscle models [39] and combinations of blendshapes [5]. While every method has a fitting application, using linear blendshape models is the most widely spread approach for high fidelity facial animation. Combining a set of blendshapes produces an arbitrary facial expression. Creating high quality blendshapes is time-consuming and tedious, requiring either high quality motion capture of real actor (and subsequent cleanup and post production) or manual modeling. However, they can be transferred from one model to another with deformation transfer [34]. This method requires triangle correspondences between source and target meshes, which is problematic if meshes have different topology. Pawaskar et al. proposed a technique to transfer blendshapes to a target mesh by first registering source mesh into target mesh using a non-rigid ICP (iterative closest point) algorithm, and then transferring deformation to a new target mesh that has direct triangle-wise correspondence [30].

2.3 Full pipeline for character creation

Malleson et al. recently proposed a pipeline for the rapid creation of VR avatars [26]. They capture a single picture of a face which is fit to a rigged template avatar. As they only use one stereo DSLR camera, parallax does not allow to finely acquire the topology of the face and reconstruct a precise mesh. They compared their results to photogrammetric scans that highlight missing geometric features such as the shape of the nose. Nagano et al. relied on a single image and deep learning (GAN) to generate a virtual face [29]. The accuracy of the geometrical reconstruction is thus limited although producing plausible results. A pipeline for a full body capture has been set up by Achenbach et al. [1]. They used two camera rigs, one for the body (40 cameras) and a second one for the face (eight cameras). A rigged template mesh is fit to the two captured point clouds. The full process takes ten minutes according to the authors. They have evaluated the realism of the captured avatar and observed that such an avatar improves the feeling of body ownership but might also look uncanny [19]. The authors pointed out that the face is a crucial part of the avatar but did not study it in detail. In a way, our approach is comparable to their pipeline, but we focus only on the face to capture a high-quality model, and to understand the artifacts that lead to an uncanny effect. Instead of using camera rigs, characters can be computed from RGB-D videos. Alldieck et al. fit a modified SMPL model to the body detected in each frame [2]. Even if the global results from a video are impressive, it turns out to be hard to determine who are the individuals are without the texture. In a close-up VR experience, that would lead to too uncanny results. While this approach requires a simple setup, the quality of the reconstructed character is limited.

2.4 Style transfer

Image style transfer has recently known a breakthrough thanks to deep neural networks. Gatys et al. make use of a classic VGG network [12], and define the content of an image as its deep features, and its style as its inter-features' correlation (Gram matrices). Using a content image, and a style image, a third image can thus be computed. Markov Random Fields in replacement of the Gram matrices allows to control the image layout at a local level and make the result more realistic [21]. This feature has allowed to extend the method for facial texture transfer [16], for which certain facial features must be preserved. It has also been showed that morphing the face of the style image to the shape of the face of the content image improves local features matching. Nevertheless, wrong matches may occur and can be solved by semantics masks [7].

These works are however limited to images. First approaches have recently investigated style transfer between 3D meshes. Ma et al. made use of a style model (exemplar), a content model (target), and a model with the style of the target but the content of the exemplar (source) [23]. The result is computed from these three meshes: i) compute the transformation from the source to the target by mapping subsets of these models with a point-to-point correspondences with minimal deformations, ii) compute the transformation from the source to the exemplar, iii) approximates the transformation from the exemplar to the result. In another method, Lun et al. input a content shape and a style shape [22]. A hierarchical segmentation of both is performed, followed by a matching of the parts. Then the style distance is minimized by a set of operations, substitution, addition, removal, and deformation, applied in that order. Additionally, a functionality constraint is used, based on the gross elements' shapes. These two approaches are however limited to simple objects (i.e. furniture).

3 Pipeline for automatic creation of facial rigs

Our pipeline inputs multiple photos of someone's face and a generic rigged character (see Figure 2). It outputs the generic character adapted to the captured face. The pipeline relies on *Meshroom*[1], an open source implementation of photogrammetry reconstruction algorithms. We have extended it to enable the mesh registration, the transfer of blendshapes and the mesh fitting to the generic body.

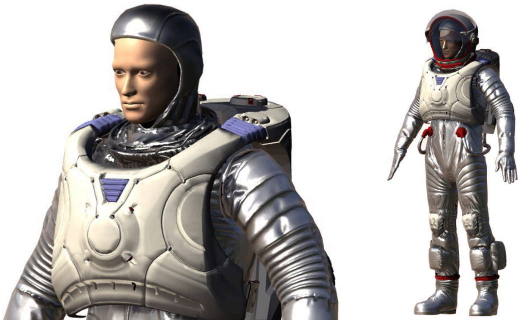

Figure 2: Example of a generic mesh: an astronaut. The face will be modified to look like the user given pictures of his or her face.

3.1 Camera setup

The first step of our pipeline is the facial acquisition. Using guidelines for close range photogrammetry [37, 40] we have built a capture setup as illustrated on Figure 3. It is composed of 14 *Canon EOS 1300D* DLSR cameras. Nine are equipped with a *Canon EF 50mm* prime (fixed focal length) lens and five with a *Canon EF 85mm*. The lighting system is composed of two *Kino Flo Tegra 455 DMX* (each composed of four neon lamps) and five LED panels. All light sources are covered with light diffuser sheets to get a more diffuse and homogeneous lighting. Triggering is hardware synchronized. One essential aspect of photogrammetry is the features matching between photos to form a single contiguous model. To support such matching, a very strong overlap (+70%) is required [37]. Our fourteen cameras ensure this overlap for capturing the face from ear to ear (see Section 3.8.1).

During the capture, the seated subject is asked to look at the frontal camera. This ensures that all captured faces are aligned in the same coordinate system where the central camera is located at its center. If needed, the height of the seat can be adjusted.

3.2 Meshing and texturing

The reconstruction process is based on the default pipeline of Meshroom for which minor elements were adjusted. First, feature extraction based on SIFT descriptors is performed. Then, images are matched based on a vocabulary tree of these descriptors. For each pair of images, the features are also matched. From this data, the rigid scene structure, as well as position and pose of the cameras, are computed (structure for motion [28]). This allows to compute

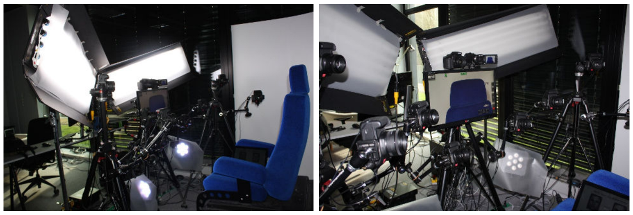

Figure 3: Our photogrammetry setup for scanning users' face composed of fourteen DSLRs and seven light sources covered with diffuser sheets.

the depth map of the viewport of each detected camera. These depth maps are filtered to ensure a global consistency. At this point the mesh is created by fusing the depth maps [15]. A filtering step is performed to clean the dense mesh and a decimation in which we limited the number of vertices to 50k. It appears to be the best balance between keeping high geometrical details and providing good performance during the registration step. Finally, the mesh is textured with a LSCM parametrization, generating a texture atlas [20].

3.3 Automatic face landmarking

An automatic landmark detection is then applied on the reconstructed textured mesh (see Figure 4). We trained 5000+ facial images annotated with 66 landmarks in the Deep Alignment Network (DAN) [18]. The facial images include Helen, LFPW, and 2300 frontal face images extracted from the Multi-PIE database [14]. The landmarks detector captures the viewport image of our 3D mesh viewer and predicts 66 facial landmarks via the retrained DAN model. To simplify computation, the viewport is captured using an orthographic camera. The predicted 2D landmarks are back projected to the facial mesh in the 3D viewer by ray-triangle (or ray-point) intersection algorithm. To get better jaw line landmarks, we also run the DAN algorithm on both side views, left and right. As the prediction of their positions is more precise and accurate on the side views, these are the values we trust. Positions of the other landmarks (eyes, eyebrow, nose, mouth and chin) are taken from the prediction of the front-view picture.

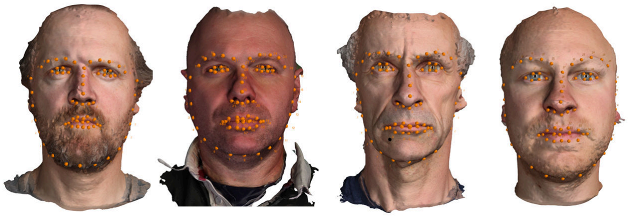

Figure 4: Automatic facial landmarking. Based on DAN, 66 landmarks are computed from the frontal view of the facial mesh.

3.4 Iris Color Extraction

Within this next step of our pipeline, and based on the previously computed landmarks, the mean color value of the eye iris is extracted. Due to its vibrant colors and its texture, the iris is the most visible and distinguishable part of the human eye [27]. We consider it as an extra geometry of the mesh and animate it using dynamic UVs. The eye texture is separated from the facial mesh one. Based on the front view of the mesh and the set of 2D landmarks, we first compute the convex hull of the six landmarks of the right eye. We use this convex hull to create a binary mask to crop the input image to isolate the eye. We convert the image in the HSV representation. As the human iris ranges from light blue to dark brown, we

[1] https://alicevision.github.io

create lower and upper color bounds to get rid of the sclera (eyes' white) and the pupil. We create a mask out of these bounds and crop the eye image. We then average the remaining pixels to get a mean value of the iris color. This mean color value is finally used to color a generic eye texture with black and white iris. Results are presented on the figure 5. From light to dark eyes, colors are correctly identified even if blue is more seen as blue-grey. The left part of each figure element is the raw rendered character on which we run the iris color extraction algorithm. The top right one is the computed color and the bottom right, the colored iris we obtain.

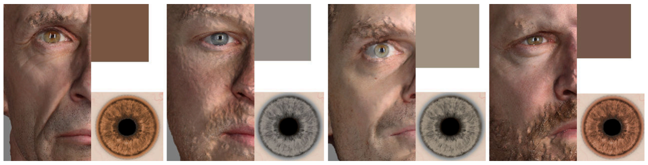

Figure 5: Results of the automatic iris color extraction.

3.5 Registration and blendshapes transfer

The goal of this step is to register the generic face mesh to the reconstructed one. This will allow to move the vertices of the generic mesh to make its geometry like the reconstructed mesh (see Figure 6). Using the approach of Sumner et al. [35], we morph the generic mesh to the reconstructed mesh by solving per-vertex affine transformation. The landmarks, computed previously, constraint the optimization process which corresponds to an iterative closest point algorithm (ICP) with regularization. The triangle correspondence is computed and for each vertex of the generic face mesh, we have the corresponding point on the photogrammetry mesh. This point, which is not a vertex, is expressed in barycentric coordinates.

Using this correspondence, we transfer the blendshapes from the original generic to the morphed generic with preservation of the connectivity between triangles [34]. Since people are more sensitive to changes around the eyes and the mouth [6], we also include the high-level facial feature lines which enable to better transfer the intensity of the blendshapes [43]. Blendshapes transfer can be performed in exactly three minutes for 102 blendshapes. This set can be reduced for a VR usage. Results are presented on Figure 7.

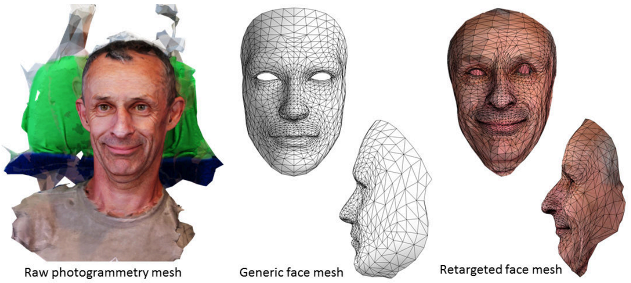

Figure 6: The generic mesh (center) is registered on to the raw photogrammetry mesh (left). Result of the retargeting is shown on the right. Texture is also transferred to the retargeted mesh using correspondence based on barycentric coordinates and continuous texture.

3.6 Extra geometry transfer

Most of the internal geometry elements, such as eyeballs, jaws, teeth, tongs and nostrils are more complex to register due to the lack of information onto the scans (i.e. only the visible external elements are reconstructed). To generate high-quality characters, these elements must be considered. To do so, we use a rigid alignment method to translate, rotate and scale these elements from our generic mesh to the morphed mesh. A binary mask is applied to the generic mesh to exclude some parts from the registration. Each masked element is retargeted individually by aligning the two generic and reconstructed outliers using the best-matching similarity transform between them. It minimizes the squared distances between source points' outlier and their corresponding target points (see Figure 8).

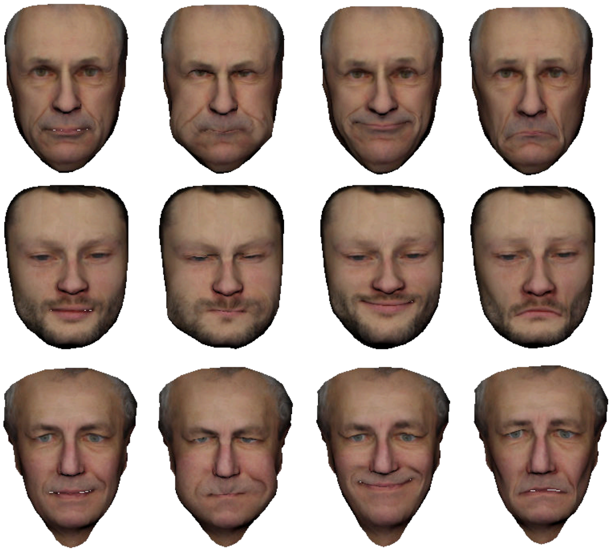

Figure 7: Results for some blendshapes transfer from the source template character to three different characters.

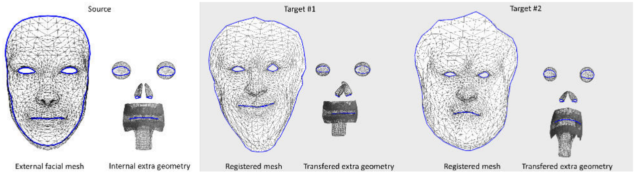

Figure 8: Results of the automatic transfer of extra geometry including eyeballs, jaws, teeth, tong and nostrils.

3.7 Face fitting to the generic mesh

Finally, the morphed facial mesh has to be merged back to the body (more precisely to the head, see Figure 2). While there still is a vertex to vertex correspondence between the two meshes (the topology has been preserved), the scale and the geometry of the face has changed. Hence a method to merge the two meshes is required. First, a rigid transformation is computed to align the reconstructed mesh to the generic face one [36]. The computation is based on the landmarks of the two meshes. The merge between the reconstructed face and the hood is based on the method proposed by Deng et al. [10]. The smoothing is however performed differently since we want to keep the border of the hood. Artifacts are often generated at the edge of the forehead because of the hairs (see Figure 9). They are smoothed by aligning the tangents of the mesh boundary to the ones of the forehead. This step may create a hole between the hood and the forehead. The hood is vertically adjusted with an FDD box to remove the distance between the forehead and the hood [31].

3.8 Results

A benchmark of the fully automatic pipeline is presented and output results are discussed.

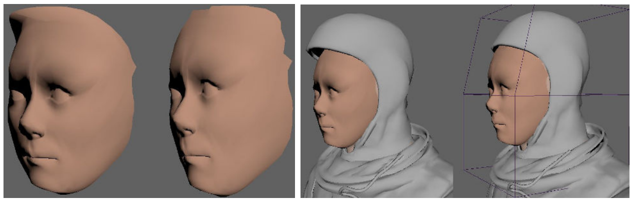

Figure 9: The reconstructed face is merged to the boundary of the hood (left). A smoothing is performed to remove artifacts due to the hairs. Because of the smoothing, a gap may appear between the forehead and the hood. The hood is adjusted with a FDD box.

3.8.1 Benchmark 3D reconstruction

To evaluate the quality of the reconstruction, we ran our pipeline under various conditions. The pipeline was evaluated until the registration and blendshape transfer step (Section 3.5). The aim of this benchmark is to determine the minimal configuration (i.e. number of cameras and image resolution) that provides the best visual facial mask that can be merged to the generic body.

We tested four camera configurations (3, 5, 9 and 14 cameras) and three resolutions: 100% (5184x3456), 50% (2592x1728) and 25% (1296x864). Pictures of nine individuals were used in this test. Starting from five cameras, success rate of reconstruction with a resolution of 5184x3456 or 2592x1728, is 100% (see Figure 10). If the number of camera or image resolution decreases, the reconstruction may fail because not enough image descriptors are found.

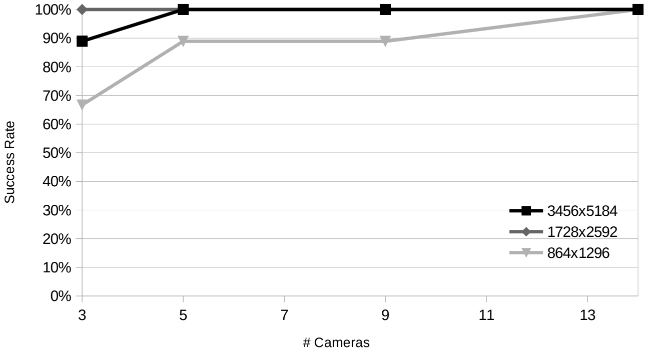

Figure 10: Success rate of the reconstruction.

Computation time increases almost linearly with the number of cameras and the image resolution (see Figure 11). The longest duration is about 25 minutes (1536s, 14 cameras and highest resolution). It may be reduced to 4 minutes (252.66s, 5 cameras and resolution of 2592x1728).

Results were visually inspected under all these conditions (see Figure 12). The *Hausdorff* and maximum distances regarding the reference mesh (14 cameras and resolution of 5184x3456) were also computed. They are estimated in millimeters by computing the ratio between the average inter ocular distance (60mm) and the mesh inter ocular distance. No difference is visible with a reconstruction with at least five cameras. With three cameras, parts of the face may be missing (i.e. the cheeks). The average maximum distance with five cameras is about 22mm for the three resolution conditions (see Figure 13). These results are acceptable for our specific scenario and therefore, in any application for which the user's face only is required.

The conclusions of this benchmark are that we strongly recommend not to use the third resolution (1296x864). It is too low to generate good meshes and textures due to bad descriptors precision

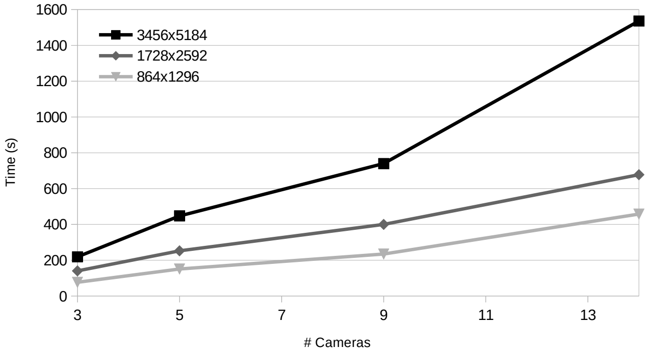

Figure 11: Computation time in seconds with four camera configurations and three different image resolutions.

on the images. The second resolution (2592x1728) has imperceptible or very low differences with the highest one. We would also strongly recommend not to use three cameras. Five and above appear to be the minimum to get precise results. In parallel, we also evaluate the use of High and Normal SIFT descriptors in Meshroom and Normal SIFT fails too often to be considered as a serious candidate. In resume, five cameras with a resolution of 2592x1728 appears to be the best good trade-off between quality and computation time.

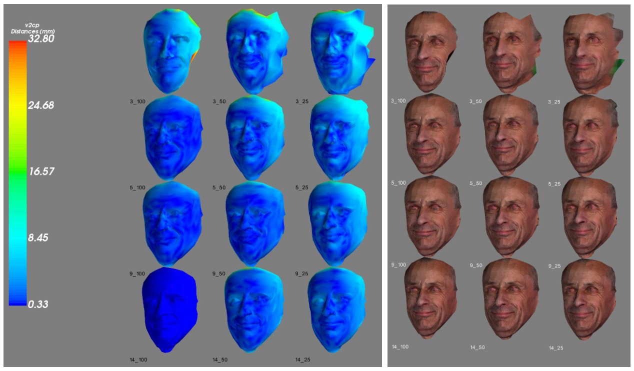

Figure 12: Example of benchmark results. From top to bottom, the number of cameras is 3,5,9,14, and from left to right the resolution is 100% (5184x3456), 50% (2592x1728) and 25% (1296x864). Hausdorff distance (left picture) is from the top bottom mesh.

3.8.2 Reconstructed character

Figure 14 shows output results of our pipeline with the configuration defined above. The full process is about seven minutes with a computer embedding a Xeon E5-2640, 32 GB of DDR3, and a Nvidia GeForce 1080 GTX. The reconstructed face is fit to the astronaut mesh suitable for any VR experience. Since the generic character is already rigged, the personalized one can be easily animated. Besides twenty blendshapes for controlling facial expression are also present (more could be added but it increases processing time). Eyes movements and blink are rendered thanks to dynamic UVs.

The pipeline is focused on facial reconstruction. While the example of the astronaut is well adapted because of the hood, the approach is suitable to any mesh. It is an artistic choice to select a mesh on which a face can be easily merged to though. Having the full head retargeted would be easier to merge back to a generic body (i.e. the boundary would be the neck). This enhancement

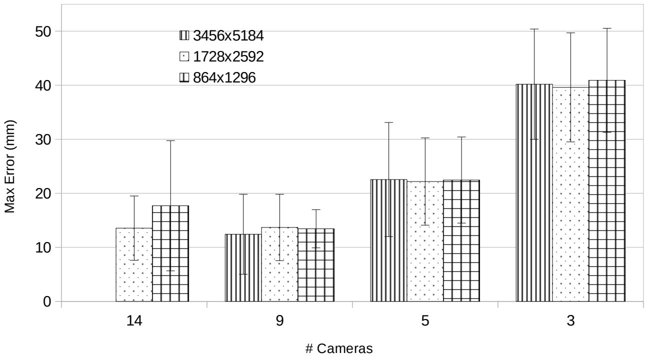

Figure 13: Maximum error in mm regarding the reference (14 cameras and maximum resolution).

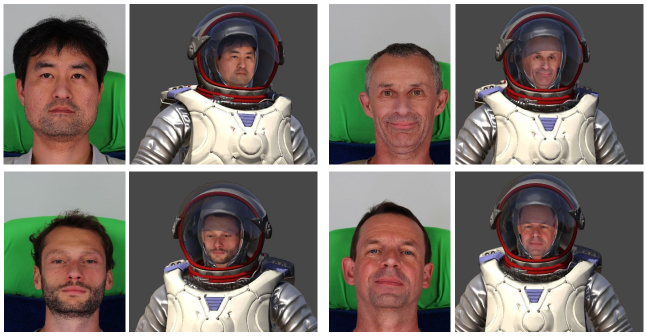

Figure 14: Results obtained from our pipeline. For each pair, the left image is the front captured picture and the right image is the final character.

is considered, but more research on the hair should be conducted. Currently facial hair is directly baked into the texture and the mesh geometry. The extension of the full body is also planned with the challenge to deal with the clothes. Indeed they will hide the actual user's morphology.

4 Facial style transfer

Depending on the target application, the photorealistic mesh from our pipeline may not be adapted to the visual style of a content or to a specific narrative. For instance, one would may look like a dwarf or an elf in a heroic fantasy world, or like an alien in a space opera. In this context, the question we want to address is, to what extent one's face can be customized? Besides, how different the target style face can be from a human face?

The point of this customization is to be able to recognize one's face in a non-human face. It is largely inspired from the James Cameron's Avatar movie in which actors can be recognized in their avatar equivalent (i.e. the Na'vi). From the literature, we identified that hair, face outline, eyes and mouth (not necessarily in this order) are important for perceiving and remembering faces [8]. Also, the most variable traits are within the triangular shape that connects the eyes, mouth and nose [33]. Our hypothesis is that these facial features allow to recognize an individual in a way similar that a caricature can be recognized ([4]).

To fulfill this goal, we propose two adaptation processes of the reconstructed facial mesh: a deformation of the geometry and a transformation of the texture. Our approach is illustrated with facial meshes reconstructed from our pipeline and non-human facial meshes extracted from *Mixamo*[2].

4.1 Geometry deformation

As mentioned above, the shape of a face is a key component of its style. Therefore, to transfer the style of one's face to another, we transfer its geometrical particularities, whether it is the size of the jaw, the angle of the nose, or the eye-to-eye distance. Since our reconstructed meshes and the non-human faces have different topologies, a correspondence must be found. This process is performed in a way like the one described in Section 3.3 and 3.5. In the case of non-human meshes, facial landmarks were manually set. Once all the meshes have the same topology, it is possible to apply vertex-to-vertex operations.

To capture the particularities of human faces, we compute their variations from an average human model. The average model was generated with *MakeHuman*[3] with the default settings (see Figure 15). This mesh was given the same topology as the others. The features of one's face are defined as the vertex-to-vertex distance between the reconstructed mesh and the average mesh. This distance is then applied to the non-human face:

$$M = M_n + w(M_h - M_a) \tag{1}$$

where M is the set of vertices of the final facial mesh, M_n is the set of vertices of the non-human mesh, M_h the set of vertices of the human mesh and M_a the set of vertices of the average human. A weight w can be applied to accentuate the geometrical features given by the distance. It is also used to compensate the size difference between the human and non-human face.

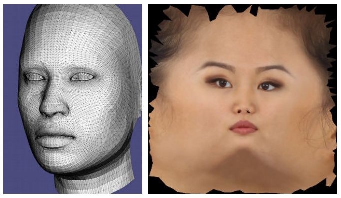

Figure 15: Average mesh (left) and texture (right)

4.2 Texture adaptation

Our approach builds upon the work of Champandard et al. [7] who make use of a semantic mask to constrain the style transfer from a specific zone of an image to another image. Since we use a common topology for all the meshes, we also convert the textures into the same representation where the flatten face is centered and continuous.

A mask is computed from the landmarks triangulation, separating face parts in different semantic zones (see Figure 16). The mask prevents wrong matches in the neural style transfer step: for instance circular facial parts such as eyes and nostrils tend to often mismatch, and the resulting error would be very noticeable.

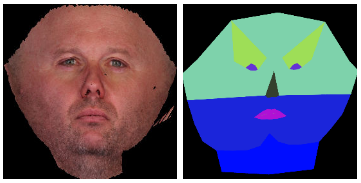

Figure 16: Masks used to constrain the texture style transfer

Directly using Champandard et al.'s network to transfer the style of the human texture to the non-human one produces a general mix

[2] https://www.mixamo.com

[3] http://www.makehumancommunity.org

of the two textures. To avoid this issue, we compute a relative style transfer, using a third texture, corresponding to an average human facial texture (see Figure 15). We use here the texture of a CG character having an artificially flawless skin. Hence facial features such as hair, scars or wrinkles are transferred. The style loss function of the neural network is modified as follow, to minimize the relative style difference.

$$argmin((style(texS) - style(texSav))w_{style} - (style(textSC) - style(texC)))^2 \quad (2)$$

With $texS$ the style texture (i.e. the non-human texture), $texC$ the content texture (i.e. the human texture), $texSav$ the style average texture, and $texSC$ the output. The non-human texture is the starting point of the output texture. Enforcing the relative style becomes a global loss and there is no longer any reason to use a content loss. Individual features are thus transferred, such as the skin tone, facial hair and wrinkles, as depicted on Figure 17.

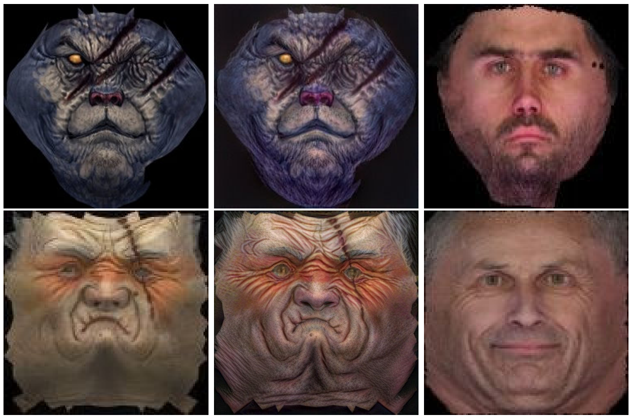

Figure 17: Texture style transfer. Left column: original non-human texture; middle: result; right: human texture.

4.3 Pilot User Study

A pilot user study has been conducted to identify the limits of our approach. Our hypothesis is that one's face transferred to a non-human mesh can be recognized.

4.3.1 Experimental data

We ran our process onto nine human faces (see Figure 18), six have been captured from our rig and three are CG faces. We also used the style of five non-human faces (bottom right row). The geometric style w was set to 1, and the textural style w_{style} to 1.75. The process took 100ms for the geometry deformation and 1.5h for the texture adaptation (1000x1000 pixels) with a Xeon E5-2687W, 32 GB of DDR3, and a Titan X Pascal. The six non-human faces were chosen to highlight the possibilities of our approach. A and B have a humanoid morphology, C have wide mouth but no nose, D is a mix between a beast and a humanoid, and E does not have any humanoid features at all.

4.3.2 Protocol

We asked each participant to recognize one's face among nine styled faces (see Figure 19). The person's face to be found is displayed as well as the non-human template mesh. Five human faces had to be recognized within the five possible styles. We did not use all the nine human faces to avoid a learning effect and to prevent participants from choosing by elimination. We also asked them to recognize people based on the geometry only (i.e. without texturing), on the texture only (i.e. with the texture applied on the average human mesh), and on both the geometry and the texture. This allows to measure the impact of geometry and texture on face recognition. Hence, they had $25x3 = 75$ faces to recognize. They were free to take the required time to accomplish the task. Besides they could control the camera to examine each model.

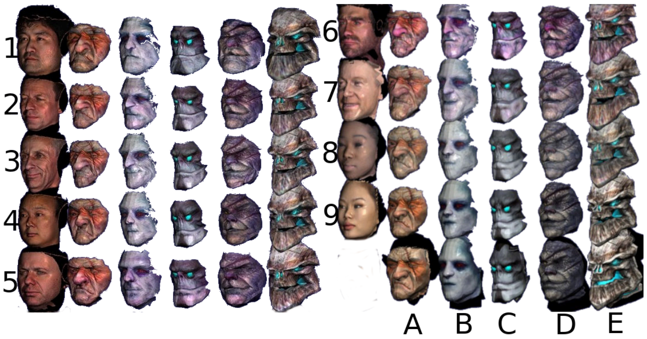

Figure 18: Experimental data. Left columns are the input human faces and the bottom row on the right is the non-human faces.

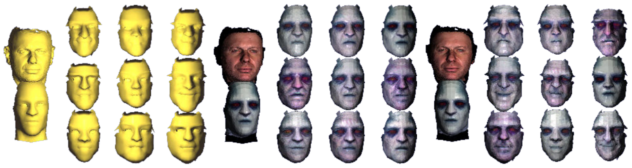

Figure 19: Experimental conditions: geometry only (left), texture only (center) and geometry with texture (right). Participants were asked to recognize one individual among the nine propositions. The template non-human face is also displayed.

4.3.3 Results

12 naive participants have taken part into the experiment (age $\bar{x} = 40, \sigma = 8.59$, 1 female). They have no expertise in computer graphics or in face recognition. Recognition rates of the human faces are plotted on Figure 20 and 21. Results were analyzed with an exact binomial test, which performs an exact test about the probability of success in a Bernoulli experiment (also used in [32]). In our context, the null hypothesis represents the probability that a correct answer has been randomly chosen with a chance of $\frac{1}{9}$.

As expected the recognition rate is higher with the style applied on both the geometry and the texture. Figure 20 shows that the task was not obvious since only face #7 was recognized by slightly more than 50% of the participants. It has to be noted that the expression of the model is not neutral, a light smile is visible. This expression is also visible on the styled mesh, which may guide the recognition.

These results can be explained by the fact that the recognition rate with some non-human meshes was particularly low. Results are more interestingly represented on Figure 21. It is clearly shown that non-human faces, too far from the humanoid shape, are hardly recognizable. Higher performance rate was achieved with mesh B (65.45%). While meshes such as C or E, for which there is no nose and the mouth is heavily deformed, cannot be recognized.

4.3.4 Discussion

As expected, the combination of both geometrical and textural style allows a better recognition. Textures seems to provide less style information that geometry with our current approach. Results also shows that recognition depends on the style of the non-human face. In our test, face B obtains better recognition results than the others,

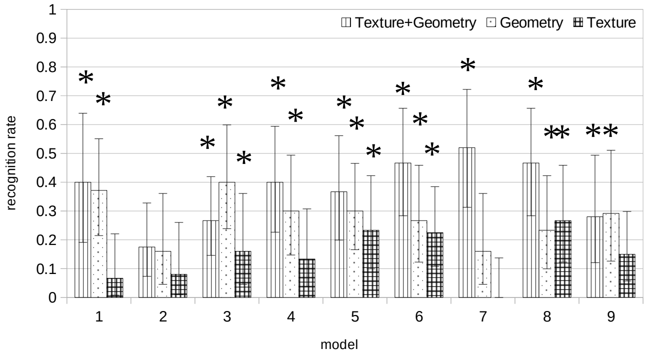

Figure 20: Recognition rate of the human faces. Black lines represent the confidence intervals (0.95), and the stars are the significance ($p < 0.05$).

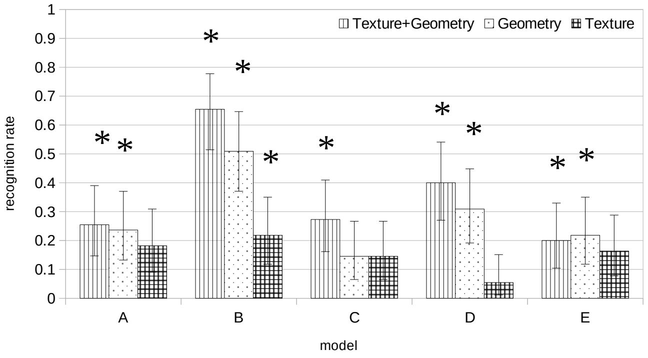

Figure 21: Recognition rate of the human faces regarding the non-human models. Black lines represent the confidence intervals (0.95), and the stars are the significance ($p < 0.05$).

which could be explained by its high similarity to a human face. On the opposite, C and E the faces whose aspect is the furthest from human ones performs the worst. Their lack of nose, and their heavily deformed mouth seems to be the reason, as they are features deemed important for facial recognition.

Although our approach is a first step toward the stylization of human faces, deeper investigation would require more user studies to reduce the confidence interval, and to test different geometric and textural style weights. Also the choice of the average human has a strong influence on the style transfer results. Average mesh and texture have to be carefully selected to not add artifacts. Yet the customization of one's character seems to be limited to humanoid faces that are not too different from a human one. This is in line with the literature in neurobiology assessing that our brain is not adapted to the fine recognition of other species [33].

5 Conclusion & Perspectives

We presented a fully automatic pipeline for generating high-quality facial rigs. From a set of input photos and a generic full-body character, this pipeline outputs a fully rigged character ready to be integrated into any real-time engine or other 3D application in less than seven minutes. Compared to existing approaches, it is strongly focused on facial feature acquisition (geometry, iris, texture) and generation (blendshapes, jaws, teeth, etc.). The benchmark we performed on our capture setup provides useful guidelines to setting up the ideal configuration and parameters for a specific target application.

We also proposed a new method to apply a style to the reconstructed face. Using a template non-human mesh as reference style, we process the geometry and texture of the reconstructed face to make it look like the non-human one. Results of a first pilot study show that this approach is suitable for humanoid faces, but it is limited for non-human faces too far from the average structure of a human one. Thus, the stylization of the character will be focused on humanoid faces for the time being.

Our future work for extending this pipeline will be twofold. First, the pipeline will be improved to capture hair and skin under multiple lighting conditions. Second, it will be extended to capture the full body in high resolution detail. Other aspects helpful in the characterization of unique character facial features will be also investigated (i.e. hair or accessories) to further extend the possible applications.

The proliferation of virtual reality and augmented reality into mainstream consumer technologies will continue to bolster use cases for personalized characters. In a world of spatialized mixed reality computing, one can foresee the utility of a relatively inexpensive, automated acquisition pipeline for every person to create and carry with them their own personal digital double for a variety of applications – from entertainment, to communication, to retail and beyond.

References

[1] J. Achenbach, T. Waltemate, M. E. Latoschik, and M. Botsch. Fast generation of realistic virtual humans. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, page 12. ACM, 2017.

[2] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[3] T. Beeler, B. Bickel, P. Beardsley, B. Sumner, and M. Gross. High-quality single-shot capture of facial geometry. *ACM Trans. on Graph.*, 29(4):40:1–40:9, 2010.

[4] P. J. Benson and D. I. Perrett. Perception and recognition of photographic quality facial caricatures: Implications for the recognition of natural images. *European Journal of Cognitive Psychology*, 3(1):105–135, 1991.

[5] P. Bergeron and P. Lachapelle. Controlling facial expressions and body movements in the computer generated animated short 'tony de peltrie'. *SigGraph '85 Tutorial Notes, Advanced Computer Animation Course*, 1985.

[6] K. S. Bhat, R. Goldenthal, Y. Ye, R. Mallet, and M. Koperwas. High fidelity facial animation capture and retargeting with contours. In *Proceedings of the 12th ACM SIGGRAPH/eurographics symposium on computer animation*, pages 7–14. ACM, 2013.

[7] A. J. Champandard. Semantic style transfer and turning two-bit doodles into fine artworks. *arXiv preprint arXiv:1603.01768*, 2016.

[8] G. M. Davies, H. D. Ellis, and J. W. Shepherd. *Perceiving and remembering faces*, volume 96. University of Illinois Press, 1981.

[9] P. Debevec. The light stages and their applications to photoreal digital actors. *SIGGRAPH Asia 2012 Technical Briefs*, 2012.

[10] Z. Deng, G. Chen, F. Wang, and F. Zhou. Mesh merging with mean value coordinates. In *2012 Fourth International Conference on Digital Home*, pages 278–282. IEEE, 2012.

[11] M. Fuchs, V. Blanz, H. Lensch, and H. P. Seidel. Reflectance from images: a model-based approach for human faces. *IEEE Trans. on Visualization and Computer Graphics*, 11(3):296–305, 2005.

[12] L. A. Gatys, A. S. Ecker, and M. Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[13] A. Ghosh, T. Chen, P. Peers, C. A. Wilson, and P. Debevec. Circularly polarized spherical illumination reflectometry. In *ACM Trans. on Graph.*, volume 29, page 162. ACM, 2010.

[14] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. In *FG*, pages 1–8, 2008.

[15] M. Jancosek and T. Pajdla. Multi-view reconstruction preserving weakly-supported surfaces. In *Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2011.

[16] P. Kaur, H. Zhang, and K. J. Dana. Photo-realistic facial texture transfer. *arXiv preprint arXiv:1706.04306*, 2017.

[17] O. Klehm, F. Rousselle, M. Papas, D. Bradley, C. Hery, B. Bickel, W. Jarosz, and T. Beeler. Recent advances in facial appearance capture. *Computer Graphics Forum*, 34(2):709–733, 2015.

[18] M. Kowalski, J. Naruniec, and T. Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. In *Proceedings of the International Conference on Computer Vision & Pattern Recognition (CVPRW), Faces-in-the-wild Workshop/Challenge*, volume 3, page 6, 2017.

[19] M. Latoschik, D. Roth, D. Gall, J. Achenbach, T. Waltemate, and M. Botsch. The effect of avatar realism in immersive social virtual realities. In *Proceedings of ACM Symposium on Virtual Reality Software and Technology*, 2017.

[20] B. Lévy, S. Petitjean, N. Ray, and J. Maillot. Least squares conformal maps for automatic texture atlas generation. In *ACM Trans. on Grap.)*, volume 21, pages 362–371. ACM, 2002.

[21] C. Li and M. Wand. Combining markov random fields and convolutional neural networks for image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2479–2486, 2016.

[22] Z. Lun, E. Kalogerakis, R. Wang, and A. Sheffer. Functionality preserving shape style transfer. *ACM Trans. on Graph.*, 35(6):209, 2016.

[23] C. Ma, H. Huang, A. Sheffer, E. Kalogerakis, and R. Wang. Analogy-driven 3d style transfer. In *Computer Graphics Forum*, volume 33, pages 175–184. Wiley Online Library, 2014.

[24] W.-C. Ma, T. Hawkins, P. Peers, C.-F. Chabert, M. Weiss, and P. Debevec. Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In *Proceedings of the 18th Eurographics Conference on Rendering Techniques*, EGSR'07, pages 183–194, 2007.

[25] N. Magnenat-Thalmann, R. Laperrière, and D. Thalmann. Joint-dependent local deformations for hand animation and object grasping. In *Proceedings on Graphics Interface '88*, pages 26–33, 1988.

[26] C. Malleson, M. Kosek, M. Klaudiny, I. Huerta, J.-C. Bazin, A. Sorkine-Hornung, M. Mine, and K. Mitchell. Rapid one-shot acquisition of dynamic vr avatars. In *Virtual Reality (VR)*, pages 131–140. IEEE, 2017.

[27] M. K. Monaco. Color space analysis for iris recognition. Master's thesis, Master of Science in Electrical Engineering West Virginia University, 2007.

[28] P. Moulon, P. Monasse, and R. Marlet. Adaptive structure from motion with a contrario model estimation. In *Proceedings of the Asian Computer Vision Conference (ACCV 2012)*, pages 257–270. Springer Berlin Heidelberg, 2012.

[29] K. Nagano, J. Seo, J. Xing, L. Wei, Z. Li, S. Saito, A. Agarwal, J. Fursund, and H. Li. pagan: real-time avatars using dynamic textures. In *SIGGRAPH Asia 2018 Technical Papers*, page 258. ACM, 2018.

[30] C. Pawaskar, W. C. Ma, K. Carnegie, J. P. Lewis, and T. Rhee. Expression transfer: A system to build 3d blend shapes for facial animation. In *2013 28th International Conference on Image and Vision Computing New Zealand (IVCNZ 2013)*, pages 154–159, 2013.

[31] T. W. Sederberg and S. R. Parry. Free-form deformation of solid geometric models. *ACM SIGGRAPH computer graphics*, 20(4):151–160, 1986.

[32] J. Seyama and R. S. Nagayama. The uncanny valley: Effect of realism on the impression of artificial human faces. *Presence: Teleoperators and virtual environments*, 16(4):337–351, 2007.

[33] M. J. Sheehan and M. W. Nachman. Morphological and population genomic evidence that human faces have evolved to signal individual identity. *Nature communications*, 5:4800, 2014.

[34] R. W. Sumner. *Mesh Modification Using Deformation Gradients*. PhD thesis, Cambridge, MA, USA, 2006.

[35] R. W. Sumner and J. Popović. Deformation transfer for triangle meshes. *ACM Trans. on Graph.*, 23(3):399–405, 2004.

[36] S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Trans. on Pattern Analysis & Machine Intelligence*, (4):376–380, 1991.

[37] P. Waldhäusl and C. Ogleby. 3 x 3 rules for simple photogrammetric documentation of architecture. *International Archives of Photogrammetry and Remote Sensing*, 30:426–429, 1994.

[38] T. Waltemate, D. Gall, D. Roth, M. Botsch, and M. E. Latoschik. The impact of avatar personalization and immersion on virtual body ownership, presence, and emotional response. *IEEE Trans. on Visualization and Computer Graphics*, 24(4):1643–1652, 2018.

[39] K. Waters. A muscle model for animation three-dimensional facial expression. *SIGGRAPH Computer Graphics*, 21(4):17–24, 1987.

[40] K. Wenzel, M. Rothermel, D. Fritsch, and N. Haala. Image acquisition and model selection for multi-view stereo. *International archives of the photogrammetry, remote sensing and spatial information sciences*, 40:251–258, 2013.

[41] T. Weyrich, W. Matusik, H. Pfister, B. Bickel, C. Donner, C. Tu, J. McAndless, J. Lee, A. Ngan, H. W. Jensen, and M. Gross. Analysis of human faces using a measurement-based skin reflectance model. *ACM Trans. on Graph.*, 25(3):1013–1024, 2006.

[42] C. Wu, B. Wilburn, Y. Matsushita, and C. Theobalt. High-quality shape from multi-view stereo and shading under general illumination. In *Conference on Computer Vision and Pattern Recognition*, pages 969–976, 2011.

[43] F. Xu, J. Chai, Y. Liu, and X. Tong. Controllable high-fidelity facial performance transfer. *ACM Trans. on Graph.*, 33(4):42, 2014.

[44] S. Zwerman and J. Okun. *Visual Effects Society Handbook: Workflow and Techniques*. Taylor & Francis, 2012.