A SURVEY OF COMPUTER GRAPHICS FACIAL ANIMATION METHODS: COMPARING TRADITIONAL APPROACHES TO MACHINE LEARNING METHODS

A Thesis

presented to

the Faculty of California Polytechnic State University,

San Luis Obispo

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Computer Science

by

Joseph Johnson

June 2021

COMMITTEE MEMBERSHIP

TITLE: A Survey of Computer Graphics Facial Animation Methods: Comparing Traditional Approaches to Machine Learning Methods

AUTHOR: Joseph Johnson

DATE SUBMITTED: June 2021

COMMITTEE CHAIR: Zoë Wood, Ph.D.
Professor of Computer Science

COMMITTEE MEMBER: Jonathan Ventura , Ph.D.
Professor of Computer Science

COMMITTEE MEMBER: Franz Kurfess, Ph.D.
Professor of Computer Science

ABSTRACT

A Survey of Computer Graphics Facial Animation Methods: Comparing Traditional Approaches to Machine Learning Methods

Joseph Johnson

Human communications rely on facial expression to denote mood, sentiment, and intent. Realistic facial animation of computer graphic models of human faces can be difficult to achieve as a result of the many details that must be approximated in generating believable facial expressions. Many theoretical approaches have been researched and implemented to create more and more accurate animations that can effectively portray human emotions. Even though many of these approaches are able to generate realistic looking expressions, they typically require a lot of artistic intervention to achieve a believable result. To reduce the intervention needed to create realistic facial animation, new approaches that utilize machine learning are being researched to reduce the amount of effort needed to generate believable facial animations. This survey paper summarizes over 20 research papers related to facial animation and compares the traditional animation approaches to newer machine learning methods as well as highlights the strengths, weaknesses, and use cases of each different approach.

ACKNOWLEDGMENTS

Thanks to:

- My family, Todd, Kelly, Jacob, Alyssa, and Jack, for being supportive in every path I decided to take.
- My advisor, Dr. Zoë Wood, for mentoring me through my thesis and degree. I consider myself lucky to be a member of the computer graphics group that you created.
- My committee, Dr. Jonathan Ventura and Dr. Franz Kurfess, for taking interest in my thesis.
- My friends and roommates, for their encouragement and humor that motivated me on a daily basis.
- Andrew Guenther, for uploading this template

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

INTRODUCTION

Facial animation is a difficult problem that continues to challenge artists and researchers who want to recreate the nuances that make faces human. Faces have many parts that must be simulated accurately to be able to create a convincing approximation of a face. Characteristics like expressions, wrinkles, and skin composition can be dead giveaways when comparing real faces to animated ones. Expressions need to be formed naturally because humans do not perfectly transition between different emotional states. Wrinkles that indicate emotional intensity must be created in the correct regions to match the mood of a character. Skin must interact correctly with different parts of its environment, such as light, to help convince viewers that the face is not detached from the world it presides in. Humans are very good at identifying these factors that make faces seem not natural and there are many theories as to why people became good at this task [32]. One theory suggests that we developed this skill to identify healthy individuals for reproduction. Another theory says that humanoid replicas remind us of death and therefor trigger an eerie feeling as a result of our anxiety about mortality. Regardless of which theories are true, humans are good at identifying faces and have developed uncanny feelings towards faces that come close to resembling humans. This phenomenon is known as the uncanny valley and it is a large factor when determining the quality of facial animations. Facial animation is hard because it must convey human emotions, with approximations, while not disturbing humans who have been trained to detect imperfections.

As a result of the complexities of human faces and a strong ability for people to detect imperfections in face approximations, there are many methods to facial animation to

try and create convincing faces for different needs of applications. Photo-realistic facial animation would not be needed for a cartoon, but it would be needed for the de-aging of an actor who is acting alongside other real actors [1]. There is an immense number of approaches to facial animation and it can be difficult for artists and developers to choose the best approach for their specific needs. This survey paper helps solve this problem by providing summaries of the implementations, strengths, and weaknesses of traditional approaches to facial animation. In addition, this paper summarizes newer machine learning approaches to facial animation and examines the problems with traditional approaches that can be solved with machine learning.

1.1 The Uncanny Valley

Artists and researchers in the facial animation field have been struggling with the uncanny valley problem which strips characters of empathy from audiences. This lost empathy can be distracting to audiences and takes away from the emotions that are trying to be portrayed by the artists. As characters more closely resemble humans, they become more relateable and empathy-worthy to the people observing them. When characters get too close resembling humans without fully resembling them, however, people begin to have an eerie feeling towards these characters and less empathetic towards them as a result [21]. The dip in empathy, when characters come close to resembling humans, gives the uncanny valley its name.

To get out of the uncanny valley, researchers are continually trying to make animated faces indistinguishable from humans. To create even more life-like characters for big-budget films, artists and engineers have looked to human anatomy to improve their approximations of human expressions and movement. By simulating the different bones, muscles, and skin that make up human faces, realistic animation can be cre-

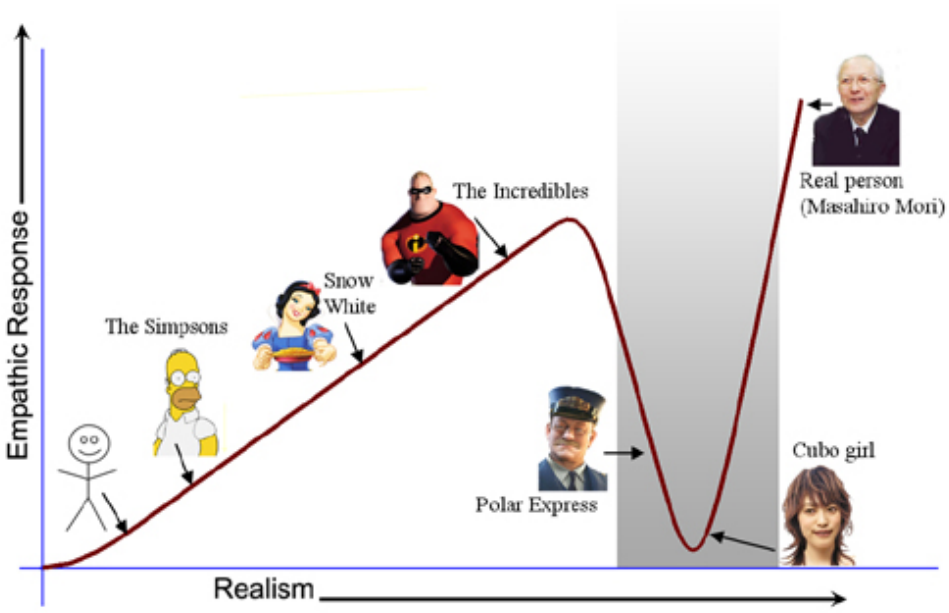

Figure 1.1: As characters become more realistic-looking, we have a higher empathetic response towards them. When these characters become too similar to humans without fully resembling them, there is a major dip in the emotional response towards these characters. This dip is known as the uncanny valley.

ated that closely resembles the nuances of human expressions. Results from these approaches have come closer to resembling human faces, but their approximations still produce uneasiness when placed next to real faces.

1.2 Applications and Goals of Facial Animation

Facial animation has the main goals of: conveying engaging emotions, expanding artistic control, and limiting the disturbances caused by the uncanny valley [26]. With these goals in mind, facial animation methods are developed to allow artists to create animations that contribute to enjoyable experiences. When these goal are not considered, disturbing results can be created that take away from the overall experience where the facial animation is being used.

The applications of facial animation are spread widely across the entertainment industry. Films, video games, and interactive avatars benefit from the expressiveness that facial animation gives to their characters. This expressiveness allows characters to tell engaging stories and gain empathy from audiences. Depending on the environment and story being told, characters need differing levels of realism and artistic control to engage their audiences. To be able to handle this range of needs, various approaches have been developed and their implementations, strengths, and weaknesses have been summarized in this survey paper.

1.3 Machine Learning in Facial Animation

In recent years, machine learning has become popular because of its ability to solve complex problems. As a result of this ability, researchers have turned to machine learning to try and improve the quality and accessibility of facial animation [3, 16, 8, 24]. With large enough data sets, machine learning is able to learn how to animate faces for a variety of applications. The improved speed, quality, and usability of machine learning results make it a great solution to solving some of the major issues with the traditional approaches to facial animation. Since the study of machine learning in facial animation is relatively new, not many studios make use of this technology and information is limited about its use cases. The current state of machine learning approaches and comparisons to traditional approaches will be summarized in this survey paper.

1.4 Methodology for Selecting Papers

Selecting relevant papers was a substantial task in the creation of this survey paper. First, we had to decide on specific topics that would be highlighted by the paper. In

the case of this paper, we decided to focus on artist interventions involved in different facial animation approaches and machine learning in facial animation. After deciding on the topics that would be highlighted, we conducted searches on Google Scholar to find relevant papers. Each paper that made it into the survey paper had to make it through multiple reading passes. The first pass would involve the reading of only the abstract and conclusion. This was important for quick identification of papers that related to the topics we wanted to research. Papers that did not mention anything about these topics were dropped at this point. The second pass involved reading each paper in its entirety. This step allowed for the assessment of the actual content of each paper and papers that did not clearly elaborate on the desired topics, or provide other meaningful insights, were dropped. Once all of the papers were gathered, similar approaches were grouped into different categories for further analysis. The result of this methodology is a refined set of papers that specifically address the desired research topics.

1.5 Contributions

Previous survey papers have done a great job at summarizing the strengths, weaknesses, and implementations of different facial animation approaches [25, 11, 26]. This survey paper extends these works by bringing attention to the less-researched topics of artistic interventions and machine learning approaches within the field of facial animation. When summarizing the relevant attributes of a specific approach, many of the previous survey papers paid close attention to the performance and realism of the approach's results without consistently highlighting the amount of artist interventions needed to create quality facial animations. This is a very important attribute to recognize because the approaches that require too much intervention from artists will not become widely used. The previous survey papers also did not include machine

learning approaches because its use in facial animation has not been researched until recently. With more available research, this survey paper is able to analyze a collection of machine learning approaches and compare them against previous methods of facial animation.

Chapter 2

RELATED WORKS

Facial animation is a complex task because of all the intricate details that make a facial expression believable to the human eye. Over the years, many different approaches have been developed to tackle the problem of facial animation because the different uses of facial animation can have drastically different goals and requirements. The algorithms that generate expressions made by a character in a video game prioritizes speed over quality because of the desire to keep frame rates high, but the algorithms used to generate character expressions for movies prioritize accuracy over speed because each frame is pre-rendered. The large variety of approaches allow a lot of freedom in how an artist can animate their characters' faces, but the increasingly large number of approaches can make it difficult for artists and engineers to decide which approach to use or implement. As a result of the need to make all of the approaches easier to understand and compare, survey papers have been written to summarize the different approaches used for making believable facial animations.

Various works have been completed to organize information about the different approaches to facial animation in ways that are easy to digest. Researchers at the University of Southern California organized the different approaches into the categories of geometric manipulations and image manipulations [25]. These categories were then broken down further to create a taxonomy that showed the different approaches to facial animation. This method of organizing the information made it easy to understand the similarities between specific techniques and where they began to differ. Other researchers, from the University of Houston, grouped approaches into more specific categories that allowed for more specialized analysis. [11]. Within

each of the categories, researchers were able provide specific information about the different strengths and weaknesses of each approach. In addition to summarizing the implementations and strengths of the different approaches, research has been completed to outline the process of rigging with the different approaches [26]. Based on the technique used for animation, rigging can be very different in terms of complexity and control. These factors can be very important depending on the experience of the artist and the level of control that is desired.

All of the previous works serve to make the information more accessible and structured so that readers could understand the different approaches that were available and make decisions about what they should be researching further. Similarly to the previous works, this paper summarizes information about the different approaches and provides details about the unique strengths and weaknesses of each approach. In this paper, we extend these works to create a summary of approaches that includes the more recent development of machine learning in facial animation.

Chapter 3

TRADITIONAL APPROACHES

3.1 General Methods

Across the different approaches to facial animation, there are a few commonly-used methods that are used as parts of solutions to many unique approaches. Each of the general methods contribute to the basic formation of facial animation and provide a great amount of artistic control when animating faces. To be able to create more accurate animations in more complex approaches, these methods are adapted to increase quality and control of the animations. Without these general methods, many of the later discussed approaches would not be possible.

3.1.1 Facial Animation Rigs

Many of the different approaches described in this paper involve facial animation rigs which allow artists to control the faces they are trying to animate. A common analogy for animation rigs is the setup of marionette dolls [26]. A marionette doll is rigged with strings so that an actor may control the movements and expressions of the doll. Facial animation rigs serve a similar purpose to allow artists to control the facial expressions created by a character. Instead of strings, control points are set up to allow artists to influence the rig. Depending on the method of animation, face rigs may be set up and operate differently, but they all have the shared goal of helping the artist animate the character.

Figure 3.1: Pinocchio is a popular example of a marionette doll.

Depending on the complexity of the rig, different levels of accuracy and control can be achieved. Complex rigs are very difficult to set up and manipulate, but they provide artists with a large variety of animations they can create. Simple rigs are easy to setup and manipulate, but they can be restrictive in the kinds of expressions that can be created. Various applications require different levels of control and expressiveness and deciding the importance of these factors will dictate the complexity needed to create adequate animations.

3.1.2 Blend Shapes

Humans are not super great at recognizing a large variety of facial expressions. It turns out that most people can only correctly recognize a small subset of facial expressions. In a study conducted with 95 participants, there were only 4 out of 21 expressions that could be correctly identified by a majority of participants [29]. These

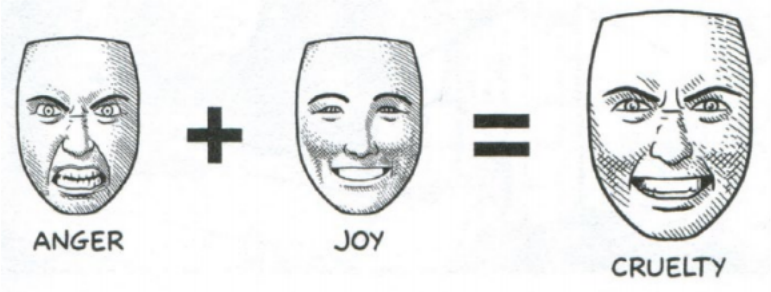

Figure 3.2: Example of combining two primary emotions to create a new emotion [29].

four expressions were: joy, cruelty, amazement, and sadness. To be able to convey the emotions of a character, it is important to prioritize displaying the expressions that the audience can understand.

There are only 6 primary expressions that are to needed to be able to convey all emotions [29]. These primaries align closely to the emotions identified above and consist of: anger, disgust, fear, joy sadness, and surprise. From these 6 primaries, every other emotion can be expressed by combining two of the primaries with varying contributions.

Blend Shapes is a facial animation approach that works similarly to how all emotions can be created from the primary emotions. Distinct base emotions are created for the blendshape rig and all other emotions can be created by combining multiple base emotions. To be able to blend the different base emotions, blendshape rigs interpolate between the multiple emotion shapes, with varying weights, to create new emotions.

Blendshape-rigged models are composed of many variations of the same shape. Each variation shape represents an emotion, or pose, and must have the same orientation as all of the other shapes in the blendshape rig. To create animations, the shapes can be combined with varying contributions to create new poses [11]. In order to create the animation of closing an eye, a blendshape pose with an eye closed would

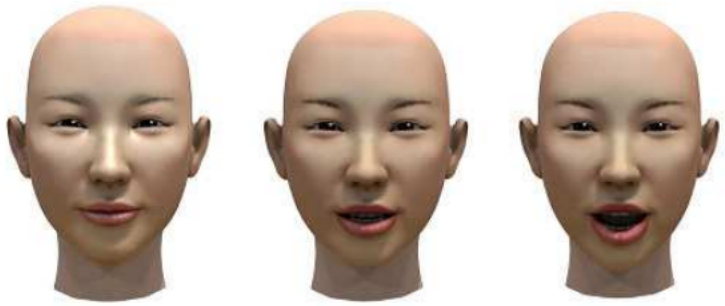

Figure 3.3: The resting blendshape pose (left) is combined with the mouth-open blendshape pose (right) to create the novel pose (middle). To animate the mouth opening, the weight of the mouth-opened blendshape would be increased over time [11].

be combined, with an increasing weight over time, with a blendshape pose of the face in a resting position to eventually create a combined pose with the eye shut. Models can be rigged using only Blend Shapes, but complex characters can require a large number blendshape poses when expressions become more extreme. Characters like Gollum, in *Lord of the Rings: The Two Towers*, required 675 blendshapes poses to be able to animate the extremely expressive facial expressions that made the character memorable [26]. Each one of these poses had to be individually modeled and would have been an extremely tedious task for artists to complete.

Blend shapes models are "the linear weighted sum of a number of topologically conforming shape primitives" [11]. To calculate the resulting pose, each vertex can be calculated by summing the contributions of each weighted blend shape pose.

$$v_j = \sum w_k b_{kj} \tag{3.1}$$

In the above equation, v_j is the j^{th} vertex of the resulting animated model, w_k is blending weight, and b_{kj} is the j^{th} vertex of the k^{th} blend shape. The weighted sum

can be applied to the vertices of polygonal models. The weights w_k are manipulated by the animator or automatically determined by different methods [11].

3.1.3 Bone-Based Rigging

Bone-Based rigging involves the process of creating a skeleton structure that drives the animation of the model. The skeleton structure is made of bones and joints. The joints define relationships between the bones and how they can move relative to each other. Possible transformations, such as translations and rotations, can be defined by the joints connecting one or more bones. Once a skeleton representation is created, the process of combining the skeleton and model for animating is called skinning. Each vertex in the model has a defined weight to specify how much it can be influenced during deformation and each bone has a map that defines the amount influence it has on each vertex [26, 12] (This can be a very tedious task when the number of bones becomes large). Higher weights of influence mean that the vertices will change more in response to changes in the orientation of the bone. After weights have been defined, the model can be animated into different expressions by changing the orientation of bones in the skeletal rig.

3.1.4 Wrinkle Mapping

Wrinkles are important to facial animation because of their ability to express intensity of emotions. Without the presence of accurately placed wrinkles, facial animations may not be interpreted correctly and the animation would seem unnatural. The lack of correct wrinkle details directly contributes to the uncanny valley problem. When someone is smiling uncontrollably, it is expected that that person would have visible crows feat at the sides of their eyes. To be able to show this level of detail in facial

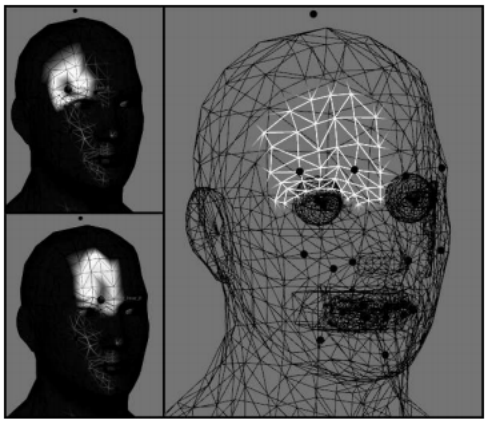

Figure 3.4: The two left images show the skinning influences of the two bones of the right eyebrow. The reference pose (right) has been created with the two bones rising up. The image (right) also shows the influence of this reference pose for each vertex attached to these bones [12].

animation, the resolution level of a mesh must be increased to contain these finer details like wrinkles and pores. If the resolution of meshes were increased to contain these details, however, the performance of the rig would be greatly impacted and no longer suitable for most real-time applications [27].

To keep the resolution of meshes lower for achieving faster performance, artists turn to wrinkle maps to store the extra details that they would like to add to their characters' faces [27, 30]. Wrinkle maps are represented by using normal maps to store more information about the surface of the face. Normal maps are a solution to storing fine-detail information about the surface of an object so that more complex lighting computations can be calculated on a lower resolution mesh. In a normal map, each pixel has an RGB value that represents the corresponding X, Y, and Z components of a normal of a surface that is being mapped. Fine details like skin pores and wrinkles can be stored in these wrinkle maps to make faces look more lifelike without having to sacrifice much performance from the added realism.

In order to keep track of which wrinkles should be drawn on a face from a wrinkle map, artists can define areas of influence that determine which wrinkles should be

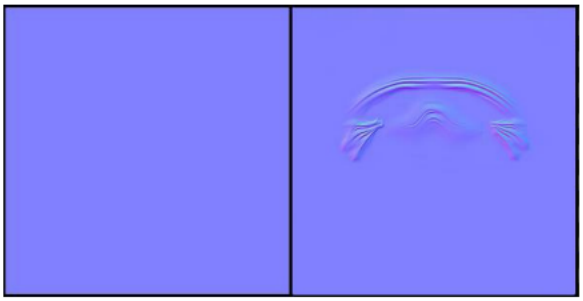

Figure 3.5: Wrinkle map without wrinkles (left) and wrinkle map with forehead, nose and crow's feet wrinkles (right) [27].

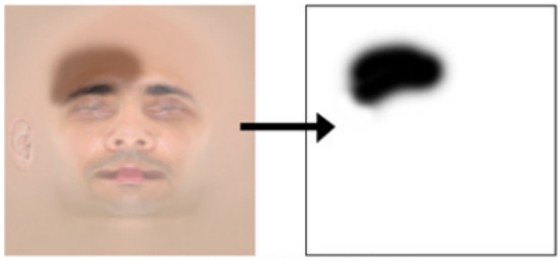

Figure 3.6: Texture map (right) that shows the wrinkle activation for the left eyebrow of the face [27].

drawn when certain regions of the face are transformed from a resting pose [27]. When a character winks with their left eye, wrinkles should form around the left eye and the right eye should not have any new wrinkles forming. By defining these different levels of influence in a map, artists could control the level of activation that each region of wrinkles would have. These maps are gray-scale textures that store the level of activation in the RGB color of corresponding pixels. Pixels that are perfectly white have 0 percent activation and pixels that are perfectly black have 100 percent activation. Differing levels of activation can be defined with differing intensities of gray that are between the two extremes.

Even though influence mapping is a great way to represent which regions of the wrinkle map should be shown, mapping these influence areas can be a time consuming task for artists. Researchers at Claude Bernard University Lyon proposed using a collection of reference poses, each with their own skeleton pose and wrinkle map, to approximate

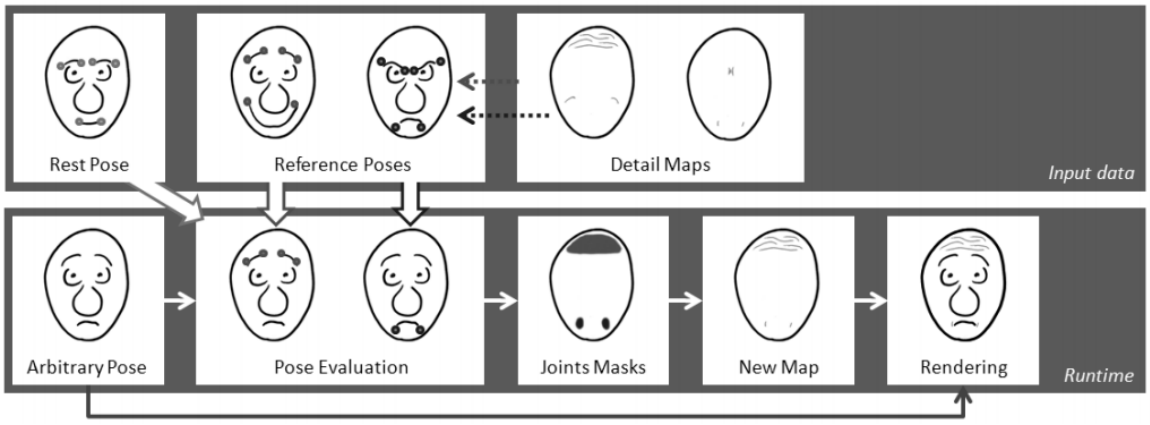

Figure 3.7: The required input data is a classic skinned mesh in a rest pose and some reference poses (reference pose = skeleton pose + wrinkle map). At runtime, each pose of the animation is compared with the reference poses, bone by bone. Then, skinning influences are used as masks to apply the bones poses evaluation. Wrinkle maps are then blended on the GPU and are added to the final render of the current frame with dynamic wrinkles and details [12].

wrinkle activation areas in skinning animation systems [12]. This approach would eliminate the need to define activation areas by using skinning weights to calculate activation areas in real time. To create new wrinkle maps, skeleton poses are compared to each bone in the reference pose to calculate the contribution of each reference wrinkle map's contribution. The more similar the orientation of the reference pose bone is to the new pose bone, the stronger the contribution to the generated wrinkle map. After all of the contributions from each of the bones in each of the reference poses are calculated, reference pose wrinkle maps are blended with the default pose wrinkle map based on their calculated contributions. This wrinkle map is then used to add the approximated wrinkle details to the arbitrary expression for the final rendering without the work of having to define activation areas.

3.1.5 Strengths and Weaknesses of General Methods

The general methods of facial animation are very successful in interactive applications that prioritize performance and provide a basis for other animation techniques. These methods having corresponding rigging methods that are fairly easy to implement and can give realistic looking results that are acceptable for their given uses. Interactive applications, like video games, greatly benefit from the ability to convey emotions without taking on a tremendous amount of computational cost. These emotions are extremely important when trying to engage audiences in a story line and help audiences empathize with characters. With more complex approaches to facial animation, interactive frame rates would not be able to be achieved and the created experiences would suffer from the addition of facial animation.

These methods struggle, however, in terms of scaling to achieve greater accuracy in creating believable human expressions and movements. To create more believable animations, more and more artist effort is required and the benefits of these approaches start to become less apparent. To be able to create believable animations, with Blend Shapes, for an expressive character like Gollum, in *Lord of the Rings: The Two Towers*, a tremendous amount of manual effort is needed [26]. When trying to achieve photo-realistic animations, these facial animation methods are not going to provide the best results.

3.1.6 Paper Highlight for General Methods

A great place to start when looking for information on general methods is the work by Dutreve et al. called *Real-time Dynamic Wrinkles of Face for Animated Skinned Mesh* [12]. This paper describes an implementation that allows for real-time facial animation with automatically added wrinkle details. These added wrinkles are derived

from wrinkle maps created by artists and allow artists to control the final look of their characters. In addition, this paper provides great descriptions of bone-based rigging and how wrinkle maps can be used to add fine details to faces.

3.2 Physics Simulations

To help tackle the problems of the uncanny valley, researchers are looking to physically-based approaches to facial animation. To create even more life-like characters for big-budget films, artists and engineers have looked to human anatomy to improve their approximations of human expressions and movement. By simulating the different bones, muscles, and skin that make up human faces, realistic animation can be created that closely resembles the nuances of human expressions. To make these simulations possible and more physically accurate, different approximations have been made to make the manipulation of the different components of faces look more realistic.

3.2.1 Mass-Spring Simulation

In order to simulate the nuanced features of facial expressions, physics-based approaches are used to approximate the different layers that make up of the skin. The composition of the human face is made up of an outer epidermal layer, an underlying muscle layer, and an inner bone layer that doesn't deform. The outer layers of the skin and muscles can be approximated as system of mass-springs [18, 31, 26]. As the springs respond to movements in the muscle layer, wrinkles form as some regions of the system of springs are compressed. When the system is finally configured, realistic animations can be created that closely approximate how real faces deform.

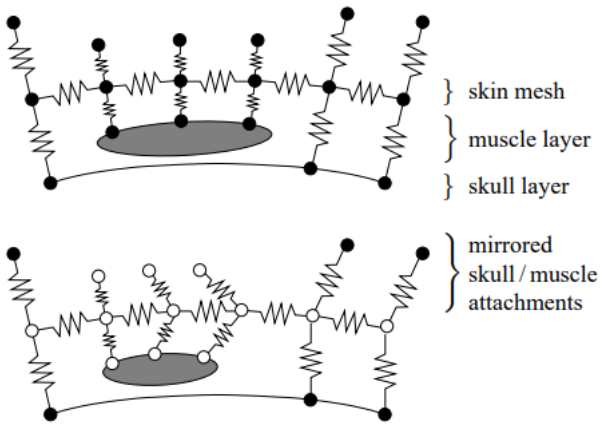

Figure 3.8: The top image shows the system of springs with a muscle that is relaxed. The bottom image shows the same system with the springs moving (indicated by empty points) when the muscle is contracted [18].

To simulate how the system will move, the forces on each point mass must be calculated for each segment of time. The equation below shows how to calculate the force, f_k, that spring k exerts on the points it connects [31]. In the equation: s_k represents the stiffness of spring k, d_k represents the current deformation of spring k, l_k represents the current length of spring k, and x_k represents the vector separation the two points that spring k connects.

$$f_k = (s_k d_k / l_k) x_k \tag{3.2}$$

To be able to set up these kinds of simulations, however, a tremendous amount of work is required to organize and tune the system of springs. Depending on which areas are supposed to wrinkle or not wrinkle, differing stiffnesses of springs must be specified in order to create a realistic looking result. In addition, increasing the resolution of the system does not scale generously. Computation time and sensitivity of the system increase significantly when more springs are added.

3.2.2 Thin Shell Simulation

The properties of thin shells make them another suitable solution for simulation of how skin wrinkles as an underlying face structure animates. Thin shells are used to approximate the multiple layers of tissue that make up the skin that people have on their face [28]. If an object has a thin surface layer with elastic properties that are stiffer than those of the underlying volume, then wrinkles will form when the object is under compression. This happens with human skin because the epidermis and dermis layers have differing thickness and elastic properties from the underlying tissues and muscles [28]. In order for wrinkles to appear in desired locations, parameters can be specified to make areas more or less prone to wrinkling under compression. When a thin shell, acting as a layer of skin, is attached to a soft underlying layer, wrinkling occurs with a critical wrinkling wavelength profile when the compressive force becomes greater than a critical value. This wrinkling occurs in the direction of the compressive force and the critical wrinkling wavelength is calculated by a combination of the thickness of the skin and elastic properties of the underlying soft layer [28]. Each region of the face must have these properties defined for the calculation of the critical values that determines when wrinkling occurs. This is a tedious task and requires a lot of trial and error to map thickness and elastic properties that would make the simulation have realistic looking results.

The equation below shows how to calculate the critical wrinkling wavelength for a given point. The equation includes: thickness of skin h, Young's modulus of the exterior surface E_s, Young's modulus of the interior E_i, Poisson ratio of the exterior surface v_s, and Poisson ratio of the interior v_i [28]. Young's modulus is a measure of elasticity of a material and Poisson ratio is a measure of how a material expands perpendicular to a compressive force.

Figure 3.9: Thin shell skin wrinkling results with differing levels of stiffness (left) and the underlying deformation mesh (right) [28].

$$\lambda = 2\pi h \left[\frac{(1 - v_i^2)E_s}{3(1 - v_s^2)E_i} \right]^{1/3} \tag{3.3}$$

3.2.3 Blend Shape Muscles Approximation

To drive the animation of these physics-based animations, blendshape muscles can be used to deform the model over the course of the animation. Instead of using Blend Shapes to deform an entire face, like what was described in Section 3.1.2, the face is deformed by many individual Blend Shape driven muscles that are controlled independently of one another [10, 31]. While animating between the different states of contraction and relaxation, the blendshape muscle system allows for the linear interpolation between the initial and final poses of the muscles. As the muscles begin to animate, the simulated skin layer follows the movement of the muscles as a result of the spring system that connects the skin to the muscles [10]. When this happens, the skin begins to deform and shows the different wrinkling and stretching affects that are expected when different expressions are being made.

Similarly to the other forms of physics-based approaches to facial animation, this type of rig takes a tremendous amount of effort to refine to create realistic looking animations. Blend Shapes allows the muscle system that drives the animation to

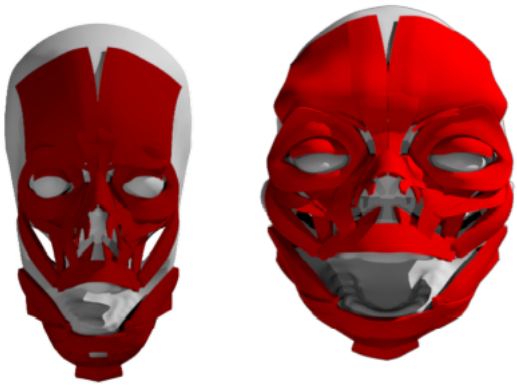

Figure 3.10: Complex rigging of blendshape muscles that drives the animation of these physics-based rigs [10].

be flexible, but identifying problematic muscle deformations can not be effectively automated and therefore requires tedious artist intervention.

3.2.4 Strengths and Weaknesses of Physics Simulations

Once a character is physically rigged for animation, realistic looking animations can be created by artists to display different expressions with little to no adjustment to the rig. Controls of the physics-based rigs are not any different from the interfaces that are used to control the other types of rigs, but the results can be far more realistic. Overall, the focus on approximating human anatomy greatly reduces many of the problems that make the uncanny valley a prevalent issue in facial animation.

The weaknesses with physically-based approaches stem from their complexity. Getting physics-based rigs configured can be a very difficult and tedious task for artists. There has been work done to automate some of the generation of these rigs for physics-based facial animation [2], but there is still a significant amount of work that has to be done by hand to make the rigs look correct and not fall into the trap of the uncanny valley. In addition to the complexity causing the rigs to be difficult to set up, these complex rigs require an immense amount of computations to calculate the ani-

mations. These intensive computations make physically-based animations unsuitable for real-time applications. Physics-based solutions to facial animation have given the entertainment industry great animations that are becoming closer and closer to being anatomically correct, but the tremendous amount of work to get these rigs completed will keep the approach from being used more widely.

3.2.5 Paper Highlight for Physics Simulations

Art-Directed Muscle Simulation for High-End Facial Animation, by Cong et al. [10], serves as a great resource for learning about physics simulations in facial animation. This paper has a large focus on maintaining artistic control, which is unique for physics simulation research because it usually prioritizes accuracy over everything else. The descriptions of how each layer of simulation can be controlled and corrected makes this research valuable to artists and developers. In addition, this paper describes a unique use-case of Blend Shapes as muscles that drive animation, which gives a further understanding of the capabilities of the general method.

3.3 Face Capture

In more recent years, artists have been using various face capture techniques that allow for real faces to be the used as the source for modeling meshes and movements for facial animation. The primary devices that are used for facial capture are cameras that are used to gather information about the subject faces. Cameras can be used to track movements of different regions of a face as a subject deforms their face to make an expression. Other cameras that capture depth have made it possible to reconstruct objects from the physical world in a virtual environment. Since recreating

actual facial expressions is the goal of facial animation, directly using actual faces to guide approximation has made it easier for artists to achieve realistic results.

3.3.1 RGB-D Cameras

A very common device used in facial capture approaches to facial animation are RGB-D cameras. These kinds of cameras allow for color and depth information to be collected about the objects in the frame being captured. The most commonly recognized RGB-D camera is the Microsoft Kinect Sensor which was used in Xbox games to track player movements, in the real world, to control characters in the game. Outside of being used as a unique way for players to control characters in video games, without the use of a controller, the depth information gathered by RGB-D cameras can be used to reconstruct real world objects in a virtual form. Despite being able to capture information about the depth of objects in a frame, raw data from RGB-D cameras can not be used exclusively to approximate facial animation. The data generated by these specialized cameras is often riddled with noise and artifacts that would not make it suitable for creating believable animations [25, 22, 34, 19]. RGB-D cameras can be used in combination with the techniques mentioned later in this section, but they can't be used alone to create quality animations.

3.3.2 Reconstruction

Through the use of camera sensors, different attributes of faces can be recreated for use in making realistic facial animation. Gathered data can be used to reconstruct large-scale features, such as shape, or fine details, such as pores and wrinkles.

When reconstructing the geometry that makes up a human face, multiple frames are captured to generate a reasonably accurate representation of the target face. When

Figure 3.11: Example of raw depth-data generated by a RBG-D camera (right) with the same capture represented in RGB form (left) [20].

using depth cameras, several frames are captured at different angles to deal with the noise that comes from the inaccuracies of the sensors. From these frames, a refined point cloud representation of the face can be constructed for each desired facial expression. With this newly gathered information about the target face, the point cloud data can be used to morph a template set of blendshape poses, that closely resemble the shape of the target face, to create a user-specific blendshape rig that can be used for animation [20, 34]. By using a template, instead of constructing the meshes solely from data generate by RGB-D scans, a cleaner mesh that resembles the target face can be constructed without having to drastically increase the number of needed captures.

If users do not have access to depth cameras, facial features can also be reconstructed using regular 2D images that can be captured using a phone [15]. Similarly to the method mentioned above, many different angles have to be captured to accurately recreate the face. From these different angles, a similarly-shaped mesh is deformed to create a user specific mesh. To help with aligning the target face with the template mesh, the user is required to mark different regions of the face such as the mouth, eyes, and eyebrows.

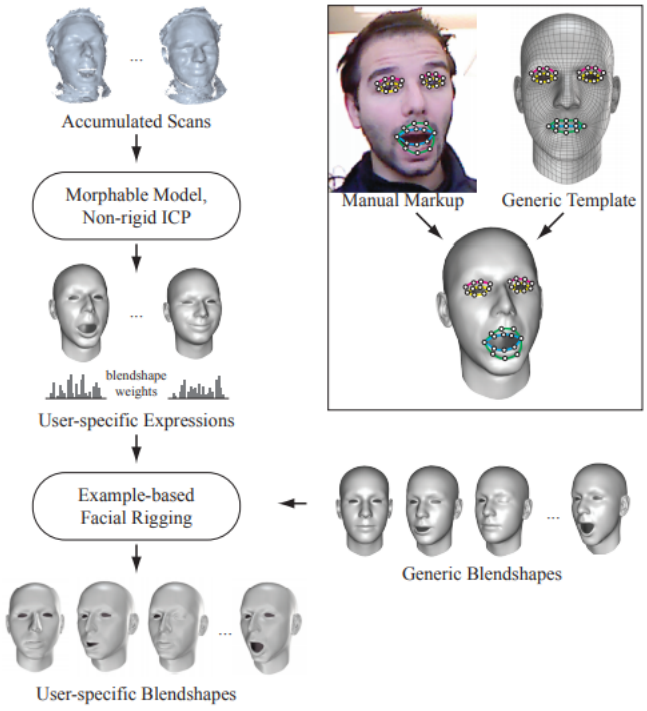

Figure 3.12: User-specific blendshapes created from morphing a generic template based on accumulated scans of a target face using a RGB-D camera [34]

In addition to reconstructing the shape of a face as it creates different expressions, fine-details that change as a face deforms, such as wrinkles, can be reconstructed using the same cameras. RGB-D captures contain the information needed to synthesize height maps that deform vertices to give meshes the appearance of having wrinkles. Before the height map can be created, the captured RGB image must be converted to a gradient image using the Sobel operator. The Sobel operator is an image filtering algorithm used to detect edges by computing an approximate gradient image. Gradient images store the direction of the change of color or intensity in an image and allow the reconstruction of wrinkles to be more robust against differences in skin color and luminance. By calculating the correspondence between the generated gradient image and depth image, a wrinkle map can be synthesized and applied to a mesh to add fine-details such as wrinkles [20].

3.3.3 Tracking

To recreate the movement of different regions of a face, tracking is used to track the deformations that are caused by the formation of an expression. Tracking can be grouped into marker and marker-less methods that require different methods for determining motion of the target.

Tracking using markers on a target face allow for the rigs to be controlled by changes in the positions of the markers. Each virtual marker on a the virtual rig has a corresponding physical marker that determines how the position the virtual marker will change. In changing the virtual marker position, the rig is deformed to match the target that is creating an expression [26]. This method is more straight forward than marker-less tracking, but the process of placing markers exactly where they need to be placed can be tedious. If they are placed in the wrong location, the recreation of the target's expression will not be accurate. To make this easier, however, a mold of

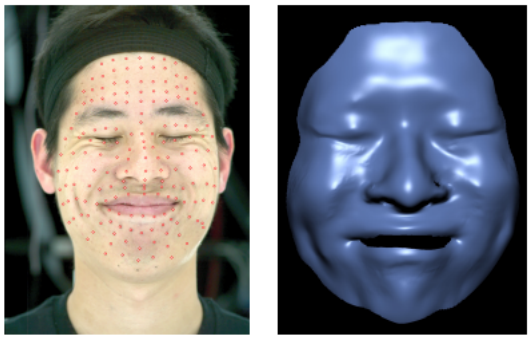

Figure 3.13: Example of marker tracking where the target face with markers (left) is deforming a mesh (right) [22].

the target face can be created and have holes drilled into it to allow for markers to placed at the same place every time [22].

Tracking without markers can also be achieved through the methods of rigid and non-rigid tracking [34, 20]. Before non-rigid tracking occurs, rigid tracking is performed to detect the translation and rotation of the target in order to estimate the transformation of the face between frames. This process is not temporally smooth because of the errors in gathering depth data, but these errors can be averaged out with the previous frames to achieve smoother coherence [34, 20]. Non-rigid tracking involves the estimation of blendshape weights needed to most accurately reproduce the target face's expression. Together, the two steps can compute the the rigid transformation of the face and estimate the mesh deformations for each expression.

3.3.4 Strengths and Weaknesses of Face Capture

Face capture has allowed facial animation to be more accessible to anyone who wants to make facial animation rigs. In almost all of the other approaches, experienced artists are required to create pleasing facial animation results. With the use of phone cameras and face capture technologies, user-specific rigs can be generated that make personalized avatars available to consumer-level applications [15].

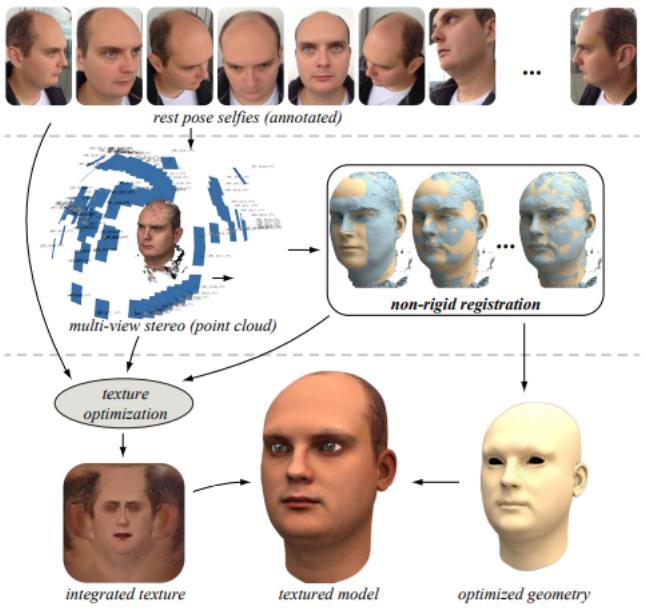

Figure 3.14: Photos captured from cameras can be used to construct personalized avatars without use of expensive camera equipment [15].

The main weakness, however, revolves around the quality of the device that is used to collect data on a target face [34, 20]. The RGB-D cameras available to common consumers can't achieve the same level accuracy as the state-of-the-art sensors used in high-end animation. Inadequate data can't be replaced and all applications can do is request more or better data to ensure the quality of reconstructions [15]. Without a way to collect high quality data, the possible animation applications are severely limited. Phone cameras have the ability to gather enough data to create low resolution avatars, but they can not produce enough data to produce facial rigs usable in applications that require more detail. Fine details, such as wrinkles and pores, would be missing from reconstructions and the results would suffer from the uncanny valley problem. Face capture thrives at making facial animation more accessible, but the reliance on high-end equipment makes high-end animations unreachable by most artists.

3.3.5 Paper Highlight for Face Capture

A helpful resource for learning about face capture techniques is *Dynamic 3D Avatar Creation from Hand-held Video Input*, by Ichim et al. [15]. The paper describes a face capture implementation that generates 3D avatars from many camera images. This implementation is incredibly powerful because it utilizes cameras that an average person would have access to. This is a great improvement in accessibility over methods that require the use of an RGB-D cameras. In addition, the paper provides great descriptions of facial reconstruction and tracking to create believable facial animations.

3.4 Artist Driven Approaches

All methods of creating facial animation are time consuming and require an immense amount of effort to make accurate expressions. Whether it is rigging a model or creating normal maps to add details to your characters, the process is not easy and the lack of standards make it difficult for artists to gain expertise in multiple animation techniques [26]. To make creating facial animation easier and to give artists more choices, researchers are designing animation systems and tools with artists as the main consideration.

3.4.1 Wrinkle Maps from Sketchy Drawings

Fine details about faces often require an artist to manually sculpt details in a complicated 3D environment. This process could be avoided with automated approaches, such as face capture, to reconstruct fine details, but artistic expression is severely limited with the use of these automated techniques. In order to give artists more

Figure 3.15: Basic face mesh (left) is altered by adding wrinkle map generated by the drawing (middle) to create a detailed face mesh with wrinkles (right) [17].

freedom in their creation of fine-scale details, researchers from Sejong University developed a tool that allows for sketchy drawings to be converted into wrinkles maps that could be applied to meshes [17]. By simply drawing on a tablet with a stylus, artists can control which regions of the face they want to add more details to. From the drawings, the center and thickness of lines are calculated to determine the areas of influence of each wrinkle that is drawn. Based on the location and thickness of strokes, varying levels of deformations are created on a surface.

During an informal user study, users reported that the method felt natural because they were used to drawing with pencils on paper and drawing thicker lines to increase intensity of wrinkles was intuitive. This research shows how animation techniques can be designed to be intuitive to artists. The most common tools for animation creation are very similar to each other and there is still a lot of room for improving the experiences of artists.

3.4.2 Stylized Facial Animation

In addition to being able to recreate realistic looking facial expressions, some artists want to be able to have more control to create stylised animations. Instead of creating another tool to generate 3D animations, researchers from Hasselt University created a face capture driven facial animation system that could animate stylized 2D cartoon faces [13]. Face capture was used to track the timing and movements of a target face that drove animation of the system. The 2D hand-drawn faces were then deformed to create animations using this collected data. By using a hybrid approach, the researchers were able to get the best out of both face capture and 2D animation that most artists are comfortable with.

Not every animation has to be photo-realistic and there are many stories that are better told by intentionally cartoon-like characters. Since the uncanny valley is such a prevalent problem in facial animation, many artists want to be able to avoid the problem entirely. By animating artistic characters, artists are given more control over the exaggerated expressions they want to portray while having to worry about disturbed reactions from audiences.

3.4.3 Strengths and Weaknesses of Artist Driven Approaches

Artists driven approaches draw their strengths from being intuitive and providing specific options to artists that could not be effortlessly reproduced using generic modeling or animation techniques [17, 13]. Many generic programs for facial animation try to be flexible to accommodate as many types of projects as possible. This flexibility hurts usability and artists struggle to be able to execute their visions for a character using these generic techniques. Artists-driven techniques are very good at providing specific assistance to artists that makes the task of animating their visions more in-

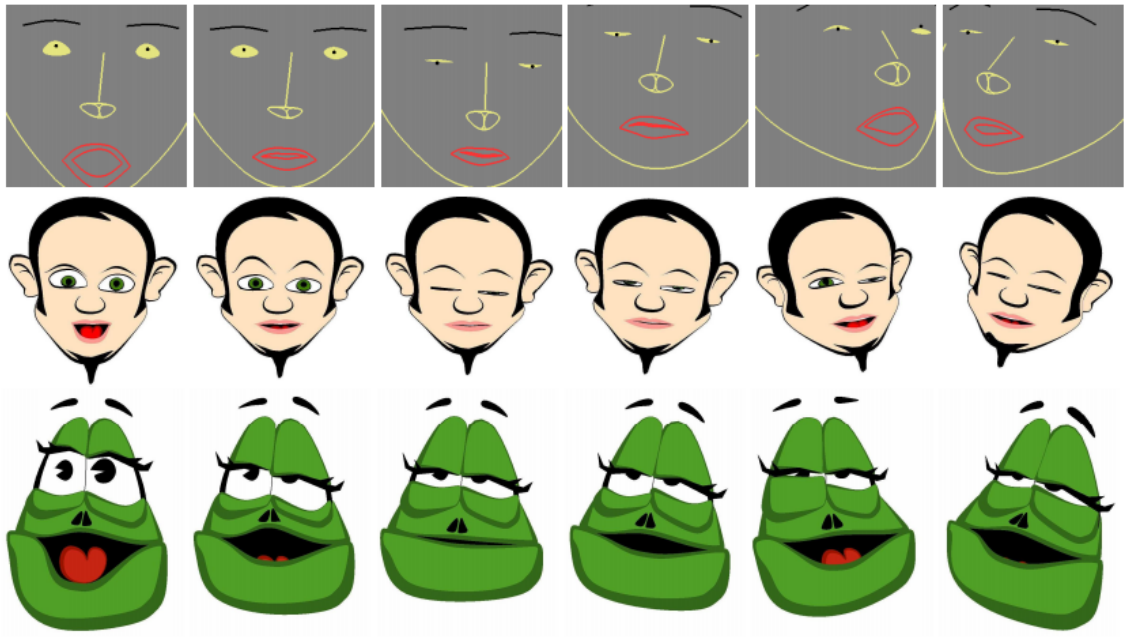

Figure 3.16: Tracked facial expression input (top) for driving animations of the 2D modeled characters (bottom) [13].

tuitive. Not only are these approaches more intuitive, but these approaches can also give artists the ability to create stylized animations that are better suited for specific types of characters. Some characters are unnaturally expressive and photo-realistic representations would fall victim to the uncanny valley. The additional options that artist-driven approaches provide allow for artists to create more fitting facial animations.

Being straightforward in terms of creation and application, however, is also what generates the biggest weakness of artist driven approaches. They are not very useful outside of their specific application and contribute to the problem of having no shared standards between facial animation pipelines [26]. Learning how to use any of these artist driven methods will not necessarily help an artist be prepared to use another method of facial animation. Without someway to translate animations from artist-driven methods to more widely used techniques, the assets created by artist-driven approaches will not be able to be used outside of their respective environments.

3.4.4 Paper Highlight for Artist-Driven Approaches

Interactive Generation of Realistic Facial Wrinkles from Sketchy Drawings, by Kim et al. [17], is a great resource for learning about artist-driven approaches to facial animation. The paper describes an implementation for generating wrinkle maps using drawings from artists. This allowed artists to directly control wrinkle intensity and placement with the thickness and location of their pen strokes. Overall, the resulting application was intuitive to use and serves as a great example of how implementations can be designed for artists.

Chapter 4

MACHINE LEARNING APPROACHES

In order to tackle the challenges associated with traditional approaches to facial animation, researchers have turned to machine learning to improve the experience of creating facial animations and expand the range of applications for high-end facial animation. Creating and perfecting facial animation rigs is a long and iterative process. Specialized artists are needed to complete rigs and each character requires their own rig for facial animation. Very little work can be reused for creating high-end animations and the process of rigging becomes the bottleneck in many CG production pipelines [26]. Machine learning provides solutions that: reduce much of the tedious work that artists must complete for accurate facial animations, optimize the computation time of complex rigs, and allow for work to be effectively reused across multiple characters.

4.1 Maintaining Artist Involvement

While creating facial animations with machine learning methods, it is important for the methods to retain artist involvement. Machine learning is often thought of as a black box because people can only observe the inputs and outputs of a given model. The lack of knowledge of how an input is transformed into an output can create a sense of separation between the user and the application. This disconnection can be detrimental to the quality of artistic creations because artists need to be able to control the fine details that represent their vision.

Dr. Alexei Efros is a Professor at University of California, Berkeley, who has completed many research projects that promote artist involvement in machine learning applications. His works contribute to the visual arts by allowing artists to directly control the outputs of his machine learning applications [7, 14, 36]. In a project to generate pictures of objects, artist inputs of sketchy drawings were used to create the boundaries of objects that would be automatically filled in to match the appearance of a specified object [14]. Another one of his projects allowed artists to automatically recolor grayscale images from colors inputted by artists at specific locations in the image [36]. These projects are examples of machine learning applications that keep artists directly involved in the creation of art. They strike an important balance between taking away tedious work without taking away artist involvement.

4.2 Approximating Mesh Deformations in Real Time

To achieve high-quality animations of faces for use in media, such as films, complex rigs and computationally expensive deformations must be calculated. These computations are intensive and require high-end hardware to be able to calculate these animations in any interactive sense. Many artists do not have access to high-end hardware and can not animate complex rigs in real-time. To be able to interactively work with complex rigs with consumer hardware, machine learning can be used to approximate the expensive deformations made by expressive characters in a fraction of the time required by non-machine learning approaches [3].

Generating the training data for the model to approximate deformations made by a complex rig is more forgiving than other forms of generating data because the data is derived directly from the same rig. The rig being approximated is deformed into many random poses by generating random values for the parameters that produce

Figure 4.1: Example poses from training data for generating a model to approximate deformations made by a character's expressions. Each of the poses are generated randomly [3].

the face deformations. For each of the models produced in the authors' experiments, 50,000 samples were generated by randomly deforming the rigs [3]. After the data sets were calculated, balancing is performed to make sure that the model is not over-fitted to more frequently generated expressions, such as a neutral pose. Balancing occurs by sorting training data into bins based on the similarity of deformation data. Training data is then selected by uniformly sampling a bin and then uniformly sampling deformations within the selected bin. Without this step, the models would perform poorly for expressions that were different from the over-fitted expressions.

The model for approximating deformations is divided into two different stages with the first being a coarse approximation of the deformation [3]. The coarse model operates on the entire mesh and creates low-resolution deformation maps that can be calculated quickly. Each of the deformation maps are used to calculate vertex offsets that are summed to produce the final vertex positions for the deformation approximation. With this approximation model, model training can occur by penalizing inaccuracies in normal approximations. This correction prevents visually disturbing errors and encourages smooth edges in the final approximation of deformation.

The second stage of the model is a refined approximation which is responsible for recovering all of the fine-details lost during the coarse approximation stage [3]. Fine-details, such as wrinkles, can not be reproduced by the coarse approximation stage. The refined approximation focuses on the vertex-dense regions of the face to produce

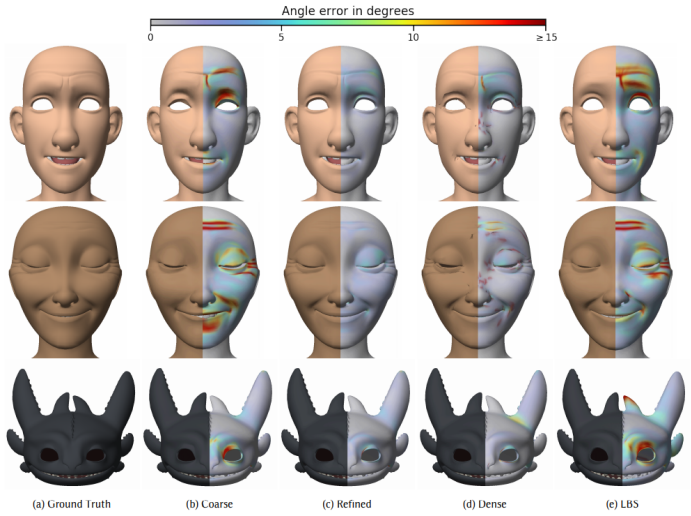

Figure 4.2: Visualization of the approximation errors produced by the different methods. The "Ground Truth" column contained the deformations that the different methods were trying to reproduce. The "Dense" and "LBS" columns showed previous approaches that tried to approximate deformations at interactive speeds. The "Refined" column was the combined coarse and refined approximations that produced the most accurate results [3].

high-resolution displacement maps which capture the missing fine-details. If refined approximations were computed for the entire face, the run-time would drastically increase making the system less useful for interactive applications. To determine the regions that need refined approximation, approximation error from coarse model is calculated per vertex. From these error calculations, clusters of vertices with high error are identified and grouped to form regions that will be subjected to refined approximation. The higher-resolution deformation maps are calculated for these high-error regions to improve the accuracy of the whole deformation without drastically increasing the computational complexity of the approximation.

By combining the coarse and refined models, an accurate and quick approximation of complex rigs is achieved. From the experiments conducted by the authors, they were able to show how this machine learning approach is able to compute approximations of

facial deformation that is 5-17 times faster than traditional methods using consumer-level hardware [3].

4.3 Audio Driven Facial Animation

When trying to create animations for a large set of characters, machine learning can be used to replace much of the effort required to make characters expressive during their conversations. Researchers at NVIDIA created a machine learning technique that can create animations for a mesh from recorded audio [16]. This can be an extremely helpful tool for applications, like video games, where there are hours of scripted dialogue that must be animated. Animations that previously required expensive face-capture equipment can now be approximated with realistic-looking results.

To drive the animations with dialogue, a voice recording must have an associated emotional state and the speech content must be processed [16]. The emotional state must be specified separately from the recording because the emotional state can not always be correctly identified. Two recordings may sound exactly the same, but they could be trying to express two distinct emotions. Having the ability to specify what emotion is being displayed gives artists more control over what animations are derived from the audio. The speech content must also be processed to identify different aspects of the recording. Specific frequencies are identified that give information on phonetic content. Phonetic content is important for distinguishing the different sounds between words like "pad" and "pat." To be able to recognize this content, phonemes, units of sound that distinguish words, can be extracted from voice recordings. The visual differences in how two words, like the ones previously mentioned, are spoken is important to consider when creating believable animations. Characteristics

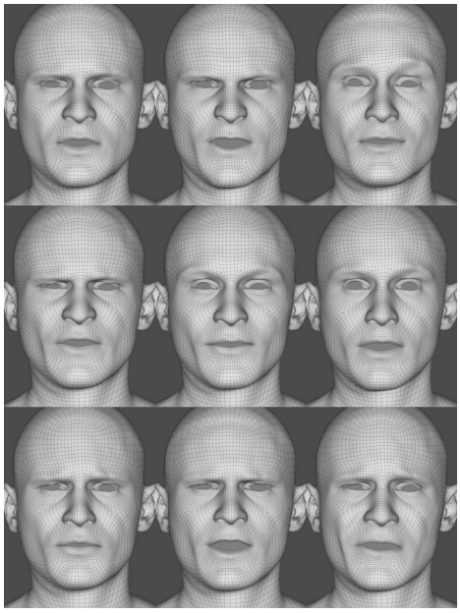

Figure 4.3: Visualization of the same recording being animated for different emotional states. This level of control could not be achieved if the emotional state was directly derived from the audio recording [16].

like pitch and timbre can also be collected from voice recordings, but they are not as important when reconstructing an accurate facial animation.

Individual models were trained for each character and training data is generated using face capture techniques. For each model created, the training target is required to pose in various poses that cover a variety of expressions and facial movements. Targets are also required to speak sentences that include many distinct phonemes that each have distinguishable facial movements. In 3-5 minutes, enough training data can be collected to create reasonably accurate machine learning models capable of deriving animations from an audio recording and a specified emotional state [16].

The training data gathered is grouped into the parts of pangrams and in-character material. Pangrams attempt to gather information on the range of motion that is unique to a character's face when speaking. This information comes from identifying the motions that result from speaking specific phonemes. In-character material

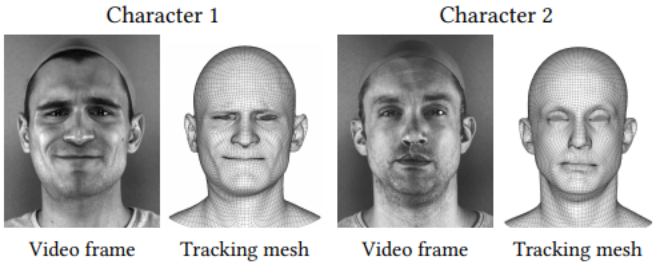

Figure 4.4: Examples of gathering training data from actors, using face capture techniques, for creating machine learning models [16].

attempts to capture the expressive and emotional range of a character so that the derived animations stay true to the character. By focusing on the expressions that are deemed important to the portrayal of the character, the inferences may not perfectly reflect the actor's expressions during recording, but the resulting animations will stay in character. The in-character material also allows for animations to remain true to the character when different voices are encountered for the same character giving more flexibility to the resulting rig.

After the researchers at NVIDIA completed their implementation, studies were conducted to test the robustness of the machine learning models they were able to create [16]. For their studies, they collected recordings from different sources that spoke in different languages with various tones. With these recordings, they created animations with their machine learning models and with face capture techniques. To compare their resulting animations, a survey was conducted that asked participants to choose the more natural animation for the same audio recording. The face capture animations were voted to be more natural than audio-driven animations by 77 percent of the participants. The audio-driven approach is not as good at creating natural-looking animations as the ones directly derived from visual data, but their animations are still of high-quality and have other potential strengths. Since the machine learning models can work with any recording, facial animations can be produced with synthetic-audio derived from text to produce realistic-looking results [16]. Ap-

plications, like games, could strongly benefit from the ability to automatically create hours of animations from hundreds of pages of text. After an actor is used to create the initial machine learning model, realistic facial animations can be created without any further human inspiration.

4.4 Non-linear Face Modeling

Many approaches that reconstruct faces using face-capture methods will create facial rigs that deform linearly. To create animations, static poses are linearly blended to deform the face into unique expressions. The problem with these approaches is faces do not move linearly. Human faces deform in very non-linear paths and trying to recreate all of these expressions with linear methods would not be possible. By blending static poses, only a limited set of poses can be created and that set will contain many unnatural expressions. Machine learning approaches to face reconstruction offer the ability to resolve these issues by providing non-linear face modeling methods that can create natural-looking expressions that maintain the same level of artistic control [8].

Machine learning models have the ability to non-linearly deform faces as they transition between different static poses and provide artistic control by separating the dimensions of identity and expression [8]. Identity is defined as the general shape and texture of a face. With the dimension of identity, different faces with unique ethnicities, genders, ages, and BMI's can be accurately modeled. Expression is defined as how the face shape deforms when the different poses are being blended to create new expressions. With the dimension of expression, different facial expressions can be non-linearly interpolated to create animations that look more natural than linear approaches.

Figure 4.5: The albedo texture maps gathered for training data store many fine details about the face and allow for effects like blood flow to be reconstructed during animation [8]. The different expressions show the differences in blood flow in the nose that are caused by scrunching (left) and stretching (right) the face.

To train these machine learning models to create non-linear face rigs, a large data set of faces, with multiple expressions for each subject, must be collected. The data set must also have an even distribution of different subject faces so that the model can synthesize a variety of unique facial types. Each of the subjects have their faces recorded as combination of a 3D mesh and a albedo texture map for each of the static blendshape poses. The albedo texture is created by taking the captured texture and dividing out the diffuse lighting that is measured by a light probe. In addition to collecting static poses, dynamic poses are captured as subjects perform face workout sequences. This data is important for training the model to recognize the non-linear deformation of faces. The machine learning model created by researchers at Disney Research required data to be collected on 224 subjects [8].

After the machine learning model is trained, novel meshes and expressions can be created and easily controlled as a result of the separation between identity and expression [8]. The major benefit of this method is the ability to reuse trained expression models with new identity data to generate new poses. This is extremely useful for tasks like avatar creation where the same set of expressions is portrayed by a wide range of characters in the same virtual application.

Figure 4.6: The error caused by linear blending (top) of static shapes is extreme when compared to non-linear blending methods made possible machine learning (bottom) [8]. The most notable errors come from using extreme blendshape weights.

4.5 Deepfakes

In more recent years, there has been an increased interest in using facial animation to alter faces of real actors to match a desired appearance. Characters in movies and shows have been tied to specific actors' faces and studios want to continue using the same faces to maintain continuity in their stories. As actors age or die, it becomes harder for stories to continue using their likeness. A common approach that allows actors to reprise an old role is called de-aging [1]. The technology overlays a 3D mask, that makes actors look similar to their younger self, making it possible for actors to continue acting as a character they no longer look like. A large problem with de-aging is the inability for the technology to avoid the uncanny valley. De-aged characters do not look human and the problem is made more severe when they are shown in frames with real people. Just like many other facial animation techniques, de-aging struggles with making facial movements and skin composition look natural. Improvements to traditional facial animation approaches are making this problem less and less of an issue, but machine learning has provided a much more convincing solution to reanimating faces. Machine learning approaches to animate target faces on top of source faces have been very successful at avoiding the uncanny valley and their

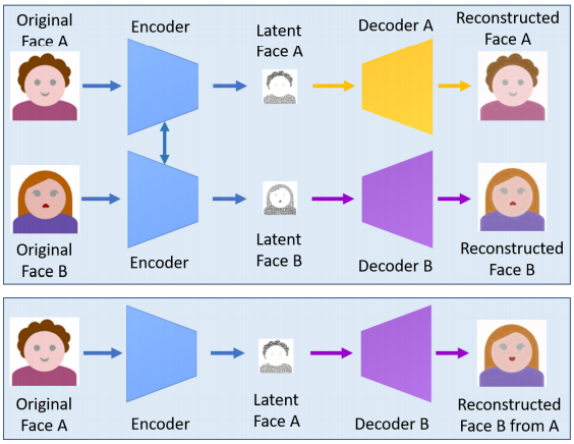

Figure 4.7: A model for creating deepfakes using two encoder-decoder pairs. Both faces share the same encoder and each face uses their own decoder for the training process (top). The source face is then encoded with the shared encoder and then decoded with the target decoder to create a deepfake (bottom) [24].

products are known as deepfakes. With deepfakes, photo-realistic facial animation can be created to replace faces in video content [24].

To be able to train a model to create deepfakes, a large quantity of pictures need to be gathered of the target face and source face that are going to be swapped[24]. For famous individuals, this data can usually be collected easily because most, if not all, of the data required can be gathered from interviews. Interviews are usually good because they can provide many unobstructed and high-quality views of the target face. These are both important qualities of good data for training an accurate deepfake model. For the source face, the data can be derived directly from the video that will contain the face swap. After the data is gathered, the data needs to be cleaned to only include the faces in the swap so that model is not confused when trying to match features of the faces that are being swapped.

Once the data has been gathered and cleaned, training can begin to start learning how to match facial features of the to faces being swapped. Common deepfake models are made up of an autoencoder-decoder pairing structure [24]. In this structure, two

Figure 4.8: A comparison of Mark Hamill reprising his role as Luke by use of de-aging technology (left) and deepfake technology (right) for the last episode of *The Mandalorian* [1].

encoder-decoder pairs are trained where each pair is trained on a face image set. One image set would be for the source face and the other image set would be for the target face that is being swapped over the source face. The encoder is shared between the two faces and learns how to encode information about the features of the two faces. The strategy of having a shared encoder is important because it trains the encoder to recognize similarities in the two faces. Features like the eyes, noes, and mouth would be shared between the two faces and would be important for the encoder to recognize. The decoder is responsible for reconstructing the encoded data back into the original image. Each face image set has its own decoder that learns how to reconstruct the encoded data back into the original image. Once both pairs have been trained, the shared encoder can be used to encode the source face and the target face's decoder can be used to reconstruct the encoded data in the form of the target face. The result is a new frame where the source face is replaced with the target face.

Tools such as "Face Swap" and "DeepFaceLab" use machine learning to make the process of creating deepfakes easy and accessible to anyone who wants to swap faces in a video [23]. The resulting frames that are produced are fairly convincing and the frames that are not temporally stable or accurate can be manually fixed. Deepfakes are so effective at producing realistic results that individuals with deepfake tools could

outperform major studios at producing photorealistic content. In the last episode of *The Mandalorian*, Mark Hamill was de-aged to be able to reprise his role as Luke Skywalker so that the episode could maintain continuity in the *Star Wars* universe. Despite the immense resources of the studio, a YouTuber, called Shamook, was able to create a more realistic recreation of Mark Hamill using deepfake technologies [1]. The produced deepfake face had more convincing features and human expressions compared to the de-aged face that had plastic-looking skin and unnatural expressions.

As a result of how convincing deepfakes are, deepfakes are being greatly criticized because of their ability to trick people into believing they are real [24, 23]. The potential to create fake news and slander reputations has led to a large amount of research into detection of deepfakes to prevent the spread of misinformation. The ability of deepfakes to produce convincing faces makes them both a very useful tool for reanimating faces in entertainment applications and a destructive weapon for spreading misinformation in the real world.

4.6 Strengths and Weaknesses of Machine Learning Approaches

Machine learning approaches to facial animation solve many of the issues found in traditional approaches to facial animation. Complex rigs that are too computationally expensive to calculate in real-time, like physics-simulations, can be approximated in real-time with models trained on data from the original rig [3]. Rigs can be animated automatically, without the expertise of an artist or an actor, to create realistic animations using unique inputs like voice data [16]. Rigs can deform in non-linear paths to correctly approximate formation of real expressions compared to linear methods used in traditional approaches [8]. Actors in videos can be swapped to produce photorealistic results that are capable of outperforming state-of-the-art de-aging techniques

used by major studios [1, 24, 23]. Machine learning approaches have created solutions that provide higher quality and accessibility than traditional approaches to facial animation. If the data is accessible, accurate models can be trained to produce high-quality facial animations.

The weaknesses of machine learning approaches stem from the difficulties to set up models and the fragility of the created models. Machine learning requires an immense of data to be able to train accurate models. This kind of data is not easily accessible, or is non-existent, for training facial animation models because the approach is fairly new and is not widely used [8]. Without enough data, models can be inaccurate and could produce results that would be trapped in the uncanny valley. Even when there is enough data, models can still struggle to produce consistent results. Machine learning models have differing levels of error and models with high levels of error struggle to maintain temporal coherence or to produce accurate facial animations [8]. Since machine learning approaches can not guarantee correct outputs, inaccurate animations may be produced that distract from the same scenes that machine learning models would be trying to make more engaging. Without addressing these problems, machine learning approaches to facial animation will continue to not be widely used.

4.7 Paper Highlight for Machine Learning Approaches

An informative resource for learning about machine learning in facial animation is *Audio-Driven Facial Animation by Joint End-to-End Learning of Pose and Emotion*, by Karras et al. [16]. The paper describes a method for training machine learning models that can create facial animations from audio inputs. Despite the use of machine learning models, artists still have control over the outputs through directing how actors move their mouths to make sounds during data collection and through

specifying emotion when the model creates animations. In addition, this paper is a great resource through its descriptions of how to collect data that trains machine learning models to learn the specific facial movements of unique characters.

Chapter 5

REFLECTION

There are many different approaches to facial animation and it can be difficult to get started with creating facial rigs or developing facial animation technologies. Since every method is so unique, the amount of knowledge that transfers between approaches is limited. Most studios have their own facial animation pipelines and the lack of universal standards makes it so there is no common approach across industry [26]. Even when trying to find help for smaller projects, useful information is hard to come by because of the large number of approaches and the specificness of each character that can not be handled perfectly any method. There is a "mixture of science, art, and craft" to creating convincing facial animation rigs [26].

To solve some of these problems that can not be solved by traditional approaches, machine learning approaches have been developed to improve the quality and accessibility of facial animation. These new methods can produce promising results, but they can be difficult to set up and there uses can be very specific. Machine learning works by creating a model that learns to make accurate inferences from data. The process of gathering enough data to create an accurate model can be very difficult and time-consuming because good data sets for training facial animation models do not exist [8]. Even when enough data is collected to train an accurate model, the created model will not usually transfer to other projects for practical reuse. Machine learning models are good at solving the specific problems they are trained for, but they are not typically flexible enough to solve the same problem with minor variations. Machine learning approaches to facial animation have provided promising results, but they are still no where near being widely adopted due to their limited use and fragility.

Regardless of the chosen method for facial animation, there is going to be a ton of work required to make the character function as needed. No method is flexible enough to produce perfect results without the intervention of an artist. Each method has its strengths and weaknesses that need to be considered when starting a new project. Knowing as many methods as possible is the best way to reduce the chance of incorrectly choosing a facial animation approach that was never meant for a specific face.

5.1 Future Directions

To be be able to substantially improve the field of facial animation with machine learning, future research will be needed to remedy the issues that make machine learning undesireable to artists and developers. Machine learning approaches to facial animation have had promising results, but difficulties with model creation and model transferability have prevented its wide-spread use. Creating machine learning models requires a large amount of labeled data and the data sets necessary for training a model are difficult and time-consuming to create. The benefits in performance are then further undercut by the insufficient ability of machine learning models to be reused for multiple characters. In order to increase the use of machine learning in facial animation, future work will be needed to research and create solutions to allow for easier model creation and transferability.

Researchers from Beijing have started to address the issues with model creation by creating a large database of faces and corresponding expression data [35]. The collection of data will be immensely helpful for the creation of new machine learning models because the cost of creation will be greatly reduced by the existence of data set that was created specifically for facial animation research. The database can still

be improved, however, because access to the database is restricted by an approval process and the diversity of the data is limited to residents of Asia. Further research, like the one previously mentioned, will be important for solving the issues of model creation and transferability and contribute to the increased use of machine learning in facial animation.

Chapter 6

CONCLUSION

Facial animation is hard and it makes sense why there are so many approaches to try and solve the same problem. Based on the application and the experience of an artist, certain approaches meet specific desired requirements better than others. Many of the traditional approaches are good at producing interactive and engaging facial animations, but they struggle with providing easy rigging and producing realistic approximations. To be able to address many of the issues of traditional facial animation approaches, machine learning has been utilized to improve the quality and capabilities of the facial animation rigs. These machine learning methods have achieved promising results, but they come with their own challenges that make them unsuitable for certain applications. Before starting a facial animation project, artists and developers should have a high-level understanding of the many different facial animation methods to ensure they follow an appropriate approach to address their specific needs. This survey paper provides artists and developers with a resource that assists with making decisions on what methods would be best for their unique projects. By extending the work of previous survey papers [25, 11, 26] to highlight artist interventions and newer machine learning methods, this survey paper provides a comprehensive summary of currently researched approaches.

BIBLIOGRAPHY

[1] Deepfake Luke Skywalker Is Way More Convincing In The Mandalorian.

[2] O. Aina and J. Zhang. Automatic muscle generation for physically-based facial animation. 01 2010.

[3] S. W. Bailey, D. Omens, P. Dilorenzo, and J. F. O'Brien. Fast and deep facial deformations. *ACM Transactions on Graphics*, 39(4), July 2020.

[4] Y. Bando, T. Kuratate, and T. Nishita. A simple method for modeling wrinkles on human skin. In *10th Pacific Conference on Computer Graphics and Applications, 2002. Proceedings.*, pages 166–175, Beijing, China, 2002. IEEE Comput. Soc.

[5] B. Bickel, M. Bächer, M. A. Otaduy, W. Matusik, H. Pfister, and M. Gross. Capture and modeling of non-linear heterogeneous soft tissue. *ACM Transactions on Graphics*, 28(3):1–9, July 2009.

[6] B. Bickel, M. Lang, M. Botsch, M. A. Otaduy, and M. Gross. Pose-Space Animation and Transfer of Facial Details. *Eurographics/SIGGRAPH Symposium on Computer Animation*, page 10 pages, 2008. Artwork Size: 10 pages ISBN: 9783905674101 Publisher: The Eurographics Association.

[7] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros. Everybody Dance Now. *arXiv:1808.07371 [cs]*, Aug. 2019. arXiv: 1808.07371.

[8] P. Chandran, D. Bradley, M. Gross, and T. Beeler. Semantic Deep Face Models. In *2020 International Conference on 3D Vision (3DV)*, pages 345–354, Nov. 2020. ISSN: 2475-7888.

[9] J. Chim and H. Kim. Dynamic skin deformation and animation controls using maya cloth for facial animation. In *ACM SIGGRAPH 2002 conference abstracts and applications on - SIGGRAPH '02*, page 175, San Antonio, Texas, 2002. ACM Press.

[10] M. Cong, K. S. Bhat, and R. P. Fedkiw. Art-Directed Muscle Simulation for High-End Facial Animation. *Eurographics/ ACM SIGGRAPH Symposium on Computer Animation*, page 9 pages, 2016. Artwork Size: 9 pages ISBN: 9783038680093 Publisher: The Eurographics Association.

[11] Z. Deng and J. Noh. Computer Facial Animation: A Survey. In Z. Deng and U. Neumann, editors, *Data-Driven 3D Facial Animation*, pages 1–28. Springer London, London, 2007.

[12] L. Dutreve, A. Meyer, and S. Bouakaz. Real-Time Dynamic Wrinkles of Face for Animated Skinned Mesh. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, G. Bebis, R. Boyle, B. Parvin, D. Koracin, Y. Kuno, J. Wang, R. Pajarola, P. Lindstrom, A. Hinkenjann, M. L. Encarnação, C. T. Silva, and D. Coming, editors, *Advances in Visual Computing*, volume 5876, pages 25–34. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. Series Title: Lecture Notes in Computer Science.

[13] F. D. Fiore and F. V. Reeth. Multi-level Performance-driven Stylised Facial Animation. page 6.

[14] A. Ghosh, R. Zhang, P. K. Dokania, O. Wang, A. A. Efros, P. H. S. Torr, and E. Shechtman. Interactive Sketch & Fill: Multiclass Sketch-to-Image Translation. *arXiv:1909.11081 [cs, eess]*, Sept. 2019. arXiv: 1909.11081.

[15] A. E. Ichim, S. Bouaziz, and M. Pauly. Dynamic 3D avatar creation from hand-held video input. *ACM Transactions on Graphics*, 34(4):1–14, July 2015.

[16] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics*, 36(4):1–12, July 2017.

[17] H.-J. Kim, A. C. Öztireli, I.-K. Shin, M. Gross, and S.-M. Choi. Interactive Generation of Realistic Facial Wrinkles from Sketchy Drawings. *Computer Graphics Forum*, 34(2):179–191, May 2015.

[18] K. Kähler, J. Haber, and H.-P. Seidel. "Geometry-based Muscle Modeling for Facial Animation". page 11.

[19] H. Li, J. Yu, Y. Ye, and C. Bregler. *Realtime Facial Animation with On-the-fly Correctives.*

[20] J. Li, W. Xu, Z. Cheng, K. Xu, and R. Klein. Lightweight wrinkle synthesis for 3D facial modeling and animation. *Computer-Aided Design*, 58:117–122, Jan. 2015.

[21] A. Lonkar. The Uncanny Valley The Effect of Removing Blendshapes from Facial Animation. page 58.

[22] W.-C. Ma, A. Jones, J.-Y. Chiang, T. Hawkins, S. Frederiksen, M. Vukovic, M. Ouhyoung, P. Debevec, and P. Peers. Facial Performance Synthesis using Deformation-Driven Polynomial Displacement Maps. page 10.

[23] B. U. Mahmud and A. Sharmin. Deep Insights of Deepfake Technology : A Review. page 12.

[24] T. T. Nguyen, C. M. Nguyen, D. T. Nguyen, D. T. Nguyen, and S. Nahavandi. Deep Learning for Deepfakes Creation and Detection. page 17.

[25] J.-y. Noh. A Survey of Facial Modeling and Animation Techniques. page 26.

[26] V. Orvalho, P. Bastos, F. Parke, B. Oliveira, and X. Alvarez. A Facial Rigging Survey. *Eurographics 2012 - State of the Art Reports*, page 22 pages, 2012. Artwork Size: 22 pages Publisher: The Eurographics Association.

[27] C. D. G. Reis, J. M. D. Martino, and H. C. Batagelo. Real-time Simulation of Wrinkles. page 8.

[28] O. Rémillard and P. G. Kry. Embedded thin shells for wrinkle simulation. *ACM Transactions on Graphics*, 32(4):1–8, July 2013.

[29] D. Stamenković, M. Tasić, and C. Forceville. Facial expressions in comics: an empirical consideration of McCloud's proposal. *Visual Communication*, 17(4):407–432, Nov. 2018.

[30] N. Tatarchuk. Advanced Real-Time Rendering in 3D Graphics and Games. page 144, 2007.

[31] D. Terzopoulos and K. Waters. Physically-Based Facial Modeling, Analysis, and Animation. page 18.

[32] S. Wang, S. O. Lilienfeld, and P. Rochat. The Uncanny Valley: Existence and Explanations. *Review of General Psychology*, 19(4):393–407, Dec. 2015.

[33] K. Waters and D. Terzopoulos. The Computer Synthesis of Expressive Faces. page 8, 2021.

[34] T. Weise, S. Bouaziz, H. Li, and M. Pauly. Realtime performance-based facial animation. In *ACM SIGGRAPH 2011 papers on - SIGGRAPH '11*, page 1, Vancouver, British Columbia, Canada, 2011. ACM Press.

[35] Y. Yan, K. Lu, J. Xue, P. Gao, and J. Lyu. FEAFA: A Well-Annotated Dataset for Facial Expression Analysis and 3D Facial Animation. *arXiv:1904.01509 [cs, eess, stat]*, Apr. 2019. arXiv: 1904.01509.

[36] R. Zhang, J.-Y. Zhu, P. Isola, X. Geng, A. S. Lin, T. Yu, and A. A. Efros. Real-time user-guided image colorization with learned deep priors. *ACM Transactions on Graphics*, 36(4):1–11, July 2017.