Multimedia Systems (2023) 29:1917–1940
https://doi.org/10.1007/s00530-023-01081-2

REGULAR PAPER

A survey on the pipeline evolution of facial capture and tracking for digital humans

Carlos Vilchis[1] · Carmina Perez-Guerrero[2] · Mauricio Mendez-Ruiz[2] · Miguel Gonzalez-Mendoza[1]

Received: 30 August 2022 / Accepted: 20 March 2023 / Published online: 1 April 2023

Abstract
With the introduction of concepts for virtual interaction and digital doubles, a rich scenario has been created for embodied avatars to strive. These avatars, more recently referred to as digital humans, have become a popular area of research, resulting in various techniques and methods that focus on improving the perception of their realism, fidelity, emphatic response, and interactivity. This survey aims to explore the literature and recent advancements on the key processes behind the creation and animation of digital human faces through the view of a general pipeline. The extensive review carried out in this study explores the usual data collection protocols, the main facial codification paradigms and databases, the approaches for digital human asset creation, facial tracking solutions for performance-driven animation, the solving process, and the final rendering delivery. Different quantitative evaluation methods, visual perception tests, and empathetic response evaluations for digital humans are also included in the survey. Additionally, the paper presents an updated summary of public and private frameworks for digital humans that go through the complete general pipeline presented. Finally, the condensed knowledge is discussed, inquiring into the possible direction of future developments in the field.

Keywords Digital humans · Photogrammetry · Deep learning · Facial codification · Facial expressions · Empathic response

1 Introduction

In the last decade, digital humans have become a relevant and consolidated subject of research as a new form of embodied conversational agents (ECAs), earning their place in the Hype Cycle of Emerging Technologies in 2021 [20]. There is a wide variety of applications for digital humans in different areas, such as customer service, government communication, healthcare, e-commerce, education, and film-making. The visual representation of a human agent requires a diverse set of technologies and an extensive list of crucial factors that must be archived, such as realistic graphics [126], emphatic response [105], and a correct model that can make replicate the facial expressions and performance of a unique human [59].

The construction of facial models and expressions for digital humans consists of recreating all of what the human face does, when talking and expressing various emotions. This process is called facial coding [97]. Common approaches in recent years make use of state-of-the-art computational techniques that involve Computer Vision and Machine Learning to improve the study, analysis, and extraction of realistic facial codifications from real-world face recognition of real subject faces. Some approaches include Artificial Networks and Convolutional Neural Networks models, using computer vision-based recognition [19, 59].

This survey explores the evolution of methods, frameworks, and solutions for facial reconstruction and expression reenactment for offline animation or real-time performances

Carlos Vilchis, Carmina Perez-Guerrero, Mauricio Mendez-Ruiz contributed equally to this work.

✉ Carlos Vilchis
carlos.vilchis@tec.mx

Carmina Perez-Guerrero
carmina@eugenia.tech

Mauricio Mendez-Ruiz
mauricio@eugenia.tech

Miguel Gonzalez-Mendoza
mgonza@tec.mx

[1] School of Engineering and Sciences, Tecnologico de Monterrey, Ave. Eugenio Garza Sada 2501, 64849 Monterrey, Nuevo Leon, Mexico

[2] Research Labs, Eugenia Virtual Humans S.A. de C.V., Garza Rios 51, 53100 Naucalpan de Juarez, Estado de Mexico, Mexico

with digital humans. This work complements various previous exploration works by categorizing and exploring the different approaches and technologies involved from the view of a general and complete pipeline, where all the components work together to deliver a realistic and emotional digital human. A general framework for the facial performance of a digital human starts by taking a real human person, analyzing their facial performance, extracting the uniqueness of their facial movements inside a discrete model, and replicating it into a 3D model or avatar. A digital 3D model with a high level of detail, which can have realistic textures and shapes with the use of photogrammetry [44], is able to collect all the detailed animation inside a particular structure named facial rig [21]. This rig will be in charge of driving all the collected information inside a rendering engine [10]. When the digital human is visualized, a particular interaction or emotional link is created between a real person and the virtual human to complete this entire process. This interaction, usually called affective computing, is necessary to get an empathic response [74], and is a crucial component that differentiates a digital human from an avatar, a character, or a 3D interactive model.

1.1 Related surveys and reviews

There are various outstanding surveys that cover the digitalization of the human face and its expressions (Table 1). In 2007, a survey written by Ersotelos and Dong [34] introduced a brief history of the computer simulation of human faces and presented a comprehensive exploration of this area through the categorization of techniques to produce 3D human face models and synthesize dynamic facial expressions. The survey talks extensively about facial codifications and refers to some of the early techniques mentioned in this work, however, since it is a little over a decade old, this survey complements it by introducing novel methods and recent technologies.

Later in 2013, Agianpuye and Minoi [2] published a survey that focuses more on the different facial animation approaches used in the literature rather than the 3D

Table 1 Summary of related surveys and reviews

Title	Year	Venue	Summary
Building highly realistic facial modeling and animation: a survey	2007	Springer Visual Computing	3D human face models Dynamic facial expressions Categorizes the approaches Summarizes important aspects Discusses the current limitations Explores the trend of future research
3D Facial expression synthesis: a survey	2013	IEEE International Conference on Information Technology in Asia	Facial animation approaches Facial expression synthesis Includes possible applications Lists advantages and disadvantages
Computer facial animation: a review	2013	International Journal of Computer Theory and Engineering	Geometric-based modeling Three modeling categories Data-driven animation Three animation categories
State of the art on monocular 3D face reconstruction, tracking, and applications	2018	Computer Graphics Forum	3D face reconstruction 2D data tracking Single RGB or RGB-D camera Optimization-based reconstruction Deformable 3D face models
3D Morphable face models-past, present, and future	2020	ACM Transactions on Graphics	Facial reconstruction Expression reproduction Focus on 3D Morphable Face Models Major contributions in last 2 decades Mentions challenges and future work
A survey of facial capture for virtual reality	2021	IEEE Access	Facial capture for Virtual Reality headsets Overview of various types of technologies Identifies research gaps Includes a realism index for analysis
Facial modelling and animation: an overview of the state-of-the art	2022	Iraqi Journal for Electrical and Electronic Engineering	Techniques for realistic facial animation Facial modeling Animation approaches Brief comparison of methods

reconstruction of the face, but it includes a wide selection of facial expression synthesis methods, including other statistical and learning-based methods. In the same year, Ping et al. [90] presented a survey that reviews geometric-based modeling for face representations and data-driven animation techniques for facial animation. These surveys talk about some of the facial animation approaches mentioned in this work, as well as the same facial codifications used for that task, however, this survey complements both works by mentioning the various available solutions that offer a complete reconstruction and animation framework with the use of one or several of the discussed animation approaches.

Then, in 2018, Zollhöfer et al. [127] presented a survey that focused on 3D face reconstruction and tracking from monocular 2D data obtained through a singular RGB or RGB-D camera. It offers an in-depth overview of different optimization-based reconstruction methods. The present survey complements this work by including not only monocular data but also mentioning technologies that use multiple camera setups and infrared cameras.

After that, in 2020, Egger et al. [31] published a detailed survey on facial reconstruction and expression reproduction using 3D Morphable Face Models. This research work focuses on such models and all the involved methods, which are just briefly mentioned in this survey but complements the facial reconstruction subject by including various other methods for facial reconstruction that are used in the industry and literature.

Next, in 2021, Wen et al. [120] created a survey that focuses on facial capture technologies and approaches for Virtual Reality (VR) headsets, introducing a realism index to evaluate and compare the explored literature. It reviews various facial capture methods also mentioned in this paper, besides, it is to the best of our knowledge the only other survey paper that mentions the recently available Metahuman Creator, which is discussed in Sect. 4.3. It is a recommended read for people interested in VR-specific solutions, however, this survey complements its information by including methods for different applications, as well as mentioning the solving, delivery, and emphatic evaluation of the facial capture results.

Finally, Shakir and Al-Azza [107] published a survey in 2022 that summarizes the most common techniques used in the industry for realistic facial animation, offering a brief comparison of the different approaches. It mainly focuses on explaining the process involved for each of the selected techniques and mentioning some methods that apply them, so this survey complements this work by also including facial reconstruction techniques and introducing complete frameworks that apply some of the methods explored.

1.2 Literature collection and analysis

The methodology used for the present survey starts by delimiting the specific subject area to focus on, which in this case was defined as the facial capture process from a development point of view. Then, the method for defining the eligibility criteria, information sources, search strategy, selection process, and synthesis methods are based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 [81]. The eligibility criteria process specifies the inclusion and exclusion rules for the review and how papers are grouped for the syntheses. The rules for the review protocol establish that the research works needed to be written in the English language, must be peer-reviewed, and the year of the publication had to be within a range of up to two decades; however, this year range does not apply to literature related to the background of relevant topics. The range of two decades was implemented so as to avoid a narrow time frame that could limit the number of eligible studies or a too-wide time frame that could hinder the synthesis of the information. The research works reviewed were grouped according to the steps of a general digital human facial development pipeline, which involve data collection, facial codification, asset creation, facial tracking, solving, delivery, and evaluation.

For information sources, the most adequate databases were defined to be ACM Digital Library, IEEE Xplore Digital Library, and Springer Link. Google Scholar was also used to find other relevant articles and websites. The sources were first accessed on 4 August 2022, and lastly revised on 19 February 2023. For the search strategy, different key concepts were identified for the initial search terms, based on the previously mentioned grouping of research works. These terms were then used to explore relevant studies mentioned in up to seven previous surveys. From those research works, candidate search terms were defined by filtering through the titles, abstracts, and keywords. After searching the sources with the acquired search terms the results yielded over 1,951,009 documents. The duplicate records were removed and the established eligibility criteria were applied, resulting in a total of 43,698 documents. Then, the selection process involved the screening of titles and abstracts, with assessments by the four authors, selecting a total of 1427 documents.

The documents that could be retrieved resulted in a knowledge base of around 768 documents. A full-text review was done on the remaining research works by three of the authors, consulting the fourth if necessary to make the final decision. Up to 428 documents were excluded, because the focus was on applications or methods outside the scope of this paper, and up to 211 documents were excluded because the terminology or methods were better explored in another more relevant document. Out of the final 129 references, a total of 34 research papers were obtained from the ACM Digital Library, 33 were obtained from the IEEE Xplore Digital Library, and 10 were obtained from Springer Link. Up to 39 other research works were obtained from Google Scholar. Additionally, 13 other relevant websites were explored. Finally, the knowledge collected from the selected documents was synthesized into the sections of this review according to their grouping. Relevant data, previous surveys, data sets, and similar technologies, were prepared to be presented in tables and figures for a suitable presentation.

1.3 Organization

In order to understand how a general production framework for digital human facial reconstruction and expression synthesis integrates different key processes, we need to split the frameworks into different areas. The processes go from the physical aspect of the digital human face, the expression set of how this digital human performs and talks, and the real-time delivery of data and 3D graphics, using cutting-edge techniques to solve a real human face into a digital face in real-time. Therefore, the structure of the survey is based on seven different steps, which can be visualized in Fig. 1, that represent a general production pipeline for digital human facial performance: (1) data collection, described in Sect. 2, including the different types of video input and the acquisition protocol, (2) facial codification, introducing in Sect. 3, the two principal methodologies used in the industry, along with the available databases, (3) digital human asset creation, describing essential terms and processes in Sect. 4, from photogrammetry, rigging controls, to recent frameworks, (4) facial tracking, following the popular methods and recent strides in the state-of-the-art, summarizing the different approaches in Sect. 5, (5) solving, a process described in Sect. 6.1, which interprets the information from the facial components to be used in a rendering engine, (6) delivery methods for digital humans, briefly introduced in Sects. 6.2 and 6.3, finalizing with (7) quantitative evaluation methods, visual perception tests and empathetic response evaluations to the final render of a digital-human, discussed in Sect. 7.

The remainder of the paper includes information regarding the current state of digital human facial performance and the path it may take in the future and is structured as follows. Section 8 introduces functional frameworks for Digital Humans. Section 9 discusses the possible future directions for Digital Humans' facial tracking. Finally, Sect. 10 offers a summarized conclusion on the overall presented material.

2 Input data collection

Input data is the base of how we want digital humans to interact in a virtual world. There are three main types of input data for a digital human: video input, audio input, and text input. These can be considered as different research branches, as they propose distinct approaches for animating the digital human. Some high-performance methods are provided by large companies, while others are available through research publications. This section briefly describes some solutions involving video input, the other two branches are out of the scope of this survey.

2.1 Video input

The use of video devices to drive face animation through video sequences is a technique used since the beginning of facial animation research and it is the most widely used approach in current solutions. Some devices used in the literature are webcams, digital cameras, and even smartphones. The most common devices used in the audiovisual industry have 1920 × 1080 image resolution with 30–60 frames per second (fps).

With the recent introduction of ARKit into the facial tracking scheme, the use of mobile iOS devices equipped with a True Depth Camera has increased. These devices can usually record videos with a resolution of 1920 × 1080 at 30, 60, 120, or 240 frames per second. The True Depth Camera has an infrared emitter capable of projecting over 30,000 invisible dots to create a face mesh and an infrared image representation of the face [77].

Infrared or hyperspectral imaging cameras have also been increasingly used for facial expression recognition, since the

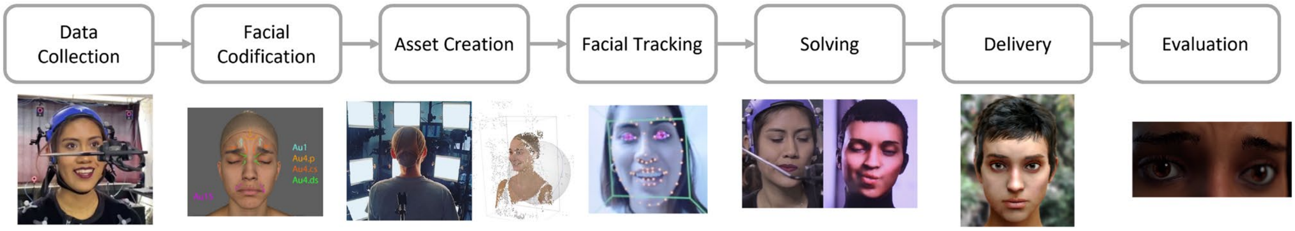


Fig. 1 The set of steps that conform a general pipeline for digital human facial performance

capture performance may degrade if there is no control over the illumination conditions or for subjects of various skin colors [108]. Some commonly used brands for this type of camera are Ximea [46, 104], with xiQ camera models that can have a resolution of 1080 pixels at 170 fps, and Teledyne FLIR [61, 124], with camera models that have a resolution of 640 pixels at 50 fps with a spectral range of 7.5–14.0 µm.

These specific devices with high-resolution cameras are mounted in helmets right in front of the face of an actor performer, in a type of rig known as Head Mounted Cameras (HMC). The development of such rigs aims to improve the quality of frame processing, avoid sudden movements due to camera shaking, and allow for easier use in a professional setting.

3 Facial codification

Different schemes exist to synthesize facial expressions for their replication through computer graphics. There are two main codification approaches, commonly used in the facial animation field. These codifications are the Face Animation Parameters (FAPs) [82] and the Facial Action Coding System (FACS) [117]. FAPs were originally designed by the Motion Picture Experts Group (MPEG) in 1996 as an effort to standardize facial animation during its fast growth in the animation industry. FACS is the product of a theory born in 1978 to observe, study, and analyze how facial expressions can describe emotions and intentions in the field of psychology [97]. Despite their difference in origin and antiquity, both are still used to replicate avatar faces in almost any professional field regarding human facial representation [75].

3.1 FAPs

Facial animation could be viewed as two separate problems, the low level that involves a parameterized facial motion implementation and the high level that involves creating streams of parameters to produce an animation sequence. This parameterization can mean different things for researchers, computer vision specialists, or artists, creating a variety of demands specific to each field. To satisfy all those demands, the Moving Picture Expert Group, also referred to as MPEG-4, focused on a set of requirements for an ideal parameterization that included [83]:

- Range of Possible facial expressions
- Ease of use
- Subtlety
- Orthogonality
- Ability to be the basis for higher level abstraction
- Predictability
- Portability
- Possibility of measuring the parameters
- Efficiency (bandwidth)

Guided by this and inspired by the Abstract Muscle Actions (AMA) [70], MPEG-4 created a model-based approach conformed by a basic data set of 68 facial animation parameters, abbreviated as FAPs. Among these parameters, two are high-level involving visual phonemes and expressions, and the others are low-level involving the movements of the facial features such as the ears, eyes, nose, cheek, lips, jaw, etc. Each FAP represents a one-dimensional measurement where a positive value represents downward motion [111]. Figure 2 shows a visualization of the feature points that conform to the FAPs.

MPEG-4 decided not to standardize a 3D geometric facial model, with the supposition that FAPs can produce good animation results with any reasonable model; however, face models could also be configured using Facial Definition Parameters, abbreviated as FDPs. These parameters allow the definition of a precise facial shape, skin texture, and animation rule if needed. This decision allowed MPEG-4 to provide a flexible solution without interoperability problems [1], becoming, over several decades, the most accepted standard for facial control of 3D avatars and digital characters. In a facial capture process with this kind of codification, the FDPs are used to initialize a geometric model of the face and the FAPs are transmitted to deform that facial model according to each of their measurements [111].

3.2 FACS

In 1872, Charles Darwin described prototypical forms for the display of six categories of emotions, however, due to the lack of systematic data collection and the anthropomorphism

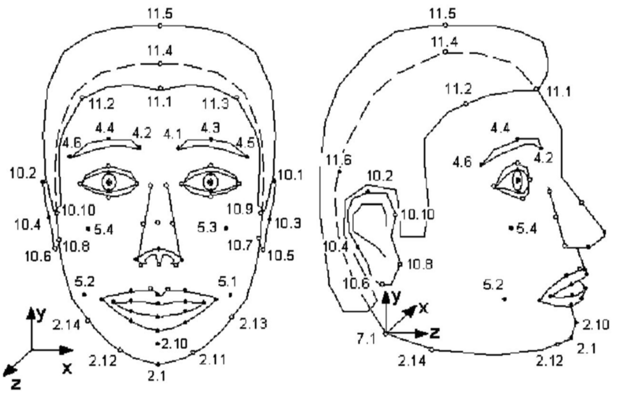


Fig. 2 The set of MPEG-4 face action parameter (FAP) key facial features points [82]

in his descriptions of nonhuman animals, his work on emotion expression did not have a significant impact until the 1960 s and 1970 s, when facial expression became relevant in psychological research. During this time Silvan Tomkins presented a theory of affect that positioned a central role to the face as a site of emotion. This led to his work with McCarter in 1964, which resulted in one of the largest judgment studies that used posed expressions based largely on Darwin's prototypes. Later, the cross-cultural work on the recognition of facial expressions of emotion done by Ekman and Izard further suggested that direct measurement of facial behavior was a fruitful approach to studying emotion [97].

In 1972, Ekman and Friesen performed an experiment to examine whether spontaneous expressions of emotion varied by culture and social context. An early observational coding system, known as the Facial Affect Scoring Technique (FAST), described the facial expressions based on the observed universality and operation of display rules on facial behavior. This approach worked with visual matches to predict configurations for particular emotions, based on the work of Darwin and Tomkins, making it a selective rather than comprehensive measurement tool with some limitations. To improve on this, Ekman and Friesen developed the Facial Action Coding System (FACS), inspired and motivated by the early work of Hjortsö in 1970, who suggested a taxonomy for facial movements in terms of elemental parts and facial muscle groups [97].

FACS describes all visually distinguishable facial activity using 44 unique action units (AUs), in addition to several categories of head and eye positions and movements. Each AU has a label with an arbitrary numeric code and a score based on a five-point intensity scale (A, B, C, D, E), for the timing of facial actions and the coding of facial expressions in terms of events. These events are the AU-based description of each facial configuration, which may consist of one or more AUs contracted as a single display [97]. An example can be observed in Fig. 3. The most common way AUs are incorporated into 3D facial animation is through blendshapes, further discussed in Sect. 4.6.2. Blendshapes can contain sets of expression geometries as defined by AUs, that are then used to interpolate with the facial mesh, using the intensity of the AUs as the alpha of the corresponding expression geometry.

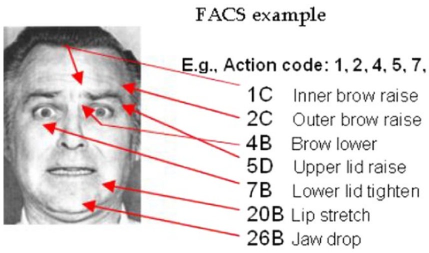


Fig. 3 An example of an expression coded by the facial action code system [13]

There exists an increased use of FACS in various fields, such as scientific research, animation, and computer science, to explore facial expressions and emotion recognition. During the last decades, researchers have also used FACS to validate digital human facial models with the aim of improving the quality of conversational agents and standardizing the rules of facial expression in their display [86].

3.3 Acquisition protocol

An acquisition protocol for facial codifications is necessary to obtain standardized results that are comparable across different studies involving expression analysis. In this sense, MPEG-4 specifies that a particular facial action sequence is generated by deforming the face model in its neutral state according to the specified FAP values for the corresponding time instant. The FAP values translate to face animation parameter units (FAPU), which represent the fractions of distances between key facial features, some visualized in Fig. 2. These distances such as eye separation, eye-nose separation, mouth nose separation, and mouth width, are defined for the face in its neutral state [80].

On the other hand, the process of applying FACS to facial behavior is officially performed by trained experts who make perceptual judgments on video sequences. To become a FACS-certified expert that makes accurate judgments, a person needs to go through approximately 100 h of training and pass a standardized test for reliability. Both these approaches are time-consuming, however, within the past decade, significant advances in computer vision have opened up the possibility of automatic coding of facial expressions at the level of detail required for behavioral studies [97].

3.4 Databases

To design and evaluate effective solutions regarding facial expression analysis, a clear overview of existing data sets is of great importance. Especially, so when exploring recent strides in automation, where the availability and quality of the data are critical. Table 2 shows a brief description of popular databases used in the field of facial expression analysis. To the best of our knowledge, there are no databases that include FAP values for facial codification. Common databases for facial expression recognition use FACS instead; among them, the most explored are MMI[84], DISFA+[71], and CK+[67]. Novel databases such as AM-FED [73] use online video content, with a large dataset of 1.8 million recordings of YouTubers in front of the camera. The work done by Benitez-Quiroz et al. [35] with Emotionet, also takes advantage of the access to web services and the cloud to annotate, classify, and code actor units from facial

Table 2 Public databases that use facial codification

Database	Database information	Affect modeling
CK+ [67]	Frontal and 30 degree images 123 Subjects 10,708 frames Controlled poses/spontaneous Includes AU intensity codes Includes 68 facial landmark points Annotated by expert FACS coders	30 AU 7 Emotion categories
MMI [84]	Frontal/Side photos 25 Subjects 79 series of face expressions Controlled poses/spontaneous Annotated by an expert FACS coder	31 AUs 6 Basic expression
DISFA+ [71]	Video with stereo cameras 27 Subjects 130,828 frames Posed/spontaneous Includes AU intensity codes Includes 66 facial landmark points Annotated by an expert FACS coder	12 AUs
FEAFA+ [41]	Video with stereo cameras 122 Subjects 230,184 frames Posed/spontaneous Includes AU intensity codes Validated by an expert FACS coder	24 AUs
AM-FED [73]	18 M Facial videos 242 Subjects 168,359 frames Spontaneous Annotated by 3 certified FACS experts and 16 trained coders Includes 22 facial landmark points	14 AUs
EmotioNet [35]	Images queried from the web 24,600 images annotated manually 950,000 images annotated automatically In the Wild images Includes AU intensity codes	12 AUs for automated labels 23 AUs for manual labels
GEMEP-FERA [116]	7000 audiovisual emotion portrayals 10 Subjects Additional use of phoneme sequences	12 AUs 18 Emotion categories
D3DFACS [26]	3DMD dynamic 3D stereo camera sequences 10 Subjects 519 Sequences Annotated by a certified FACS expert	44 AUs 20 Action Descriptors (ADs)
UNBC-McMaster Pain archive [68]	Video with two digital cameras 25 Subjects 48,398 frames Spontaneous Includes AU intensity codes Includes 66 facial landmark points Annotated by expert FACS coders	10 AUs
BP4D+ [124]	3D Facial expressions 140 Subjects Spontaneous Includes AU intensity codes Includes head pose and 28 facial landmarks Includes thermal and physiological data Annotated by expert FACS coders	34 AUs

Table 2 (continued)

Database	Database information	Affect modeling
The Bosphorus Database [102]	3D Facial expressions 105 subjects 4652 face scans Includes AU intensity codes Includes 24 facial landmark points Validated by expert FACS coders	28 AUs 6 Emotion categories

expressions obtained from wild origins using search engines. Another approach has been the use of 3D acquisition. Such is the case of D3DFACS [26], BP4D [124], and Bosphorus Database [102]. Furthermore, 3D data can also be extracted from single-view facial databases, for example, through the estimation of 3D facial feature points or the approximation of the face with a 3D Morphable Model [23, 101].

4 Digital human asset creation

The most important step in defining the physical aspect of a digital human face is asset creation. There are different approaches to creating a digital face, as well as different methodologies to control face deformations so that it can create coherent facial expressions. The following subsections explore some of the most prevalent approaches and methodologies used in the field. Sections 4.1 and 4.2 describe techniques used to digitally represent the physical aspect of a real human to use as a digital human face. Sections 4.3, 4.4, and 4.5 present different tools and engines that aid in the creation of digital human assets. Finally, Sect. 4.6 summarizes different methodologies used to define how human assets can be manipulated and deformed during animation and motion-tracking performances.

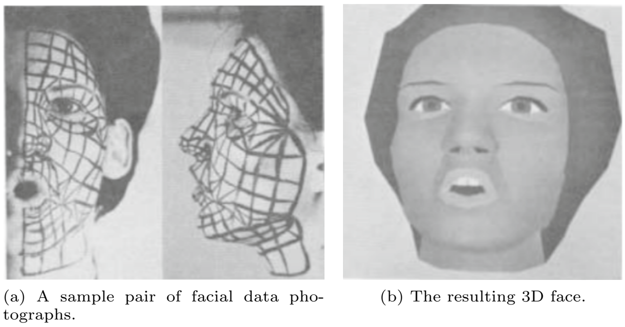
4.1 Photogrammetry for digital humans

Photogrammetry is the science of obtaining reliable information about the properties of objects and surfaces without physical contact with the objects, along with measuring and interpreting this information [103]. Photogrammetry dates back to 1839, with the invention of photography by Daguerre and Niépce, and has evolved with the introduction of new technology. Analog photogrammetry was born due to the invention of stereophotogrammetry in 1901. Analytical photogrammetry came about with the emergence of computers. The recent advent of digital photogrammetry is thanks to storage devices with rapid access to digital images and microprocessor chips.

Nowadays, creative practitioners can generate accurate 3D models through the use of photographic equipment for any digital application. One of the many applications of photogrammetry is the capture of photorealistic human digital doubles. Different perspectives on the evolution of digital face cloning come from the film industry and academia [88].

The film industry has been motivated to create digital doubles of actors to be used for difficult stunts or for the representation of people that are no longer living. One of the first Computer Graphics (CG) humans appeared in the movie Futureworld in 1976. Digital doubles began to become more common during the early 2000 s with state-of-the-art digital stuntpeople in movies, applying different techniques for lightning capture, subsurface scattering, and dense motion capture. In Spiderman 2 [100], four film cameras were placed at various angles around the main characters (Tobey Maguire and Alfred Molina) and synchronized to the strobes for simultaneous image capture. The images were color corrected and projected onto a 3D model of each subject. Finally, colorspace analysis captured the specular and diffuse components. In The Matrix Reloaded [18], an array of five synchronized cameras captured the actor's performance in ambient lighting. The optical flow aided in tracking the motion of each pixel over time in each camera view. Each camera has a vertex of the model projected into it to track its motion in 2D and to estimate the 3D position at each frame using triangulation.

The Academia has also been working on facial photogrammetry research since the 70 s. One of the first documented works is the research done by Parke in 1972 [85], which describes the development of a pioneering system made by two orthogonal photographs and patterns painted


(a) A sample pair of facial data photographs.

(b) The resulting 3D face.

Fig. 4 Parke et al. pioneering method for computer-generated faces [85]

Fig. 5 A model scanned by photogrammetry for the Digital Emily project [4]

on the face of a performer to recover 3D facial geometry (Fig. 4). Later on, Pighin et al. [89], extended this method to an arbitrary number of photographs and texture extraction. Researchers explored the use of facial scans to automatically model structured face meshes augmented with a physically based model of skin and muscles [48, 62, 63, 112]. The process evolved into a technique that allows researchers to create 3D models from multiple photographs taken from different angles and synchronized to capture specific expressions [69]. Most of the systems have up to 200 high-speed and high-resolution cameras using photogrammetry reconstruction software [98] that can read perspective variations and create three-dimensional point clouds.

There is research work with the aim to improve how to create digital humans using inverse engineering based on real humans into digital humans, like the process carried out in Alexander et al. [4] (Fig. 5). Different components of images are extracted from a real human face and processed to acquire a 3D mesh with full surface description maps such as skin reflectance, polygon normals, diffuse base colors, etc. [40]. These components aid in the construction of photorealistic skin with various algorithms and techniques such as raytracing, sub-surface Scattering [122], etc.

The photogrammetry process consists of creating a 3D model from photographs of the same subject doing different expressions (AUs). By taking pictures with constant light and specific landmarks, to obtain the actual size and orientation of the subject, it is possible to measure the distance between every pixel in the image to recreate it in a point cloud voxel model and process that data into a 3D scanned model.

After processing the 3D scanned model, it is normal to have noise all over the mesh. It is the task of a 3D Artist to clean and fill noise from the mesh to keep it as clean and decimated as possible. Photo scanned textures can be projected to the mesh in order to create a texture, then, a scanned texture cleaning task can be added to the pipeline to ensure a quality texture to the 3D mesh.

The last step is wrapping, where algorithms use the 3D topology from the scanned mesh to fit a base geometry to the closest surface points for the 3D scanned model. By having a neutral pose topology of the subject, it is possible to repeat the task for the remaining scanned expressions in order to share the same polycount, vertices, UV maps, etc., so the blendshapes can be compatible with a rigging methodology.

Cross-polarization, a technique used to improve scans, becomes relevant due to reflectance scanning techniques that use a group of photographs under different lighting conditions (or positions). Since skin has a reflectivity component, specular highlights are present in images to be processed in several areas based on the surface normal [43]. The cross-polarization process depends on the linear polarization of the emitting light (flashes or LED lights) and the respective polarization filter in each camera of the system. The images obtained with this technique are accurate in the colors of the skin and the removal of undesired artifacts in the 3D mesh processing. Further methods can improve the quality of how photogrammetry scans can get skin appearance faster and more realistically using a single shot [96].

4.2 3D morphable models and generative adversarial networks 3D face reconstruction

Face reconstruction is the estimation from single images of the facial shape, texture, and other intrinsic components, such as albedo or normals. Blanz and Vetter [16], proposed a statistical model that fits to an image of the face to estimate its 3D shape and texture. This statistical model became known in the literature as the 3D Morphable Model (3DMM), visualized in Fig. 6. This model and its variants

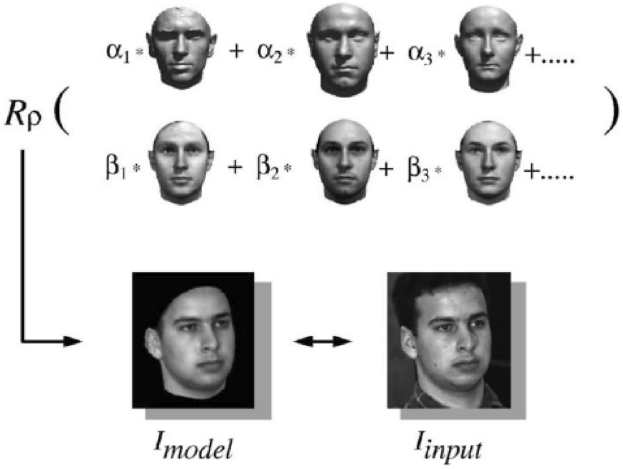


Fig. 6 A visualization of the 3D Morphable Model (3DMM) where the goal of the fitting process is to find the shape and texture coefficients (α_i and β_i) required for the rendering (R_ρ) to produce an estimation (I_{model}) that is similar to the input image (I_{input}) [16]

have been the method of choice to address the problem of face reconstruction for a long time, a more in depth exploration of this method and its application for expression reproduction can be found in the survey article by Egger et al. [31]. However, in recent years, following the advent of Deep Learning, novel solutions that use Generative Adversarial Networks (GANs) have become available and greatly improve the fidelity of facial reconstructions. Some recent examples are Fast-GANFIT [42] and AvatarMe++ [60].

4.3 Metahumans creator

In March of 2021, Epic Games released the initial version of the MetaHuman Creator [55], a tool that aids in the creation of digital human assets with photorealistic results. This tool represents a huge leap in the democratization of digital-human tools, as it reduces the creation workflow to an easy-to-use online creation tool. The detail of textures, quality of models, and realistic face deformations in the MetaHumans platform are the result of the combined knowledge generated by Epic Game's recent acquisition of 3Lateral and Cubic Motion, becoming a new option in the horizon of tools for digital humans. The initial idea that MetaHumans pose is to democratize access to high-quality, customizable digital humans inside the ecosystem of Unreal Engine, which is easy to use, intuitive, and free.

4.4 Digital humans in unity

During the Game Developers Conference (GDC) in March of 2022, Unity featured a demo, titled Enemies, to showcase the power and capabilities of their platform to create digital humans with visual quality and realism. Their work expands on the previous systems created for another demo, titled The Heretic, which featured Unity's first realistic digital human. The system included facial animation systems for sequences of meshes captured over time, skin attachment systems, and shaders for skin, eyes, teeth, and hair. The recent improvements included a new skin shader, more realistic eyes with caustics on the iris, a GPU skin attachment system for high-density meshes, a hair solution for authoring, importing, simulating, and rendering strand-based hair, and tension technology for blood flow simulation and wrinkle maps [114]. These systems can bring more realism to characters, giving full control of the customization options; however, the creation of a digital human in Unity still requires a lot of previous work and resources to generate the initial assets.

4.5 MakeHuman project

MakeHuman [113] is an open-source tool developed to simplify the creation of virtual humans, through the manipulation of controls that blend the 3D model attributes. These attributes are categorized into two groups, the macro-targets that deal with characteristics such as gender, age, height, weight, and ethnicity, and the detail targets that focus on the low-level details such as eye shape, finger length, etc. The MakeHuman project seeks to include other tools in the future that control poses, animation cycles, facial expressions, hair, and clothes. However, the current system only allows the creation of simple 3D human characters.

4.6 Rigging control

To control the content acquired, there are custom tools that manipulate the digital human's articulated parts. Construction of a full facial rig would include bones, blendshapes, and detailed correction adjustments to archive a flexible facial model. These structures work together through a parallel process that performs tasks such as smoothing the data, handling multipliers for the expressions, providing manual control of facial movements, and delivering information into the final facial solving tool. Facial rigging methodologies may change from studio to studio, however, the most common approaches are described below.

4.6.1 Bones

The bones methodology improves the movement of the skeleton in the characters along with the mesh (or muscles). Joints or bones follow a hierarchical structure from the head to the fingers and toes. These structures are relevant for facial animation to move the jaw, eyes, checks, or skin. The facial animation of a digital human usually uses such methodologies for motion capture input, more than with keyframe animation, since it is an easy and low-cost way to move real-time bodies [78].

4.6.2 Blendshapes

The method involving blendshapes has changed and evolved over the years, applying different approaches for their creation, being driven, and replicated inside the digital human process. Blendshapes are 3D meshes created to help drive complex movements of parent meshes. Traditional methods for facial rigging include about 40–50 blendshapes (eyebrows, mouth positions, phonemes, eyelids, etc). The constant change in the blendshapes technology has come along with the evolution of video games, which seek to improve the quality of characters in real-time. The number of blendshapes has increased in recent years and has started to be related to the number of expressions based on facial codifications, with projects having about 200 total blendshapes. With such a large number of blendshapes, the process of controlling this information has left space for contemporary technologies to improve with deep learning [5, 91].

4.6.3 Hybrid approximation using advanced methods

Real production environments need complex solutions for realistic and better facial deformations. Therefore, professionals use hybrid solutions with a mixture of either geometric deformations or blendshapes [58]. Hybrid solutions are more than just locators, blendshapes, and bones, some new methods are developed following FACS rigging.

Skincluster, released in 2016, was created as a specific tool inside Autodesk Maya [25]. This tool offers improved binding between joints and geometry, where every vertex in the 3D mesh needs to be affected by a specific number of joints, making it possible to create more realistic character deformations.

Skin displacement and wrinkle maps are specific textures that can be generated from photogrammetry scans with the main objective of magnifying and reinforcing the wrinkles coming from facial expressions [109]. Deformation is not reflected in the original mesh of the face but is done by the shading material description thanks to specific rendering techniques developed in rendering engines. Color map blending is a recent technique used for Digital Humans to improve the realism of a 3D rigged face. The colors of the skin, particularly the diffuse map, can help create the perception of blood flow variation and circulation due to skin stretching or compressing.

Similarly, beyond the rigging components that help drive the face, there are control processes that manage the intensity, parameters, and predictions of the face data coming from facial tracking. This information is usually smoothed or equalized so that the expressions portrayed in the digital human are more natural [110].

5 Facial tracking

Facial motion capture or facial tracking has a key role in emotion and expressive acquisition in 3D characters, which is born from the need for representation of the performance of actors to drive the roles of the characters [17]. To improve the quality, realism, and to drive emotional characters, such as Thanos [32] in the Avengers Endgame movie, the film industry invests millions of dollars into specialized hardware and software [50], keeping democratic options very far and unreliable.

To complete this task, high-performance facial expression recognition needs to take information from several face regions. The most common method to do this is by obtaining information directly from marks, which are placed on specific areas of the skin of the face [75] and are called markers. Other methods take information directly from the contours of the lips, eyes, and eyebrows using computer vision recognition [95]. All the visual information is recorded directly using camera rigs, like the Head-Mounted Cameras (HMC), and the most advanced methods include the usage of Machine Learning tools to recognize expressions, forecast facial movements, and drive accurate data into virtual characters [15].

Facial tracking to get an actor's performance into 3D animation has been researched for over three decades, with some of the earliest documented works being the research by Lance [121] and the research by Valente and Dugelay in 2000 [115]. These initial approaches used basis functions on the digital face and basic linear interpretation of distances between points of reflective markers in the performer's eyebrows, lips, etc. The translation movement was tracked and processed for several hours, sometimes manually. Some commercial and professional tools in the industry used this method in the early the 2010 s, such as Vicon [57] and Faceware [87]. Following those techniques, real-time tracking began improving without human interaction, using image-based tracking and computer vision [115]. The most relevant tool during 2016 was Dynamyxz [118], but since 2021 it is no longer available due to the company being sold. Professional applications for digital characters improved the portrayal of realistic facial expressions; however, that application still needed several hours of human interaction to correct interpretation errors without a reliable option to use in real time. Bigger contributions in facial tracking and expression recognition started to appear in the middle of the 2010 decade due to the access to Kinect and RGBD cameras. The possibility of evaluating 3D surfaces and not just 2D images drastically improved how faces can be recognized and tracked in real-time for animation purposes. Several available methodologies used facial codifications as base to their approach, some of which are explored in the following Sect. 5.1 and 5.2.

5.1 FAP-based methods

5.1.1 Candide-3

Candide [99] is a deformable 3D wireframe model that describes a parameterized face with around 113 facial feature points widely used for research. This model had an extension with Candide-2 [119], through the representation of the entire head. The most recent actualization is Candide-3 [3], which corresponds better to the FAP codification.

Some research work that uses the Candide model for facial tracking is the following. The work done by Lefevre and Odobez [64], that uses the face model along a hybrid set of features composed of adaptive and trained features. The work done by Horain et al. [52], where a statistical method dynamically fits the Candide-3 model to the subject's face and the face parameters, encoded as FAPs, are sent to a remote player that animates the face avatar. And finally, the work done by Dornaika and Davoine [29], uses the Candide model and a trained auto-regressive models to track and estimate temporal FAPs as a tool for facial expression analysis and recognition.

5.1.2 Dynamic expression model

Early research in 2014 with RGBD images included the work of Bouaziz et al. [19] with a Dynamic Expression Model (DEM), also used by Cao et al. [22], which helped to push the boundaries of facial tracking with an impressive speed at 28–30 frames per second. A PCA model helped Bouaziz to generate a parametrized deformation of 3D models and during tracking, the generic DEM progressively adapted to the facial features of the user; the process is summarized in Fig. 7. Later in 2016, Cao integrated both 2D and 3D inputs, observing more robustness and efficiency in RGBD-based algorithms. Both researches works demonstrated the advantage of a generic face, without pre-training, compensating the inaccuracy with a 2D displacement projection of 3D facial landmarks.

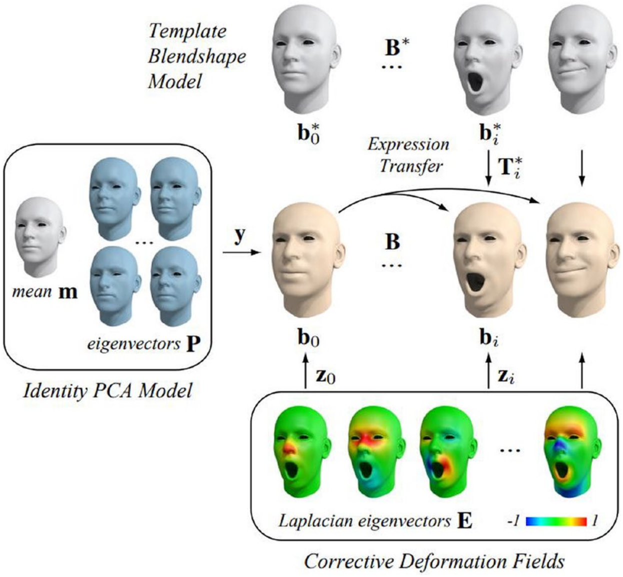


Fig. 7 Visual representation of the dynamic expression model used in the research of facial tracking with RGBD cameras [19]

5.1.3 Deep learning

The most relevant contributions to facial expression recognition in the computer vision field, came along with the popularization of deep learning methods and the computational availability to process multiple data.

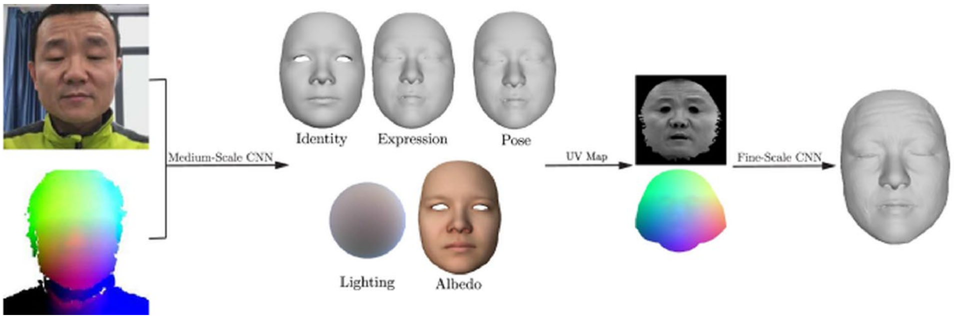
Some research work that uses a Deep Learning based solution is the work of Jia et al. [56], where a trained Artificial Neural Network (ANN) estimates 3D head orientations from five facial features to capture 3D head motions from video input. The orientations are expressed in Euler angles and then translated to FAPs in order to animate a talking avatar.

Another research work that uses Deep Learning as an approach for facial capture is the proposal of Guo et al. [47], to perform real-time RGBD-based 3D face capture. This approach uses a Convolutional Neural Network (CNN) framework composed of two different models that regress a face model and recover the surface details, as visualized in Fig. 8.

5.1.4 ARKit

The landscape of tools and methodologies in the industry lacked reliable democratized solutions, however, that started to change with the launch of Apple's ARKit [77], thanks to the accessibility of Apple smartphones to the general public. This attractive solution can recognize facial expressions and generate animations with a facial parameterization loosely based on FAPs. Several researchers have used this technology as the main tool for facial animation, tracking, and recognition but lacked emphatic results due to its limitations [123], so Apple's ARKit is most reliable when dealing with cartoonish faces.

Fig. 8 A visualization of the CNN-based method for RGBD 3D Face Capture proposed by Guo et al. [47]



5.2 FACS-based methods

5.2.1 DeepExpr

DeepExpr [9], is a deep learning framework that transfers human expressions to stylized characters. The framework is composed of two CNN that recognize the expression of humans and stylized characters independently. Then, using a transfer learning method by Oquab [79], the framework learns the mapping from humans to characters and creates a shared embedding feature space. Finally, with the learned information, human expressions are displayed in the stylized characters using human geometry and a perceptual model mapping.

5.2.2 ExprGen

Following the work of DeepExpr by Aneja et al. [9], ExprGen [8] was developed as a system to automatically generate 3D stylized character expressions from humans. The system uses a multi-stage deep learning approach that uses the latent variables of human and character expression recognition to control a 3D animated character rig.

The ExprGen pipeline is visually summarized in Fig. 9. The process starts with the joint embedding obtained from DeepExpr, followed by a similarity analysis performed by a regression network, 3D-CNN, that maps the human expression onto parameters of a primary 3D character rig. Finally, a lightweight mechanism, Character Multi-Layer Perceptron (C-MLP), transfers the primary character's expressions to secondary characters in a semi-supervised fashion.

This process improved how FACS expressions were delivered to Chartoonish characters without a real-time end-to-end process or a realistic digital human orientation.

5.2.3 FaceLab

FaceLab is a system created for the 2019 film Cats [7] to drive the animation of a 3D facial rig. The system used a facial reconstruction methodology that extracts the shape, pose, and reflectance of the face. After that, standard linear delta blendshapes interpolation represents the expressions, and iterative optimization alters the target weights. Finally, the process culminates with the integration of a robust rendering stage.

5.2.4 Other machine learning approaches

Navarro et al. [76], proposed a deep learning method for real-time facial animation. The architecture takes a video sequence as input and outputs a set of animation controls based on FACS for each frame. The framework is composed of two stages, face detection, and regression. Face detection is done with a fast variant of the Multi-task Cascaded Convolutional Networks (MTCNN) algorithm that localizes and aligns the face. The regression model uses a multitask setup that co-trains landmarks and FACS weights using a shared backbone. The regressed values finally help create synthetic animation sequences.

Another regression model, based on globally-optimized modular boosted ferns (GoMBF), is proposed by Lou et al. [66]. The model first locates the face and 66 facial landmarks, then the facial shape parameters are predicted by fitting a parametric face model to the landmarks. Then, a cascade version of the GoMBF regresses the facial motion. Finally, the facial motion is mapped to expression vectors that update the facial mesh of the 3D model.

6 Digital humans solving and delivery methods

The final solving and rendered display of a digital human is the last step in the workflow. Solving, as further detailed in Sect. 6.1, is where all the information from the facial codification translates to blend shapes and the bones are rotated and positioned [106]. Then, rendering of the 3D mesh with realistic materials and lights can be done for a real-time display or an offline display, as explained in Sects. 6.2 and 6.3.

6.1 Solving

The role of the Solving step is to translate the information from both the face, and body components into a common and easy-to-interpret language for 3D rendering engine to handle, such as facial information based on Action Units that

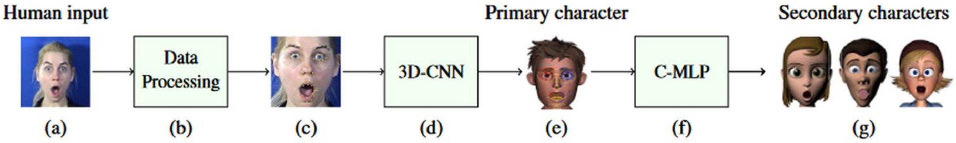


Fig. 9 Overview of ExprGen [8]: **a** 2D images of human facial expressions are pre-processed (**b**, **c**). **d** A CNN generates rig parameters corresponding to the human expression for primary characters (**e**). **f** A separate neural network performs primary-character-to-secondary-character expression transfer (**g**)

need to be delivered as blendshape rig rotations or translations. The final relationship between facial motion capture methodology and the blendshapes or movements is determined by the respective codification (discrete, simple units, combo units, etc) and the way the facial rig is constructed.

The facial animation translator is key to having a simple and customizable interpretation of the facial movements in the facial rig, therefore, several commercial tools, such as Faceware [37], Unreal's Live Link Face [33], and FaceGood [36], offer their tools to make this possible. Table 3 makes a quick comparison of the characteristics of these solutions.

6.2 Real-time delivery

A real-time display can improve the emphatic response since it incorporates the notion of interactivity to the experimental workflow, which is possible through recent real-time tools that archive high-quality results with low latency[10].

The industry has several options for real-time rendering engines, with characteristics favorable for digital humans. The most notable engines are Unreal Engine by Epic Games, Octane by OTOY Inc., Eevee by Blender, and Unity by Unity Tech. Furthermore, Nvidia has pushed forward RTX technologies capable of performing raytracing, a rendering technique for three-dimensional graphics that simulates light interactions with various materials [92]. Raytracing has improved hair simulation, subsurface scattering shaders, and other benefits for the render of digital humans. Performing retargeting before streaming into a rendering engine is also possible through standardized connections, like the ones in Motion Builder by Autodesk, that connect across several real-time technologies from companies such as Vicon, Optitrack, or Xsens [14]. Some options like Shogun Live, and Unreal Engine 5 IK retargeter [54], allow skipping Motion Builder by directly connecting to Unreal Engine. Finally, with the launch of MetaHumans [55] as a democratized tool that allows customization, manipulation, and real-time rendering with motion-capture, the use of Unreal Engine for the delivery of digital humans has become a standard and the main rendering engine.

Table 3 Comparison of the current public solutions for facial motion capture

Facial mocap tool	Characteristics
Faceware studio	Real-time Single subject calibration No machine learning FAP 24,000 USD
Unreal's Live Link Face	Real-time No subject calibration No machine learning FAP Free
FaceGood	Offline Infrared camera Artificial intelligence tracking FAP 469 USD

6.3 Offline delivery

Offline rendering does not allow the same interactivity that a real-time display would, but it offers higher quality graphics that could improve the visual realism of a digital human. As such, offline rendering is mostly used for cinematography, as a replacement for photoshoots or to present products through digital media [93].

In this regard, the previously mentioned engines (Unreal Engine, Octane, Eevee, and Unity) are useful as well, however, other popular rendering engines are better for offline rendering, some of them, to mention a few, are V-Ray by Chaos Group, Arnold by Autodesk, and Redshift by Maxon.

Given the recent technological advances in graphics cards and render engines, which have shown great improvements in performance, the gap between real-time and offline rendering continues to shrink. Because of this, in the last decade, researchers have started using both traditional rendering engines and game engines to explore digital humans with increased quality and decreased rendering cost [30].

7 Evaluation of the empathic response and visual perception

The final step in a digital human research workflow relies on perception tests and experiments designed to construct a solid and representative proof of concept. Digital faces have been able to express emotions by following realistic movements with anatomical precision, or through simplified forms and exaggerated actions. However, when a human drives a hyper-realistic avatar, many other variables need consideration when mimicking emotions. Emotions drive us, so we do not call emotions on command, so their recreation through micro-expressions is not perfect and capturing authentic emotions in a controlled environment is complex.

There are certain quantitative evaluation methods for facial reconstruction and expression reenactment. A standard metric is the geometry fitting error, defined as the point-to-point distance between the reconstructed 3D face and a corresponding groundtruth face geometry [47]. Other benchmarking methods depend on intermediate tasks, for example, facial landmark estimation or emotion recognition.

The evaluation through landmark estimation can measure the difference in distance between landmarks of the input face or expression and the 3D-generated results. The evaluation through emotion recognition tasks can quantify with metrics, such as correlation coefficients, sign agreement, or accuracy, how much of the input emotion is conveyed in the synthesized 3D face [28]. Other metrics that evaluate image quality and structural similarities can be used for evaluation, but the output needs to be rendered into an image with the same alignment and lightning as the input.

These objective measures can help improve the geometry of facial reconstruction and expression reenactment, however, the human perception remains the ultimate benchmark for the facial results of digital humans [28]. Since the alikeness of digital humans to real humans is still a subjective and perceptive topic, the most common validation approaches in this aspect are described in Sects. 7.1 and 7.2.

7.1 Validation of emotion elicitation with image or video stimuli

FACS can be mixed in several combinations and reflected in emphatic emotion combos, named emFACS [39]. The evaluation of such emFACS consists of the exhibition of several multimedia elements, like static or dynamic images of digital humans, to controlled groups of volunteers that relay their perception of the emotions displayed. Some common emotions used in research are neutral, happiness, sadness, surprise, anger, fear, contempt, embarrassment, and pride. A matrix with the resulting perceived and modeled expressions can show the agreements and mismatches in the population.

Stimuli design testing usually employs videos and images to evaluate how humans perceive the emFACS model. Recently, interactive experiments are possible with the aid of real-time rendered digital humans. One such experiment called the “Wizard of Oz” [105], gives a set of test subjects the ability to talk and interact with a digital human in real-time. Such approaches create an interesting way to expose a new dimension to evaluate digital human interactivity. Results in this field showed that interactivity improves the emphatic response and makes the digital human a feasible model to display realistic emotions [6].

7.2 Uncanny valley perception

Another concept related to the validation of a digital human is the need to measure how the human eye perceives the uncanny valley. The uncanny valley is a concept born from Masahiro Mori’s hypothesis in 1970, that a person’s response to a human-like character would abruptly shift from empathy to revulsion as it approached, but failed to attain, a lifelike appearance [74]. A graphical representation of this descent into eeriness is visualized in Fig. 10.

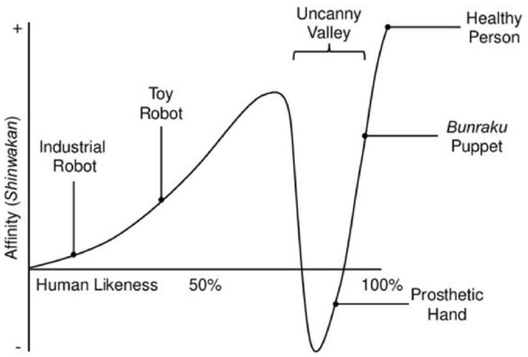


Fig. 10 A graph that depicts the uncanny valley, the proposed relation between the human likeness of an entity, and the perceiver’s affinity for it [74]

The evaluation method for the uncanny valley usually involves a Likert scale of perception and an Analysis of Variance (ANOVA) of metrics like Realism, Appeal, Re-assuring, Familiar, Friendly, Trustworthy, and motion pleasant [72]. Another measuring approach is proposed by Ho and MacDorman [51], whose method involves indices developed and validated for the perceptual-cognitive dimension of humanness, along with three affective dimensions composed of inter-personal warmth, attractiveness, and eeriness. A card sorting task, a laddering interview, an adjective evaluation, and a validation representative survey were applied for subjects to revise the humanness, attractiveness, and eeriness indices. The revised indices enable empirical relations among characters to be plotted similarly to Mori’s graph of the uncanny valley.

The quality or style used to render the digital human has an important role in the uncanny valley perception. If the digital human presents reduced photo-realism, there can be a direct negative impact on the emphatic perception [125]. However, any level of realism can be paired with some degree of interactivity, which could make the difference in the final uncanny valley perception measured.

8 Fully functional frameworks

A fully functional framework for digital humans is the one that covers the whole pipeline of digital human facial tracking, going from asset creation, facial expression tracking, solving, and display of the digital human. The solutions available as open-source or private services are further explored in Sects. 8.1 and 8.2. A comparison of each of these frameworks can be seen on Table 4.

Table 4 A comparison of open-source and private frameworks

Framework	Year	Characteristics
HapFACS	2015	Open-source Research Avatars created with Haptek PeoplePutty Software Manual control of FACS AUs on the character
FACSvatar	2018	Open-source Research Digital Human generated by the MakeHuman Project Facial expressions captured using OpenFace Solving the AU into blendshape values Render on Unity or Blender and FACSHuman Real-time and offline GUI on Jupyter Notebook
AvatarSim	2019	Open-source Research Avatar created within the AirSim environment Bones, FACS and Phonemes values for face motion Unreal Engine Environment Real-time Windows desktop app
EMOCA	2022	Open-source Research Generated textured 3D Mesh Expressions based on emotions Real-time Python Interface
Project Vincent	2019	Private Research Digital human created with photogrammetry Deep Learning based facial capture Render on Unreal Engine Real-time
Renderpeople	2021	Private Commercial Digital Human generated with photogrammetry FACS and phonemes based expressions Render with Unreal Engine Offline
Medusa	–	Private Commercial Delivers high-resolution 3D faces Reconstruct faces in full motion using a mobile rig of cameras Recovers per-frame dynamic appearance, such as blood flow
iClone	–	Private Commercial Large library of characters Facial performance and full-body motion capture Real-time Rendering on game engines
Masquerade	–	Private Research Face model generated based on FACS Performance capture with facial markers and stereo camera Offline
Cubic Motion	–	Private Commercial Custom made digital human FACS based solving Real-time and offline

Table 4 (continued)

Framework	Year	Characteristics
Renderpeople	2021	Private Research Custom made digital human FACS based solving Real-time and offline
Animatomy	2022	Private Research Custom made digital human Custom muscle-based parameterization Real-time and offline

8.1 Open-source solutions

8.1.1 FACSvatar

FACSvatar [110] presents an open-source modular framework that processes and animates FACS-based data in real-time. A modified version of OpenFace 2.0 [12] provides FACS-based input for the framework. Then, a simple Gated Recurrent Unit neural network processes the information and enables generative data-driven facial animation. After that, a converted module turns the action unit data into blendshape values of a virtual face. Finally, a visualization engine animates the resulting action unit, gaze, and head rotation values. The framework supports visualization in renderers such as Blender or Unity and allows the use of models generated by the MakeHuman Project, previously discussed in Sect. 4.5, which has an integrated tool to create facial expressions, namely FACSHuman [45].

8.1.2 AvatarSim

AvatarSim [10] is a framework that performs facial expression and lip syncing over an avatar using a video of human expressions and the phonemes present on a speech audio. There are two pipelines available to drive the expressions of the avatar, as observed in Fig. 11. One pipeline uses FACS recognition, where the video input first passes through face detection, then a Facial Action Unit Recognizer analyzes the facial regions of interest, and finally, the resulting data feeds the Expression Synthesizer to display the expressions on the avatar's face. The other pipeline uses bone position controls, where the video input also passes first through face detection, then a multi-stage Deep Learning system, previously described in Sect. 5.2.2, retargets the facial expressions to a primary 3D avatar and then to the final human avatar. For the lip syncing, the voice is converted into a sequence of phonemes to synchronously play the audio and drive the avatar lips.

8.1.3 EMOCA

EMOCA or Emotion Capture and Animation is a method proposed by Danecek, Black, and Bolkart [28], that reconstructs a 3D face from a single image while conveying the emotional state of the input. The method is built on the DECA [38] framework, used to reconstruct a detailed animatable 3D face model from a single image, and the FLAME [65] 3D statistical head model, which has parameters for identity shape, facial expression, pose, and rotations. Figure 12 shows a detailed visualization of the processes within EMOCA.

8.2 Private solutions

8.2.1 iClone

iClone is a real-time 3D animation software that includes a large library of characters and motion, as well as a variety of tools for full-body motion capture and real-time production [11]. For facial performance, the iClone workflow includes voice lip-sync, puppet emotive expressions, muscle-based face key editing, and iPhone Live Face-based facial capture. The latter introduces a real-time smoother, tracking data multiplier, and live retargeting tools, which allow for jitter-free and balanced facial triggers [53].

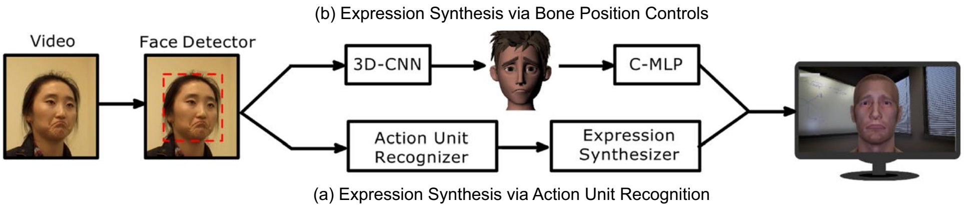

(b) Expression Synthesis via Bone Position Controls
Video
Face Detector
3D-CNN
C-MLP
Action Unit Recognizer
Expression Synthesizer
(a) Expression Synthesis via Action Unit Recognition

Fig. 11 The two pipelines that compose the AvatarSim framework [10] for retargeting an expression from a human to a 3D avatar

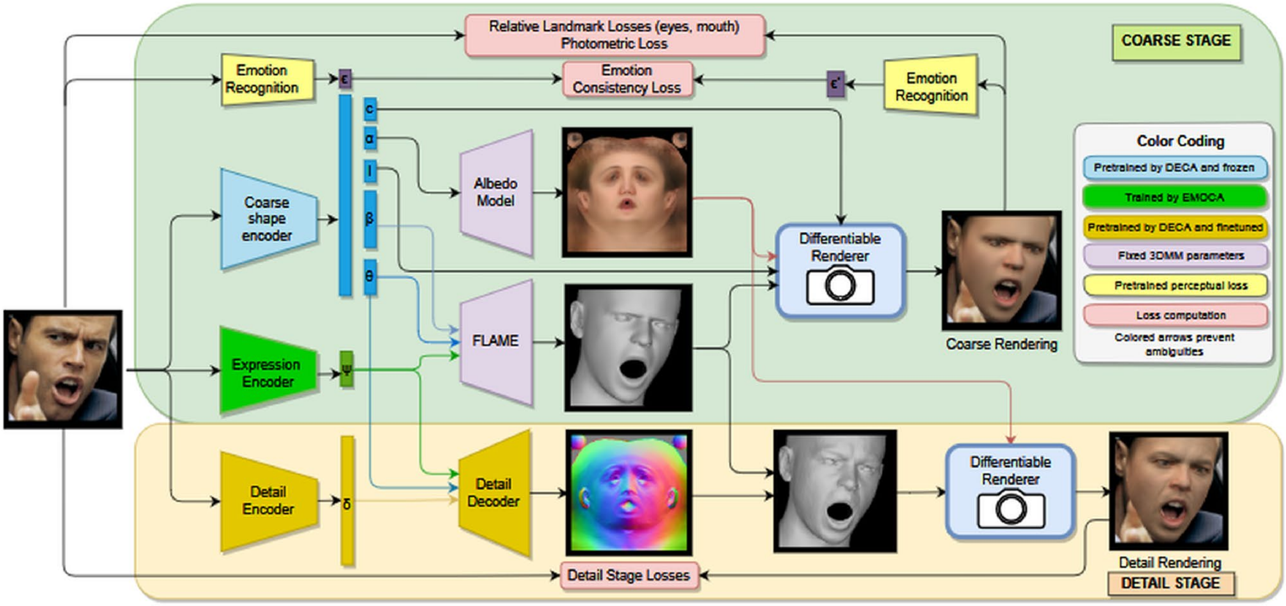


Fig. 12 An overview of the EMOCA framework [28]

8.2.2 Project vincent

GianStep designed project Vincent as an experimental project around a digital human based on a commercial research framework created by a private research group in Korea named GXLab. Sakamoto et al. [44] developed an internal framework that uses photogrammetry to acquire a set of expressions based on the FACS methodology. The capture of the facial expressions is separated into 6 areas of the face, producing up to 35 expressions for each area. The scanned facial expressions are then processed into 210 blend-shapes while maintaining the realism of the subject.

Two deep learning models are responsible for the real-time facial performance. The first one, a facial marker tracker, locates and predicts the coordinate system of 3D markers with marker-less 2D images of facial expressions. The second Network, designed as a Blend-shape mapper, connects with the 6 areas to find an appropriate blend-shape intensity for the AU with the information of the 3D marker coordinates. Final results are delivered in real-time to the 3D model inside Unreal Engine, allowing the achievement of a real-time representation of a digital human, with convincing realism [44].

8.2.3 Disney's Medusa

The Medusa Facial Capture system, developed by Disney Research Studios, consists of a mobile rig of cameras and lights coupled with proprietary software that can reconstruct actors' faces in full motion, without using traditional motion-capture markers [49].

Medusa has the capability of delivering high-resolution 3D faces, with the ability to track individual pores and wrinkles over time. The software can also recover per-frame dynamic appearance, such as the blood flow or the shininess of the skin, providing a very realistic virtual face that is ideal for creating digital humans. Medusa can be used to build an expression shape library or to reconstruct a performance dialog.

8.2.4 Masquerade

Masquerade is a modular and expandable in-house facial capture system built by Digital Domain. It is capable of adding fine-scale details to facial motion capture data from low-resolution capture using head-mounted cameras. High-resolution 4D scans become the training data on how the actor's face moves, and the base from which to extract facial data to create a detailed face model. The Facial Action Coding System serves as the base for a module called Shape Propagation that can generate a 3D version of an actor's head to provide up to 1500 different shapes in case a FACS scanning session, specific to the actor. Then for the facial performance capture, an actor uses 150 facial markers and a vertical stereo helmet camera rig [75].

Masquerade has different modules aimed at identifying, accurately predicting, and tracking all of the markers during the performance. Deformation gradients represent the pose of the face, which extract information regarding the bending and deformation of the face surface, relative to the rest of the pose. This also makes it possible to reuse training data from different marker sets. Local vertex offsets create fine-scale detail encoding, which avoids scaling issues related to large-scale deformations [75].

The output from Masquerade is a moving mesh that accurately matches the imagery of the actor, which improves through corrections during training of any missing details or features that are not coming through sufficiently detailed. The newest 2.0 version includes fully animated eyes to the moving mesh and also solves the mesh onto some form of the facial rig as part of the process. Solving is done through Radial Basis Function interpolation with a biharmonic kernel, applied individually to small segments of the face relative to each marker, and blended using geodesic weights for each vertex. This makes it independent of the output resolution and has a low cost of computation and memory [75].

Masquerade is responsible for the representation of Thanos in Marvel's Avengers: Infinity War, along with another tool from Digital Domain for multistage facial re-targeting, called Direct Drive. This tool builds a deformation stack that employs gradient-based deformation transfer and general-purpose mesh deformers, which constraint points on the mesh, resolve skin-to-bone collisions, and control the rigidity of the face. This process is possible through the mapping of correspondences between key facial features on the deformed actor's mesh and the creature's face, as well as the animated facial performance from Masquerade. This combination of techniques results in realistic facial performances and reduced facial animation time [50].

8.2.5 Cubic motion

Cubic Motion is a company that specializes in producing facial animation for video games and other media. It collaborated with Epic Games, Tencent, 3Lateral, and Vicon, for the creation of Siren, a high-fidelity digital human that can be driven in real-time [27]. Cubic Motion's role in this collaboration was to translate the live performance of an actress into a live, real-time rendered character. This process entailed the capture of the performer, the tracking of facial features, the solving of digital character controls, and the data streaming to the game engine.

A side and forward camera rig that uses Vicon technology captures the performance. However, Cubic Motion's technology and pipeline are compatible with multiple capture types, such as single or stereo head-mounted cameras, 4D data, and depth cameras. A trained model that captures the face in separate segments and digitally marks the required facial elements handles the facial feature tracking. The obtained data then goes through a solving phase into the FACS-based facial rigs constructed via joints, blend shapes, scanned data, or wrinkle maps. Lateral was responsible for the facial rig in project Siren. The result was an impressive high-fidelity, real-time rendered digital human who can interact with audiences and push the boundaries of current digital humans [27].

8.2.6 Digital human by renderpeople

Renderpeople created a hyper-realistic copy of a real human actor in 2021 to showcase their technologies through a digital human called Fred [94]. They use a photogrammetry scanner, composed of over 300 aligned Digital Single Lens Reflex (DSLR) cameras, to create a set of highly detailed 3D meshes with textures that capture even the smallest facial features of the actor. Different poses and phonemes were recorded to digitize realistic motion sequences that cover all the deformations of different facial muscle areas. All of which are necessary to create a hyper-realistic face rig with realistic blendshapes. Additional to the facial data, there are recordings of the body motion through a motion capture suit, called Xsens MVN Link. The final render of the digital human is in Unreal Engine, where all the meshes, textures, rigs, and animations are combined and unified to create the digital human Fred.

8.2.7 Animatomy by Wētā FX

Wētā FX proposed Animatomy [24], an end-to-end modular deformation architecture to approximately represent the deformation of muscle fibers and obtain anatomically plausible animation controls. The Animatomy system enables the automatic optimization of specific face parameters based on dynamic facial scans, animation driven by performance capture, dynamic simulation, and animation transfer. From the 3D reconstruction of an actor's face through photogrammetry, a flesh mask is constructed by tetrahedrons. Then, a muscle-based parameterization is embedded in the mask by inversely simulating a representative set of skeletal face muscles. The facial expressions of the architecture are parameterized by a vector of strains corresponding to 178 muscle fiber curves defined for a human face. This brings about a departure from the FACS-based blendshape systems and offers fine-grained, anatomically plausible animation control and straightforward animation transfer.

9 Future directions

The improvement in the realism of digital humans that has been achieved in recent years by private and open projects has been astounding, however, overcoming the uncanny valley is still a work in progress. Facial motion capture can now achieve higher frame rates, of up to around 480 frames, which makes it possible to add a layer of real-time interactivity that can help overcome the emphatic shortcomings. Additionally, as the processing power and resources get faster and more bountiful, higher resolutions and the use of multiple other technologies is feasible. This also makes realistic physical interaction possible with the real-time composition of a digital human with live-action elements.

Deep Learning solutions can obtain increasingly better results for various fields and diverse problems, such as previously discussed for facial reconstruction, facial expression recognition, and facial expression transfer, which can lead, in the near future, to new groundbreaking facial performances in digital humans. Furthermore, the usage of 4D scans for the asset creation of digital humans seems to be leading future trends in the direction of more realistic anatomic simulations and representations. Overall, private solutions tend to be more robust and accurate, however, with the advent of democratized platforms, such as MetaHumans by Epic Games or ARKit by Apple, the creation of digital humans and their performance aided by facial tracking has become available to a wider group of creators and researchers, which in turn will enrich the current landscape of tools and areas of application for digital humans, having a positive impact on other fields.

10 Conclusions

Given the recent interest in a digital world where humans interact or exchange virtual experiences through the use of embodied avatars, digital humans have become a popular topic of research, bringing about recent strides in diverse technologies aimed at improving their quality, realism, emphatic response, and interactivity. These improvements are explored in this survey by following a conventional pipeline for the creation and testing of a digital human, along with recent and relevant frameworks developed as a complete solution for digital human facial performance.

Input data collection is possible through video, audio, or text input. This survey, however, focused solely on video input for facial animation. RGBD cameras give relevant depth information that aided in the implementation of various facial capture solutions. Infrared or hyperspectral cameras have also positively impacted facial tracking since their characteristics allow for illumination-independent solutions.

Facial expression synthesis is possible by following a facial codification, therefore, this work focuses on two of the most popular codifications, FAPs and FACS. Their characteristics and acquisition protocol are described, along with relevant databases that include FACS labeling, whose characteristics are summarized.

This survey introduces different approaches to human asset creation. With the description of the photogrammetry process, including the exploration of solutions based on Generative Adversarial Networks, and frameworks such as the MetaHumans Creator and the MakeHuman Project. Information regarding the components of a rigging control for the animation of digital humans is also detailed.

Facial motion capture or facial tracking is introduced, presenting frameworks based on the previously described facial codifications. The works explored include past models and recent advancements, showing a general description of the evolution of technologies in the field.

The text offers a brief description of the solving step, while also comparing some of the available tools that fulfill solving tasks. Real-time and offline rendering are two types of delivery methods for digital humans, so this survey describes their advantages and disadvantages, presenting some of the most popular engines.

This survey includes a general overview of the possible evaluation approaches to measuring the empathic response and perception of digital humans, while also introducing relevant concepts, such as the Uncanny Valley.

Fully functional frameworks are explored and classified as open-source or private solutions so that a complete landscape of tools is present. Finally, a brief description of the perceived future directions for the creation and usage of digital humans is explored, based on the review of methods, frameworks, and technology described in the survey.

Author contributions CV, CP-G, and MM-R contributed equally to the main manuscript text and figures. MG-M wrote the Introduction and Conclusion sections of the manuscript. All authors reviewed the manuscript.

Funding This work was supported through a scholarship for Carlos Vilchis by the Mexican National Council of Science and Technology (CONACYT). This work was also supported by the Epic MegaGrants program under Grant FACS DEEP LEARNING TOOL for Eugenia Virtual Humans S.A. de C.V., which has a financial interest in the subject matter or materials discussed in the survey.

Data availability The authors declare that data sharing is not applicable to this survey as no datasets were generated; however, the information regarding the datasets mentioned in this study is available within the article.

Declarations

Conflict of interest This work was supported through a scholarship for Carlos Vilchis by the Mexican National Council of Science and Technology (CONACYT). This work was also supported by the Epic MegaGrants program under Grant FACS DEEP LEARNING TOOL for Eugenia Virtual Humans S.A. de C.V., which has a financial interest in the subject matter or materials discussed in the survey.

References

1. Abrantes, G.A., Pereira, F.: Mpeg-4 facial animation technology: survey, implementation, and results. IEEE Trans Circuits Syst Video Technol **9**(2), 290–305 (1999)
2. Agianpuye, S., Minoi, J.L.: 3d facial expression synthesis: a survey. In: 2013 8th International Conference on Information Technology in Asia (CITA), pp. 1–7 (2013). https://doi.org/10.1109/CITA.2013.6637552
3. Ahlberg, J.: Candide-3: an updated parameterised face (2001)
4. Alexander, O., Rogers, M., Lambeth, W., Chiang, M., Debevec, P.: Creating a photoreal digital actor: the digital emily project. In: 2009 Conference for Visual Media Production, pp. 176–187. IEEE (2009)
5. Alkawaz, M.H., Mohamad, D., Basori, A.H., Saba, T.: Blend shape interpolation and facs for realistic avatar. 3D Res. **6**(1), 6 (2015)
6. Amini, R., Lisetti, C., Ruiz, G.: Hapfacs 3.0: Facs-based facial expression generator for 3d speaking virtual characters. IEEE Trans. Affect. Comput. **6**(4), 348–360 (2015)
7. Andrus, C., Ahn, J., Alessi, M., Dib, A., Gosselin, P., Thébault, C., Chevallier, L., Romeo, M.: Facelab: Scalable facial performance capture for visual effects. In: The Digital Production Symposium. DigiPro ’20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3403736.3403938
8. Aneja, D., Chaudhuri, B., Colburn, A., Faigin, G., Shapiro, L., Mones, B.: Learning to generate 3d stylized character expressions from humans. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 160–169. IEEE (2018)
9. Aneja, D., Colburn, A., Faigin, G., Shapiro, L., Mones, B.: Modeling stylized character expressions via deep learning. In: Lai, S.H., Lepetit, V., Nishino, K., Sato, Y. (eds.) Computer Vision: ACCV 2016, pp. 136–153. Springer International Publishing, Berlin (2017)
10. Aneja, D., McDuff, D., Shah, S.: A high-fidelity open embodied avatar with lip syncing and expression capabilities. In: 2019 International Conference on Multimodal Interaction, pp. 69–73 (2019)
11. Aseeri, S., Marin, S., Landers, R.N., Interrante, V., Rosenberg, E.S.: Embodied realistic avatar system with body motions and facial expressions for communication in virtual reality applications. In: 2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), pp. 580–581. IEEE (2020)
12. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.: Openface 2.0: facial behavior analysis toolkit. In: 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018), pp. 59–66 (2018). https://doi.org/10.1109/FG.2018.00019
13. Bartlett, M., Littlewort, G., Vural, E., Lee, K., Cetin, M., Ercil, A., Movellan, J.: Data mining spontaneous facial behavior with automatic expression coding. In: Esposito, A., Bourbakis, N.G., Avouris, N., Hatzilygeroudis, I. (eds.) Verbal and Nonverbal Features of Human-Human and Human–Machine Interaction, pp. 1–20. Springer, Berlin (2008)
14. Bennett, G., Kruse, J.: Teaching visual storytelling for virtual production pipelines incorporating motion capture and visual effects. In: SIGGRAPH Asia 2015 symposium on education. SA ’15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2818498.2818516
15. Bhat, K.S., Goldenthal, R., Ye, Y., Mallet, R., Koperwas, M.: High fidelity facial animation capture and retargeting with contours. In: Proceedings of the 12th ACM SIGGRAPH/eurographics symposium on computer animation, pp. 7–14 (2013)
16. Blanz, V., Vetter, T.: Face recognition based on fitting a 3d morphable model. IEEE Trans. Pattern Anal. Mach. Intell. **25**(9), 1063–1074 (2003). https://doi.org/10.1109/TPAMI.2003.1227983
17. Blascovich, J.: Social influence within immersive virtual environments. In: Schroeder, R. (ed.) The social life of avatars, pp. 127–145. Springer (2002)
18. Borshukov, G., Piponi, D., Larsen, O., Lewis, J.P., Tempelaar-Lietz, C.: Universal capture-image-based facial animation for “the matrix reloaded”. In: ACM Siggraph 2005 Courses, pp. 16–es (2005)
19. Bouaziz, S., Wang, Y., Pauly, M.: Online modeling for realtime facial animation. ACM Trans. Graph. (2013). https://doi.org/10.1145/2461912.2461976
20. Burke, B.: Hype cycle for emerging technologies, 2021. Tech. rep., Gartner, Inc, Stamford, CT 06902 USA (2021)
21. Cañamero, L., Aylett, R.: Animating Expressive Characters for Social Interaction, vol. 74. John Benjamins Publishing, Amsterdam (2008)
22. Cao, C., Hou, Q., Zhou, K.: Displaced dynamic expression regression for real-time facial tracking and animation. ACM Trans. Graph. (TOG) **33**(4), 1–10 (2014)
23. Chang, F.J., Tuan Tran, A., Hassner, T., Masi, I., Nevatia, R., Medioni, G.: Expnet: Landmark-free, deep, 3d facial expressions. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 122–129 (2018). https://doi.org/10.1109/FG.2018.00027
24. Choi, B., Eom, H., Mouscadet, B., Cullingford, S., Ma, K., Gassel, S., Kim, S., Moffat, A., Maier, M., Revelant, M., Letteri, J., Singh, K.: Animatomy: An animator-centric, anatomically inspired system for 3d facial modeling, animation and transfer. In: SIGGRAPH Asia 2022 Conference Papers. SA ’22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3550469.3555398
25. Community, A.H.: Autodesk help center. https://knowledge.autodesk.com/support/maya/learn-explore/caas/CloudHelp/cloudhelp/2016/ENU/Maya/files/GUID-2E292C8A-388A-4E77-B42D-165F1C9E1E5F-htm.html (2021)
26. Cosker, D., Krumhuber, E., Hilton, A.: A facs valid 3d dynamic action unit database with applications to 3d dynamic morphable facial modeling. In: 2011 International Conference on Computer Vision, pp. 2296–2303 (2011). https://doi.org/10.1109/ICCV.2011.6126510
27. CubicMotion, U.: Siren case study. https://cubicmotion.com/case-studies/siren/ (2018)
28. Daněček, R., Black, M., Bolkart, T.: Emoca: Emotion driven monocular face capture and animation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 20279–20290 (2022). https://doi.org/10.1109/CVPR52688.2022.01967
29. Dornaika, F., Davoine, F.: Facial expression recognition using auto-regressive models. In: 18th International Conference on Pattern Recognition (ICPR’06), vol. 2, pp. 520–523 (2006). https://doi.org/10.1109/ICPR.2006.539
30. Eckschlager, M., Lankes, M., Bernhaupt, R.: Real or unreal? An evaluation setting for emotional characters using unreal technology. In: Proceedings of the 2005 ACM SIGCHI

International Conference on Advances in Computer Entertainment Technology, pp. 375–376. ACE '05, Association for Computing Machinery, New York, NY, USA (2005). https://doi.org/10.1145/1178477.1178556
31. Egger, B., Smith, W.A.P., Tewari, A., Wuhrer, S., Zollhöfer, M., Beeler, T., Bernard, F., Bolkart, T., Kortylewski, A., Romdhani, S., Theobalt, C., Blanz, V., Vetter, T.: 3d morphable face models-past, present, and future. ACM Trans. Graph. (2020). https://doi.org/10.1145/3395208
32. Ennis, C., Hoyet, L., Egges, A., McDonnell, R.: Emotion capture: emotionally expressive characters for games. In: Proceedings of Motion on Games, pp. 53–60 (2013)
33. Epic Games.: Recording facial animation from an ios device. https://docs.unrealengine.com/4.27/en-US/AnimatingObjects/SkeletalMeshAnimation/FacialRecordingiPhone/ (2022)
34. Ersotelos, N., Dong, F.: Building highly realistic facial modeling and animation: a survey. Vis. Comput. **24**(1), 13–30 (2007). https://doi.org/10.1007/s00371-007-0175-y
35. Fabian Benitez-Quiroz, C., Srinivasan, R., Martinez, A.M.: Emotionet: an accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5562–5570 (2016)
36. FACEGOOD.: Facegood website. https://www.facegood.cc/ (2022)
37. Faceware Technologies.: Faceware website. https://facewaretech.com/ (2022)
38. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3d face model from in-the-wild images. ACM Trans. Graph. (2021). https://doi.org/10.1145/3450626.3459936
39. Friesen, W.V., Ekman, P.: Emfacs-7: Emotional facial action coding system. University of California at San Francisco **2**(36), 1 (1983) (**Unpublished manuscript**)
40. Fyffe, G., Graham, P., Tunwattanapong, B., Ghosh, A., Debevec, P.: Near-instant capture of high-resolution facial geometry and reflectance. In: Madeira, J., Patow, G. (eds.) Computer Graphics Forum, vol. 35, pp. 353–363. Wiley Online Library (2016)
41. Gan, W., Xue, J., Lu, K., Yan, Y., Gao, P., Lyu, J.: FEAFA+: an extended well-annotated dataset for facial expression analysis and 3d facial animation. CoRR arXiv:abs/2111.02751 (2021)
42. Gecer, B., Ploumpis, S., Kotsia, I., Zafeiriou, S.P.: Fast-ganfit: generative adversarial network for high fidelity 3d face reconstruction. IEEE Trans. Pattern Anal. Mach. Intell. (2021). https://doi.org/10.1109/TPAMI.2021.3084524
43. Ghosh, A., Fyffe, G., Tunwattanapong, B., Busch, J., Yu, X., Debevec, P.: Multiview face capture using polarized spherical gradient illumination. In: Proceedings of the 2011 SIGGRAPH Asia Conference, pp. 1–10 (2011)
44. GiantStep Studio Co.: The process behind "project vincent". https://www.giantstep.com/work/vincent/ (2019)
45. Gilbert, M., Demarchi, S., Urdapilleta, I.: Facshuman a software to create experimental material by modeling 3d facial expression. In: Proceedings of the 18th International Conference on Intelligent Virtual Agents, pp. 333–334 (2018)
46. Gotardo, P., Riviere, J., Bradley, D., Ghosh, A., Beeler, T.: Practical dynamic facial appearance modeling and acquisition. ACM Trans. Graph. (2018). https://doi.org/10.1145/3272127.3275073
47. Guo, Y., Zhang, J., Cai, L., Cai, J., Zheng, J.: Self-supervised cnn for unconstrained 3d facial performance capture from an rgb-d camera. arXiv preprint arXiv:1808.05323v1 (2018)
48. Haber, J., Kähler, K., Albrecht, I., Yamauchi, H., Seidel, H.P.: Face to face: from real humans to realistic facial animation. In: Proceedings Israel-Korea Binational Conference on Geometrical Modeling and Computer Graphics, vol. 2001, pp. 73–82. Citeseer (2001)
49. He, Y., Choi, Cv.: A study of facial expression of digital character with muscle simulation system. Int. J. Adv. Smart Converg. **8**(2), 162–169 (2019)
50. Hendler, D., Moser, L., Battulwar, R., Corral, D., Cramer, P., Miller, R., Cloudsdale, R., Roble, D.: Avengers: capturing thanos's complex face. In: ACM SIGGRAPH 2018 Talks, pp. 1–2 (2018)
51. Ho, C.C., MacDorman, K.: Measuring the uncanny valley effect: refinements to indices for perceived humanness, attractiveness, and eeriness. Int. J. Soc. Robot. **9**, 129–139 (2017). https://doi.org/10.1007/s12369-016-0380-9
52. Horain, P., Marques Soares, J., Zhou, D., Li, Z., Gomez Jauregui, D.A., Allusse, Y.: Perceiving and rendering users in a 3D interaction. In: IHCI 2010 : Second IEEE International Conference on Intelligent Human Computer Interaction. pp. 42–53. Springer (2010). https://hal.archives-ouvertes.fr/hal-00836606
53. iClone: 3d facial animation | iclone | reallusion. https://www.reallusion.com/iclone/3d-facial-animation.html (2022)
54. games International, E.: Unreal 5 full body ik system. https://docs.unrealengine.com/5.0/en-US/AnimationFeatures/PBIK/ (2021)
55. International Epic Games.: Epic games metahuman creator. https://metahuman.unrealengine.com/ (2021)
56. Jia, J., Wu, Z., Zhang, S., Meng, H., Cai, L.: Head and facial gestures synthesis using pad model for an expressive talking avatar. Multimed. Tools Appl. (2013). https://doi.org/10.1007/s11042-013-1604-8
57. Kim, H.Y., Park, D.J., Lee, T.G.: Comparative analysis of markerless facial recognition technology for 3d character's facial expression animation-focusing on the method of faceware and faceshift. Cartoon Anim. Stud. **37**, 221–245 (2014)
58. Komorowski, D., Melapudi, V., Mortillaro, D., Lee, G.S.: A hybrid approach to facial rigging. In: ACM SIGGRAPH ASI LeA 2010 Sketches. SA '10, Association for Computing Machinery, New York, NY, USA (2010). https://doi.org/10.1145/1899950.1899992
59. Krumhuber, E.G., Tamarit, L., Roesch, E.B., Scherer, K.R.: Facsgen 2.0 animation software: Generating three-dimensional facs-valid facial expressions for emotion research. Emotion **12**(2), 351 (2012)
60. Lattas, A., Moschoglou, S., Ploumpis, S., Gecer, B., Ghosh, A., Zafeiriou, S.P.: Avatarme++: Facial shape and brdf inference with photorealistic rendering-aware gans. IEEE Trans. Pattern Anal. Mach. Intell. (2021). https://doi.org/10.1109/TPAMI.2021.3125598
61. Lee, J., Kim, S., Kim, S., Sohn, K.: Multi-modal recurrent attention networks for facial expression recognition. IEEE Trans. Image Process. **29**, 6977–6991 (2020). https://doi.org/10.1109/TIP.2020.2996086
62. Lee, Y., Terzopoulos, D., Waters, K.: Constructing physics-based facial models of individuals. In: Graphics Interface, pp. 1. Canadian Information Processing Society (1993)
63. Lee, Y., Terzopoulos, D., Waters, K.: Realistic modeling for facial animation. In: Proceedings of the 22nd annual conference on Computer graphics and interactive techniques, pp. 55–62 (1995)
64. Lefevre, S., Odobez, J.M.: Structure and appearance features for robust 3d facial actions tracking. In: 2009 IEEE International Conference on Multimedia and Expo, pp. 298–301 (2009). https://doi.org/10.1109/ICME.2009.5202494
65. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. ACM Trans. Graph. (2017). https://doi.org/10.1145/3130800.3130813
66. Lou, J., Cai, X., Dong, J., Yu, H.: Real-time 3d facial tracking via cascaded compositional learning. IEEE Trans. Image Process.

30, 3844–3857 (2021). https://doi.org/10.1109/TIP.2021.3065819

67. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression. In: 2010 IEEE computer society conference on computer vision and pattern recognition-workshops, pp. 94–101. IEEE (2010)
68. Lucey, P., Cohn, J.F., Prkachin, K.M., Solomon, P.E., Matthews, I.: Painful data: the unbc-mcmaster shoulder pain expression archive database. In: 2011 IEEE International Conference on Automatic Face Gesture Recognition (FG), pp. 57–64 (2011). https://doi.org/10.1109/FG.2011.5771462
69. Ma, W.C., Jones, A., Hawkins, T., Chiang, J.Y., Debevec, P.: A high-resolution geometry capture system for facial performance. In: ACM SIGGRAPH 2008 talks, pp. 1 (2008)
70. Magnenat-Thalmann, N., Primeau, E., Thalmann, D.: Abstract muscle action procedures for human face animation. Vis. Comput. **3**(5), 290–297 (1988)
71. Mavadati, M., Sanger, P., Mahoor, M.H.: Extended disfa dataset: Investigating posed and spontaneous facial expressions. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1452–1459 (2016). https://doi.org/10.1109/CVPRW.2016.182
72. McDonnell, R., Breidt, M., Bülthoff, H.: Render me real? Investigating the effect of render style on the perception of animated virtual humans. ACM Trans. Graph. **31**, 91:1-91:11 (2012)
73. McDuff, D., el Kaliouby, R., Senechal, T., Amr, M., Cohn, J.F., Picard, R.: Affectiva-mit facial expression dataset (am-fed): naturalistic and spontaneous facial expressions collected "in-the-wild". In: 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 881–888 (2013). https://doi.org/10.1109/CVPRW.2013.130
74. Mori, M., MacDorman, K.F., Kageki, N.: The uncanny valley [from the field]. IEEE Robot. Autom. Mag. **19**(2), 98–100 (2012). https://doi.org/10.1109/MRA.2012.2192811
75. Moser, L., Hendler, D., Roble, D.: Masquerade: fine-scale details for head-mounted camera motion capture data. In: ACM SIGGRAPH 2017 Talks, pp. 1–2. ACM, Los Angeles (2017). https://doi.org/10.1145/3084363.3085086
76. Navarro, I., Kneubuehler, D., Verhulsdonck, T., Du Bois, E.D., Welch, W., Verma, V., Sachs, I., Bhat, K.: Fast facial animation from video. In: ACM SIGGRAPH 2021 Talks. SIGGRAPH '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3450623.3464681
77. Nhan, J.: Face Tracking, pp. 293–307. Apress, Berkeley (2022). https://doi.org/10.1007/978-1-4842-7836-9_16
78. O'Hailey, T.: Rig it right!: Maya animation rigging concepts. Taylor & Francis, CRC Press (2019)
79. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1717–1724 (2014)
80. Ostermann, J.: Animation of synthetic faces in mpeg-4. In: Proceedings Computer Animation '98 (Cat. No.98EX169), pp. 49–55 (1998). https://doi.org/10.1109/CA.1998.681907
81. Page, M.J., Moher, D., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hróbjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., McGuinness, L.A., Stewart, L.A., Thomas, J., Tricco, A.C., Welch, V.A., Whiting, P., McKenzie, J.E.: Prisma 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ **372** (2021). https://doi.org/10.1136/bmj.n160. https://www.bmj.com/content/372/bmj.n160
82. Pandzic, I.S., Forchheimer, R.: MPEG-4 Facial Animation: The Standard Implementation and Applications. Wiley, New York (2003)
83. Pandzic, I.S., Forchheimer, R.: MPEG-4 Facial Animation: The Standard, Implementation and Applications. Wiley, New York (2003)
84. Pantic, M., Valstar, M., Rademaker, R., Maat, L.: Web-based database for facial expression analysis. In: 2005 IEEE international conference on multimedia and Expo, pp. 5. IEEE (2005)
85. Parke, F.I.: Computer generated animation of faces. In: Proceedings of the ACM annual conference-Volume 1, pp. 451–457 (1972)
86. Pelachaud, C., Maya, V., Lamolle, M.: Representation of expressivity for embodied conversational agents. In: Workshop Balanced Perception and Action, Third International Joint Conference on Autonomous Agents & Multi-Agent Systems, vol. 10. Citeseer, New York (2004)
87. Pettersson, E.: Social interaction with real-time facial motion capture (2017)
88. Pighin, F., Lewis, J.: Digital face cloning. Siggraph course (2005)
89. Pighin, F., Hecker, J., Lischinski, D., Szeliski, R., Salesin, D.H.: Synthesizing realistic facial expressions from photographs. In: Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques, pp. 75–84. SIGGRAPH '98, Association for Computing Machinery, New York, NY, USA (1998). https://doi.org/10.1145/280814.280825
90. Ping, H.Y., Abdullah, L.N., Sulaiman, P.S., Halin, A.A.: Computer facial animation: a review. Int. J. Comput. Theory Eng. **5**(4), 658–662 (2013). https://doi.org/10.7763/IJCTE.2013.V5.770
91. Purushothaman, R.: Morph animation and facial rigging. Character Rigging and Advanced Animation: Bring Your Character to Life Using Autodesk 3ds Max, pp. 243 (2019)
92. Rademacher, P.: Ray tracing: graphics for the masses. XRDS **3**(4), 3–7 (1997). https://doi.org/10.1145/270955.270962
93. Ramamoorthi, R.: Precomputation-Based Rendering. NOW Publishers Inc, Baltimore (2009)
94. Renderpeople.: Digital human—behind the scenes. Developer diary and the making of the digital human. https://digitalhuman.io/behind-the-scenes/ (2021)
95. Reverdy, C., Gibet, S., Larboulette, C.: Optimal marker set for motion capture of dynamical facial expressions. In: Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games, pp. 31–36 (2015)
96. Riviere, J., Gotardo, P., Bradley, D., Ghosh, A., Beeler, T.: Single-shot high-quality facial geometry and skin appearance capture. ACM Siggraph 2020 (2020)
97. Rosenberg, E., Ekman, P.: What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS) Series in Affective Science. Oxford University Press, Oxford (2020)
98. Ruan, G., Wernert, E., Gniady, T., Tuna, E., Sherman, W.: High performance photogrammetry for academic research. In: Proceedings of the Practice and Experience on Advanced Research Computing, pp. 1–8 (2018)
99. Rydfalk, M.: Candide, a parameterised face. Tech. Rep. LiTH-ISY-I-866, Linköping University, Sweden (1987)
100. Sagar, M.: Reflectance field rendering of human faces for "spider-man 2". In: ACM SIGGRAPH 2005 Courses, pp. 14–es (2005)
101. Sanyal, S., Bolkart, T., Feng, H., Black, M.J.: Learning to regress 3d face shape and expression from an image without 3d supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
102. Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., Akarun, L.: Bosphorus database for 3d face analysis. In: Schouten, B., Juul, N.C., Drygajlo, A., Tistarelli, M.

(eds.) Biometrics and Identity Management, pp. 47–56. Springer, Berlin (2008)

103. Schenk, T.: Introduction to Photogrammetry, vol. 106. The Ohio State University, Columbus (2005)

104. Serra, J., Moser, L., McLean, D.A., Roble, D.: Simplified facial capture with head mounted cameras. In: ACM SIGGRAPH 2021 Talks. SIGGRAPH '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3450623.3464637

105. Seymour, M., Riemer, K., Kay, J.: Interactive realistic digital avatars-revisiting the uncanny valley. Proceedings of the 50th Hawaii International Conference on System Sciences (2017)

106. Seymour, M., Evans, C., Libreri, K.: Meet mike: epic avatars. In: ACM SIGGRAPH 2017 VR Village, pp. 1–2 (2017)

107. Shakir, S.D., Al-Azza, A.A.: Facial modelling and animation: an overview of the state-of-the art. Iraqi J. Electr. Electron. Eng. **18**(1), 28–37 (2022). https://doi.org/10.37917/ijeee.18.1.4

108. Socolinsky, D., Wolff, L., Neuheisel, J., Eveland, C.: Illumination invariant face recognition using thermal infrared imagery. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, vol. 1, pp. I (2001). https://doi.org/10.1109/CVPR.2001.990519

109. Spring, A.: Adam spring website. https://adamspring.co.uk/2020/05/25/facs-rigging-texture-blending-digital-humans/ (2021)

110. van der Struijk, S., Huang, H.H., Mirzaei, M.S., Nishida, T.: Facsvatar: an open source modular framework for real-time facs based facial animation. In: Proceedings of the 18th International Conference on Intelligent Virtual Agents, pp. 159–164. IVA '18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3267851.3267918

111. Tao, H., Chen, H., Wu, W., Huang, T.: Compression of mpeg-4 facial animation parameters for transmission of talking heads. IEEE Trans. Circuits Syst. Video Technol. **9**(2), 264–276 (1999). https://doi.org/10.1109/76.752094

112. Terzopoulos, D., Waters, K.: Techniques for realistic facial modeling and animation. In: Computer Animation'91, pp. 59–74. Springer (1991)

113. The MakeHuman Community: Makehuman. http://www.makehumancommunity.org/ (2022)

114. Unity: Enemies. https://unity.com/es/demos/enemies (2022)

115. Valente, S., Dugelay, J.L.: Face tracking and realistic animations for telecommunicant clones. IEEE Multimed. **7**(1), 34–43 (2000). https://doi.org/10.1109/93.839309

116. Valstar, M.F., Jiang, B., Mehu, M., Pantic, M., Scherer, K.: The first facial expression recognition and analysis challenge. In: 2011 IEEE International Conference on Automatic Face Gesture Recognition (FG), pp. 921–926 (2011). https://doi.org/10.1109/FG.2011.5771374

117. Villagrasa, S., Susín Sánchez, A.: Face! 3d facial animation system based on facs. In: IV Iberoamerican symposium in computer graphics, pp. 203–209 (2009)

118. Weise, T., Li, H., Van Gool, L., Pauly, M.: Face/off: Live facial puppetry. In: Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, pp. 7–16. SCA '09, Association for Computing Machinery, New York, NY, USA (2009). https://doi.org/10.1145/1599470.1599472

119. Welsh, W.J.: Model-based coding of videophone images. Ph.D. thesis, British Telecom Research Lab (1991)

120. Wen, L., Zhou, J., Huang, W., Chen, F.: A survey of facial capture for virtual reality. IEEE Access **10**, 6042–6052 (2021). https://doi.org/10.1109/ACCESS.2021.3138200

121. Williams, L.: Performance-driven facial animation. SIGGRAPH Comput. Graph. **24**(4), 235–242 (1990). https://doi.org/10.1145/97880.97906

122. Xie, T., Olano, M., Karis, B., Narkowicz, K.: Real-time subsurface scattering with single pass variance-guided adaptive importance sampling. Proc. ACM Comput. Graph. Interact. Tech. **3**(1), 1–21 (2020)

123. Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., Stolcke, A.: The microsoft 2017 conversational speech recognition system. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 5934–5938. IEEE (2018)

124. Zhang, Z., Girard, J.M., Wu, Y., Zhang, X., Liu, P., Ciftci, U., Canavan, S., Reale, M., Horowitz, A., Yang, H., Cohn, J.F., Ji, Q., Yin, L.: Multimodal spontaneous emotion corpus for human behavior analysis. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3438–3446 (2016). https://doi.org/10.1109/CVPR.2016.374

125. Zibrek, K., Martin, S., McDonnell, R.: Is photorealism important for perception of expressive virtual humans in virtual reality? ACM Trans. Appl. Percept. (2019). https://doi.org/10.1145/3349609

126. Zibrek, K., McDonnell, R.: Does render style affect perception of personality in virtual humans? In: Proceedings of the ACM Symposium on Applied Perception, pp. 111–115 (2014)

127. Zollhöfer, M., Thies, J., Garrido, P., Bradley, D., Beeler, T., Pérez, P., Stamminger, M., Nießner, M., Theobalt, C.: State of the art on monocular 3d face reconstruction, tracking, and applications. Comput. Graph. Forum **37**(2), 523–550 (2018). https://doi.org/10.1111/cgf.13382