







A Study of the Frameworks for Digital Humans: Analyzing Facial Tracking evolution and New Research Directions with AI

Carlos Vilchis¹^a, Miguel Gonzalez-Mendoza¹^b, Leonardo Chang¹^c, Sergio A Navarro-Tuch¹^d,
Gilberto Ochoa Ruiz¹^e and Isaac Rudomin²^f

¹*School of Engineering, Computer Science Department Tecnologico de Monterrey, Nuevo Leon, Mexico*

²*Departamento de Tecnologías de la Información, Universidad Autónoma Metropolitana-Cuajimalpa, Mexico*

Keywords: Motion Capture, Facial simulation, Human-Computer Interfaces, Avatars, FACS, Digital Humans.

Abstract: There actual scenario of techniques and methods to improve our perception of digitals humans is mostly oriented to increase likeness and interactivity inside a helpful workflow. Breakthrough milestones have been achieved to improve facial expression recognition systems thanks to the effectiveness of Deep Neural Networks. This survey analyzes all the open and fully integrated frameworks (deep learning-based or not) available today. All of those can be replicated by peers to analyze their effectiveness and efficiency in real-time environments. Also, we present an overview analysis of present-day environments for digital humans that use state-of-the-art facial tracking and animation retargeting to settle a direction on the future steps in our research objectives in this field.

1 INTRODUCTION


Digital humans have become a consolidated area of research in the last five or six years. Also called embodied conversational agents (ECA) or virtual avatars, they are now part of the Hype Cycle of Emerging Technologies in 2021 (Burke, 2021). This visual representation of a digital human requires a diverse set of technologies and an extensive list of factors that must be achieved, such as realistic graphics (Zibrek and McDonnell, 2014), emphatic response (Seymour et al., 2017b), or a correct model that can make replicate the facial expressions and performance coming from a specific real human (Krumhuber et al., 2012).


A wide array of applications in the industry makes Digital Humans part of the options to create computer-human interfaces; these include e-commerce, agent services, government communication, customer services, etc. Even the simple option of seeing a future including automated AI-based humans, with a smile and all, serving customers 24/7,


makes it a relevant area of research.


A small group of companies in the field has access to a fully developed framework to create Digital Humans with Deep Learning-based tracking. (Antoniades, 2016) (Motion, 2017) (Moser et al., 2017). Such systems are based on taking a real human person, analyzing their facial performance, extracting the uniqueness of its movements inside a discrete model, and replicating it into a 3D model or avatar. A digital 3D model with a high level of detail (most of the cases requiring the use of photogrammetry) (GiantStep, 2019), is able to collect all the detailed animation inside a set of constraints and controllers enabling to drive a system named rig (Cañamero and Aylett, 2008). This rig will be in charge of driving all the collected information inside the traditional animation pipeline or, for purposes of this analysis: a real-time engine (Aneja et al., 2019). Lastly, the way humans build a particular observed link between anthropomorphism and emotional response is contemplated by the concept of the Uncanny Valley (Mori et al., 2012), and it is a crucial component inside this process to improve digital humans.


All the processes mentioned earlier can be summarized as a simple phrase: "framework for digital humans". There is a limited number of them that are public and open research. As the tools needed to repli-


^a <https://orcid.org/0000-0001-5289-6070>

^b <https://orcid.org/0000-0001-6451-9109>

^c <https://orcid.org/0000-0002-0703-2131>

^d <https://orcid.org/0000-0002-3551-7689>

^e <https://orcid.org/0000-0001-6451-9109>

^f <https://orcid.org/0000-0002-1672-1756>

cate it (hardware and software) are not easy to integrate, fully operational frameworks are rare.

In order to understand how a framework for digital humans integrates different key processes, we need to split the frameworks into different areas. The first is the physical aspect of the digital-human, the expression set of how this digital human performs and talks (most of the time, extracted from a real human being). The process ends in 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; commonly, this is done using deep learning for prediction (Roble et al., 2019a).

The way facial expressions of a digital human are constructed to recreate all of what the human face does, when talking, expressing positive emotions or disgust, is called facial codification. The initial facial codification was standardized in the industry back in 1995 (MPEG, 2021), and while more than two decades have passed, both academics and industry are still using it as the most popular method. Facial Animation Parameters (FAP), based on control points codification, lacks a relation between muscles and expressions. Therefore, the use of facial expression methodologies, such as Facial Action Code System (FACS), proposed by Paul Ekman (Ekman, 1997), includes this. Popular research, databases, and methods use this facial codification; some of them also include a validation process to certify the codification of FACS with an expert demonstrating a success rate of more than 72% (Amini et al., 2015).

Common approaches are used to improve the study, analysis, and extraction of a realistic facial codification from real-world face recognition of real subject faces. This helps to improve the validation and in obtaining information from the wild. To analyze this imagery, state-of-the-art science and computational techniques such as including AI-based tools is essential. This trend includes Convolutional Neural Networks using computer vision-based recognition (Bouaziz et al., 2013)(Krumhuber et al., 2012).

Several challenges were found when developing these frameworks, and due to the increased need of the industry to achieve better, faster, and reliable digital humans, most of the key factors that the main players in the industry consider important to overcome are:

- The more significant challenge is to improve the quality of rendered the 3D model. This means achieving photorealism. (Zibrek et al., 2018)
- To reduce the Uncanny Valley effect, some components considered are improved animation, lip-sync capabilities, facial deformations, and expressions.

- Interactivity increases the chances of success in convincing the user of the uniqueness and individual identity of a digital human; this means, the possibility of not rendering performances linearly when using real-time graphics. (Hyde et al., 2015)(Seymour et al., 2017b)

In perspective, with past studies and research papers trying to describe how to integrate an open framework (van der Struijk et al., 2018)(Aneja et al., 2019), we explore and study the specifications, efficiency rate, and quality of the results when using state-of-the-art frameworks to complete this task. The main aspects analyzed for each framework are the facial expression recognition method, the integration of the best facial codification available, real-time control, and adequate evaluation of the emphatic response. This survey also covers the most recent tools, methodologies, and new research horizons in facial expression recognition with deep learning in the digital-human field.

As the main objective, this paper contributes as follows:

- Providing in-depth analysis and formalization of the processes required of a framework for real-time digital humans.
- Providing a complete study of the elements covered in existing frameworks, as well as what is lacking to approach new methodologies for real-time digital humans.
- Providing a perspective for the future in the real-time facial tracking field of digital humans.

2 DIGITAL HUMANS FACIAL ANIMATION FRAMEWORK

2.1 Facial Codification and Rigs

Understanding facial animation based on expression recognition, means talking about a facial codification method. The two main codification approaches are FAP (MPEG, 2021) and FACS (Ekman, 1997). These are the most accepted methodologies in the facial animation field. FACS was proposed almost 42 years ago, FAP almost 26 and, until now, are still used to replicate avatar faces in almost any professional field inside human representation (Moser et al., 2017). FAP was originally designed as an industry standard to approach the fast growth of the animation industry to standardize it, and FACS was originally achieved as a theory in the field of psychology to observe, study and analyze how facial expressions describe emotions

and intentions. The facial codification era in animation started in 2000s to parametrize facial controls in animated characters (commonly named *rigs*) by videogames/animation companies and studios. Several authors (Pandzic and Forchheimer, 2003) considered a closer relationship between FAP and FACS. FAP is oriented towards parametrizing the controls to manipulate the face like a puppet, useful for speech-driven avatars and performance animation by motion capture. FACS, on the other hand, was originally developed to parametrize emotions and their relation with muscle groups by observation.

2.2 FAP Codificaion / Traditional Rig

Over several decades the Facial Animation Parameters (FAP) methodology has become the most accepted standard for facial control of 3D avatars and digital characters. Originally inspired by AMA (Abstract Muscle Actions), was later developed by request by MPEG (Moving Picture Expert Group) and became standard codification.

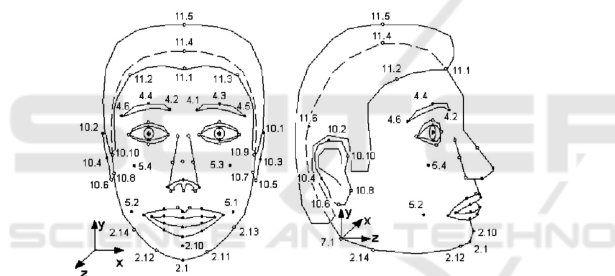


Figure 1: The set of MPEG-4 Face Action Parameter (FAP) feature points (MPEG, 2021).

In the field of emotion recognition, FAP has been used as the main coding system by several researchers, not just in the animation field (Pardas et al., 2002) were is more popular due to being easy to integrate as a traditional rig as Figure 1. Therefore, a few contents in this field have been published to codify facial recognition or to be used inside a contemporary AI framework.

In order to get successful and realistic human animatable faces, some authors have proposed better modeling techniques based on FAP (Fratarcangeli and Schaerf, 2004). This includes a mixture of hand-modeled morph targets and considering how this linear animation was related to facial muscle structure, where the controls of the face were closely related to muscular descriptions. Several research results were published into digital humans (Pasquariello and Pelachaud, 2002), with a particular pioneering approach to handle skin deformation with wrinkles and

bulges in combination with FAP codification.

Facial Expression Synthesis with FAP have their research origins in the animation field thanks to some published papers (Malatesta et al., 2009), which looked for an improved way to create better affective computing interfaces with Embodied Conversational Agents (ECA). The need to generate emotion and emphatic results while using FAP codification drove the researchers into the need to model universal expressions; this is how the Ekman theory of FACS appears in the facial animation research field for the first time. Facial Action Tracking with parametrized FAP (2001) was fundamental to complete several frameworks proposed to improve the use of ECA in web and mobile platforms back in 2002. As the web contents were rapidly increasing, the use of Active Appearance Models (AAMs) (Pandzic and Forchheimer, 2003) as a statistical model-based method for estimating the 3D pose and deformation of faces in real-time. This model included a Geometry, Pose, and Texture parametrization tool inside an analysis-synthesis process. Along with the 2000s, FAP codification has become increasingly popular as the main facial animation and mocap methodology (Li et al., 2010) with a mix of blendshapes and joints until today.

2.3 FACS Codification / 3D based Scanned Rigs

Early FACS codification in animation approaches became popular at the end of the 2000's in research fields due to interest take human 3D models from real human inverse engineering. The real muscular system of the face needed to be coded to split information of subject faces to de-code it again and represent like Figure 2 shows. As soon as the methodology was reborn in the animation field, more researchers used it to validate 3D human facial models (Cosker et al., 2011) in order to improve the quality of Conversational Agents and standardize the rules of facial expression into it.

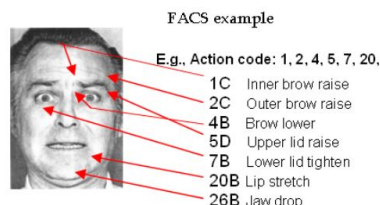


Figure 2: An example of expression coded by Facial Action Code System (Ekman, 1997).

FACS became essential thanks due to the increasing importance of 3D facial scanning in the field as it helped to capture single shots of facial geometry and

obtain an accurate representation of muscles, wrinkles, and skin folding. Now is a well-established technique used in movies to represent real actors in VFX. Better achievements in photogrammetry scan helped researchers separate different groups of textures such as albedo, normals, specular, etc. However, the emphatic response related to feeling less uncanny digital humans (Tinwell et al., 2011) pushes the research to continuously retake the FACS methodology under new tools and new methods.

Alongside better quality methods to scan FACS from real people, better techniques to represent the facial movements were needed. The first attempts to include FACS codification in 3D facial rig construction (Villagrasa and Susín Sánchez, 2009) allowed getting a better understanding of how the face moves, using 3D model blendshapes to generate custom controls. Hybrid methods combining scanned blendshapes and bones become the newest standard in the digital-humans field (Ravikumar et al., 2016). Better techniques for the interpolation of blendshapes based on FACS (Alkawaz et al., 2015) include some newer algorithms to interpolate traditional face models between expressions inside a contemporary framework. Some results showed a strong relationship between the color of skin, FACS and emphatic performance (Alkawaz et al., 2014) on digital humans, this way, the different human expressions can be achieved realistically with easily replicated workflows. New techniques related to Machine Learning training models (Radzihovsky et al., 2020) have proven the improvement of facial rig construction using information previously trained in deep learning models, making it easier to integrate into frameworks and to reuse facial animation easily.

FACS Validation by an expert coder becomes real (Rothkrantz et al., 2009) in animation framework to validate rig blendshapes and expressions (Cosker et al., 2011). In the beginning, just a small sample of 20 expressions from upper and lower face Actor Units (AU) was enough to let experts validate the precision of 3D face models. Later the need to code validated FACS in linear and nonlinear motion faces became a challenging goal in computer graphics (Cosker et al., 2010), results about how nonlinear blending of expressions is more convincing to facial expert coders than simple linear ones.

FACS databases are used to study, train and evaluate expression inside a digital human framework (Aneja et al., 2018)(Aneja et al., 2019), were composed by libraries such as MMI(Pantic et al., 2005) with 10,000 images, DISFA(Mavadati et al., 2013) with 50,00 images, CK+[53] with 593 videos, and KDEF(Calvo and Lundqvist, 2008) with 4900 Im-

ages. Novel databases such as AMFED(McDuff et al., 2016) are based on YouTube content, with a large dataset of 1.8 million recordings of YouTubers in front of the camera. Thanks to web services and the cloud. Researchers such as Benitez-Quiroz et al. (Fabian Benitez-Quiroz et al., 2016) with Emotionet, and Mollahosseini et al. (Mollahosseini et al., 2017) in AffectNet, annotate, classify and code actor units from facial expressions obtained from the Web to get bigger and better datasets to improve new algorithms of facial research.

3 FACIAL EXPRESSION RECOGNITION AND REAL TIME RETARGETING INTO A 3D ENGINE

The facial tracking process and the facial expression recognition process, has been evolving over the past years. Facial tracking to get an actor's performance into 3D animation has research papers dated from 2000(Valente and Dugelay, 2000) and was a basic linear interpretation of distances between points or reflective markers in the eyebrows, lips, etc. The translation movement was tracked and processed by several human hours as Vicon and Faceware. The next milestone was related to research about improving tracking in real-time (Valente and Dugelay, 2000) without human interaction, based on image-based tracking and computer vision to complete high-performance FER. The process is made taking information from physical markers, skin marks (Moser et al., 2017), or directly from contours of the lips, eyes, and eyebrows using computer vision recognition (Reverdy et al., 2015) and advanced methods like Machine Learning to recognize expressions, forecast facial movements, and retarget accurate data into the digital humans(Bhat et al., 2013).

Connection to real-time animation retargeting is part of the complete framework for digital humans under interactive circumstances. Experiments in the last years using both traditional rendering engines or game engines due to flexibility to perform high-quality graphics for a fraction of the cost of rendering digital humans (Eckschlager et al., 2005). The achievement made by Ninja Theory with Hellblade (Antoniades, 2016), and collaboration in 2017 between Epic Games, Vicon, Cubic Motion, 3lateral to get Siren (CubicMotion, 2018) showed the path of real-time expression recognition and real-time retargeting into a 3D engine. Research made by Mike Seymour under Unreal Engine(Seymour et al., 2017a),

has strengthened the research into cross the uncanny valley with real-time interaction.

3.1 Democratized Facial Tracking Tools for Digital Humans.

Understanding a democratized FER with FACS for digital humans means to not depending on commercial tools like mentioned before. This means to process all the actor's facial expressions or databases from the acquisition step, commonly related to using a 3D-Convolutional Neural Network or Generative Adversarial Networks for this type of AI training. We will talk about tools based on different Artificial Intelligence methodologies, STRUIJK with FACSvatar (van der Struijk et al., 2018), Aneja with AvatarSim, Project Vincent (GiantStep, 2019), Doug Roble Masquarade (Moser et al., 2017), but several tools like Amini HapFACS (Amini et al., 2015), Gilbert FACShuman (Gilbert et al., 2018), Krumhuber with FACSGen2.0 (Krumhuber et al., 2012), are not considered due to their lacking AI or their not being created to be integrated into a full FACS framework.

3.2 Truly AI Tools for Digital Human Face Tracking in Real Time with FACS.

DeepExpr. Aneja's research into virtual humans started in 2016 with DeepExpr (Aneja et al., 2016), a deep learning method to create static facial expressions in 3D characters using FACS codification. The method was focused on creating emphatic cartoonish characters expressions based on two basic CNN with Gabor Filters, first trained with standard databases (CK+, DISFA, MMI) to recognize features, the second CNN to do the same process into a 3D digital cartoon character, then, using transfer learning by Oquab (Oquab et al., 2014), to create a shared embedding feature space. They fine-tuned the pre-trained CNN on the human dataset with the character dataset by continuing the backpropagation step. Finally, a fully connected layer of the human-trained model was fixed, so earlier layers were kept fixed to avoid overfitting problems.

ExprGen. ExprGen, later by Aneja et al. in 2018, was into 3D Character FACS expression generation (Aneja et al., 2018), but the method was optimized into a 3D CNN. This end-to-end system worked in a perceptually valid and geometrically consistent manner. They trained a Pseudo-Siamese network named fused-CNN (f-CNN) composed of 2 branches, Human CNN (HCNN) and Shared CNN (SCNN). In the last

step, with a similar transfer to their past work, they trained the f-CNN based on the distance between two image encodings. The final match of geometry is perceptually method matched. The last work of Aneja et al. (Aneja et al., 2019) in 2019 was related to a full framework for virtual characters, composed of a real-time recognition tool for FACS, translation of values into a facial expression tool, and direct transfer into a real-time 3D engine.

FACSvatar. Struijk et al. (Hasani and Mahoor, 2017) presented FACSvatar at 2018 as the first full framework to have real-time facial recognition based on FACS as an open source for digital humans. FACSvatar allows the integration of models generated by MakeHuman Project (MakeHuman, 2021), a full 3D environment for creating digital characters in 3D, which is integrated with the plugin FACSHuman (Gilbert et al., 2018). Several researchers developed all this suite of tools and made it open for research. This framework takes a modified version of OpenFace2.0 (Baltrusaitis et al., 2018) as the main tool to recognize AU, eye gaze, and head rotation in real-time via RGB simple camera (no special setup is needed), and process data inside a simple Gated Recurrent Unit Neural Network to enable generative data-driven facial animation with Keras. It uses 17 AU intensities and training with MAHNOB Mimicry Database. Specially designed modules in Python with help in this framework's operation as a GUI, which controls AU's parameters and values in real-time. The information between all these modules is driven using a distributed messaging library ZeroMQ. The development of extra features in FACSvatar allows the user to drive facial expressions with a CVS Offline file, so a live human performer is not always needed in the framework.

AvatarSim. Aneja et al. work concluded in a full framework end-to-end named AvatarSim (Aneja et al., 2019). This real and contemporary framework was proposed in 2019. Setup made with Unreal Engine and a CNN, FACS, and a recognition model trained with Emotionet (Fabian Benitez-Quiroz et al., 2016), including a phoneme extraction based on the audio wave. AvatarSim works with a simple RGB camera, recognizes 12 FACS AU from the user-detected face, and then synthesizes the expression. The Action Unit Recognizer is the same as ExprGen; just with a small difference, these frameworks do not retarget the movements into a midway character but directly send the data to an expression synthesizer. A Python interface made in PyTorch transforms the 12-dimensional feature vector representing the probability of each AU with an average F1 score of 0.78. There is an option to use an

external off-the-shelf FACS AU detection SKD like AFFDEX(McDuff et al., 2016). Finally, the Expression Syntethizer, made in Python, drives 38 Bone, 24 FACS, and 19 phonetic controls in the final target digital avatar. The final character is real-time driven inside Unreal Engine with traditional blendshapes. The AI model was based on a CNN of the input image with 49 FACS detection and trained with MMI, CK+, KDEP, and DISFA. The result of this training model (HCNN) was connected to a second CNN in a pseudo-siamese network (fused CNN). The second CNN (Shared CNN) was designed to recognize 10 cartoon expressions. All character expressions were validated to 70% with Mechanical Turk (50 test subjects) but not a FACS expert. Both CNN, HCNN, and SCNN, inside the f-CNN, were trained to get a similarity score between human and primary character expressions.

Project Vincent by GIANTSTEP. Like the work done by Digital Domain and Cubic Motion for Siren, is the case of GianStep(SouthKorea,) from project Vincent, an experimental project of a Digital human-based in a private research framework but with better results than all the ones previously mentioned in this document. Sakamoto et al.(GiantStep, 2019) developed an internal framework based on facial 3d scans to acquire the set of expressions with the FACS methodology. Sakamoto developed a Machine Learning Facial Capture tool based on Keras and Python with a CNN model. No external database was used for the Neural Networks, as information was taken from just one face and all the possible expressions. The framework uses two Neural networks. First, a Facial Marker Tracker was designed to find and predict the coordinate system of 3D markers with marker-less 2D images of facial expressions. The face was segmented into six areas, and every area learns separately with 35 expressions to have a total of 210 different blendshapes of the face. A second Neural Network, designed as a Blend-shape mapper, was connected 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.

Masquerade by Digital Domain. are one of the newest unsupervised AI methods. Private solutions made by cutting-edge companies like Digital Domain (Roble et al., 2019b) to improve the quality, realism and to drive emotional characters such as Thanos (Ennis et al., 2013) in the Avengers Endgame movie(Hendler et al., 2018). They have been testing GANs to train and facial recognition. Above the CNN methods, some researchers have chosen a Generative Adversarial Network to complement their tools. Few

details are public about this project than a couple of neural networks (generative and discriminator) take the role of recognizing markers painted in the actor's face. Later, the solution is complemented with another set of AI tools to train data like CNN and label it and predict like the other solutions before studied.

4 RESULTS OF THE STATE OF THE ART

This section discusses, and analyzes the state-of-the-art frameworks reviewed, the importance of facial codification methodologies, and how they become important inside tracking faces to drive digital humans in real-time.

4.1 Discussion on Codification for Real-time Tracking on Digital Humans

Native FAP codification is not considered in the field of study of this article, but we considered as a common solution used in traditional keyframe animation. Important solutions like ARKit from Apple, the most adopted solution due to the democratization of RGB-D cameras inside Apple phones, have 51 basic coded expressions but are not considered a tool to customize or modify. Professional solutions in digital humans like Faceware and Dynamixyz, based on 48 and 47 coded expressions, cannot allow custom expressions or modify them. These tools, being commercial-born, have become extremely popular around the world.

As we analyzed previously, the use of an expressive-emotional solution-oriented to drive facial performance in digital humans adds extra fidelity because it matches with real 3D scanned humans, whether the accessibility to access facial database to train the AI model results in the digital-human. Using FACS as a codification method to classify facial expressions has become common in the computer graphics field. But there is no customization and parameterization in commercial solutions. FACS generation tools and solutions are available using it, but the lack of full frameworks options makes it more interesting to further research.

4.2 Discussion on Digital Humans and Open Frameworks

We studied options that begin as simple facial expression recognition and transferred knowledge to digital avatars to complete solutions with real-time tracking

Table 1: Analyzed frameworks (FW) and expression generators (GEN), their respective and recognition rates. *Frameworks discarded in the survey because are not FACS based or AI techniques used.

Facial Resarch / Type	Recognition Rate / Certified by coder
AvatarSim (FW)	Unknown / No
HapFACS* (FW)	72.5% / Yes
FACSHuman* (FW)	Unknown / No
FACSGen* (Gen)	72% / Yes
ExprGen (Gen)	75.5% / No
DeepExpr (Gen)	89% / No
FACSVatar (FW)	88% / No
Masquerade (FW)	N/A
Vincent (FW)	N/A

into live 3D digital humans. All the options analyzed were based on FACS codification as a basic way to understand the human face. Several frameworks and solutions were discarded because they were not oriented to this methodology. As well discarded solutions just made to extract FACS like ExprGen, FACSGen (Krumhuber et al., 2012) and DeepExpr were not related to this survey as you can see at Table 1.

An important part of the research was to cross-match frameworks, FACS and Digital humans, showing the importance of validating FACS working with certified coders in some parts of the process. Different Frameworks were subject of emFACS(Friesen et al., 1983) recognition test protocols with test subjects to compare the recognition rate presented in all the frameworks and tools. Not all the frameworks studied included the results of this test. A certified coder was used to validate the codification of input/output on the models in HapFACS (Amini et al., 2015), FACSGen, with the highest recognition rate (72.5% and 72%), but was also discarded as a being a Generator and incomplete framework, respectively. Some of these, like FACSVatar, DeepExpr had results up to 89 percent of recognition rate for the emFACS model, but, is important to clarify is not Certified by Coders.

Suppose we filter all the options to determine the complete open frameworks with an updated state today. In that case, we will find a niche interesting to research with certified coder validation, deep learning-based, and the highest recognition rate.

4.3 Discussion on Future Work

The high demand for solutions for conversational assistants, chatbots, virtual influencers, streaming video hosts, etc., combined with the expectation to democratize the use of these tools, will cause more options to push the industry to have improved solutions. The need to research to have democratic options of frame-

works, with documented solutions for the right way to get scanned persons, AI models, trained databases will be predominant.

By March 2021, Epic Games had released the initial version of MetaHuman Creator(Games, 2021), a tool created to help independent and small studios easily create digital human assets with photorealistic results. This tool represents a huge leap in the democratization of digital-human tools, but the lack of to input custom face scans to create a digital human based on a real human shows the limitation of this tool and how there is still a long road to place custom FACS in easy-to-use tools such as MetaHuman.

Research and development in real-time facial tracking solutions, with easier and affordable options, aligned the same way to improve the democratization of these technologies is an essential direction. The starting point is better portable devices as HMC with computer vision, from a fraction of the cost and open to be customized (different than iPhone ARKit) are the starting point. If we consider the last standardization for a facial codification was the one made by the MPEG group, without an updated version or new alternatives, we can deduce that it is time to generate newly standardized protocols for FACS methodology. Standardizing modern facial codifications based on emotional expression and skin/muscle activity is another direction of research. Proven methods should be analyzed, while correctly documented protocols and data labeling homologation are other topics to research.

5 CONCLUSIONS

We have reviewed digital humans' main concepts and elements inside a real-time rendering framework to introduce deep learning facial tracking. The analyzed frameworks for real-time digital humans provide a wide study of the elements covered. The review of every step to distinguish limitations and advantages let us check and compare actual options. Future steps will include the test of each solution to compare effectiveness with Unreal Metahumans, and prototype a custom FACS and CNN-based tracking inside a validated FACS codification. Future research and experiments in this field are priority to our research group.

REFERENCES

Alkawaz, M. H., Basori, A. H., Mohamad, D., and Mohamed, F. (2014). Realistic facial expression of vir-

- tual human based on color, sweat, and tears effects. *The Scientific World Journal*, 2014.
- Alkawaz, M. H., Mohamad, D., Basori, A. H., and Saba, T. (2015). Blend shape interpolation and faces for realistic avatar. *3D Research*, 6(1):6.
- Amini, R., Lisetti, C., and Ruiz, G. (2015). Hapfacs 3.0: Facs-based facial expression generator for 3d speaking virtual characters. *IEEE Transactions on Affective Computing*, 6(4):348–360.
- Aneja, D., Chaudhuri, B., Colburn, A., Faigin, G., Shapiro, L., and Mones, B. (2018). Learning to generate 3d stylized character expressions from humans. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 160–169. IEEE.
- Aneja, D., Colburn, A., Faigin, G., Shapiro, L., and Mones, B. (2016). Modeling stylized character expressions via deep learning. In *Asian conference on computer vision*, pages 136–153. Springer.
- Aneja, D., McDuff, D., and Shah, S. (2019). A high-fidelity open embodied avatar with lip syncing and expression capabilities. In *2019 International Conference on Multimodal Interaction*, pages 69–73.
- Antoniades, T. (2016). Creating a live real-time performance-captured digital human. In *ACM SIGGRAPH 2016 Real-Time Live!*, SIGGRAPH '16, New York, NY, USA. Association for Computing Machinery.
- Baltrusaitis, T., Zadeh, A., Lim, Y. C., and Morency, L. (2018). Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 59–66.
- Bhat, K. S., Goldenthal, R., Ye, Y., Mallet, R., and Koperwas, M. (2013). High fidelity facial animation capture and retargeting with contours. In *Proceedings of the 12th ACM SIGGRAPH/eurographics symposium on computer animation*, pages 7–14.
- Bouaziz, S., Wang, Y., and Pauly, M. (2013). Online modeling for realtime facial animation. *ACM Trans. Graph.*, 32(4).
- Burke, B. (2021). Hype cycle for emerging technologies, 2021. Technical report, Stamford, CT 06902 USA.
- Calvo, M. G. and Lundqvist, D. (2008). Facial expressions of emotion (kdef): Identification under different display-duration conditions. *Behavior research methods*, 40(1):109–115.
- Cañamero, L. and Aylett, R. (2008). *Animating Expressive Characters for Social Interaction*, volume 74. John Benjamins Publishing.
- Cosker, D., Krumhuber, E., and Hilton, A. (2010). Perception of linear and nonlinear motion properties using a facs validated 3d facial model. In *Proceedings of the 7th Symposium on Applied Perception in Graphics and Visualization*, APGV '10, page 101–108, New York, NY, USA. Association for Computing Machinery.
- Cosker, D., Krumhuber, E., and Hilton, A. (2011). A facs validated 3d human facial model. In *Proceedings of the SSPNET 2nd International Symposium on Facial Analysis and Animation*, FAA '10, page 12, New York, NY, USA. Association for Computing Machinery.
- CubicMotion, U. (2018). Siren case study. <https://cubicmotion.com/case-studies/siren/>.
- Eckschlager, M., Lankes, M., and Bernhaupt, R. (2005). 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*, ACE '05, page 375–376, New York, NY, USA. Association for Computing Machinery.
- Ekman, R. (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.
- Ennis, C., Hoyet, L., Egges, A., and McDonnell, R. (2013). Emotion capture: Emotionally expressive characters for games. In *Proceedings of Motion on Games*, pages 53–60.
- Fabian Benitez-Quiroz, C., Srinivasan, R., and Martinez, A. M. (2016). 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*, pages 5562–5570.
- Fratarcangeli, M. and Schaefer, M. (2004). Realistic modeling of animatable faces in mpeg-4. In *Computer Animation and Social Agents*, pages 285–297.
- Friesen, W. V., Ekman, P., et al. (1983). Emfacs-7: Emotional facial action coding system. *Unpublished manuscript*, University of California at San Francisco, 2(36):1.
- Games, I. E. (2021). Epic games metahuman creator. <https://metahuman.unrealengine.com/>.
- GiantStep, S. (2019). Project vincent. <http://www.giantstep.co.kr/>. (2F, 8, HAKDONG-RO 37-GIL, GANGNAM-GU, SEOUL 06053).
- Gilbert, M., Demarchi, S., and Urdapilleta, I. (2018). Facshuman a software to create experimental material by modeling 3d facial expression. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 333–334.
- Hasani, B. and Mahoor, M. H. (2017). Facial expression recognition using enhanced deep 3d convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.
- Hendler, D., Moser, L., Battulwar, R., Corral, D., Cramer, P., Miller, R., Cloudsdale, R., and Roble, D. (2018). Avengers: capturing thanos's complex face. In *ACM SIGGRAPH 2018 Talks*, pages 1–2.
- Hyde, J., Carter, E. J., Kiesler, S., and Hodgins, J. K. (2015). Using an interactive avatar's facial expressiveness to increase persuasiveness and socialness. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1719–1728.
- Krumhuber, E. G., Tamarit, L., Roesch, E. B., and Scherer, K. R. (2012). Facs-gen 2.0 animation software: Generating three-dimensional facs-valid facial expressions for emotion research. *Emotion*, 12(2):351.

- Li, H., Weise, T., and Pauly, M. (2010). Example-based facial rigging. *ACM Trans. Graph.*, 29(4).
- MakeHuman (2021). Make-human. <http://www.makehumancommunity.org/>. (Accessed on 11/02/2020).
- Malatesta, L., Raouzaoui, A., Karpouzis, K., and Kollias, S. (2009). Mpeg-4 facial expression synthesis. *Personal and Ubiquitous Computing*, 13:77–83.
- Mavadati, S. M., Mahoor, M. H., Bartlett, K., Trinh, P., and Cohn, J. F. (2013). Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160.
- McDuff, D., Mahmoud, A., Mavadati, M., Amr, M., Turcot, J., and Kaliouby, R. e. (2016). Affdex sdk: a cross-platform real-time multi-face expression recognition toolkit. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*, pages 3723–3726.
- Mollahosseini, A., Hasani, B., and Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31.
- Mori, M., MacDorman, K. F., and Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, 19(2):98–100.
- Moser, L., Hendler, D., and Roble, D. (2017). Masquerade: fine-scale details for head-mounted camera motion capture data. In *ACM SIGGRAPH 2017 Talks*, pages 1–2.
- Motion, C. (2017). Realtime live [cubic motion web site].
- MPEG (2021). Movie picture expert group. <https://mpeg.chiariglione.org/standards/mpeg-4>.
- Oquab, M., Bottou, L., Laptev, I., and Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724.
- Pandzic, I. S. and Forchheimer, R. (2003). *MPEG-4 facial animation: the standard, implementation and applications*. John Wiley & Sons.
- Pantic, M., Valstar, M., Rademaker, R., and Maat, L. (2005). Web-based database for facial expression analysis. In *2005 IEEE international conference on multimedia and Expo*, pages 5–pp. IEEE.
- Pardas, M., Bonafonte, A., and Landabaso, J. L. (2002). Emotion recognition based on mpeg-4 facial animation parameters. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages IV–3624. IEEE.
- Pasquariello, S. and Pelachaud, C. (2002). Greta: A simple facial animation engine.
- Radzihovsky, S., de Goes, F., and Meyer, M. (2020). Facebaker: Baking character facial rigs with machine learning. In *ACM SIGGRAPH 2020 Talks*, SIGGRAPH '20, New York, NY, USA. Association for Computing Machinery.
- Ravikumar, S., Davidson, C., Kit, D., Campbell, N. D., Benedetti, L., and Cosker, D. (2016). Reading between the dots: Combining 3d markers and face classification for high-quality blendshape facial animation. In *Graphics Interface*, pages 143–151.
- Reverdy, C., Gibet, S., and Larboulette, C. (2015). Optimal marker set for motion capture of dynamical facial expressions. In *Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games*, pages 31–36.
- Roble, D., Hendler, D., Buttell, J., Cell, M., Briggs, J., Reddick, C., Iannazzo, L., Li, D., Williams, M., Moser, L., et al. (2019a). Real-time, single camera, digital human development. In *ACM SIGGRAPH 2019 Real-Time Live!*, pages 1–1.
- Roble, D., Hendler, D., Buttell, J., Cell, M., Briggs, J., Reddick, C., Iannazzo, L., Li, D., Williams, M., Moser, L., Wong, C., Kachkovski, D., Huang, J., Zhang, K., McLean, D., Cloudsdale, R., Milling, D., Miller, R., Lawrence, J., and Chien, C. (2019b). Real-time, single camera, digital human development. In *ACM SIGGRAPH 2019 Real-Time Live!*, SIGGRAPH '19, New York, NY, USA. Association for Computing Machinery.
- Rothkrantz, L., Datcu, D., and Wiggers, P. (2009). Facs-coding of facial expressions. In *Proceedings of the International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing*, CompSysTech '09, New York, NY, USA. Association for Computing Machinery.
- Seymour, M., Evans, C., and Libreri, K. (2017a). Meet mike: epic avatars. In *ACM SIGGRAPH 2017 VR Village*, pages 1–2.
- Seymour, M., Riemer, K., and Kay, J. (2017b). Interactive realistic digital avatars-revisiting the uncanny valley.
- SouthKorea, G. Gx-lab. <http://www.gxlab.co.kr/>. (Accessed on 11/02/2020).
- Tinwell, A., Grimshaw, M., Nabi, D. A., and Williams, A. (2011). Facial expression of emotion and perception of the uncanny valley in virtual characters. *Computers in Human Behavior*, 27(2):741–749.
- Valente, S. and Dugelay, J.-L. (2000). Face tracking and realistic animations for telecommunicant clones. *IEEE MultiMedia*, 7(1):34–43.
- van der Struijk, S., Huang, H.-H., Mirzaei, M. S., and Nishida, T. (2018). Facsvatar: An open source modular framework for real-time facs based facial animation. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 159–164.
- Villagrasa, S. and Susín Sánchez, A. (2009). Face! 3d facial animation system based on facs. In *IV Iberoamerican symposium in computer graphics*, pages 203–209.
- Zibrek, K., Kokkinara, E., and McDonnell, R. (2018). The effect of realistic appearance of virtual characters in immersive environments-does the character's personality play a role? *IEEE transactions on visualization and computer graphics*, 24(4):1681–1690.
- Zibrek, K. and McDonnell, R. (2014). Does render style affect perception of personality in virtual humans? In *Proceedings of the ACM Symposium on Applied Perception*, pages 111–115.