22/Dec/2021, Ver 1.52

Digital Object Identifier

A Survey of Facial Capture for Virtual Reality

LIHANG WEN[1], JIANLONG ZHOU[1], WEIDONG HUANG[2], and FANG CHEN[1]
[1]HCAI Lab, Faculty of Engineering and IT, University of Technololgy Sydney, 15 Broadway Ultimo, NSW 2007, Australia
[2]TD School, University of Technololgy Sydney, 15 Broadway Ultimo, NSW 2007, Australia

Emails: Lihang.Wen@student.uts.edu.au, {Jianlong.Zhou, Weidong.Huang, Fang.Chen}@uts.edu.au

This work has been funded by the Research Training Program (RTP) Fees Offset Scholarship from the Australian Government.

ABSTRACT The holograms in Star Wars have inspired many researchers to capture the whole human body in real-time and present it as an avatar into Virtual Reality (VR). Facial capture is important to achieve this because facial expressions are essential for social interaction. However, facial capture works well only when the face is not occluded. While a VR headset has been widely used as a display device for VR, it occludes half of the face and becomes an obstacle. This paper presents a systematic literature review on facial capture for VR. The survey aims to: review the current state-of-the-art facial capture technologies for VR; identify the types of technologies and context of the use, the methodologies, and theoretical groundings; identify research gaps in facial capture for VR; and identify solutions of facial capture for VR headsets. A realism index is defined to evaluate and compare the collected papers. The results show several technology trends in the facial capture for VR: tracking facial motion with markers, facial capture in headsets using cameras or sensors, facial performance capture, and hologram/volumetric capture. It is shown that the Modular Codec Avatar is the best facial capture method in a VR headset, whereas Metahuman has the best output effects. This paper also proposes that an open-design VR headset is an effective approach to lower the Total Cost of Ownership (TCO).

INDEX TERMS Facial Capture, Facial Motion Capture, Facial Performance Capture, Headset, Virtual Reality

I. INTRODUCTION

VIRTUAL Reality(VR) is a simulated experience that enables the user to feel and interact with objects and characters in a virtual environment in real-time. It replaces the user's senses (sight, hearing, etc.) with computer-generated feeds. Merkx and Nawijn [1] stated that VR experiences have multiple purposes and benefits in tourism. Caulfield [2] demonstrated a digital twin of a future car factory that revolutionized the planning process. In [3], it was demonstrated that VR headsets helped the car design team boost efficiency and productivity. In the App "Tilt Brush" [4], an artist can paint a 3D volcano model unlimited colors from the left-hand palette, changing the brush shape and size, painting in 3D space, and drawing fog and snowflakes, which is not possible in traditional painting media. VR systems are also used in education, architecture, and medical applications. Telepresence refers to a set of technologies that allow a user in a remote location to be seen, heard, and sensed by other people in a local environment, and the user can feel this local environment. Telepresence enables people to talk, see and interact with each other remotely. In this way, they can save the cost of long-distance travel. A well-known example is the holograms in the science fiction Star Wars. Heller [5] stated that business users of virtual meetings "can use their valuable time much more efficiently by using telepresence solutions from their particular offices". It also stated that in cases of natural disasters, epidemics or war, telepresence could provide the opportunity to virtually communicate with colleagues or business partners while avoiding traveling to dangerous places. Telepresence is more critical in the COVID-19 pandemic because it retains the interaction between people without physical contact, reducing inter-person disease transmission. VR is often used in telepresence because it is realistic, three-dimensional, and offers many control methods.

VR headset (also known as Head Mounted Display, or

HMD) systems have many advantages over CAVE systems and are good VR solutions for individuals. Mallaro [6] used a Vive VR headset and a CAVE system to study pedestrian road-crossing behavior, and found that both systems were effective in presenting the road-crossing task, and the headset offered some benefits: inexpensive, portable, easy to use, and simple to maintain. Manuelraj [7] compared VR headsets and CAVE and stated that CAVE systems had higher resolutions, had a larger Field Of View (FOV), were comfortable for users, and could be viewed by multiple users, while the headsets were more immersive, low cost and easy to set up, carry and store, offering a much smaller footprint and more ways for interaction. The advantages for the choice of the headset system by the individual end-users are obvious: It is more immersive, small enough for a household, and affordable at hundreds of dollars, while a CAVE system is too large and costs tens of thousands. Because a CAVE system also has a pair of head-tracking glasses, many principles and techniques reviewed in this paper can be applied to a CAVE system as well.

VR headsets have gained popularity. Some headsets can track hands and controllers to improve interactions. VR headsets are now moving from the early majority stage towards full market adoption. According to Bill Myers from S3 Technologies [8], in 2017, a VR arcade station that used an HTC Vive headset and a high-end computer cost around $2,500. Now, consumers can have the same experience with a standalone headset for just a few hundred dollars. This dramatically improves user experience and the market size.

There are various concepts related to facial capture. The movements of a person's face are called facial motions. When we investigate its psychological aspects (neutral, happy, angry, disappointed, sad, etc.), they are called facial expressions. Facial capture is also known as facial motion capture, and it is the process of electronically converting facial motions into a digital database using cameras or other sensors. When it is used to analyze the expressions of a subject, it is called facial expression capture. When the purpose is for filming or performance, it is called facial performance capture.

Facial capture is essential for communication, telepresence, and social VR. For presenting an avatar properly in VR, it is necessary to track the body, head, gesture, and facial motions. Body, head, and gesture tracking are relatively mature, and some commercial products are available. However, facial capture and tracking are still under development. Xu et al. [9] states that facial expressions can display personal emotions and intentions, which are critical in a social situation. In [10], the Latin proverb states: "the face is the portrait of the mind; the eyes, its informers". Gunkel et al. [11] investigated social VR use cases involving 91 users and 4 types of applications. The research found that comic-like avatars were not beneficial for business meetings. It was also found that "the two most interesting applications to users (i.e., education and video-conference) are those that involves a lot of face-to-face conversations or interactions in non-remote/real-world settings".

Facial capture can help increase the adaptation of VR headsets. Herz and Rauschnabel [12] found that consumers consider these four factors important when purchasing VR headsets: Wearable comfort, making life more efficient, entertainment, and data privacy. Facial capture can help in the efficiency and entertainment aspects. It can enable face-to-face conversation in business VR meetings and those between family and friends. It can enable students to see the teacher's expressions while the teacher is demonstrating complex concepts using interactive VR models. This can make life more efficient. In VR games, with facial capture, players can see the excitement of their team members and the pain on the opponents' faces. This will make entertainment more enjoyable.

However, the current VR headsets in the market make facial capture very difficult. First, the headset occludes the user's upper face, preventing it from being seen from outside. Second, the headset touches the user's face. When the user smiles, it has constraints, and is not comfortable. On the other hand, feeling uncomfortable, the user is less likely to smile or make facial motions. Sometimes the pressure on the face affects the circulation of blood in the facial skin, which makes it very uncomfortable to wear. When the user takes it off, red marks are left on the face. Mainstream VR headsets are trying to tackle these problems. HTC recently announced an accessory device called "Vive Facial Tracker", which only supports 38 motions and cartoon characters. Oculus (Facebook) is doing facial tracking through "Codec Avatars" which will be discussed in more detail below. However, this technology is still far from consumers.

Therefore, we seek to answer the following three research questions in this survey. **RQ1:** How can we capture the facial model while the headset occludes the user's face? **RQ2:** How can we make the headset more comfortable and allow the face free to make any expressions? **RQ3:** How can we lower the cost of facial capture as well as VR headsets and make them affordable to the public?

II. LITERATURE COLLECTION AND ANALYSIS

We used these keywords to search different databases: facial motion capture, facial expression capture, facial performance capture, and volumetric video capture. Facial performance capture and volumetric capture are added because: First, they can be applied to the lower part of the face in HMD settings, which is not occluded at all. Second, they can increase the number of papers that contain similar types of questions, methods, and principles related to facial capture. Third, they can be used to collect data as a baseline or ground truth, or to train models for further use in HMD settings. Furthermore, volumetric capture is ideal for VR content creation, which is an important aspect of VR.

The databases mainly used in this literature review are: ACM Digital Library; IEEE Xplore; Google Scholar; Scopus; Elsevier; ProQuest; Springer; and Google Search Engine. When reading these papers, we also reviewed the key

references and added them to our collection if they were related to facial capture. We collected more than 100 papers and other materials for this research topic. Some of them were eliminated because: they were too old or obsolete; they were already widely accepted; duplication; demonstration or a simplified version of another paper; they were not related to facial capture. After elimination, 82 papers were included in this study. Among them, 25 papers were kept but not analyzed because: the data or statements were good for quotation purposes; the main topic was not facial capture. 57 papers were analyzed for methods and types in the tables. Note that for simplicity, not all papers in the tables were described in the text. They were published between 2004 and 2021. Figure 1. shows the distribution of these papers by year.

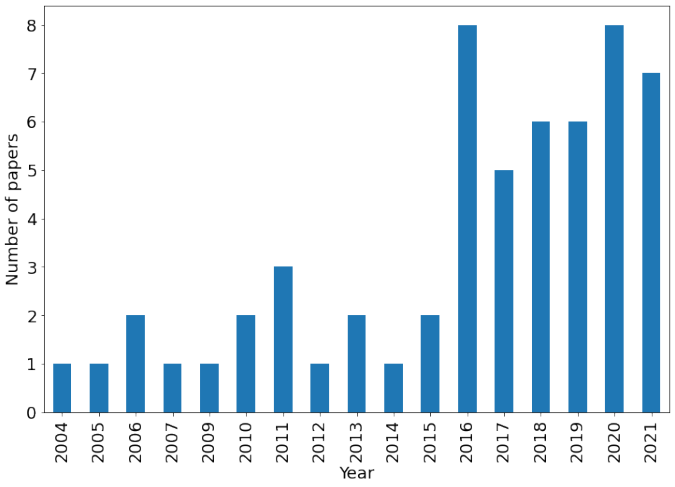

FIGURE 1. Number of papers by year

Chronologically there are three stages of facial capture. Before 2008 (inclusive), using markers were the most effective way to track facial motion and performance. From 2009 to 2017, because of the advancement of hardware, software, algorithms, and machine learning, many new technologies have been applied in facial capture. From 2018 until now, Codec Avatars and volumetric capture have matured. There are several different technology trends in facial capture for VR. The first trend is head-mounted devices (sensors and cameras). With these devices, the cameras are closer to the face, and the face is always in the exact location in the video. The second trend is facial performance capture, mainly without a headset. The third trend is volumetric capture, which captures everything in 3D within a specific range, including the face and the whole body.

To identify how realistic the facial capture results are and to compare between papers, we define a realism index ranging from 1 to 9 (see Figure 2) as follows: 1–Tracked dots on the face; 2–Lines connecting dots; 3–Cartoon; 4–Smooth model without texture; 5–Detailed model without texture; 6–Facial model with partial texture; 7–Facial model with full texture; 8–Facial model with realistic texture, and detailed features such as eyes, teeth, tongue, wrinkles, and

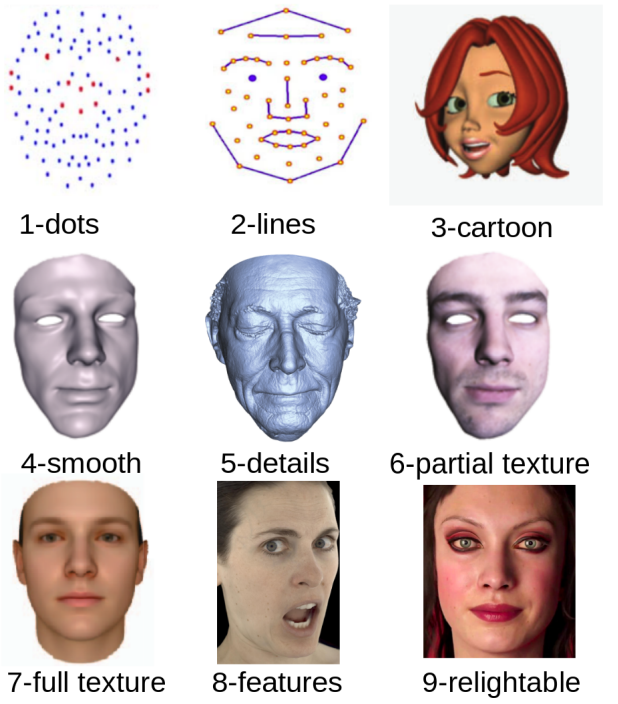

FIGURE 2. Realism Index (images from [13]–[20])

pores; 9–Detailed facial model with texture and other maps, fully relightable. Note that the following features can be valued for additional points: 0.5–Unlimited facial expressions; 0.3–Realistic dynamic movement; and 0.2–Translucent skin. Some of these methods and processes are not real-time and are denoted by a "-" sign beside the score.

III. FACIAL CAPTURE USING MARKERS OR HEADSETS

Markers are fast and effective. Sensors and cameras mounted on a headset have the following advantages: the signals are stable, and the relative position of the face in the video is permanently fixed. Table 1 shows the methods of facial capture using markers or headsets.

A. USING MARKERS

Markers are effective in tracking facial motion. These markers refer to special markers attached to the actor/subject, which are tracked by video cameras. In the early years, markers were used to track key points of the face for facial performance capture. For example, in [13], 102 markers were used to capture faces and were aligned with videos to record the performance. In [21], 80-90 markers were placed on the face and captured by eight cameras at different scales. The wrinkles were also marked and tracked. This marker-tracking technology is quite mature and is still widely used in the film industry.

However, these markers are not suitable for VR end users. Colyer et al. [41] stated that marker-based methods have these shortcomings: long participant preparation time; the potential for erroneous marker placement or movement; the

TABLE 1. Facial Capture using Markers or Headsets

Type	Paper and Description	Context	Method	Advantages	Disadvantages	realism Index
Markers	[13]:102 markers, align with video [21]:80-90 markers, track wrinkles	Entertainment, Animation	[13]:PCA,RBF [21]:RBF	Fast and accurate	long time to setup	[13]:3.5- [21]:6-
Sensors	[22]:5 IR sensors [23], [24]:14+2 optical IR [25]:8 EMG sensors	HMD	[22] Semi-empirical equation [23]Classification [24]: Multi-class classification, ReLU [25] LS-SVM	Low cost, simple and fast	Only 4-6 expressions	3
Sensors + HMC	[26]:IR eye-tracking cameras	HMD, VR broadcast	Translucent-composition	Simple and fast	No expressions	3.2
Sensors + HMC	[27]:8 strain gauges + RGB+D	HMD	GMM	Support many expressions. Low cost, simple and fast	Training per user; Calibration per use.	4
HMC	[28]: 2 IR on the whole face	VR broadcast	HMC + Camera dome	Rigged, realistic	Not VR HMD	9
HMC	[29]:ToF IR + Depth, For lower part of face only.	HMD	Blendshape	Simple	Driving a cartoon character, unrealistic	3.5
HMC	[30], [31]: 2IR(eyes)+ RGB(lower face)	HMD	[30]: SfM, eye and face synthesis [31]: CNN, weighted blendshapes	[30]: Output to video [31]: 3D render	[30]: Video but not 3D [31]: No texture	4
HMC	[32]–[34]: 3 IR HMC, 2 IR for eyes, 1 lower face. [35]: Output method [19]:Eye tracking [36]: Relightable models	HMD	AAM, CVAE, Codec Avatar	Realistic, [35]: Decoder is faster for multiple avatars	Expensive to build	[32]:8 [33]–[35]:8.5 [19]:8.6, [36]:9
HMC	[37]: IR for eye-tracking, 3D Stereo Camera for lower face(not HMC)	HMD	AAM, regression, classification	Render in stereo video	Not rendered to 3D	7
HMC+filters	[38]: Fisheye view of whole face	HMD	Polarizing filter	RGB video of whole face	Half transmittance, distortion	N/A
HMC+filters	[39]: IR for each eye, RGB for lower face.	HMD	IR-cut filter, colorizing neuron network	Good view of eyes	Fitting per user	7
HMC+filters	[40]: IR view for whole face	HMD	IR-pass filter	Whole face	Greyscale, filter edges	N/A

PCA: Principle Component Analysis; RBF: Radial Basis Function; GMM: regression Gaussian Mixture Model; SfM: Structure from Motion; AAM: Active Appearance Model; CVAE: Conditional Variational Auto-Encoder; ReLU: Rectified Linear Unit

unfeasibility of attaching markers in certain settings; physical and psychological constraints. These are also true for the case of facial capture for VR end users. It takes 30 to 60 minutes to attach markers to the face. That may be acceptable for commercial purposes because their usage may take hours or even a whole day. However, many VR applications (gaming, conference, and education) last 10 to 40 minutes. In these cases, it seems that the 30-minute setup is a high cost and is not acceptable. Thus, we focused on methods without markers.

B. USING SENSORS

Sensors are low cost and are often used to detect skin movement, but they lack details, and the output is too simple and limited. The correspondence between the sensor signals and facial expressions can be established. Most of these sensors require contact with the facial skin to detect signals. In [22], IR light was chosen because it could pass through several layers of the human skin and had lateral propagation characteristics. These IR emitters and receivers were contactless. Li et al. [27] used eight ultra-thin strain gauges and one RGB+D camera. For each subject (individual), the headset must be trained once. A short calibration was required for each use. A Gaussian Mixture Model and a linear regression model were trained to process the data and combine them. The result was a realistic blend shape model. Suzuki et al. [23] placed 14 contactless IR optical sensors around the eyes, and another two were attached to the bottom of the headset for the cheek. A 3-layer perceptron network was trained to classify the facial expressions. Later, Suzuki et al. [24] described the above process in more detail and applied a three-layered feed-forward neural network. The training was required for each user. Lou et al. [25] used eight electromyography (EMG) sensors and built the correspondence from signals to FACS codes. For each user, a single photo was taken with a neutral expression and inverse-rendered into a multi-linear PCA face model. The training data were collected from 15 volunteers with five common facial expressions. The Root Mean Square (RMS) and Least-Square Support Vector Machine (LS-SVM) were used as the classifier. It could output six of the Action

Unit (AU) and drive a facial model.

Frueh et al. [26] presented "Headset Removal", a modified headset with an eye-tracking feature that yields improved output. First, the user's face was captured using an RGB-D camera. The eye gaze direction and blinks were recorded in a database. Second, QR code markers were placed on the headset for tracking. Third, the eye-gaze parameters from the headset were used to retrieve a model from the database. Fourth, the corespondent upper-face model was aligned with the actual face in this video, along with a transparent headset to eliminate the "uncanny valley". In this video, the headset appeared transparent. No facial expressions were used in the upper facial model.

C. USING HEADSET MOUNTED CAMERAS

In recent years, due to the development of hardware, especially advanced cameras and faster processing power, it is possible to capture facial motion in video and perform real-time processing. Edwards et al. [42] used the Active Shape Model(ASM) to determine the boundary of a face in an image. In [43], the Active Appearance Model(AAM) was defined. These two laid the foundation for many later studies. In March 2013, the startup company Oculus shipped its first developer kit(DK1). Moreover, in mid-2014, developer kit two(DK2) was shipped out. The VR headset became a big hit, and researchers began to add facial capture functions to this device. Li et al. [27] used both sensors and an RGB+D camera to achieve many expressions. Yu and Park [29] used only one ToF IR+D HMC to capture the lower half of the face and drive a cartoon face. This required one calibration for each user. Zhao et al. [30] used two IR HMCs to track the eyes and one RGB camera (not mounted but away from the user) to capture the lower face. Three cascaded learning networks were used to track the eyes and lower face. The online processing time for one frame was approximately 560ms on an Intel i7-4710 CPU (3.4GHz). Thies et al. [37] used multi-view (RGB+IR) to drive the model for the same person to be displayed in VR. Olszewski et al. [31] used two IR cameras(eyes) and one RGB camera(lower face). A microphone captured the voice to synchronize the mouth movement. Convolutional Neural Network (CNN) regressors were trained separately for the mouth and eyes. The output mouth and eye control weights were used to drive an animated character online.

Some papers from Facebook Reality Labs showed significant results. In [32], 40 cameras were mounted on a hemisphere and could capture the user's face with pore-level details. Multiview stereo reconstruction was used to construct the facial model. Each research subject/individual was asked to perform 122 facial expressions and recite 50 sentences. This large database formed the basis for further development. Each facial model with texture and mesh was converted to a code and vice versa. A Conditional Variational Auto Encoder (CVAE) was used. One key component of the CVAE is view-dependent texture synthesis. In the headset, only three IR HMCs were used. Views from three HMCs were synthesized to match the real images. When a match was found, the corresponding code was used to synthesize the facial model. The output was compared with linear and bi-linear models and was better in quality and had fewer parameters. Wei et al. [33] improved the Deep AAM by using more HMCs for training, and IR images were pseudo-rendered before the matching. The matching performance was more precise than that in [32]. The result was significantly better with detailed nuances. The algorithm was improved and supported funny expressions that were not in the samples. Although it could interfere with background flash and sometimes weaken the emotion, it showed improvements in expressiveness and robustness. In [34], the Deep AAM was named Codec Avatar, and Modular Codec Avatar (MCA) was defined by several high-level formulas. Schwartz et al. [19] added eye-tracking features. Ma et al. [35] improved the decoder to display multiple avatars faster. Chen et al. [36] improved the building process by including different lighting, and thus made the avatar relightable.

D. USING FILTERS TOGETHER WITH CAMERAS

Sometimes special filters are used in the headset for better capture. These filters work on a different wavelength, thus will not interfere with the visual light from the RGB screen. Yamada et al. [38] used polarized filters to enable one-way video capture in the "Selfie-Mask" headset. The user could only see the display but not outside, while the outside camera could see the whole face of the user. However, display lights were reduced, and the full-face video was a bit distorted. Rekimoto et al. [39] used one RGB camera for the lower face and two IR cameras for the eyes. IR-reflective screens were placed in front of each eye, resulting in a better viewing angle. Chiba et al. [40] used IR-pass filters to block the visible light around the display. An outside IR camera could capture the entire face with just natural lighting.

IV. FACIAL PERFORMANCE CAPTURE

Facial performance capture is often used in filmmaking. Usually, the cameras are more than 30 cm away from the face, so the video is less distorted. The facial motions are then transferred to a computer-rendered character. According to the type of input device, facial performance capture can be divided into monocular depth sensor, monocular RGB video(Mono-RGB), and multi-view. The former two are important because they are available to the public through smartphones. They provide the baseline to build the initial facial models for an individual to implement facial capture in HMD settings. Table 2 shows papers related to facial performance capture, mostly without a headset.

A. MONOCULAR DEPTH SENSOR

A monocular RGB+D senor can capture the facial model in detail and precision but is not very realistic. Wang et al. [44] captured an area of 260×244 mm with an error of less than 0.05mm. It could capture a facial model at a coarse level (1000 node mesh) or a refined higher level (8000 node mesh).

TABLE 2. Facial Performance Capture

Type	Paper and Description	Context	Method	Advantages	Disadvantages	realism Index
Mono-Depth Sensor	[44]RGB+D	Entertainment and Animation	LLE	High precision	Narrow field of view (FOV), low resolution	7
Mono-Depth Sensor	[16] RGB+D, expression transfer to another model [45] RGB+D(Kinect),39 blendshapes	Entertainment and Animation	[16] PCA, Rigid Reconstruction, Optical Flow [45] Rigid Tracking, EM, MPPCA	[16] Rigid, mouth and eyelid tracking, chin aligned [45] Rigid	[16]Projector too bright [45]Low resolution, lack detail	6
Mono-Depth Sensor	[46]RGB+D(Kinect), 150 subjects aged from 7 to 80, 20 expressions each	General CG	ASM, PCA, Facial rigging, blendshape	Low cost, mobility	Low resolution severe noise	7
Mono-RGB	[47]Re-targeting and relighting	Entertainment and Animation	SHBMM, de-lighting, relighting	Few images as input	Not real-time	7-
Mono-RGB	[48]Inverse Rendering	General CG	3DMM, PCA	Complete framework, multilinear system	Complex pipeline	7-
Mono-RGB	[49]: 3D regression [50]:Face Tracking [51]: 46 FACS AUs	General CG	[49]: 3D Regression [50]: DDE [51]: Local regressor	[49]: Easy,robust [50]: General training [51]: Wrinkle details	[49]: Training per user [50]: No wrinkles [51]: Lighting dependent	[49]:6 [50]:4 [51]:5
Mono-RGB	[52]:Video reenactment	Entertainment and Animation	Mouth database and synthesis, non-rigid model-based bundling	Photo-realistic	Not 3D output	7
Mono-RGB	[53]: Eye gaze animation, multi-linear expression deformation model	General CG	LBF, Randomized Forest, double eye-gaze constraints	Eye and pupil tracking	Cartoon charactor	4.5
Mono-RGB	[54]:4-level hierarchical reconstruction	General CG	AAM, SfS, hierarchical reconstruction	Wrinkle details	Relies on landmarks	7.5
Mono-RGB	[55]: 3DFaceNet, generated training set	General CG	CNN, Inverse Rendering, Refinement	Fine detail with texture	inaccurate reflection and self-shading	6
Mono-RGB	[56], [57]: Video reenactment	Entertainment and Animation	DCNN, NMFC, mouth synthesis [57]: DenseFaceReg, eye synthesis	Realistic	2D output	6-
Mono-RGB	[58]: Geometry and texture streams	General CG	ASM, SSRPM	Semantic Region Stylization	Low to medium resolution	6
Mono-RGB	[59]: Monocular videos of actor	Entertainment and Animation	Deep CNN	Fast, 287fps(real-time)	Training per user	5
Mono-RGB	[60]: Relightable model from single image	General CG	GANFIT, pix2pixHD, CNN	Open-source, Specular map	Not well on dark skin, minor alignment errors, quality dependent	7
Multiview	[61]: 7 pairs of stereo RGB [62]: merged and enhanced [17]: anchor-based reconstruction	Entertainment and Animation	[61]: Frame Propagation, mouth tracking [62]: Mesoscopic Augmentation [17]: Frame anchoring, image-space tracking	Pore-level details [62]: Versatile [17]: Robust	[61]:Fast motion error, [62]: static model, lighting sensitive	[61]:7.6- [62]:8.0- [17]:8.2-
Multi-view	[63]: 5 high-speed cameras synchronized with lighting	Entertainment and Animation	Active lighting, heuristic diffuse / specular separation	Automatic process, novel viewpoints, relightable	occasional spike, computing intensive	8.8-
Multi-view	[64]: 2 RGB HMCs, 3 cameras on rig	Entertainment and Animation	Compared HMCs to static rig	Multi-dimensional regression	Actor specific rig and regressor	6.5
Multi-view	[65]: Starline virtual meeting	Communication	Active depth sensing, 3D display	Portable, realistic	Expensive, prototype level	8.5

LLE: Locally Linear Embedding framework; SHBMM: Spherical Harmonic Basis Morphable Model; EM: Expectation Maximization; CNN: Convolutional Neural Network; DCNN: Deep Convolutional Neural Network; MPPCA: Mixtures of Probabilistic Principal Component Analyzers; DDE: Displaced Dynamic Expression; NMFC: Normalized Mean Face Coordinates; ASM: Attribute Spatial Maps; SSRPM: Shared Semantic Region Prediction Module (SSRPM) .

Then a model was designed to decompose content (non-user-specific expression) and style (user-specific) from the 3D facial model. From this model, new expressions could be synthesized for an individual, or dynamic morphing could be generated from one individual to another. In [16], a personalized linear facial model was built for each actor. Weise et al. [45] used an RGB+D(Kinect V1) camera to capture the user's face, convert it into blendshape weights, and drive a digital avatar in an animation. Fifteen user-specific expressions and 39 blendshapes were used. In [46], a Kinect V1 was used to capture expressions from subjects. Four example applications were demonstrated using these models: facial image manipulation, face component transfer, real-time performance-based facial image animation, and facial animation retargeting from video to an image.

B. MONOCULAR RGB CAMERA

Cao et al. conducted extensive investigations on real-time performance. In [49], a novel 3D shape regression algorithm was developed. In the first frame, a 2D shape regressor was applied to locate the landmark positions. A shape regression model was trained for each actor and was used to generate blendshapes through iterations. In [50], the process was improved so that user-specific training was not required. A Displaced Dynamic Expression (DDE) model was developed for this purpose. In [51], local regression was used to augment a global mesh model with winkle details.

In [52], the actor's facial performance was transferred to another subject in a photo-realistic manner. Wang et al. [53] used a pre-trained user-dependent iris and pupil pixel classifier to perform eye tracking together with the Maximum A Posterior (MAP) framework. Multi-linear expression deformation models were used to reconstruct the 3D facial model. In [54], a high-resolution RGB camera (Logitech C922x) was used to capture the actor's face. The same algorithm from [53] was used to reconstruct the facial model with 5.6k vertices and 33k triangle faces. A hierarchical approach was used to subdivide the mesh model into a 4-8 subdivision scheme at each level to enhance the mesh. With the normal map from the mesh, an albedo map was calculated for the lighting estimation. Wrinkles and folds were interpreted as albedo changes. The system was run at 50 fps on a PC with a GPU. In [55], a framework named 3DFaceNet was proposed, which consisted of three CNNs: a CoarseNet for the first frame and a single image, a Tracking CoarseNet to track between frames, and a FineNet to enhance the coarse mesh with fine-scale details. An optimization-based inverse rendering process was used to generate training data from the existing datasets. From the RGB video, this 3DFaceNet could recover detailed geometry, albedo, lighting, pose, and projection parameters in real-time. Lattas et al. [60] captured 7 expressions from 200 individuals into a dataset called "RealFaceDB". The final model used GANFIT on an input image to estimate the initial 3DMM, then used RCAN to up-sample the texture map. This was then delighted into a diffused albedo map. Subsequently, a diffused normal map and two specular maps (albedo and normal) were generated from the trained networks. The result was a view-independent facial model that could be realistically relighted under different lighting conditions.

C. MULTI-VIEW PERFORMANCE CAPTURE

Multi-view performance capture can retrieve more details than monocular performance capture. However, the extensive input data takes longer to calculate, making some of these methods not real-time. Derek Bradley et al. from the University of British Columbia (Canada) published a series of papers on 3D and facial capture. In [66], they enhanced the Multi-View Stereo algorithm and achieved a high performance and well-shaped mesh. Due to its low resolution, it was suitable for fabric and sculptures. Later on in [61], this technique was used with high-resolution cameras to capture seven parts of the face, and merge them into a detailed model. In [62], seven stereo cameras (Fuji Real 3D W1) were used to construct a passive stereo vision system. The resulting error was less with mesoscopic augmentation and was more accurate than the existing solution (PMVS [67]). It only required an initial manual focus. Beeler et al. [17] improved [62] from the previous single-shot mode to video mode. The cameras were upgraded to Dalsa Falcon 4M60. and synchronized. One reference frame and some anchor frames were used to track other frames and refine the motions. Fyffe et al. [63] used five high-speed and high-resolution cameras (Phantom v640, up to 1500fps) to capture the actor's face under controlled lighting. The reflectance and geometry were estimated under different lighting conditions. The resulting model included the facial model, diffuse albedo, surface normal, and specular albedo. This could be rendered under different lighting conditions with realistic outcomes. In [64], two cameras were installed on a head-mounted rig to capture the user's face. A cascaded regression scheme was trained for the user to estimate the facial model in real-time. The result was a high-quality real-time facial model. Laine et al. [59] trained a deep learning network (12-level deep CNN) with multi-view RGB videos. The 5-10 minutes of video footage consisted of extreme expressions, FACS-like expressions, pangrams, and in-character material. The output facial model was better than the other two state-of-the-art methods.

V. VOLUMETRIC CAPTURE

Volumetric capture usually involves multi-view set-ups of RGB, RGB+D, or depth cameras, and sometimes even pattern projectors. Due to a large amount of processing, these systems are mostly non-real-time. Table 3 lists papers related to volumetric capture. In [68], Microsoft Research presented the real-time "Holoportation" system which consisted of 8 pods, 8 GPUs on 4 PCs. To solve the interference, random dots were used and treated as reference points for further details. A prototype called "Visor Removal" was implemented by capturing each eye from a camera installed at the corner of the visor, projecting the video as a texture onto a reconstruction of the user's face. In [69], Holoportation was

TABLE 3. Volumetric Capture and Other Trends

Type	**Paper and Description**	**Context**	**Method**	**Advantages**	**Disadvantages**	**realism Index**
Volumetric	[68], [69]: 8 RGB+D (Stereo IR), Holoportation	Communication	Visor removal, active stereo, [69]: NLP	Real time Volumetric	High band-width, expensive*, no eye-contact	8
Volumetric	[70]: Distributed system with 1 centre and 4 RGB+D units	Entertainment and Animation	CNN, Multi-task learning	Open-source	Low resolution	7-
Volumetric	[71]: Huge (10,000 square feet) dome, 90 servers	Entertainment and Animation	Multi-view capturing, synchronized cameras	High resolution	Huge cost*	8.8-
Volumetric	[72]: 58 RGB + 32 structured IR, 331 programmable lights, room-scale	Entertainment and Animation	Active depth sensing, Deep Learning Based Segmentation, Reflectance Maps Generation	High resolution, relightable	High cost*	9-
Volumetric	[73]: 46 4K RGB camera dome, 70 cm viewing baseline	Entertainment and Animation	Multi-Sphere Images, Layered mesh model	Portable, broadcast level	Limited views	8.5-
Statistics	[74]: Synthesis of facial detail	Entertain-ment and Animation	Markovian techniques, steerable pyramid, Gaussian noise	Pore-level details, aging and de-aging	Plausible output	5-
Statistics	[75]: Facial assets	Entertain-ment and Animation	FACS, Influence Map	Details, teeth/tongue modeling	Large training data	8-
Output /Render	[20], [76]: Real-time facial rendering	Entertainment and Animation	Opacity map and pore mask	Hyper-realistic	High processing needs	9.2-
Output /Render	[77]: The Heretic demo short film	Entertainment and Animation	Cavity map, eyes and teeth modeling	Realistic	Lack opacity	9-

NLP: Natural Language Processing. *: The costs are for the capture system.

improved by combining Natural Language Processing (NLP) and Text To Speech so that the hostess could present the talk in any language. Shrestha et al. [78] solved the interference problem by modulating the ToF cameras at different frequencies to cancel the interference out. Sterzentsenko et al. [70] created a low-cost system with four RGB-D units, which send data to a center "orchestrator" through LAN. In [71], a huge dome system was able to capture volumetric video for film making at a rate of over one terabyte of data every 10 seconds. Guo et al. [72] developed the "The Relightables" system, which has a geodesic sphere fitted with LED lights, cameras, and custom-designed depth sensors. The final model had a specular map and was relightable. Broxton et al. [73] presented a portable VR/AR broadcast system with 46 low-cost cameras (Yi 4K) on a 92-cm diameter hemisphere facing outwards. The acquired multi-view video was processed into a layered mesh representation.

VI. OTHER TRENDS

Statistics were used for facial capture. In [74], statistical models were used to enhance details and generate aged or de-aged geometries regarding ages, genders, and races. In [75], a whole set of dynamic facial assets could be generated from a single facial scan. A high-fidelity facial scan database (178 subjects, each with 19 to 26 different expressions) was used to train a Blendshape Generation network and a Texture Generation network. The personalized blendshapes were augmented with facial components and template blendshapes, and combined with the dynamic textures to make the final output that contains the expressive models of the actor.

In addition to facial capture, the output (rendering) method is also vital for VR to persuade people's eyes that this is a real human being instead of a cartoon character. Real-time game engines are quite advanced in rendering human faces. Pedersen et al. [77] described the render pipeline of facial model in Unity 3D. Humphreys [20] used "MetaHuman Creator" (based on Unreal Engine) to create a hyper-realistic virtual human called "Dana". It is state-of-the-art and far beyond competitors. According to [76], the skin is translucent and contains pore-level details.

VII. DISCUSSION

This systematic review of facial capture for VR is based on a sample of 57 papers and materials from typical databases such as ACM Digital Library and IEEExplore, as well as through search engines and the Internet until June 2021. Facial tracking markers have been used for a long time. They are mature, fast, and often used for film production. However, the long preparation time and physical/psychological constraints make them unsuitable for VR end users. Sensors can be used in VR headsets to track facial motions, either in contact or contactless. They are low cost and easy to implement. However, the output expression is sparse, limited, and unrealistic. This is because only several (4 to 6) expressions can be supported by most sensors.

Monocular RGB cameras can be used to capture facial models and motions, and are widely available. Since 2014, more than 1.2 billion smartphones were sold every year [79], all equipped with RGB cameras. They are almost accessible to everyone. The monocular RGB video of a user contains

much redundant information. Basically, regardless of what expression the user makes, the facial skin textures mostly remain the same. Only the shape and blood circulation may change. If the user holds the same expression, then the shape does not change, and only the posture and lighting conditions may change. Using the methods in [60], we can retrieve facial shape, texture (albedo map), specular map, etc. from these videos. With some close-up shots, it is possible to retrieve pore-level details.

Monocular RGB+D cameras can also be used for facial capture. In a market report [80], the smartphone depth sensor market size was valued at $561 million in 2017 and is projected to reach $9,280 million by 2025, registering a compound annual growth rate(CAGR) of 42.3% from 2018 to 2025. Apple iPhone 12 is equipped with a depth sensor, and sales were $47 billion in Q2 2021 [81].

Other RGB+D sensors are also available on the market, such as Intel Realsense, Structure Sensor, and Microsoft Kinect Azure. Some of them are used for facial recognition, while others are used for robot vision. The depth channel can provide more details, which makes the process faster and more accurate. Binocular RGB video can be used for facial capture, but the setup requires calibration and is suitable for professionals and studios.

The most promising technology for facial capture for VR is multi-view capture in HMCs. Some HMCs for the eyes are from a narrow-angle, which do not provide enough details, especially when the eyes are looking away from the camera. This needs to be addressed in future studies. According to our comparison of the realism index values among these HMC techniques, the Modular Codec Avatar presented in [34], [35] and [36] is state-of-the-art. However, building an initial model for a Codec Avatar(CA) is expensive. Volumetric capture or the camera dome is too expensive for the public. This burden can be lowered by using facial performance capture technologies with AI and statistics. Many photos and videos can be collected on social media for the same user, which can be used as a database for AI to build CA for the same user. Facial performance capture using monocular videos is a mature technology. It is possible to use one RGB camera or one RGB+D camera, which is widely available on mobile phones, to collect data for the CA. In addition, building a CA requires many samples. Currently, we are in the COVID-19 pandemic, and video conferences are widely used for safety reasons. If these videos can be used in the facial capture and development of the CA, much time can be saved. A key component of CA is view-dependent texture synthesis. However, the texture is view-dependent mostly because of specular reflections. It could possibly be simplified as a view-independent model using specular maps [60]. From the computing point of view, mobile phones and laptop computers have limited computing power, so some of the calculations can be shifted to the cloud. These are future research topics to be addressed.

Volumetric capture is an expensive setup that is used only in studios. Most of these high-fidelity systems are too expensive and difficult to access for the public, such as the Light Stage system and "the Relightables" [72].

One aspect of facial capture is data privacy. Herz et al. [12] stated that data privacy is a key factor when consumers purchase VR headsets. A database of facial capture includes the details of a user's face at different ages and times. It knows and remembers the user's face better than anyone else. Strong authentication should be applied to ensure that a facial model is not manipulated by a person other than the owner.

The ergonomics of the VR headset must be improved. As stated before, the physical constraints of the headset on the skin make it uncomfortable for the user. The solution is to keep the headset contactless from the face so that the facial motions are not constrained. Yan et al. [82] found that any headset above 200g would cause at least slight discomfort on certain parts of the head. Given that the mainstream headsets are above 500g, there is a lot to do to reduce the weight.

Finally, the VR headsets' Total Cost of Ownership(TCO) is still high. A high-quality VR headset system costs from $400 to $3,000. Once purchased, the headset value decreases rapidly. When the next version emerges, the old headset quickly becomes absolute. It is better to have an open-design, to make hardware and software compatible with different manufacturers. Therefore, the upgrade process will be more gradient and smooth, and the VR headsets will be affordable to everyone.

VIII. CONCLUSION

For answering **RQ1**, this survey proposed the use of multiple inward-facing HMCs and deep learning methods such as the Modular Codec Avatar to capture and present the facial model. For **RQ2**, this survey suggested enabling the headset and sensors to be contactless from the face and improve their ergonomics. For **RQ3**, it is possible to use a single high-resolution RGB camera to collect data to build the Codec Avatar model. This would make Codec Avatars widely available. An open-designed VR headset can lower the TCO.

Several potential research gaps have been identified. In the Codec Avatar, facial expressions are weakened for some extreme emotions, which still cause an uncanny valley. More data can be collected, and the model can be improved. To build the Codec Avatar, we can use RGB video or reuse existing photos and videos from social media and video conferencing.

REFERENCES

[1] Celine Merkx and Jeroen Nawijn, "Virtual reality tourism experiences: Addiction and isolation," Tourism Management, vol. 87, pp. 104394, 2021.
[2] Brian Caulfield, "NVIDIA, BMW Blend Reality, Virtual Worlds to Demonstrate Factory of the Future," [Online].Available:https://blogs.nvidia.com/blog/2021/04/13/nvidia-bmw-factory-future/, 2021.
[3] Magna International Inc., "Virtual reality: From vehicle concept to production," [Online].Available:https://www.magna.com/innovation/driven-people-driving-change/article/virtual-reality-from-vehicle-concept-to-production, 2021.
[4] Google, "Tilt brush: Painting from a new perspective," [Online].Available: https://www.youtube.com/watch?v=TckqNdrdbgk, 2016.
[5] Rosi Maria Heller, Telepresence: a modern way for collaborative work, Diplomica Verlag, 2010, OCLC: 897049557.

[6] Sophia Mallaro, "A comparison of head-mounted displays vs. large-screen displays for an interactive pedestrian simulator," 2018, p. 10, University of Iowa.

[7] Reghill J. Manuelraj, "Hmd vs cave in the world of vr," [Online].Available:https://medium.com/xrpractices/hmd-vs-cave-in-the-world-of-vr-a0c9cbfb435a, 2020.

[8] Logan Kugler, "The state of virtual reality hardware," Commun. ACM, vol. 64, no. 2, pp. 15–16, 2021.

[9] Qiang Xu, Yaping Yang, Qun Tan, and Lin Zhang, "Facial expressions in context: Electrophysiological correlates of the emotional congruency of facial expressions and background scenes," Front. Psychol., vol. 8, pp. 2175, 2017.

[10] Kerstin Ruhland, Sean Andrist, Jeremy Badler, Christopher Peters, Norman Badler, Michael Gleicher, Bilge Mutlu, and Rachel Mcdonnell, "Look me in the eyes: A survey of eye and gaze animation for virtual agents and artificial systems," in Eurographics 2014 - State of the Art Reports, 2014, pp. 69–91.

[11] Simon Gunkel, Hans Stokking, Martin Prins, Omar Niamut, Ernestasia Siahaan, and Pablo Cesar, "Experiencing Virtual Reality Together: Social VR Use Case Study," in Proceedings of the 2018 ACM International Conference on Interactive Experiences for TV and Online Video, New York, NY, USA, 2018, TVX '18, p. 233âĂŞ238, Association for Computing Machinery.

[12] Marc Herz and Philipp A. Rauschnabel, "Understanding the diffusion of virtual reality glasses: The role of media, fashion and technology," Technological Forecasting and Social Change, vol. 138, pp. 228–242, 2018.

[13] Zhigang Deng, Pei-Ying Chiang, Pamela Fox, and Ulrich Neumann, "Animating blendshape faces by cross-mapping motion capture data," in Proceedings of the 2006 symposium on Interactive 3D graphics and games - SI3D '06. 2006, p. 43, ACM Press.

[14] Deepak Ghimire, Joonwhoan Lee, Ze-Nian Li, and Sunghwan Jeong, "Recognition of facial expressions based on salient geometric features and support vector machines," vol. 76, no. 6, pp. 7921–7946.

[15] Eunjung Ju and Jehee Lee, "Expressive Facial Gestures From Motion Capture Data," Computer Graphics Forum, vol. 27, no. 2, pp. 381–388, Number: 2.

[16] Thibaut Weise, Hao Li, Luc Van Gool, and Mark Pauly, "Face/off: live facial puppetry," in Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation - SCA '09. 2009, p. 7, ACM Press.

[17] Thabo Beeler, Fabian Hahn, Derek Bradley, Bernd Bickel, Paul Beardsley, Craig Gotsman, Robert W. Sumner, and Markus Gross, "High-quality passive facial performance capture using anchor frames," in ACM SIGGRAPH 2011 papers on - SIGGRAPH '11. 2011, p. 1, ACM Press.

[18] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter, "A 3D Face Model for Pose and Illumination Invariant Face Recognition," in 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance. pp. 296–301, IEEE.

[19] Gabriel Schwartz, Shih-En Wei, Te-Li Wang, Stephen Lombardi, Tomas Simon, Jason Saragih, and Yaser Sheikh, "The Eyes Have It: An Integrated Eye and Face Model for Photorealistic Facial Animation," ACM Trans. Graph., vol. 39, no. 4, July 2020.

[20] Alex Humphreys, "How MetaHuman Creator helped me create 'Dana'," [Online].Available:https://www.bbc.com/news/av/technology-57569224, 2021, BBC News.

[21] Bernd Bickel, Mario Botsch, Roland Angst, Wojciech Matusik, Miguel Otaduy, Hanspeter Pfister, and Markus Gross, "Multi-scale capture of facial geometry and motion," ACM Trans. Graph., vol. 26, no. 3, pp. 33, 2007, Number: 3.

[22] Jaekwang Cha, Jinhyuk Kim, and Shiho Kim, "An IR-based facial expression tracking sensor for head-mounted displays," in 2016 IEEE SENSORS. 2016, pp. 1–3, IEEE.

[23] Katsuhiro Suzuki, Fumihiko Nakamura, Jiu Otsuka, Katsutoshi Masai, Yuta Itoh, Yuta Sugiura, and Maki Sugimoto, "Facial expression mapping inside head mounted display by embedded optical sensors," in Proceedings of the 29th Annual Symposium on User Interface Software and Technology - UIST '16 Adjunct. 2016, pp. 91–92, ACM Press.

[24] Katsuhiro Suzuki, Fumihiko Nakamura, Jiu Otsuka, Katsutoshi Masai, Yuta Itoh, Yuta Sugiura, and Maki Sugimoto, "Recognition and mapping of facial expressions to avatar by embedded photo reflective sensors in head mounted display," in 2017 IEEE Virtual Reality (VR), 2017, pp. 177–185.

[25] Jianwen Lou, Yiming Wang, Charles Nduka, Mahyar Hamedi, Ifigeneia Mavridou, Fei-Yue Wang, and Hui Yu, "Realistic facial expression reconstruction for VR HMD users," IEEE Trans. Multimedia, vol. 22, no. 3, pp. 730–743, 2019.

[26] Christian Frueh, Avneesh Sud, and Vivek Kwatra, "Headset removal for virtual and mixed reality," in ACM SIGGRAPH 2017 Talks on - SIGGRAPH '17. 2017, pp. 1–2, ACM Press.

[27] Hao Li, Laura Trutoiu, Kyle Olszewski, Lingyu Wei, Tristan Trutna, Pei-Lun Hsieh, Aaron Nicholls, and Chongyang Ma, "Facial performance sensing head-mounted display," ACM Trans. Graph., vol. 34, no. 4, pp. 1–9, 2015.

[28] Mike Seymour, Chris Evans, and Kim Libreri, "Meet mike: epic avatars," in ACM SIGGRAPH 2017 VR Village. 2017, pp. 1–2, ACM.

[29] Jihun Yu and Jungwoon Park, "Real-time facial tracking in virtual reality," in SIGGRAPH ASIA 2016 VR Showcase on - SA '16. 2016, pp. 1–1, ACM Press.

[30] Yajie Zhao, Qingguo Xu, Weikai Chen, Chao Du, Jun Xing, Xinyu Huang, and Ruigang Yang, "Mask-off: Synthesizing face images in the presence of head-mounted displays," in 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), 2016, pp. 267–276.

[31] Kyle Olszewski, Joseph J. Lim, Shunsuke Saito, and Hao Li, "High-fidelity facial and speech animation for VR HMDs," ACM Trans. Graph., vol. 35, no. 6, pp. 1–14, 2016, Number: 6.

[32] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh, "Deep appearance models for face rendering," ACM Trans. Graph., vol. 37, no. 4, pp. 1–13, 2018.

[33] Shih-En Wei, Jason Saragih, Tomas Simon, Adam W. Harley, Stephen Lombardi, Michal Perdoch, Alexander Hypes, Dawei Wang, Hernan Badino, and Yaser Sheikh, "Vr facial animation via multiview image translation," ACM Trans. Graph., vol. 38, no. 4, July 2019.

[34] Hang Chu and Shugao Ma, "Expressive telepresence via modular codec avatars," arXiv:2008.11789 [cs.CV], p. 16, 2020.

[35] Shugao Ma, Tomas Simon, Jason Saragih, Dawei Wang, Yuecheng Li, Fernando De La Torre, and Yaser Sheikh, "Pixel codec avatars," arXiv:2104.04638 [cs.CV], 2021.

[36] Lele Chen, Chen Cao, Fernando De la Torre, Jason Saragih, Chenliang Xu, and Yaser Sheikh, "High-fidelity face tracking for AR/VR via deep lighting adaptation," arXiv:2103.15876 [cs.CV], p. 12, 2021.

[37] Justus Thies, Michael Zollhoefer, Marc Stamminger, Christian Theobalt, and Matthias Niessner, "FaceVR: Real-time facial reenactment and eye gaze control in virtual reality," arXiv:1610.03151 [cs], 2018.

[38] Wataru Yamada, Hiroyuki Hakoda, Hiroyuki Manabe, Daizo Ikeda, and Jun Rekimoto, "Selfie mask: A face-capturing HMD with polarizing plates," in Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. 2020, pp. 1–4, ACM.

[39] Jun Rekimoto, Keishiro Uragaki, and Kenjiro Yamada, "Behind-the-mask: a face-through head-mounted display," in Proceedings of the 2018 International Conference on Advanced Visual Interfaces - AVI '18. 2018, pp. 1–5, ACM Press.

[40] Mariko Chiba, Wataru Yamada, and Hiroyuki Manabe, "Transparent mask: Face-capturing head-mounted display with IR pass filters," in The 31st Annual ACM Symposium on User Interface Software and Technology Adjunct Proceedings - UIST '18 Adjunct. 2018, pp. 149–151, ACM Press.

[41] DP Cosker DP AIT Salo SL Colyer, M Evans, "A Review of the Evolution of Vision-Based Motion Analysis and the Integration of Advanced Computer Vision Methods Towards Developing a Markerless System," Sports Medicine - Open, vol. 2018 Dec; 4(1):24., 2018.

[42] G.J. Edwards, A. Lanitis, C.J. Taylor, and T.F. Cootes, "Statistical models of face images - improving specificity," Image and Vision Computing, vol. 16, no. 3, pp. 203–211, 1998, Number: 3.

[43] T.F. Cootes, G.J. Edwards, and C.J. Taylor, "Active appearance models," IEEE Trans. Pattern Anal. Machine Intell., vol. 23, no. 6, pp. 681–685, 2001, Number: 6.

[44] Yang Wang, Xiaolei Huang, Chan-Su Lee, Song Zhang, Zhiguo Li, Dimitris Samaras, Dimitris Metaxas, Ahmed Elgammal, and Peisen Huang, "High resolution acquisition, learning and transfer of dynamic 3-d facial expressions," Computer Graphics Forum, vol. 23, no. 3, pp. 677–686, 2004.

[45] Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly, "Realtime performance-based facial animation," in ACM SIGGRAPH 2011 papers on - SIGGRAPH '11. 2011, p. 1, ACM Press.

[46] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou, "FaceWarehouse: A 3d facial expression database for visual computing,"

IEEE Trans. Visual. Comput. Graphics, vol. 20, no. 3, pp. 413–425, 2013, Number: 3.

[47] Lei Zhang, Yang Wang, Sen Wang, Dimitris Samaras, Song Zhang, and Peisen Huang, "Image-driven re-targeting and relighting of facial expressions," in International 2005 Computer Graphics. pp. 11–13, IEEE.

[48] Oswald Aldrian and William Smith, "Inverse Rendering of Faces with a 3D Morphable Model," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 5, pp. 1080–1093.

[49] Chen Cao, Yanlin Weng, Stephen Lin, and Kun Zhou, "3d shape regression for real-time facial animation," ACM Trans. Graph., vol. 32, no. 4, pp. 1–10, 2013, Number: 4.

[50] Chen Cao, Qiming Hou, and Kun Zhou, "Displaced dynamic expression regression for real-time facial tracking and animation," ACM Trans. Graph., vol. 33, no. 4, pp. 1–10, 2014, Number: 4.

[51] Chen Cao, Derek Bradley, Kun Zhou, and Thabo Beeler, "Real-time high-fidelity facial performance capture," ACM Trans. Graph., vol. 34, no. 4, pp. 1–9, 2015, Number: 4.

[52] Justus Thies, Michael Zollhoefer, Marc Stamminger, Christian Theobalt, and Matthias Niessner, "Face2face: real-time face capture and reenactment of RGB videos," Commun. ACM, vol. 62, no. 1, pp. 96–104, 2016, Number: 1.

[53] Congyi Wang, Fuhao Shi, Shihong Xia, and Jinxiang Chai, "Realtime 3d eye gaze animation using a single RGB camera," ACM Trans. Graph., vol. 35, no. 4, pp. 1–14, 2016, Number: 4.

[54] Luming Ma and Zhigang Deng, "Real-time hierarchical facial performance capture," in Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games. 2019, pp. 1–10, ACM.

[55] Yudong Guo, Juyong Zhang, Jianfei Cai, Boyi Jiang, and Jianmin Zheng, "CNN-based real-time dense face reconstruction with inverse-rendered photo-realistic face images," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 6, pp. 1294–1307, 2019, Number: 6.

[56] Mohammad Rami Koujan, Michail Christos Doukas, Anastasios Roussos, and Stefanos Zafeiriou, "Head2head: Video-based neural head synthesis," arXiv:2005.10954 [cs, eess], 2020.

[57] Michail Christos Doukas, Mohammad Rami Koujan, Viktoriia Sharmanska, Anastasios Roussos, and Stefanos Zafeiriou, "Head2head++: Deep facial attributes re-targeting," IEEE Trans. Biom. Behav. Identity Sci., vol. 3, no. 1, pp. 31–43, 2021.

[58] Xiaoyu Chai, Jun Chen, Chao Liang, Dongshu Xu, and Chia-Wen Lin, "Expression-aware face reconstruction via a dual-stream network," IEEE Trans. Multimedia, pp. 1–1, 2021.

[59] Samuli Laine, Tero Karras, Timo Aila, Antti Herva, Shunsuke Saito, Ronald Yu, Hao Li, and Jaakko Lehtinen, "Production-level facial performance capture using deep convolutional neural networks," in Proceedings of the ACM SIGGRAPH / Eurographics Symposium on Computer Animation. 2017, pp. 1–10, ACM.

[60] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou, "AvatarMe: Realistically Renderable 3D Facial Reconstruction "in-the-wild"," 2020.

[61] Derek Bradley, Wolfgang Heidrich, Tiberiu Popa, and Alla Sheffer, "High resolution passive facial performance capture," ACM Trans. Graph., vol. 29, no. 4, pp. 1–10, 2010, Number: 4.

[62] Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross, "High-quality single-shot capture of facial geometry," ACM Trans. Graph., vol. 29, no. 4, pp. 1–9, 2010, Number: 4.

[63] Graham Fyffe, Tim Hawkins, Chris Watts, Wan-Chun Ma, and Paul Debevec, "Comprehensive facial performance capture," Computer Graphics Forum, vol. 30, no. 2, pp. 425–434, 2011, Number: 2.

[64] Martin Klaudiny, Steven McDonagh, Derek Bradley, Thabo Beeler, and Kenny Mitchell, "Real-time multi-view facial capture with synthetic training," Computer Graphics Forum, vol. 36, no. 2, pp. 325–336, 2017, Number: 2.

[65] Paresh Dave, "ANALYSIS-google's starline shows promise and perils of 3d chats," The Economic Times, p. 3, 2021.

[66] Derek Bradley, Tamy Boubekeur, and Wolfgang Heidrich, "Accurate multi-view reconstruction using robust binocular stereo and surface meshing," in 2008 IEEE Conference on Computer Vision and Pattern Recognition. 2008, pp. 1–8, IEEE.

[67] Yasutaka Furukawa and Jean Ponce, "Accurate, dense, and robust multiview stereopsis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 8, pp. 1362–1376, 2007.

[68] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L. Davidson, Sameh Khamis, Mingsong Dou, Vladimir Tankovich, Charles Loop, Qin Cai, Philip A. Chou, Sarah Mennicken, Julien Valentin, Vivek Pradeep, Shenlong Wang, Sing Bing Kang, Pushmeet Kohli, Yuliya Lutchyn, Cem Keskin, and Shahram Izadi, "Holoportation: Virtual 3d teleportation in real-time," in Proceedings of the 29th Annual Symposium on User Interface Software and Technology. 2016, pp. 741–754, ACM.

[69] Microsoft Research, "Demo: The magic of AI neural TTS and holograms at microsoft inspire 2019," [Online].Available:https://www.youtube.com/watch?v=auJJrHgG9Mc.

[70] Vladimiros Sterzentsenko, Antonis Karakottas, Alexandros Papachristou, Nikolaos Zioulis, Alexandros Doumanoglou, Dimitrios Zarpalas, and Petros Daras, "A low-cost, flexible and portable volumetric capturing system," in 2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS). 2018, pp. 200–207, IEEE.

[71] Intel Corporation, "Huge geodesic dome is world's largest 360-degree movie set," [Online].Available:https://www.businesswire.com/news/home/20181218005178/en/Huge-Geodesic-Dome-is-World%E2%80%99s-Largest-360-Degree-Movie-Set, 2018.

[72] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, Danhang Tang, Anastasia Tkach, Adarsh Kowdle, Emily Cooper, Mingsong Dou, Sean Fanello, Graham Fyffe, Christoph Rhemann, Jonathan Taylor, Paul Debevec, and Shahram Izadi, "The relightables: volumetric performance capture of humans with realistic relighting," ACM Trans. Graph., vol. 38, no. 6, pp. 1–19, 2019, Number: 6.

[73] Michael Broxton, John Flynn, Ryan Overbeck, Daniel Erickson, Peter Hedman, Matthew Duvall, Jason Dourgarian, Jay Busch, Matt Whalen, and Paul Debevec, "Immersive light field video with a layered mesh representation," ACM Trans. Graph., vol. 39, no. 4, 2020, Number: 4.

[74] Aleksey Golovinskiy, Wojciech Matusik, Hanspeter Pfister, Szymon Rusinkiewicz, and Thomas Funkhouser, "A statistical model for synthesis of detailed facial geometry," ACM Trans. Graph., vol. 25, no. 3, pp. 1025âĂŞ1034, July 2006.

[75] Jiaman Li, Zhengfei Kuang, Yajie Zhao, Mingming He, Karl Bladin, and Hao Li, "Dynamic facial asset and rig generation from a single scan," arXiv:2010.00560 [cs], 2020.

[76] Epic Games Inc., "Creating Human Skin," [Online].Available:https://docs.unrealengine.com/4.26/en-US/RenderingAndGraphics/Materials/HowTo/Human_Skin/, 2021.

[77] Lasse Jon Fuglsang Pedersen, "Making of The Heretic: Digital Human tech package," [Online].Available:https://blog.unity.com/technology/making-of-the-heretic-digital-human-tech-package, Google Keywords: Unity making heretic digital human.

[78] Shikhar Shrestha, Felix Heide, Wolfgang Heidrich, and Gordon Wetzstein, "Computational imaging with multi-camera time-of-flight systems," ACM Trans. Graph., vol. 35, no. 4, pp. 1–11, 2016, Number: 4.

[79] Statista S. O'Dea, "Number of smartphones sold to end users worldwide from 2007 to 2021," [Online].Available:https://www.statista.com/statistics/263437/global-smartphone-sales-to-end-users-since-2007/, 2020, Google Keywords: statista smartphones sales.

[80] Rahul Kumar, "Smartphone 3d camera market by technology (stereoscopic camera and time-of-flight (tof)) and resolution (below 8 mp, 8-16 mp, and above 16 mp): Global opportunity analysis and industry forecast, 2018 - 2025," [Online].Available:https://www.alliedmarketresearch.com/smartphone-3d-camera-market, 2017.

[81] David Lumb, "iphone 12 led a record-breaking $47 billion in iphone sales over q2 2021," [Online].Available:https://www.techradar.com/au/news/iphone-12-led-a-record-breaking-dollar47-billion-in-iphone-sales-over-q2-2021, 2021.

[82] Yan Yan, Ke Chen, Yu Xie, Yiming Song, and Yonghong Liu, "The Effects of Weight on Comfort of Virtual Reality Devices," in Advances in Ergonomics in Design, Francisco Rebelo and Marcelo M. Soares, Eds., Cham, 2019, pp. 239–248, Springer International Publishing.

LIHANG WEN received a B.S. degree in Computer Science in Sun Yat-sen University. He worked as Software Developer, DevOps Engineer for many years, and held several patents. He's pursuing the Ph.D. degree in University of Technology Sydney. His research interests include Virtual Reality, Computer Vision, and IoT.

JIANLONG ZHOU is a Senior Lecturer in Data Science Institute, University of Technology Sydney, Australia. His research work focuses on ethics of AI, AI fairness, AI explainability, data analytics, visual analytics, behaviour analytics, human-computer interaction, VR, and related applications. Dr. Zhou is a leading senior researcher in trustworthy and transparent machine learning, and has done pioneering research in the area of linking human and machine learning.

WEIDONG HUANG is currently an Associate Professor at University of Technology Sydney, Australia. He holds a PhD in Computer Science from University of Sydney. He also has formal training in experimental psychology and professional experience in psychometrics. His main research interests are in Human-Computer Interaction and Visualization. He is the author of over 150 publications. Dr Huang is a founding chair of the Technical Committee on Visual Analytics and Communication for IEEE SMC Society and a guest editor of a number of SCI indexed journals. He has served as a conference chair, a PC chair, or an organization chair for a number of international conferences and workshops.

FANG CHEN is a Distinguished Professor in AI and data science at the University of Technology Sydney (UTS), Australia. She is the Executive Director Data Science and Distinguished Professor, the University of Technology Sydney. Prof. Chen's research interests include machine learning, AI ethics, behaviour analytics and their applications. She has created many world-class AI innovations while working in Intel, Motorola, NICTA, CSIRO and now UTS, and helped governments and industries utilising data and significantly increasing productivity, safety and customer satisfaction.