A Study on the Quality of Facial Expression in Digital Humans

Shiori Kikuchi, Hisayoshi Ito, Oky Dicky Ardiansyah Prima
Graduate School of Software and Information Science
Iwate Prefectural University
Takizawa, Iwate, Japan
e-mail: g231u017@s.iwate-pu.ac.jp, {hito, prima}@iwate-pu.ac.jp

***Abstract*—There have been many examples recently of the use of digital humans for interactive communication, such as in customer support and digital health care. The use of digital humans is expected to reduce communication barriers caused by differences in personal appearance, behavior, and facial expressions. Despite the potential for digital humans to represent realistic nonverbal communication, there is a lack of evidence regarding the extent to how effective they are in communication. This study attempts to evaluate the six basic emotions (anger, fear, disgust, happiness, sadness, and surprise) of digital humans with the following two experiments. First, the quality of facial expression is evaluated by an actor's facial expression and the expression of the digital human created from the actor. Second, we evaluate the quality of the facial expressions of 17 subjects captured from various angles and those of digital humans created from the corresponding angles. Two deep learning-based facial expression recognitions: DeepFace and HSEmotion were used for the evaluation. Experimental results showed that HSEmotion demonstrated a more stable recognition rate than DeepFace for the same facial expressions of subjects captured from different directions. However, when the facial expressions of these subjects were transferred to the digital humans, both tools failed to properly recognize their facial expressions. Future work will include a facial expression recognition library that considers both real people and digital humans.**

Keywords-expression; digital human; facial expression; basic emotions; avatar.

I. INTRODUCTION

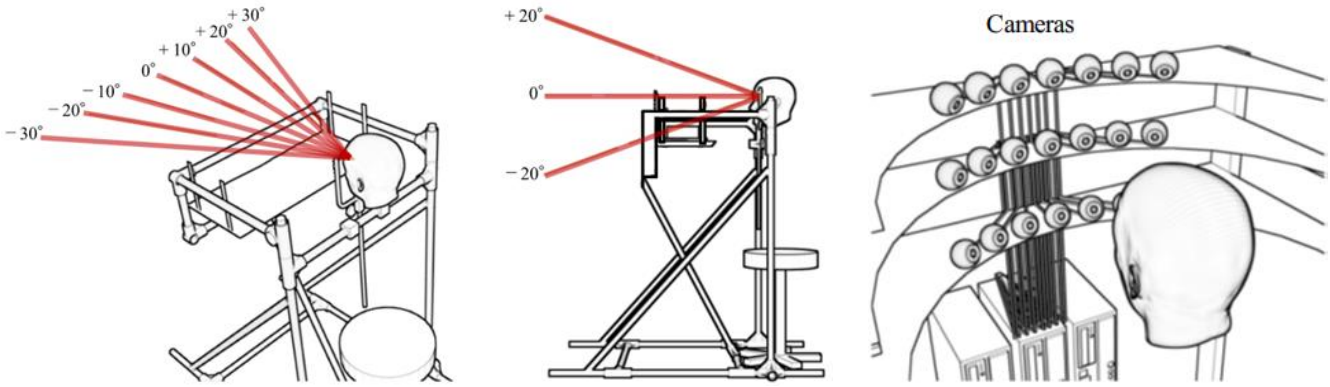
There have been changes in communication methods for forming interpersonal relationships due to the spread of the new corona virus (COVID-19) outbreak, including the use of videoconferencing for meetings. In addition to voice messages, video calls convey visual and non-verbal information, making more effective communication possible. Recent video calls allow participants control over the content to be delivered, such as the ability to remove the video background and manipulate their faces. It is even possible to make video calls using avatars or digital humans, making it easier to communicate with others without worrying about their own appearance. However, facial expressions are not always conveyed adequately by these tools. This paper extends our previous work evaluating facial expressions by digital humans [1].

Significant emotional expression is necessary for communication, which is especially evident in human facial expressions. The quality of digital human facial expressions is therefore considered important for communication in virtual space as well. Interest for communication in these spaces is growing worldwide as well as in Japan, such as "Virtual Shibuya" by KDDI Corporation [2] and "Medical Metaverse Joint Research Chair" by IBM Japan and Juntendo University [3].

Communication in virtual conference applications and games is conducted in real time using virtual characters that have been designed to look like real people in appearance. With the development of artificial intelligence and computer vision, facial expression recognition from facial images is becoming more practical. By transferring the recognized facial components of the user to the virtual character, the user can play various roles through the virtual character. This character is expected to stimulate communication in the medical and business fields.

Digital humans are more realistic than virtual characters, and many are being used in business to enable natural conversations with customers. In addition, digital humans are also equipped with the ability to speak multiple languages, which further expands their applications [4]. Digital humans can be generated from Three-Dimensional (3D) human pose information obtained by capturing a person with a monocular depth camera or a stereo camera [5]. In addition, the development of 3D human pose using a monocular camera, such as the MediaPipe library [6], has become popular, making it possible to create a digital human using only a webcam.

Recently, software tools have been developed to enable the expression of detailed facial expression changes in digital humans using actors' expressions. MetaHuman by Epic Games provides a framework for creating realistic human characters [7]. This framework works by transferring 3D human pose information from the Motion Capture (MoCap) device to the digital human. Similar frameworks include "Character Creator 4" (Reallusion) [8] and "Buddy Builder" (Hologress) [9]. Both frameworks offer a wide variety of resources for 3D clothing and accessories. Buddy Builder features real-time cross-physics, which allows for more natural-looking moving characters. These frameworks are expected to enable sophisticated digital humans to express nonverbal information accurately, which is currently the subject of further research and development. However, the extent to which the quality of digital human facial expressions is comparable to that of humans has not been fully verified.



(a) Horizontal angles (*yaw*). (b) Vertical angles (*pitch*). (c) Location of 21 cameras for capturing.

Figure 1. Multi-angle camera device used in this study.

The purpose of this study is to evaluate the quality of human and digital human facial expressions. The following two attempts were made for the evaluation. First, a digital human that resembles an actor was created, and both facial expressions were compared. Second, basic facial expressions by several subjects were captured using multi-angle camera devices, and these facial expressions were transferred to the digital human. For the evaluation of facial expressions, existing deep learning-based libraries for facial expression recognition are used.

The rest of this paper is organized as follows. Section II discusses the related works of facial expression analyses and digital human. Section III describes the generation of facial expression data and tools used for this purpose in this study. In Section IV, we describe our experiments to evaluate the quality of facial expressions for the digital human. Finally, Section V summarizes the results of this study and discusses future perspectives.

II. Related Works

A. Facial Expressions in Communication

The understanding of one's emotions is important in communication. Particularly, since facial expressions represent human emotions, smooth communication can be achieved by being aware of changes in the facial expressions of conversational partners. Facial expressions consist of movements of small muscles in the face that are used to infer a person's emotional state.

Emotions are difficult to measure since they are often fleeting, hidden, and conflicted. Ekman et al. proposed the Facial Action Coding System (FACS), which classifies Action Units (AUs) of facial parts to identify emotions from facial expressions [10]. Their work pointed out that there are "display rules" based on cultural norms, making the intensity of facial expressions differs from culture to culture. The Japanese, for example, tend to suppress facial expressions [11]. Ohta et al. observed facial expressions in Japanese nursing practice and concluded that the results were generally consistent with Ekman's analysis, but that differences were observed in the distinctness of facial expressions due to the way Japanese people move their facial muscles and their ability to express themselves being weaker than Westerners [12]. Baltrušaitis et al. developed OpenFace [13], a toolkit that can recognize AUs in real time based on facial landmarks taken from the user's face.

B. Facial Expression Recognition based on Deep Learning

As the applications of facial expression recognition have expanded, the development of deep learning-based facial expression recognition has progressed rapidly. With facial expression recognition, an individual's emotional state can be predicted from the appearance of facial deformations. Furthermore, these techniques enable real-time analysis of facial expressions.

Many studies have been conducted in the field of facial expression recognition to realize feasible tasks. In addition, lightweight models have been proposed to enable recognition in web applications and mobile devices. Serengil et al. developed Deepface, a lightweight model of facial expression recognition [13]. Andrey et al. developed a model that utilizes several eEfficientNet-based models to classify emotions of static facial images [14]. This model has been published as the HSEmotion (High-Speed Face Emotion Recognition) [15].

C. Reality of the Digital Human

Digital humans have recently been developed that can resemble humans by appearance and can reflect their body movements [4]. Kang et al. surveyed and simplified the research on digital human reality, introducing two types of reality: visual realism, which is the similarity between the rendering of visual information of a person, and behavioral realism, which is the similarity between human behavior and the reality of a person [16]. Visual realism is high as the digital human looks more like a person, and behavioral realism is high as the digital human performs natural movements. They also stated the importance of the influence of digital human reality on communication. Grewe et al. compared the reality of facial expression animations created by experts with the reality of facial expression animations created statistically

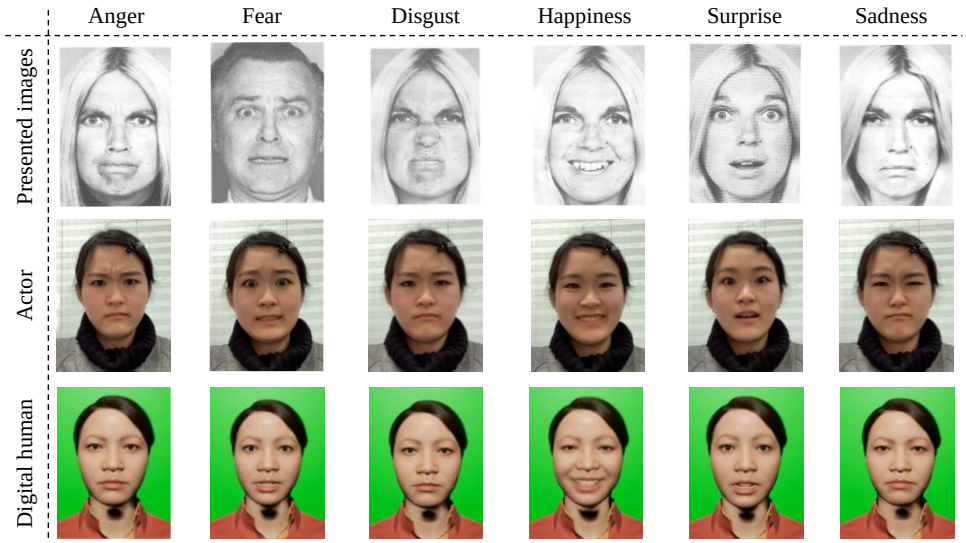


Figure 2. The actor and the generated digital human facial images for each expression.

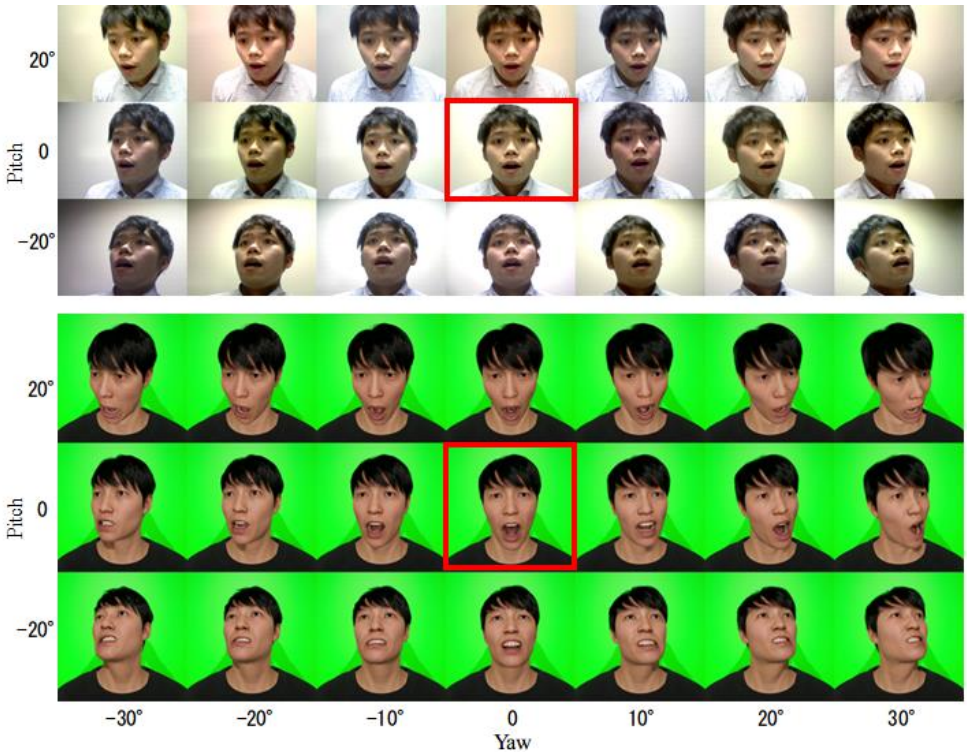


Figure 3. MACD-captured subjects' faces expressing "Surprise," and the digital humans corresponding to each face in each capture direction.

from a database of images. They found that the statistically created animations were perceived as more realistic [17].

III. Generation of Facial Expression Data

This study generated human digital facial expression data using two different capturing methods: frontal and multi-angle photography and evaluated them respectively. Research collaborators photographed in the dataset gave us permission to use the dataset in this study. The methods and tools used to create the dataset and the details of the dataset created are described as follows.

A. Tools Used in This Study

MetaHuman Creator (MHC) and Unreal Engine (UE) by Epic Games were used to create digital humans and reflect the actor's facial expressions on the digital human [18]. MHC is a free cloud-based tool that facilitates the creation of photorealistic digital humans and can be used in conjunction

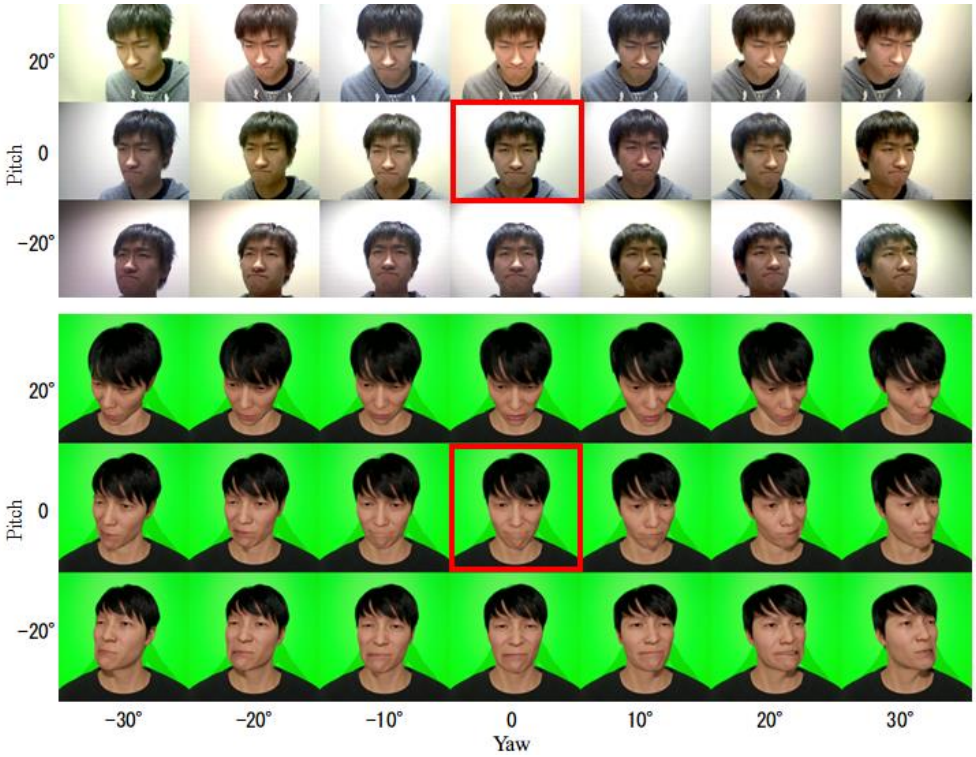


Figure 4. MACD-captured subjects' faces expressing "Disgust," and the digital humans corresponding to each face in each capture direction.

with the 3D object rendering engine of UE. Users can create facial shapes, hairstyles, clothes, and other accessories easily, all of which can be manipulated intuitively through the Graphical User Interface (GUI). This allows average users to create realistic digital humans like computer graphics designers. In addition, MetaHuman data such as meshes, skeletons, facial rigs, animation controls, and materials can be downloaded and exported to other CG software.

MoCap is needed to transfer small facial muscle movements to the MHC. Live Link Face (LLF) and MeFaMo [19] were employed as the MoCap in this study. The former is an iOS application developed by Epic Games that uses the mobile device's depth sensor to extract facial features. The latter uses a monocular camera to estimate and extract facial features based on the MediaPipe library. MeFaMo is used to extract facial muscles in face images obtained from multiple cameras, described in the next subsection.

B. Multi-Angle Camera Device (MACD)

In this study, a MACD was constructed to capture facial images simultaneously from multiple angles. Figure 1 shows the device consisting of 21 webcams (640 x 480 pixels, 30 fps). Each camera is arranged in three rows and seven columns. These cameras are networked which allow to simultaneously capture the subject's face at ±30° horizontally (*yaw*) and ±20° vertically (*pitch*) in 10° and 20° increments, respectively. 21 face facial images can be obtained in a single capture.

C. Facial Expression Data

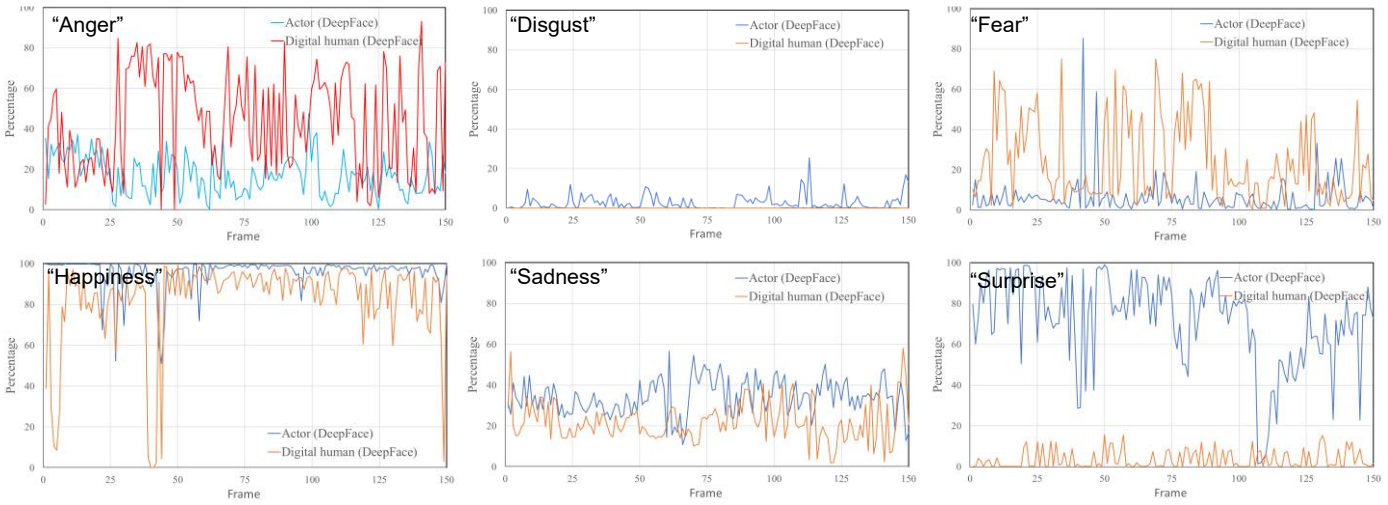
1) Data Taken with LLF

To generate human digital facial expression data, six basic facial expression images were presented in sequence to a 22-year-old Japanese female actor, who was asked to mimic each facial expression for 5 seconds. The facial images were acquired with their features recorded at 30 fps using LLF in order to analyze the changes in facial expressions over time. To minimize the effects of cultural differences and the actor's experience with facial expression, a set of facial images representing Ekman's six basic emotions [10] was presented to the actor. The characteristic of facial muscles for each emotion was described for the actor to mimic appropriately. We collected 150 frames of facial images from a 5-second video of each expression, resulting in 900 frames each of facial images of the actor and the digital human imitating the six basic facial expression.

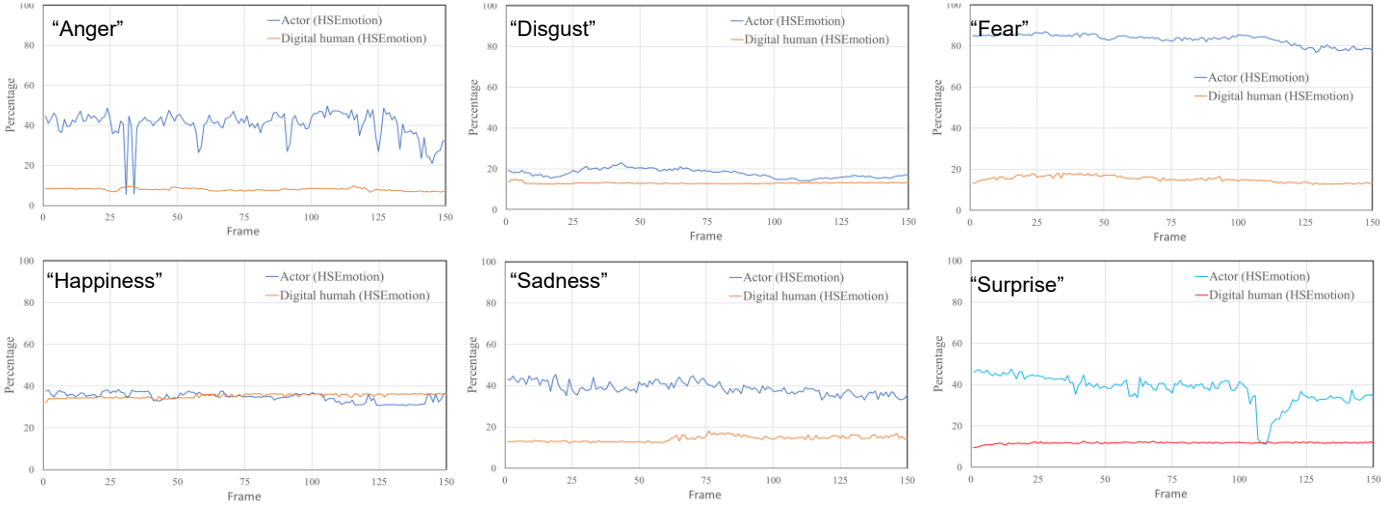
The actor and the generated digital human facial images for each expression are shown in Figure 2. The appearance of the digital human was made to resemble the actor as close as possible. The generated digital humans generally adequately represent the actor's facial expressions, but the detailed expressions tend to look significantly poor, resulting in a weaker negative facial expression.

2) Data Taken with MACD

The MACD was used to acquire a total of 2,142 face images of 17 Japanese students aged 18-22 from the Faculty of Software and Information Science at Iwate Prefectural University, Japan. These facial images are representing six basic facial expressions plus neutral, captured at 21 different angles. As we focused on evaluating how the facial



(a) Facial expression predicted by DeepFace.



(b) Facial expression predicted by HSEmotion.

Figure 5. Recognition rates for the actor's and its corresponding digital human's facial expressions for each frame.

expressions of each student are represented by the digital human, we did not transfer the facial expressions to a digital human that looked exactly like each student, but instead used one common digital human as the transfer destination. MeFaMo was used to capture facial images and their features.

Figures 3 and 4 show two students expressing anger and disgust, respectively, and the digital human corresponding to each capture direction. The red rectangle in the center indicates the direction of the frontal view (*pitch* = 0, *yaw* = 0). Here, different students' facial expressions are transferred to the same digital human, showing that each student's expression is well represented.

IV. Evaluation of the Quality of Facial Expression

The facial expression data created in Section III is evaluated using the deep learning-based facial expression recognition libraries: DeepFace [13] and HSEmotion [15]. While the architecture of DeepFace simply consists of three convolutional layers and two fully-connected layers, HSEmotion uses eEfficientNet as a backbone, which was fine-tuned on a face identification task using the VGGFace2 dataset. Both libraries perform class classification and use softmax functions in the output layer to normalize the inferred results for each expression from 0 to 1, where the sum of all inferred expression proportions accounts for 1. Hereinafter, this value is called the recognition rate.

DeepFace was trained on the Fec2013 [20] containing 32,298 facial images. In contrast, HSEmotion was trained on the AffectNet [21], a large facial expression dataset containing more than 1,000,000 facial images, and evaluated on datasets such as EmotiW, AFEW (Acted Facial Expression In The Wild), VGAF (Video level Group AFfect) and EngageWild. The trained model "enet_b2_8.pt", which is highly accurate on these datasets, was employed in this study.

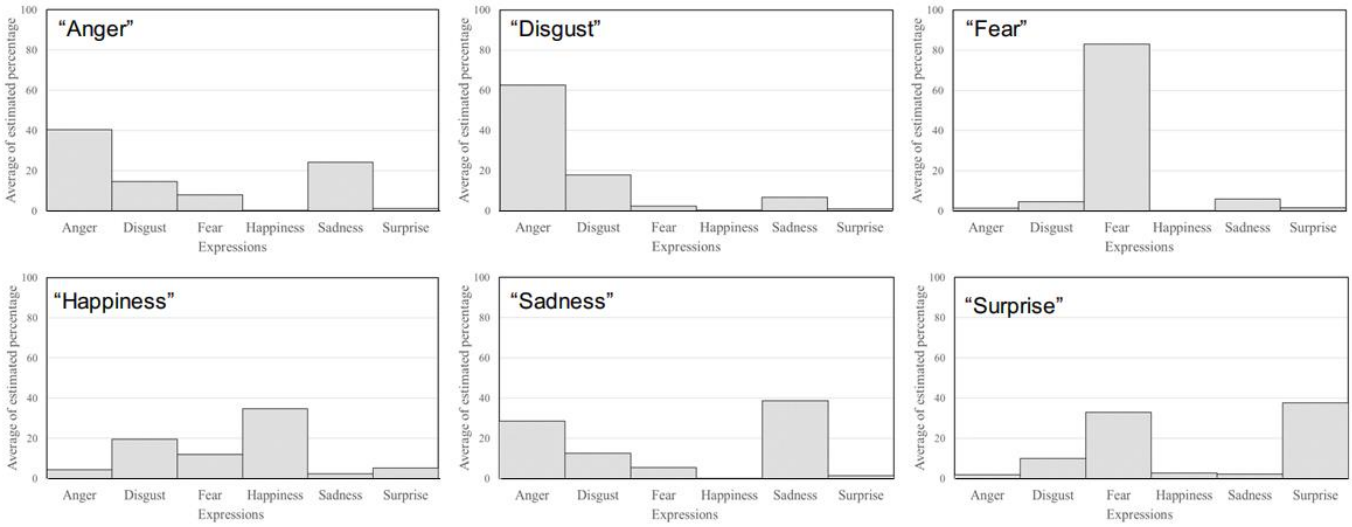


Figure 6. Average recognition rate of the actor's facial expressions by the HSEmotion.

TABLE I. ACTOR'S EXPRESSIONS PREDICTED BY DEEPFACE

		Predicted						Recall
		Anger	Disgust	Fear	Happiness	Sadness	Surprise	
TRUE	Anger	**6**	0	0	0	56	0	0.10
	Disgust	20	0	0	0	128	0	0.00
	Fear	0	0	**2**	7	1	0	0.20
	Happiness	0	0	0	**150**	0	0	1.00
	Sadness	2	0	0	0	**74**	0	0.97
	Surprise	0	0	11	0	0	**135**	0.92
Precision		0.21	NaN	0.15	0.96	0.29	1.00	

TABLE II. DIGITAL HUMAN EXPRESSIONS PREDICTED BY DEEPFACE

		Predicted						Recall
		Anger	Disgust	Fear	Happiness	Sadness	Surprise	
TRUE	Anger	**83**	0	7	0	6	0	0.86
	Disgust	3	0	0	0	19	0	0.00
	Fear	1	0	**38**	0	0	0	0.97
	Happiness	0	0	0	**139**	5	0	0.97
	Sadness	33	0	0	0	**3**	0	0.08
	Surprise	0	0	45	2	0	0	0.00
Precision		0.69	NaN	0.42	0.99	0.09	NaN	

TABLE III. ACTOR'S EXPRESSIONS PREDICTED BY HSEMOTION

		Predicted						Recall
		Anger	Disgust	Fear	Happiness	Sadness	Surprise	
TRUE	Anger	**131**	2	0	0	18	0	0.87
	Disgust	151	0	0	0	0	0	0.00
	Fear	0	0	**151**	0	0	0	1.00
	Happiness	0	0	0	**151**	0	0	1.00
	Sadness	6	0	0	0	**145**	0	0.96
	Surprise	0	0	39	0	0	**106**	0.73
Precision		0.45	0.00	0.79	1.00	0.89	1.00	

TABLE IV. DIGITAL HUMAN EXPRESSIONS PREDICTED BY HSEMOTION

		Predicted						Recall
		Anger	Disgust	Fear	Happiness	Sadness	Surprise	
TRUE	Anger	0	0	0	0	0	0	NaN
	Disgust	0	0	0	0	0	0	NaN
	Fear	0	0	0	0	0	0	NaN
	Happiness	0	0	0	**150**	0	0	1.00
	Sadness	0	0	0	0	0	0	NaN
	Surprise	0	0	0	0	0	0	NaN
Precision		NaN	NaN	NaN	1.00	NaN	NaN	

A. LLF-captured Facial Espressions

Facial expression recognition was performed for each video of facial expression of the actor captured by the LLF and the corresponding digital human. Figure 5 shows the facial expression recognition rates obtained for a 5-second facial expression. The blue and red lines are the results for the actor and the digital human, respectively.

Overall, the results of facial expression recognition for DeepFace were more variable than those for HSEmotion, suggesting that its recognition is unstable. For a more detailed analysis, we investigate results based on their recognition rates using the following criteria.

1) 50% rate for a Correct Recognition

In a given frame, an expression is considered to be correctly judged by the expression recognition library if the recognition rate is 50% or more. According to this criterion, DeepFace was found to recognize "Fear," "Happiness," "Sadness," and "Surprise" of the actor and "Anger," "Fear," "Happiness," and "Sadness" for the digital human from the 5-second video, as shown in Figure 5(a). On the other hand, as shown in Figure 5(b), HSEmotion could only recognize the actor's "Fear."

At first glance, the results for HSEmotion appear to be very poor. However, the results for the actor show that the average recognition rate of its facial expressions is higher than that of the others, with the exception of "Disgust," as shown in Figure 6. The results show that HSEmotion fairly

TABLE V. AVERAGE OF DEEPFACE RECOGNITION RATES FOR 17 SUBJECTS' FACIAL EXPRESSIONS CAPTURED BY MACD.

		Predicted						
		Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
TRUE	Anger	**18.3%**	0.2%	9.1%	1.4%	44.2%	26.5%	0.1%
	Disgust	18.8%	**0.2%**	9.2%	1.4%	43.1%	27.2%	0.1%
	Fear	19.7%	0.2%	**9.4%**	1.1%	41.0%	28.4%	0.1%
	Happiness	19.9%	0.2%	9.5%	**1.7%**	39.9%	28.5%	0.1%
	Neutral	20.0%	0.2%	9.1%	2.3%	**39.9%**	28.4%	0.2%
	Sadness	19.5%	0.2%	8.3%	2.2%	42.6%	**27.2%**	0.1%
	Surprise	17.6%	0.2%	8.1%	2.3%	45.4%	26.3%	**0.1%**

TABLE VI. AVERAGE OF DEEPFACE RECOGNITION RATES FOR DIGITAL HUMAN FACIAL EXPRESSIONS FROM 17 SUBJECTS CAPTURED BY MACD.

		Predicted						
		Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
TRUE	Anger	**9.7%**	1.2%	18.0%	13.9%	33.7%	19.1%	4.0%
	Disgust	8.9%	**1.2%**	17.9%	13.8%	33.7%	19.9%	4.1%
	Fear	8.4%	0.8%	**17.9%**	13.3%	34.4%	20.2%	4.3%
	Happiness	8.2%	0.8%	18.3%	**13.7%**	34.2%	19.8%	4.3%
	Neutral	8.4%	1.3%	18.7%	13.9%	**33.6%**	19.4%	4.2%
	Sadness	9.2%	1.4%	19.1%	13.4%	33.0%	**19.4%**	4.1%
	Surprise	9.4%	1.4%	19.6%	13.1%	32.7%	20.1%	**3.1%**

TABLE VII. AVERAGE OF HSEMOTION RECOGNITION RATES OF 17 SUBJECTS' FACIAL EXPRESSIONS CAPTURED BY MACD.

		Predicted						
		Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
TRUE	Anger	**30.4%**	21.8%	6.1%	2.7%	8.9%	14.9%	2.1%
	Disgust	20.1%	**29.4%**	5.4%	3.9%	8.8%	16.4%	2.3%
	Fear	8.3%	19.2%	**16.5%**	5.6%	15.3%	10.9%	11.3%
	Happiness	4.7%	17.2%	9.4%	**24.2%**	9.3%	5.0%	4.0%
	Neutral	12.5%	14.9%	6.9%	2.4%	**28.2%**	14.3%	3.3%
	Sadness	16.3%	16.5%	6.2%	5.4%	14.1%	**21.2%**	2.5%
	Surprise	4.2%	13.6%	15.6%	7.5%	10.9%	3.7%	**33.8%**

TABLE VIII. AVERAGE OF HSEMOTION RECOGNITION RATES OF DIGITAL HUMAN FACIAL EXPRESSIONS FROM 17 SUBJECTS CAPTURED BY MACD.

		Prediction						
		Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
TRUE	Anger	**6.7%**	44.0%	7.1%	1.7%	3.0%	23.4%	2.0%
	Disgust	6.3%	**42.6%**	9.0%	2.6%	2.5%	23.5%	2.8%
	Fear	7.1%	29.5%	**22.5%**	1.2%	2.9%	21.9%	5.4%
	Happiness	5.6%	32.1%	15.2%	**2.4%**	2.7%	25.8%	2.6%
	Neutral	4.0%	34.0%	11.9%	0.7%	**3.4%**	33.2%	2.3%
	Sadness	4.0%	43.5%	8.2%	1.4%	2.5%	**28.3%**	1.9%
	Surprise	7.2%	25.1%	29.3%	1.1%	3.2%	16.7%	**8.4%**

recognized most of the actor's facial expressions. Therefore, we believe that judging the largest recognition rate of each expression as the dominant expression would be preferable, as described in the next sub-section.

2) Highest rate for a Correct Recognition

Facial expression was applied to each frame of the 5-second facial expression video, and the expression with the highest recognition rate was identified as the final recognition result. These results are summarized as confusion matrices in Tables I to IV. The numbers underlined in bold indicate the number of frames in which the actor's facial expression was correctly recognized.

Table I shows that DeepFace has high Precision and Recall for the actor's "Happiness" and "Surprise." Recall is high for the actor's "Sadness," but Precision is low. On the other hand, in digital humans, Table II shows that DeepFace has high Precision and Recall in "Anger" and "Happiness." In addition, the Recall is high in "Fear." In terms of high Precision and Recall, we observed that digital humans only accurately represent actors expressing "Happiness."

TABLE IX. RECOGNITION RATES IN DEEPFACE FOR SUBJECTS'FACIAL EXPRESSIONS CAPTURED IN THE PITCH- AND YAW-DIRECTIONS.

	Pitch			St.Dev
	-20°	-0°	20°	
Anger	15.5%	21.8%	17.7%	3.2%
Disgust	0.4%	0.2%	0.0%	0.2%
Fear	6.7%	9.3%	15.7%	4.6%
Happiness	44.6%	43.6%	40.8%	2.0%
Sadness	42.4%	19.3%	18.4%	13.6%
Surprise	7.0%	30.6%	28.8%	13.1%

	Yaw							St.Dev
	-30°	-20°	-10°	-0°	10°	20°	30°	
Anger	23.9%	12.9%	13.0%	7.2%	13.2%	24.1%	33.9%	9.2%
Disgust	0.0%	0.3%	0.5%	0.3%	0.3%	0.0%	0.0%	0.2%
Fear	17.0%	10.6%	10.2%	9.0%	10.8%	7.5%	8.7%	3.1%
Happiness	35.8%	37.3%	27.7%	38.6%	46.8%	59.9%	56.2%	11.7%
Sadness	26.4%	19.3%	16.1%	29.0%	38.1%	33.5%	24.3%	7.7%
Surprise	18.0%	24.0%	25.0%	22.0%	19.3%	26.3%	20.3%	3.1%

Table X. RECOGNITION RATES IN DEEPFACE FOR FACIAL EXPRESSIONS OF DIGITAL HUMANS CAPTURED IN THE PITCH- AND YAW-DIRECTIONS.

	Pitch			St.Dev
	-20°	-0°	20°	
Anger	6.5%	2.8%	19.7%	8.9%
Disgust	1.5%	0.5%	2.3%	0.9%
Fear	18.8%	27.4%	15.3%	6.2%
Happiness	11.5%	20.8%	24.5%	6.7%
Sadness	34.4%	9.7%	6.0%	15.5%
Surprise	3.9%	31.5%	27.1%	14.8%

	Yaw							St.Dev
	-30°	-20°	-10°	-0°	10°	20°	30°	
Anger	10.4%	11.6%	15.6%	8.8%	9.6%	4.5%	7.3%	3.5%
Disgust	1.0%	5.2%	2.5%	1.2%	0.0%	0.1%	0.0%	1.9%
Fear	30.7%	21.7%	20.3%	25.1%	19.6%	11.6%	14.6%	6.3%
Happiness	29.3%	25.8%	26.0%	15.4%	26.4%	22.3%	19.0%	4.8%
Sadness	8.4%	17.1%	25.1%	27.6%	17.4%	12.9%	8.4%	7.6%
Surprise	14.0%	8.7%	8.7%	23.4%	30.6%	35.6%	24.8%	10.6%

Table XI. RECOGNITION RATES IN HSEMOTION FOR SUBJECTS' FACIAL EXPRESSIONS CAPTURED IN THE PITCH- AND YAW-DIRECTIONS.

	Pitch			St.Dev
	-20°	-0°	20°	
Anger	29.8%	33.9%	27.6%	3.2%
Disgust	24.6%	33.8%	29.9%	4.6%
Fear	15.2%	16.8%	17.4%	1.2%
Happiness	24.0%	24.8%	23.8%	0.6%
Sadness	30.6%	19.8%	13.3%	8.8%
Surprise	29.4%	36.3%	35.7%	3.8%

	Yaw							St.Dev
	-30°	-20°	-10°	-0°	10°	20°	30°	
Anger	29.5%	31.1%	30.9%	31.0%	29.3%	31.5%	29.7%	0.9%
Disgust	29.3%	29.0%	29.8%	29.8%	27.4%	30.0%	30.7%	1.1%
Fear	18.1%	16.8%	17.4%	17.5%	16.0%	15.1%	14.4%	1.4%
Happiness	23.9%	24.3%	23.9%	24.2%	24.6%	24.5%	24.0%	0.3%
Sadness	22.6%	21.8%	23.0%	21.0%	22.8%	18.5%	18.8%	1.9%
Surprise	34.3%	35.5%	34.0%	34.9%	30.3%	33.8%	33.5%	1.7%

Table XII. RECOGNITION RATES IN HSEMOTION FOR FACIAL EXPRESSIONS OF DIGITAL HUMANS CAPTURED IN THE PITCH- AND YAW-DIRECTIONS.

	Pitch			St.Dev
	-20°	-0°	20°	
Anger	6.2%	6.6%	7.3%	0.6%
Disgust	27.3%	43.9%	56.5%	14.6%
Fear	17.2%	24.6%	25.7%	4.6%
Happiness	1.9%	2.0%	3.4%	0.8%
Sadness	47.2%	27.7%	9.9%	18.7%
Surprise	7.2%	9.1%	9.0%	1.1%

	Yaw							St.Dev
	-30°	-20°	-10°	-0°	10°	20°	30°	
Anger	5.0%	4.9%	6.7%	7.7%	6.6%	7.7%	8.4%	1.4%
Disgust	45.6%	45.7%	48.2%	43.4%	39.7%	37.5%	37.8%	4.2%
Fear	28.0%	27.2%	24.4%	22.5%	21.2%	17.2%	16.8%	4.5%
Happiness	2.9%	2.4%	2.8%	2.2%	2.7%	2.2%	1.8%	0.4%
Sadness	30.2%	29.2%	23.6%	22.0%	27.2%	31.6%	34.3%	4.4%
Surprise	7.3%	8.3%	10.5%	11.5%	6.9%	7.7%	6.6%	1.9%

In contrast to DeepFace, HSEmotion produces high Precision and Recall for the actor's "Anger," "Fear," "Happiness," "Sadness," and "Surprise," as shown in Table III. However, since only "Happiness" is recognized by the digital human (Table IV), we can consider that the digital human only expresses the actor's "Happiness" accurately, as mentioned above. This result is consistent with the findings of our previous study [1].

B. MACD-captured Facial Expressions

Tables V to VIII shows the average of the facial expression recognition rates of the 17 subjects and their corresponding digital humans captured by MACD, respectively. The numbers underlined in bold indicate the rate that the expressed facial expression was judged correctly, and shaded areas indicate high rates of incorrect recognition of the expressed facial expression.

The recognition results of DeepFace showed that most facial expressions of the subjects were judged as Neutral with the high rates (Table V). The same trend can be seen in the recognition results of the digital humans for these subjects (Table VI). These results differ significantly from the recognition results of the actor captured by LLF and its corresponding digital human.

On the other hand, HSEmotion fairly recognized the facial expressions of the subjects (Table VII). Only the subjects' "Fear" was relatively misrecognized as "Disgust." However, HSEmotion failed to recognize facial expressions of the digital humans (Table VIII).

To summarize, HSEmotion has more stable recognition rates compared to DeepFace for the same facial expressions of subjects who were captured from different directions. However, when these subjects' facial expressions were transferred to digital humans, both libraries failed to recognize their facial expressions.

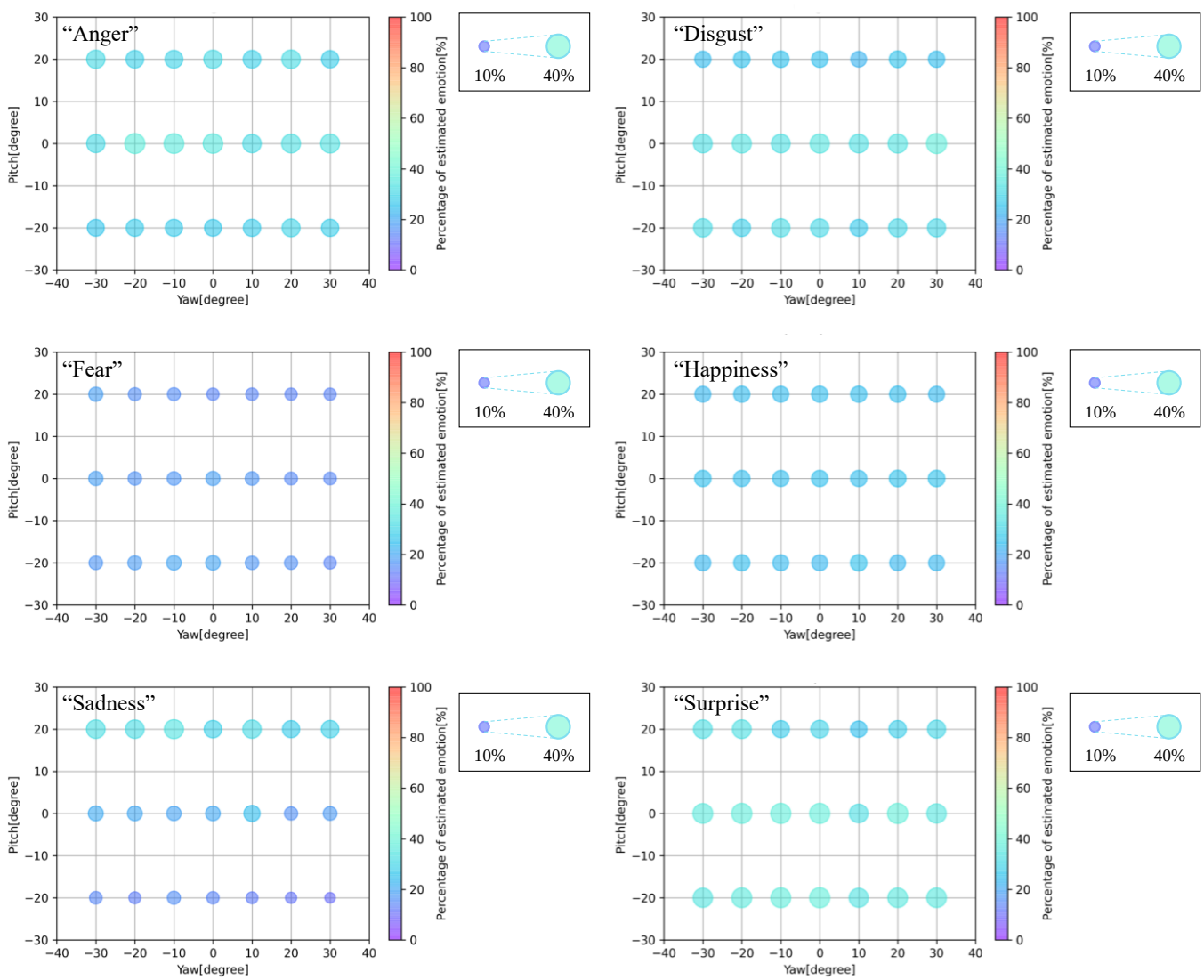


Figure 7. MACD-captured subjects' facial expressions recognized by the HSEmotion.

For a more detailed analysis, we averaged the facial expression recognition rates for three pitch- and six yaw-directions and examined the variation in facial expressions recognized for each direction (Tables IX to XII). Here, we used the Standard Deviation (St.Dev) to represent the variation in the recognition rates. Shaded areas indicate areas where variations in recognition rates differs by 5% or more depending on the camera angles, despite the same facial expression.

The changes of recognition rates in pitch- and yaw-directions are greater for DeepFace. On the other hand, HSEmotion showed less variation in recognition rates for facial expressions captured in different yaw directions. Finally, as shown in Figure 7, the facial expressions of the subjects captured by MACD were recognized by HSEmotion, and the average recognition rate for each angle was visualized. To intuitively see the recognition rate for each angle, the rate is represented by a circle and a color. From this figure, less variation in the recognition rates in the yaw-direction can be observed. These results of HSEmotion benefit from the use of eEfficientNet as its backbone and the large training dataset.

To obtain better facial expression recognition for our experiment, it is necessary to include digital human facial expression images in the training data. However, the limitations of this experiment were due to the fact that the MHC face images could not be used as training data by its agreement. Alternatives to MHC that can be used as training data are needed for further experiments.

V. Conclusion

The widespread use of digital humans has led to a demand to evaluate the facial expressions of them. In this study, we focused on the reality of facial expressions of digital humans

and evaluated the results of their facial expression using existing deep learning-based automated facial expression recognitions.

Two strategies were employed in the creation of the dataset to allow for a more detailed analysis. First, we created a dataset of facial expression images by an actor whose facial expressions were clearly expressed. Second, we created a dataset for 17 subjects' facial expressions using multi-angle camera devices. We then created a digital human based on this dataset, and objectively evaluated the reality of their facial expressions using two deep learning-based facial expression recognition libraries.

The evaluation using DeepFace and HSEmotion revealed that the accuracy of facial expression recognition by these libraries was problematic. HSEmotion was effective in evaluating the facial expressions of actual people. The library also absorbed differences in capture direction, providing stable recognition of facial expressions. However, we found that these libraries are not sufficient for evaluating expressions by digital humans, indicating the need for the development of better expression recognition libraries in the future. The terms of use of digital human in this study do not permit training on facial expressions by digital human, which prevented us from creating our own training model based on this dataset.

Future work will include a facial expression recognition library that considers both real people and digital humans. We would like to improve the facial expression recognition model using trainable digital humans to develop a facial expression recognition library that takes both actual people and digital humans into consideration.

REFERENCES

[1] S. Kikuchi, O. D. A. Prima, and H. Ito, “Do Digital Human Facial Expressions Represent Real Human's?,” The Fifteenth International Conference on Advances in Computer-Human Interactions, ACHI2022, pp. 1-5, 2022.

[2] Virtual Shibuya, https://news.kddi.com/kddi/corporate/newsrelease/2020/05/15/4437.html [retrieved: December, 2022]

[3] Juntendo Virtual Hospital, https://jp.newsroom.ibm.com/2022-04-13-Juntendo-Virtual-Hospital [retrieved: December, 2022]

[4] UNEEQ, https://digitalhumans.com [retrieved: October, 2022]

[5] Y. Iwayama, “Real Avatar Production - Raspberry Pi Zero W Based Low-Cost Full Body 3D Scan System Kit for VRM Format,” 10th International Conference and Exhibition on 3D Body Scanning and Processing Technologies, pp. 22-23, 2019.

[6] MediaPipe, https://google.github.io/mediapipe/ [retrieved: December, 2022]

[7] Digital Andy Serkis, https://www.unrealengine.com/en-US/blog/epic-games-and-3lateral-introduce-digital-andy-serkis [retrieved: December, 2022]

[8] Character Creator, https://www.reallusion.com/ [retrieved: December, 2022]

[9] Buddy Builder, https://www.hologress.com/ [retrieved: December, 2022]

[10] P. Ekman and W. V. Friesen, “Unmasking the Face: A Guide to Recognizing Emotions From Facial Expressions,” Malor Books, 2003.

[11] T. Kudoh and D. Matsumoto, “The Emotional World of the Japanese - Uncovering the Mysteries of Their Mysterious Culture,” Seishinshobo, 1996.

[12] T. Ohta, M. Tamura, M. Arita, N. Kiso, and Y. Saeki, “Facial Expression Analysis : Comparison with Results of Paul Ekman,” Special Issue for the 10th Anniversary of the Faculty of Nursing), 3(1), pp. 20-24, 2005. (in Japanese)

[13] S. I. Serengil and A. Ozpinar, "LightFace: A Hybrid Deep Face Recognition Framework," 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), pp. 1-5, 2020.

[14] A. V. Savchenko, “High-Speed Emotion Recognition Library,” Software Impacts, 14, pp. 1-4, 2022. https://doi.org/10.1016/j.simpa.2022.100433

[15] HSEmotion , https://github.com/HSE-asavchenko/face-emotion-recognition [retrieved: December, 2022]

[16] S. H. Kang and J. H. Watt, “The Impact of Avatar Realism and Anonymity on Effective Communication Via Mobile Devices,” Computers in Human Behavior, 29(3), pp. 1169-1181, 2013.

[17] M. Grewe et al., “Statistical Learning of Facial Expressions Improves Realism of Animated Avatar Faces,” Frontiers in Virtual Reality, 2, pp. 1-13, 2021.

[18] Meta Human Creator, https://www.unrealengine.com/en-US/metahuman-creator [retrieved: December, 2022]

[19] MeFaMo - MediaPipeFaceMocap, https://github.com/JimWest/MeFaMo [retrieved: December, 2022]

[20] Challenges in Representation Learning: Facial Expression Recognition Challenge, https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data [retrieved: December, 2022]

[21] A. Mollahosseini, B. Hasani, and M. H. Mahoor, “AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild,” IEEE Transactions on Affective Computing, 10(1), pp. 18-31, 2017.