ISSN 1870-4069

Impulsando los rostros del futuro: Evaluación comparativa de tecnologías de captura de movimiento facial para humanos digitales

Sharon Ramírez Lechuga[1], Carlos Vilchis[1],
Miguel Gonzalez Mendoza[1], Armando Rodríguez Mendoza[2],
Carmina Pérez Guerrero[2]

[1] Tecnológico de Monterrey,
Escuela de Ingeniería y Ciencias,
México

[2] Eugenia Virtual Humans S.A. de C.V.,
Laboratorio de Investigación,
México

{A01379035,carlos.vilchis, mgonza}@tec.mx,
{armando,carmina}@eugenia.tech

Resumen. El creciente universo de creadores de contenido virtuales, avatares del metaverso, y humanos digitales en general, ha creado una oportunidad para integrar soluciones de captura facial en un amplio panorama de nuevas aplicaciones para la industria de creación de contenido. Junto con este crecimiento se ha incrementado la demanda por humanos digitales que generen empatía y cuenten con un mejor desempeño en sus expresiones faciales. Es por esto que en el presente artículo se exploran las principales codificaciones faciales empleadas para la captura de movimiento facial y las diversas soluciones existentes para dar vida a humanos digitales. Adicionalmente, se presenta un experimento realizado con un humano digital dentro de un ambiente de realidad virtual para medir el vínculo empático creado a partir de algunas tecnologías recientes de captura facial, Faceware, Live Link UE, y Avatary. Los resultados exploran la percepción de determinadas expresiones emocionales, la respuesta empática, y la semblanza de familiaridad que reflejan las soluciones disponibles. Finalmente, se discute la necesidad de alternativas nuevas y accesibles con una codificación más expresiva, como medio para abrir el panorama a un amplio campo de investigación que busca mejorar la captura facial.

Palabras clave: Realidad virtual, captura de movimiento facial, interfaces humano-computadora, humanos digitales, codificación facial.

Driving the Faces of the Future: Benchmarking Facial Motion Capture Technologies for Digital Humans

Abstract. The growing universe of virtual content creators, metaverse avatars, and digital humans in general has created an opportunity to integrate facial capture solutions into a broad landscape of new applications for the content

creation industry. Along with this growth, the demand has increased for digital humans who generate empathy and have a better performance in their facial expressions. That is why this article explores the main facial codifications used for facial movement capture and the various existing solutions to bring digital humans to life. Additionally, an experiment carried out with a digital human within a virtual reality environment is presented to measure the empathic link created from some recent facial capture technologies, Faceware, Live Link UE, and Avatary. The results explore the perception of certain emotional expressions, the empathic response, and the semblance of familiarity that reflect the available solutions. Finally, the need for new and accessible alternatives with a more expressive coding is discussed, as a means to open the panorama to a wide field of research that seeks to improve facial capture.

Keywords: Virtual reality, facial motion capture, human-computer interfaces, digital humans, facial codification.

1. Introducción

Desde hace varios años, los avatares forman parte de un área de investigación experimental que busca explorar las interfaces humano-computadora [7]. Ahora, la tendencia de humanos digitales tiene expectativas importantes para los próximos 5 a 8 años del mercado de consumo [5]. Las posibles aplicaciones incluyen servicio al cliente, asistentes conversacionales, y soporte técnico virtual [6].

Recientemente, soluciones como Metahumans de Epic Games [11], han creado nuevas oportunidades, ya que pone a disposición del público una amplia gama de humanos digitales realistas y gratuitos, listos para ser utilizados dentro de procesos profesionales. Esto convierte a los humanos digitales realistas en herramientas sencillas y asequibles, que con tecnologías como la captura de movimiento y los gráficos 3D en tiempo real, se logran resultados de interacción mejorados [16].

Todos estos avances tienen la capacidad de cambiar la percepción de un ser humano digital gracias a la construcción de un vínculo emocional [24]. Este vínculo es necesario para impulsar el realismo, a fin de superar el Valle Inquietante [22], un término creado para describir el punto donde la respuesta emocional a representaciones humanas que tienen una apariencia y comportamiento similar al de un ser humano, causan una reacción negativa de extrañeza e inquietud.

Se han empleado humanos digitales para determinar la respuesta empática, la aceptabilidad, y la calidad de la interacción entre las computadoras y los humanos [27, 21]. Utilizando metodologías de seguimiento modernas, los resultados han mejorado [16] en comparación con experimentos realizados unos años atrás [1].

Entonces, la captura facial y su rendimiento es mejor ahora, pero ¿cómo mejoran las tecnologías de seguimiento facial disponibles, la respuesta empática de los humanos digitales democratizados de última generación? Para dar respuesta a esta pregunta y para aprovechar las interfaces empáticas y realistas modernas, este documento propone un experimento de percepción que involucra la realidad virtual (RV).

Tabla 1. Comparación de las principales soluciones analizadas para la captura de movimiento facial en tiempo real usadas en humanos digitales.

Solución de Captura Facial	Número de Blendshapes	Capacidad de Tiempo Real	Calibración Específica al Sujeto	Basado en Inteligencia Artificial	Codificación Facial	Inversión Aproximada
Faceware Studio	59	✓	×	×	FAPs	$2,340 US / Year
Live Link Face UE	51	✓	×	×	FAPs	Gratuito
Avatary	Ilimitadas	✓	✓	✓	FAPs & FACS	$2,388 US / Year

Los experimentos se basan en la interacción con un MetaHumano [11], impulsado con un conjunto contemporáneo de sistemas de captura facial: Faceware Studio de Faceware Tech, Live Link Face UE de Epic Games, y Avatary de Facegood. Todas son soluciones de vanguardia disponibles para la investigación y la creación de contenido, sin embargo, según nuestros conocimientos, este es el primer trabajo que los compara entre sí.

Por lo que los resultados obtenidos son importantes para medir el funcionamiento de estas tecnologías de seguimiento facial para ofrecer un rendimiento realista y de calidad. El resto del documento está estructurado de la siguiente manera: la sección 2 introduce el concepto de codificación facial utilizado para el seguimiento facial. La sección 3 presenta las soluciones de seguimiento facial contemporáneas.

La sección 4 detalla los métodos utilizados, los datos demográficos de los sujetos y el análisis estadístico aplicado a los resultados. La sección 5 muestra los resultados de la investigación. Finalmente, la sección 6 ofrece una discusión basada en los hallazgos, las áreas de oportunidad para futuras investigaciones y resume el trabajo de investigación presentado en este documento.

2. Codificación facial para la captura de movimiento

Las herramientas de seguimiento facial deconstruyen los rostros humanos para replicar su función, y la industria respalda este proceso con metodologías y estándares llamados Codificación facial. Hay dos corrientes principales, el Sistema de Codificación de Acción Facial (FACS por sus siglas en inglés) y los Parámetros de Animación Facial (FAPs por sus siglas en inglés).

La metodología FACS fue propuesta por Paul Ekman en los años 70s, para entender cómo las emociones y las expresiones faciales se relacionan con los huesos y músculos de nuestro rostro [10]. El modelo clasifica las expresiones facial por medio de etiquetas numeradas con diferentes niveles de intensidad (A, B, C, D o E), denominados Unidades de Acción. Estas unidades, con ciertas configuraciones, pueden representar emociones específicas denominadas FACS emocionales (emFACS) [13].

Hasta el día de hoy, esta metodología se utiliza como una de las formas más fiables de comprender las expresiones humanas. Más tarde, en los años 90, Moving Pictures Experts Group (MPEG-4) creó un estándar internacional para representar el habla y los gestos humanos en la animación, y un componente de ese estándar es el modelo de FAPs [23], que describe los movimientos faciales a medida que la unidad cambia desde una cara neutra.

Este modelo ha sido el estándar más común utilizado en escenarios de animación tradicional durante casi dos décadas debido a la simplicidad de su implementación en dibujos animados y modelos 3D [4]. Sin embargo, a medida que el realismo se convirtió en una prioridad, la industria comenzó a alcanzar los límites de esta codificación facial.

El modelo FACS se ha utilizado para experimentos con humanos digitales desde 1995 [19, 26], con investigaciones que continúan hasta el siglo XXI [20], evocando emociones de manera confiable en los rostros de los avatares.

Además, el modelo emFACS tiene un conjunto específico de combinaciones de movimientos musculares creadas para determinar las expresiones más comunes junto con la emoción humana correspondiente.

El modelo emFACS original incluye emociones como tristeza, neutral, disgusto, enojo, felicidad, sorpresa, y miedo. A base de este estándar, es posible verificar la efectividad de las expresiones faciales humanas digitales.

3. Estado del arte en soluciones de captura facial

Las herramientas de captura facial se hicieron populares a mediados de la década del 2010 [3], tras la mejora del procesamiento de video en tiempo real para hacer seguimiento en vivo. Las primeras opciones se basaban en Head-Mounted Cameras (HMC), que solo graban vídeo para ser procesado en postproducción, como Vicon Cara o primeras versiones de Faceware con cámaras GoPro.

Más tarde, las herramientas de seguimiento en vivo basadas en el reconocimiento facial estuvieron disponibles con las cámaras de los teléfonos inteligentes, como la cámara TrueDepth incluida en los iPhone y utilizada por Live Link Face UE. Hoy en día, las opciones más populares y disponibles para el seguimiento facial son Faceware [2], Live Link Face UE [14] y Avatary [12].

Faceware y Live Live Link Face UE utilizan el método de 51 o más blendshapes porque se ha convertido en el estándar de conjuntos de expresiones en la animación facial debido a su similitud con la codificación FAP. Solo Avatary de Facegood tiene la opción de impulsar conjuntos de expresiones personalizados más grandes, como lo requiere el modelo FACS. Cada una de estas soluciones están compiladas en la tabla 1, teniendo en cuenta características específicas que se compararon para los fines de esta investigación.

4. Configuración experimental

El análisis para evaluar la percepción de muestras emotivas visuales en la interacción humano-computadora, por lo general involucra la medición del Efecto del Valle Inquietante, que se puede realizar a través de la evaluación de las dimensiones de la personalidad con la encuesta Big–Five [8] como en Hyde et al. [18], a través de la percepción de las dimensiones afectivas definidas por Ho y MacDorma [17] o a través de ensayos interactivos como la modificación al experimento del Mago de Oz, aplicado con humanos digitales por Seymour, Riemer y Kay [24], donde un humano digital se expone a sujetos humanos, mitigando el efecto del Valle Inquietante con interactividad.

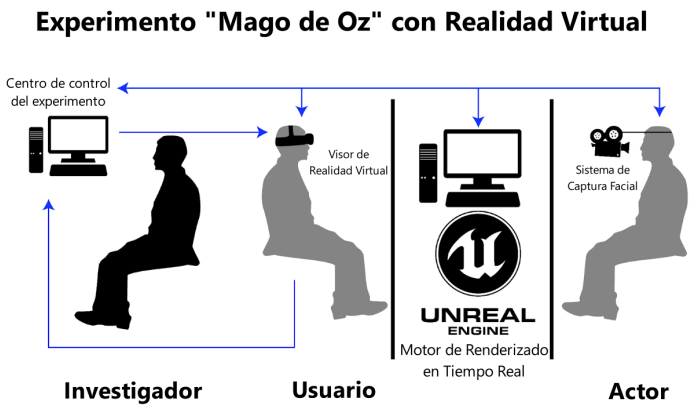

Fig. 1. El diseño del Experimento Mago de Oz utilizado en este documento, basado en el trabajo de Seymour [24] para integrar la interactividad en tiempo real como una variable clave, que se complementa con el uso de un visor de RV en el presente trabajo.

Para lograr esto se necesita un conjunto específico de elementos tecnológicos y la definición de un sistema de evaluación que permita explorar cómo se percibe a un Humano Digital, junto con la efectividad de los sistemas de seguimiento facial. Estos requisitos llevaron a la creación de pruebas específicas que mezclan aspectos psicológicos, de familiaridad e interactividad.

El experimento propuesto en esta investigación agrega una nueva capa de interactividad al experimento modificado del ”Mago de Oz”, al involucrar el uso de la realidad virtual, como se ilustra en la Figura 1. El estudio se realiza entre varios sujetos para evaluar diferentes tecnologías de captura facial a través de la interacción con un Metahumano [11] en un entorno de realidad virtual.

El diseño del estudio propuesto expone a cada sujeto humano a una sola tecnología de captura facial y una sola exposición interactiva, evitando cualquier sesgo debido a experiencias previas o expectativas emocionales. El experimento cuenta con dos salas diferentes que permiten un espacio suficiente para la experiencia del usuario y el rendimiento de los sistemas de captura facial. Las salas son la sala experimental y la sala de control.

La sala experimental es donde el usuario está expuesto a un humano digital a través de un visor de RV. Para el experimento descrito en este documento, los visores contienen animaciones de emFAC pre-grabadas con los sistemas de captura facial obtenidas de un actor, que luego se transmiten a través de un motor de 3D en tiempo real, en este caso, Unreal Engine 5.0.2.

La sala de control, por otro lado, contiene la estación de trabajo donde las expresiones grabadas a partir de los diferentes sistemas de seguimiento facial se transmiten en diferentes sesiones y en orden aleatorio. Un Metahumano [11] descargado con la más alta calidad, se carga en Unreal Engine y se procesa en una computadora con un procesador Intel i7-8700K de 12 núcleos, 40 GB DDR4 de RAM y una tarjeta GPU RTX 3070 con 12 GB GDDR6X.

La computadora entrega la representación final en tiempo real de las secuencias animadas a la sala experimental. El sistema de comunicación conecta ambas estaciones de trabajo en una dirección de transmisión unidireccional.

4.1. Diseño de experimentos

El experimento expone a un sujeto humano con un visor de RV, el usuario, a un entorno virtual con un humano digital y le asigna la tarea de resolver una encuesta dentro del ambiente de RV. El usuario de RV se encuentra aislado dentro de la experiencia para no sentir presión ni estímulos externos. El experimento toma el tiempo necesario para cubrir un conjunto aleatorio de emFACS creados para construir un puente empático entre los sujetos, además de una actuación adicional de 60 segundos, donde el humano digital habla sobre su vida, sus cosas favoritas, pensamientos, entre otras cosas, para crear una experiencia empática con el sujeto.

En la Figura 2 se puede observar un ejemplo del emFACS de enojo capturado por las diferentes soluciones de seguimiento facial exploradas en este experimento.Las preguntas de la encuesta se dividen en dos grupos. El primer grupo de preguntas está relacionado con la identificación del conjunto de emFACS específicos representados en el humano digital (tristeza, neutral, disgusto, enojo, felicidad, sorpresa, y miedo), que consiste en la presentación de una expresión emFACS seleccionada aleatoriamente, y el sujeto tiene la tarea de seleccionar el emFACS percibido.

El segundo grupo de preguntas está relacionado con la encuesta Big–Five de empatía y familiaridad [18], que consta de 5 preguntas en las que el usuario comparte su opinión sobre cuán confiable, amigable, familiar, atractivo y realista parece el humano digital, con opciones para cada cualidad presentadas en una Escala Likert de 1 a 7 para su análisis, donde 1 significa fuerte desacuerdo, 7 significa fuerte acuerdo y 4 es una respuesta neutral.

4.2. Participantes

El experimento se realizó con un grupo de 42 personas que fueron seleccionadas al azar de la comunidad que deambula cerca del laboratorio de ingeniería. Los sujetos fueron expuestos a solo una tecnología de seguimiento facial cada uno, lo que resultó en 3 grupos de 14 sujetos, un grupo por tecnología. Se les pidió que interactuaran con el experimento durante cinco minutos sin información previa sobre la experiencia interactiva.

Cuando se les presentó el experimento, se preguntó a los sujetos sobre su género, edad y si estaban familiarizados con humanos digitales o RV. La distribución del grupo fue un 58,1 % de hombres y un 41,9 % de mujeres. La edad media del grupo es de 25.7 años con una desviación estándar de 6.9.

Los sujetos que conocían el concepto de humanos digitales antes del experimento representan el 45,2 %, y los que no conocían el concepto de humanos digitales representan el 54,8 %. Los sujetos que estaban familiarizados con la RV antes del experimento representan el 71 %, mientras que los que no estaban familiarizados con la RV representan el 29 %.

Fig. 2. Una muestra del emFACS de enojo. De izquierda a derecha se muestra una captura del video del actor, la expresión por Live Link Face UE, la expresión por Faceware y la expresión por Avatary.

4.3. Análisis estadístico

Para validar los resultados del experimento presentado, se emplean diferentes métodos de prueba de hipótesis para la identificación emFACS y la encuesta empática y de familiaridad Big–Five con un valor de nivel de significacia de $0{,}05$. Se tomó en cuenta que cada sujeto fue expuesto aleatoriamente a una sola tecnología y experimentó una sola ejecución del experimento, así como el hecho que los emFACS se presentaron en un orden aleatorio por tecnología. Más información sobre los enfoques estadísticos utilizados se detalla en las siguientes subsecciones.

Identificación emFACS Para las pruebas de percepción resultantes, el enfoque propuesto es una prueba U de Mann-Whitney para analizar la diferencia entre los porcentajes obtenidos a partir de una matriz de confusión de los resultados. También se propone una prueba de bondad de ajuste Chi-Cuadrado de Pearson, con las observaciones percibidas y las observaciones esperadas donde todos los emFACS serían correctamente identificados.

La prueba U de Mann-Whitney se puede utilizar para comprobar si dos muestras independientes tienen una diferencia estadísticamente significativa. Esta prueba también se considera el equivalente no paramétrico de la prueba t de independencia de dos muestras. Los supuestos para la prueba U de Mann-Whitney incluyen muestras aleatorias e independientes, así como un tamaño de muestra pequeño con menos de 30 muestras.

Dado que las muestras se obtienen por tecnología, donde un sujeto se expone una vez a una sola tecnología, con un orden aleatorio de emFACS por ejecución, la suposición de muestras aleatorias e independientes se aplica al experimento. Dado que el grupo de muestra por tecnología es de 14 sujetos, esta condición también se aplica al experimento.

Para esta prueba, la hipótesis nula supone que ninguno de los modelos comparados funciona mejor que el otro, y la hipótesis alternativa supone que los rendimientos de los modelos comparados no son iguales. El valor crítico U en el nivel de significancia $0{,}05$ es 8.

La prueba de chi-cuadrado de Pearson es una prueba estadística para datos categóricos. Se puede usar para probar la bondad de ajuste, la independencia o la homogeneidad. La prueba de bondad de ajuste chi-cuadrado se puede usar cuando se trata de una variable categórica. Le permite probar si la distribución de frecuencias de la variable categórica es significativamente diferente de las expectativas de proporciones iguales.

Para esta prueba, la hipótesis nula asume que los emFACS percibidos obtenidos a partir de tecnologías de seguimiento facial están en proporciones iguales a los emFACS observados, y la hipótesis alternativa asume que los emFACS percibidos a partir de tecnologías de seguimiento facial están en diferentes proporciones a los emFACS observados.

Cuestionario de empatía y familiaridad Big–Five. Los resultados obtenidos a partir de los resultados de la escala de Likert se comparan a través del análisis estadístico mediante la prueba de hipótesis. El enfoque propuesto es una prueba U de Mann-Whitney, que se ha descrito en la subsección anterior. Dado que las muestras se obtienen por tecnología, donde un sujeto se expone una vez a una sola tecnología, con un rendimiento de 60 segundos, la suposición de muestras aleatorias e independientes es adecuada.

Finalmente, dado que el tamaño de muestra por tecnología es de 14, inferior a 30, se puede considerar un tamaño de muestra pequeño. Para esta prueba, la hipótesis nula asume que las dos tecnologías comparadas tienen respuestas empáticas y similares, y la hipótesis alternativa asume que existe una diferencia estadísticamente significativa en las respuestas empáticas y de familiaridad entre las dos tecnologías comparadas. El valor crítico U en el nivel de significancia $0{,}05$ es 7.

5. Resultados

5.1. Identificación de emFACS

El experimento de identificación emFACS consistió en el uso de 7 conjuntos diferentes de videos en orden aleatorio de humanos digitales que expresan emociones. Se encargó a un grupo de sujetos que observaran y reconocieran los emFACS que podrían ser tristeza, neutral, disgusto, enojo, felicidad, sorpresa, o miedo.

Matriz de confusión. Los resultados del reconocimiento general se pueden observar en la matriz de confusión ilustrada en la Tabla 2. Dado que en otros casos de investigación de reconocimiento de emociones [15, 25, 9], una precisión considerada fiables en escenarios aleatorios podría ir del 60 %-86 %, el umbral utilizado para evaluar los resultados de estos experimentos como confiables van desde el 60 % en adelante.

Con eso en consideración, se puede observar que la mayoría de los sujetos pueden reconocer de manera confiable ciertas expresiones a través de las tres tecnologías, como Neutral (Faceware: 100 %, Avatary: 100 %, Live Link Face UE: 100 %), Felicidad (Faceware: 71,43 %, Avatary: 92,86 %, Live Link Face UE: 78,57 %), y Sorpresa (Faceware: 78,57 %, Avatary: 92,86 %, Live Link Face UE: 92,86), con Avatary y Live Link Face UE generalmente mostrando mejor desempeño que Faceware.

Tabla 2. Porcentaje de precisión de las emociones reconocidas por el grupo de prueba. FW representa Faceware; AV representa Avatary; LLF representa Live Link Face UE.

	emFACS Reconocidas																					
	Tristeza			**Neutral**			**Disgusto**			**Enojo**			**Felicidad**			**Sorpresa**			**Miedo**		
	FW	**AV**	**LLF**	**FW**	**AV**	**LLF**	**FW**	**AV**	**LLF**	**FW**	**AV**	**LLF**	**FW**	**AV**	**LLF**	**FW**	**AV**	**LLF**	**FW**	**AV**	**LLF**
Tristeza	7.14	64.28	28.57	0	0	0	0	21.43	0	0	7.14	0	0	0	14.29	35.71	0	50	57.14	7.14	7.14
Neutral	0	0	0	100	100	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Disgusto	7.14	78.57	0	0	0	0	64.29	7.14	50	21.43	0	14.29	7.14	0	21.42	0	14.29	0	0	0	14.29
Enojo	0	0	0	0	0	0	71.43	21.43	42.86	14.29	78.57	42.86	7.14	0	14.29	0	0	0	7.14	0	0
Felicidad	0	0	0	0	0	7.14	7.14	0	7.14	0	0	7.14	71.43	92.86	78.57	21.43	7.14	0	0	0	0
Sorpresa	0	0	0	0	0	0	0	7.14	0	0	0	0	0	0	0	78.57	92.86	92.86	21.43	0	7.14
Miedo	7.14	7.14	0	0	0	7.14	42.86	0	28.57	7.14	0	14.29	0	0	0	14.29	92.86	42.86	28.57	0	7.14

En Tristeza, solo Avatary obtuvo un resultado fiable del 64,28 %, mientras que con Faceware se confundía mayoritariamente con Miedo (57,14 %) y Sorpresa (35,71 %), y con Live Link Face UE se confundiá mayoritariamente con Sorpresa (50 %). Para Disgusto, solo Faceware obtuvo un resultado fiable del 64,29 %, mientras que con Avatary se confundió mayoritariamente con Tristeza (78,57 %), y con Live Link Face UE se confundió un poco con Felicidad (21,42 %).

Para Enojo, solo Avatary obtuvo un resultado fiable del 78,57 %, mientras que Faceware y Live Link Face UE se confundieron mayoritariamente con Disgusto (71,43 % y 42,86 % respectivamente).

Finalmente, para Miedo, las tres tecnologías funcionaron mal (Faceware: 28,57 %, Avatary: 0 %, Live Link Face UE: 7,14 %), con Faceware se confundió mayoritariamente con Disgusto (42,86 %), con Avatary fue completamente confundido con Sorpresa (92,86 %), y con Live Link Face UE se confundió mayoritariamente con Sorpresa también (42,86 %).

Prueba U de Mann-Whitney. Con base en los valores obtenidos de la prueba U de Mann-Whitney sobre los resultados de percepción emFACS, se encontró que entre Link Face UE y Faceware, el valor U es de 22 y el valor p es de $0{,}7949$; entre Link Face UE y Avatary, el valor U es $22,5$ y el valor p es $0,8493$; entre Avatary y Faceware el valor U es 21 y el valor p es $0{,}7039$.

Dado que el valor p de todos los sistemas de seguimiento facial es mayor que nuestro umbral de significancia asumido ($\alpha = 0{,}05$), y todos los valores U son mayores que el valor crítico en ese nivel de importancia ($U = 8$), no se puede rechazar la hipótesis nula y se concluye que no hay pruebas suficientes para afirmar que existe una diferencia estadísticamente significativa entre cualquiera de las tecnologías de seguimiento facial.

Sin embargo, según los valores p y dado que cada tecnología comparte el mismo tamaño de muestra, el orden de las tecnologías desde el valor p más pequeño hasta el valor p más grande es Avatary y Faceware, Link Face UE y Faceware, y Link Face UE y Avatary.

Prueba de Chi-Cuadrado. Con base en los valores obtenidos de la prueba Chi-Cuadrado en los resultados de percepción de emFACS, se encontró que para Faceware, el valor p es $0{,}00153$ y el estadístico de prueba X^2 es $21{,}44$; para Link Face UE, el valor p es $0{,}00201$ y la estadística de prueba X^2 es $20{,}78$; para Facegood, el valor p es $0{,}00435$ y la estadística de prueba X^2 es $18{,}89$.

Dado que el valor p de todos los sistemas de seguimiento facial es más pequeño que nuestro umbral de importancia asumido ($\alpha = 0{,}05$), y todas las estadísticas de prueba X^2 no están en la región de aceptación de $95\,\%$, rechazamos nuestra hipótesis nula y asumimos que existe una diferencia estadísticamente significativa entre el emFACS percibidas de las tecnologías de seguimiento facial y las observaciones esperadas de emFACS. Sin embargo, según los valores p y dado que cada tecnología comparte el mismo tamaño de muestra, el orden de las tecnologías del valor p más pequeño al valor p más grande es Faceware, Link Face UE y Avatary.

5.2. Encuesta empática y de familiaridad big-five

La encuesta Big–Five consistió en 5 preguntas en las que los sujetos compartían su percepción de cuán confiable, amigable, familiar, atractivo y realista parecía el ser humano digital durante una actuación adicional de 60 segundos. Las respuestas se representaron en una escala de Likert de 1 a 7, donde 1 significa fuerte desacuerdo, 7 significa fuerte acuerdo y 4 es una respuesta neutral.

Escala Likert. Los resultados se resumen en la Fig. 3. Se puede observar que en la categoría de Realista, Faceware y Avatary presentan un mayor porcentaje que Live Link Face UE, donde Faceware es el que presenta mayor porcentaje por una pequeña diferencia.

En la categoría Atractivo, Avatary presenta un porcentaje visiblemente menor en comparación con las otras dos tecnologías, sin embargo, en la categoría Familiar se presenta el comportamiento contrario. Finalmente, Live Link Face UE tiene una clara ventaja sobre las otras tecnologías en las categorías de Amigable y Confiable.

Prueba U de Mann-Whitney. Con base en los valores obtenidos de la Prueba U de Mann-Whitney en los resultados de la Encuesta Big–Five, se encontró que entre Faceware y Link Face UE, el valor U es 8 y el valor p es $0{,}13362$; entre Avatary y Link Face UE, el valor U es 6 y el valor p es $0{,}02574$; entre Avatary y Faceware, el valor U es 12 y el valor p es $0{,}27572$.

Solo el valor p de la comparación entre Avatary y Live Link Face UE es menor que nuestro umbral de importancia asumido ($\alpha = 0{,}05$) y el valor U es menor que el valor crítico en ese nivel de importancia ($U = 7$), por lo que es el único caso en el que rechazamos nuestra hipótesis nula y asumimos que existe una diferencia estadísticamente significativa en las respuestas empáticas y de similitud entre Avatary y Link Face UE.

Sin embargo, en base a los otros valores p y dado que comparten el mismo tamaño de muestra, el orden del par de tecnologías desde el valor p más pequeño hasta el valor p más grande es Faceware y Link Face UE, seguidos por Facegood y Faceware.

6. Discusión y conclusiones

Este documento explica las diferencias entre las codificaciones faciales, presenta una comparación de algunas soluciones comerciales para el seguimiento facial en humanos digitales, expone un diseño experimentos empáticos que usan RV y compara Faceware, Live Link Face UE y Avatary, sistemas contemoráneos de captura facial.

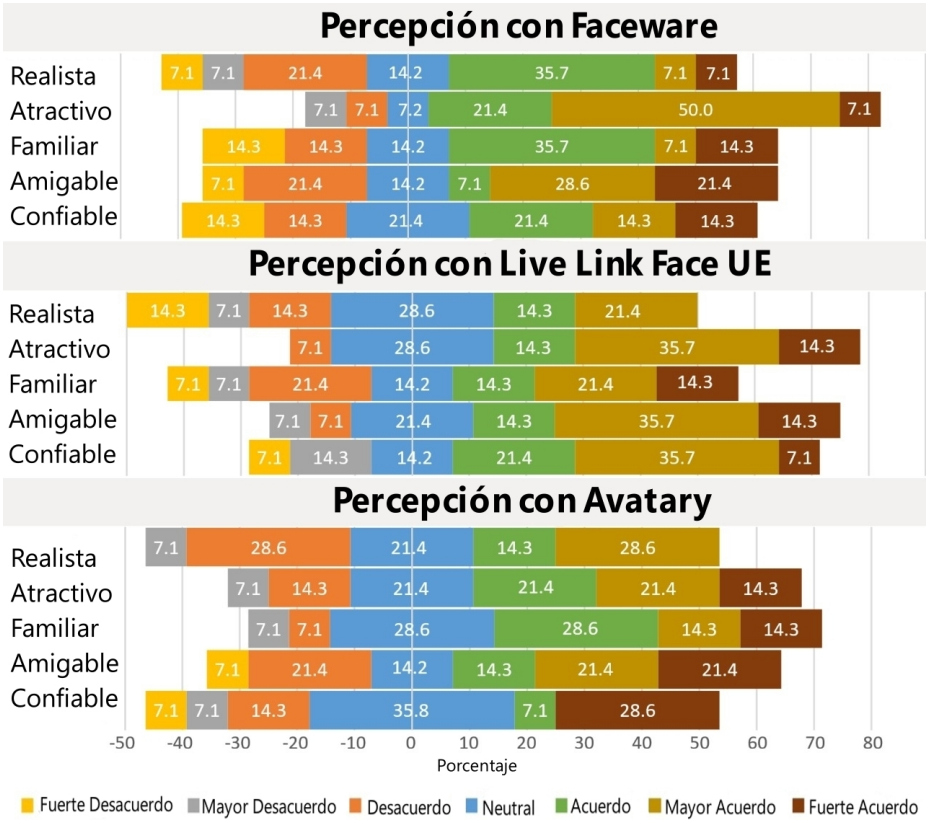

Fig. 3. Visualización de los resultados de la encuesta Big–Five representados en una escala Likert.

El objetivo de los experimentos es evaluar las capacidades de representación de emociones faciales de estas tecnologías a través de una prueba de percepción emFACS, y hasta qué punto las expresiones mostradas con estas tecnologías se alejan del Valle Inquietante, a través de la encuesta Big-Five.

De igual manera, durante cada paso del experimento se evaluaron las limitaciones y ventajas de las opciones disponibles. Las tres tecnologías muestran una representación confiable de emFACS Neutral, Felicidad y Sorpresa, con Avatary y Live Link Face UE, generalmente funcionando mejor que Faceware. Con respecto a los emFACS de Tristeza, Disgusto y Enojo, generalmente una tecnología muestra mejor desempeño que otra, sin embargo, ninguna de las soluciones pudo representar de manera confiable el emFACS de Miedo, lo que muestra un área de oportunidad enfocada en la representación realista del Miedo.

El análisis estadístico de los resultados no pudo encontrar una diferencia estadísticamente significativa entre las tecnologías o que alguna de ellas mostrara similitudes estadísticamente significativas con la percepción esperada de emFACS, por lo que una investigación adicional debe incluir un grupo más grande de sujetos para mostrar potenciales diferencias y similitudes estadísticamente significativas.

En comparación con otras tecnologías, Faceware presenta un realismo y atractivo superiores, Avatary presenta una familiaridad superior y Live Link Face UE presenta una amigabilidad y confiabilidad superiores. El análisis estadístico de los resultados de la encuesta Big–Five solo encontró una diferencia estadísticamente significativa entre Avatary y Live Link Face UE. por lo tanto, una investigación adicional debe incluir un grupo más grande de sujetos para mostrar potenciales diferencias estadísticamente significativas entre Faceware y Link Face UE o Avatary y Faceware.

Dado que la codificación FAP se utiliza en la mayoría de las soluciones existentes, esta investigación demuestra que las soluciones de última generación tienen un área de oportunidad relacionada con la investigación con otras codificaciones faciales que pueden aprovechar aún más las expresiones.

Además, Live Link Face UE representa una solución sencilla, asequible y democratizada para la investigación y el desarrollo con resultados estandarizados similares a las costosas herramientas de seguimiento facial. Las ventajas de usar opciones democratizadas pueden abrir nuevas direcciones para la investigación y la innovación en el campo, pero aún existe la necesidad de mejorar la respuesta empática y el desempeño de las expresiones faciales.

Estas necesidades pueden conducir a una investigación predominante sobre opciones democráticas de codificaciones más expresivas, soluciones novedosas para realizar la captura de movimiento de personas, y bases de datos con humanos digitales. Finalmente, el enfoque de evaluación presentado, podría usarse para evaluar futuras soluciones de seguimiento facial en términos de percepción.

Agradecimientos. Este trabajo fue apoyado a través de una beca para Carlos Vilchis por parte del Consejo Nacional de Ciencia y Tecnología de México (CONACYT). Este trabajo también fue apoyado por el programa Epic MegaGrants bajo el nombre de Grant FACS DEEP LEARNING TOOL.

Referencias

1. Amini, R., Lisetti, C., Ruiz, G.: Hapfacs 3.0: Facs-based facial expression generator for 3d speaking virtual characters. IEEE Transactions on Affective Computing, vol. 6, no. 4, pp. 348–360 (2015) doi: 10.1109/TAFFC.2015.2432794
2. Bausch, P.: Faceware website. Faceware, (2021)
3. Bennett, G., Kruse, J.: Teaching visual storytelling for virtual production pipelines incorporating motion capture and visual effects. In: Special Interest Group on Computer Graphics and Interactive Techniques Asia Symposium on Education (2015) doi: 10.1145/2818498.2818516
4. Briggs, C.: An essential introduction to maya character rigging. Chemical Rubber Company Press (2021)
5. Burke, B., Davis, M., Dawson, P.: Hype cycle for emerging technologies. Gartner Research (2021)
6. Caballar, R. D.: Are digital humans the next step in human-computer interaction? Spectrum IEEE (2021)
7. Cassell, J.: Embodied conversational interface agents. Communications of the Association for Computing Machinery, vol. 43, no. 4, pp. 70–78 (2000)

8. Costa, P., McCrae, R.: A five-factor theory of personality. The Five-Factor Model of Personality: Theoretical Perspectives, vol. 2, pp. 51–87 (1999)
9. Dzedzickis, A., Kaklauskas, A., Bucinskas, V.: Human emotion recognition: Review of sensors and methods. Sensors, vol. 20, no. 3, pp. 592 (2020)
10. Ekman, P., Rosenberg, E. L.: What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system, Oxford University Press (2005)
11. Epic Games: Epic games metahuman creator. Metahuman Unreal Engine (2021)
12. FACEGOOD Co. Ltd.: Avatary by facegood (2021)
13. Friesen, W. V., Ekman, P.: Emfacs-7: Emotional facial action coding system. University of California at San Francisco, vol. 2, no. 36, pp. 1 (1983)
14. Games, I. E.: Live link face for UE (2021)
15. Gavrilescu, M.: Recognizing emotions from videos by studying facial expressions, body postures and hand gestures. In: 23rd Telecommunications Forum Telfor, pp. 720–723 (2015) doi: 10.1109/TELFOR.2015.7377568
16. Higgins, D., Egan, D., Fribourg, R., Cowan, B., McDonnell, R.: Ascending from the valley: Can state-of-the-art photorealism avoid the uncanny? In: Association for Computing Machinery Symposium on Applied Perception 2021 (2021) doi: 10.1145/3474451.3476242
17. Ho, C. C., MacDorman, K. F.: Measuring the uncanny valley effect. International Journal of Social Robotics, vol. 9, pp. 129–139 (2017) doi: 10.1007/s12369-016-0380-9
18. Hyde, J., Carter, E. J., Kiesler, S., Hodgins, J. K.: Using an interactive avatar's facial expressiveness to increase persuasiveness and socialness. In: Proceedings of the 33rd Annual Association for Computing Machinery Conference on Human Factors in Computing Systems, pp. 1719–1728 (2015)
19. Ko, H., Kim, J. H., Kim, J.: Searching for facial expression by genetic algorithm. In: Virtual Environments '95, pp. 87–98 (1995) doi: 10.1007/978-3-7091-9433-1_8
20. Malatesta, L., Raouzaiou, A., Karpouzis, K., Kollias, S.: MPEG-4 facial expression synthesis. Personal and Ubiquitous Computing, vol. 13, pp. 77–83 (2009) doi: 10.1007/s00779-007-0164-1
21. McDonnell, R., Breidt, M., Bulthoff, H.: Render me real?: Investigating the effect of render style on the perception of animated virtual humans. Association for Computing Machinery, vol. 31, pp. 1–91 (2012)
22. Mori, M., MacDorman, K. F., Kageki, N.: The uncanny valley [from the field]. IEEE Robotics & Automation Magazine, vol. 19, no. 2, pp. 98–100 (2012)
23. Pandzic, I. S., Forchheimer, R.: MPEG-4 facial animation: The standard, implementation and applications. John Wiley and Sons, Inc (2003)
24. Seymour, M., Riemer, K., Kay, J.: Interactive realistic digital avatars-revisiting the uncanny valley. In: Proceedings of the 50th Hawaii International Conference on System Sciences, pp. 547–556 (2017)
25. Shan, C., Gong, S., McOwan, P. W.: Beyond facial expressions: Learning human emotion from body gestures. In: Proceedings of the British Machine Vision Conference, pp. 43–44 (2007) doi: 10.5244/C.21.43
26. Terzopoulos, D.: Modeling living systems for computer vision. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, vol. 1, pp. 1003–1013 (1995)
27. Zibrek, K., Martin, S., McDonnell, R.: Is photorealism important for perception of expressive virtual humans in virtual reality? Association for Computing Machinery, vol. 16, no. 3 (2019) doi: 10.1145/3349609