Metaverso y XR. Una visión profesional

Daniel Rivas-Perpen
Netflix Eyeline

Referencia de este artículo

Rivas-Perpen, Daniel (2023). Metaverso y XR. Una visión profesional. En: *adComunica. Revista Científica de Estrategias, Tendencias e Innovación en Comunicación*, n°26. Castellón de la Plana: Departamento de Ciencias de la Comunicación de la Comunicación, 331-334.

El metaverso es, sin duda, una de las tendencias actuales que brinda la tecnología con mayor proyección y que más expectativas genera.

En este texto presento un breve análisis de la situación actual del metaverso y algunas de las tecnologías en las que se sostiene. Pero antes paso a señalar algunos hitos de mi trayectoria profesional. Desde hace 24 años estoy relacionado profesionalmente con los gráficos por ordenador. Primero desde una perspectiva más artística pero gradualmente sintiéndome más interesado por el lado técnico. Después de trabajar por unos años en televisión, Localia TV, en Fuengirola, Marbella y Málaga, y tras pasar por Irlanda y Suiza donde ejercí de director técnico en gráficos en movimiento, acabé en EEUU donde participé en proyectos para

Sony, Paramount, NFL, ABC, y Disney entre otros. En 2017 entré a formar parte del equipo inicial de Intel Studios, el mayor estudio de captura volumétrica en Manhattan Beach, California. Allí nos enfrentamos a problemas para los que no había herramientas, teniendo que construirlas nosotros y fue ahí donde se aceleró mi transición al lado técnico de los gráficos por ordenador. Nuestro equipo fue premiado con un Lumiere Award al mejor uso de la tecnología AR en una experiencia musical y a título personal, Intel patentó uno de mis algoritmos de captura volumétrica. Actualmente formó parte del equipo de investigación y desarrollo de captura volumétrica de Paul Debevec en Eyeline Studios, la rama de captura volumétrica para cine y televisión de Netflix.

Me gustaría aclarar que voy a referirme a XR como el conjunto de las tecnologías VR (realidad virtual) y AR (realidad aumentada) y sus derivados.

A diferencia, por ejemplo, del cine, el *contenido real* para XR es notablemente más complicado de capturar, manipular y reproducir. En el cine, la captura y manipulación lleva décadas perfeccionándose, con aparatos y herramientas que han ido evolucionando en paralelo con las expectativas de la audiencia.

En el mundo XR, la captura de *contenido real* depende de una serie de factores que multiplican la dificultad y la escala de la tarea. Uno de esos factores es la necesidad de un elevado número de cámaras que graben la acción desde diferentes puntos de vista. Otro es la reconstrucción en 3d de la escena a partir de las imágenes capturadas. Este proceso se realiza generalmente con herramientas de fotogrametría que fueron pensadas, efectivamente, para reconstruir una escena 3d a partir de decenas o cientos de imágenes, pero necesitando numerosos ajustes manuales, más un proceso artesanal que automatizado. La clientela de estas herramientas eran principalmente estudios topográficos y de arquitectura. En cambio, en el proceso de reconstrucción de escenas 3d para XR necesitamos movimiento, eso es como mínimo generar 24 de esas escenas 3d por cada segundo de grabación. Aunque a raíz de varias pruebas con diversos visores VR que realizamos en Intel Studios, concluimos que, tanto para el *contenido real* como *sintético* se necesita un mínimo de 60 imágenes por segundo para evitar situaciones cercanas a la cinetosis.

En el caso de *contenido sintético*, eliminamos de la ecuación el primer paso de captura, que es sustituido por otros métodos que sí han venido evolucionando desde hace tiempo, principalmente herramientas de creación de contenido en 3d y los motores para creación para videojuegos. Pero aunque esto libera al proceso de uno de sus más grandes lastres, seguimos teniendo el problema de ajustar el contenido y la calidad de este a las especificaciones de cada dispositivo en el que va a ser reproducido. ¿Cómo hacer que un contenido XR de gran detalle y calidad visual corra tanto en un PC de última generación como en un dispositivo portátil sin que nadie se maree?

Empresas como Meta, Microsoft, Google, Apple o Sony han contratado a los mejores ingenieros, con una inversión milmillonaria que ha hecho avanzar la calidad del contenido XR y del reconocimiento gestual. Para dar algo de perspectiva me gustaría destacar que el departamento Reality Labs, la sección de Meta relacionada de una manera u otra con su metaverso, llegó a contar con más de 17.000 empleados. Ese número es mayor que el de la plantilla completa de empresas como AMD o Levi Strauss. Pero al final todo pasa por el embudo del *hardware*, y estos equipos tienen actualmente que lidiar con dos procesos: uno, generar y mover un mundo virtual en 3D y dos, y en paralelo, ejecutar los algoritmos que detectan nuestros gestos, movimientos y acciones. Ese equipo, quizás mejor dicho, la miniaturización de ese equipo, de ese *hardware*, no van tan rápido como el desarrollo del *software*.

Esa disonancia no sería tan profunda si, como ocurrió con el nacimiento de los videojuegos, no hubiera ninguna referencia en la que sostener expectativas. Ahora por el contrario, los primeros ejemplos de metaversos se han topado con una audiencia educada en imágenes generadas por ordenador indistinguibles de la realidad, completamente fotorrealistas. Y parece que esta audiencia no está dispuesta a hacer el esfuerzo imaginativo que se hacía con los primeros videojuegos, donde el cerebro y su lado creativo rellenaban los huecos que la tecnología todavía no podía.

Desde mediados de 2022 ha habido tandas encadenadas de despidos que han afectado profundamente a las grandes tecnológicas. En la mayoría de los casos no tanto por pura necesidad financiera, como por realizar el sacrificio demandado por los grandes accionistas, lo que hace curioso que las secciones relacionadas con XD hayan sido de las más afectadas: Intel, Google, Tenent, y sobre todo Meta y Microsoft han echado el freno al desarrollo en XR. Una de las lecturas posibles es que, aunque innegablemente siguen viendo el potencial del metaverso, se han dado cuenta de que hay que esperar al *hardware*.

En los últimos años, las plataformas de *streaming* han acelerado la creación de algoritmos muy eficaces para la compresión y reproducción de video y audio por internet. Todo indica que cuando se pongan ese tipo de recursos para comprimir y reproducir contenido XR, tendremos el primer ingrediente de la receta que algunos vemos como la solución a corto plazo para reducir esa distancia entre software y hardware. Realizar la mayor parte del cómputo en la nube, y hacer de los visores algo más cercano a un reproductor de contenido cuyo procesador se vea liberado de las tareas más costosas, que se externalizaran para luego volver de manera comprimida. Esquematizando, mi visor manda mis señales gestuales y de posición a la nube, de vuelta le llega todo el contenido creado a partir de esa información de forma comprimida de manera que el visor solo necesita descomprimir el contenido y reproducirlo, tal como hace reproduciendo cualquier video actualmente. El segundo ingrediente necesario para llevar a buen puerto una iniciativa de este tipo sería una tecnología que permita reducir la latencia

entre el visor y la nube de forma que nuestros movimientos sean capturados, enviados, interpretados, comprimidos, devueltos, descomprimidos y reproducidos a una velocidad mucho mayor a la que, a día de hoy, se comunican los dispositivos.

En este campo está jugando un papel muy importante la probabilística aplicada a través de IA, que ayuda a predecir los movimientos y gestos que el usuario va a realizar. El sistema crea una base de datos específica para cada usuario, y este se actualiza continuamente con nuevas combinaciones de movimientos y gestos. El modelo de IA se entrena constantemente con esa base de datos actualizada, y así puede predecir cada vez con mayor eficacia cuál será el próximo movimiento que va a realizar el usuario. De esta manera el servidor puede reducir los posibles escenarios que tiene que interpretar (renderizar) y así tenerlos listos antes de que el usuario los intente acceder.

Cuando ocurra esa sincronización entre software y hardware, habrá que pasar dos últimos obstáculos. El primero, que surjan aplicaciones que hagan del metaverso una nueva necesidad, como ocurrió con *Whatsapp* o *Google Maps* y los *smartphones*. El segundo, que no haya sido desarrollada una nueva tecnología que haga obsoleto el concepto actual de visor, y con ello gran parte del desarrollo de *software* que se hizo a medida para estos. Uno de los campos experimentales que más atención está atrayendo es el HMI (Human-Machine-Interface), el siguiente paso evolutivo del UX, donde los movimientos y gestos serán sustituidos por impulsos cerebrales, digamos por el pensamiento de esos movimientos y acciones. Recientemente investigadores de la Universidad de Osaka publicó un estudio donde un individuo era invitado a mirar unas imágenes mientras se le realizaba una resonancia magnética. Los resultados de esa resonancia eran interpretados por un modelo IA para generar una imagen a través de un modelo *stable diffusion*. Incluso teniendo en cuenta lo experimental de la técnica, las imágenes generadas guardaban sin duda un perceptible parecido con las originales.

Referencias

Takagi, Yu y Nishimoto, Shinji (2022). High-resolution image reconstruction with latent diffusion models from human brain activity. En: *bioRxiv*, 2022-11. Nueva York: Cold Spring Harbor Laboratory