(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 951 587**

(51) Int. CI.:

H04N 5/272 (2006.01)
G06T 7/73 (2007.01)
H04N 13/204 (2008.01)
H04N 13/156 (2008.01)
H04N 13/275 (2008.01)
H04N 5/265 (2006.01)
H04N 5/222 (2006.01)
G06T 19/00 (2011.01)
G06T 15/20 (2011.01)
H04N 13/239 (2008.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **09.05.2013 PCT/GB2013/051205**

(87) Fecha y número de publicación internacional: **14.11.2013 WO13167901**

(96) Fecha de presentación y número de la solicitud europea: **09.05.2013 E 13726815 (7)**

(97) Fecha y número de publicación de la concesión europea: **03.05.2023 EP 2847991**

(54) Título: **Un sistema para mezclar o componer en tiempo real objetos 3D generados por ordenador y una señal de video de una cámara cinematográfica**

(30) Prioridad:

09.05.2012 GB 201208088

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
23.10.2023

(73) Titular/es:

NCAM TECHNOLOGIES LIMITED (100.0%)
8-9 Carlisle Street
London W1D 3BP, GB

(72) Inventor/es:

BOIVIN, SAMUEL y
MICHOUD, BRICE

(74) Agente/Representante:

ELZABURU, S.L.P

DESCRIPCIÓN

Un sistema para mezclar o componer en tiempo real objetos 3D generados por ordenador y una señal de video de una cámara cinematográfica

Antecedentes de la invención

1. Campo de la invención

Un sistema para mezclar o componer en tiempo real objetos 3D generados por ordenador y una señal de video de una cámara cinematográfica, tal como una cámara de video, para generar un video de realidad aumentada en tiempo real para difusión de TV, cine o videojuegos.

2. Descripción del estado de la técnica

Durante los últimos 20 años, ha habido una considerable actividad comercial y de investigación en este campo; se puede hacer referencia a sistemas de inserción de video o de realidad aumentada de empresas, como Sportvision, Inc, que ha desarrollado mejoras en la visualización de televisión para eventos deportivos, como el fútbol americano, añadiendo una primera línea virtual que los espectadores ven superpuesta sobre el campo. Otras empresas que han desarrollado sistemas en esta área incluyen PVI, Inc. Una característica común de los sistemas conocidos es que se basan principalmente en el análisis de características de la señal de video de la cámara, para determinar a qué parte de la escena del mundo real apunta la cámara; el sistema habrá creado previamente un mapa 3D de esa escena de manera que, una vez que sepa a qué parte de la escena apunta, puede añadir o componer objetos generados por ordenador (como la primera línea descendente virtual) en la señal de video, de manera que la posición y la orientación del objeto lo hacen parecer una parte natural de la escena. Una desventaja de responder únicamente al flujo óptico de esta manera es que tales sistemas pueden no ser fiables.

Otros sistemas se basan en enfoques basados en marcadores puros (Lightcraft Technologies, por ejemplo). Estos requieren que un operador coloque marcadores físicos reales (1 m x 1 m de largo) en el plató para que el sistema los detecte. Esto es muy ineficiente dado que requiere horas o días para montar un escenario, algo que es muy poco probable que funcione para la producción de películas. También tiene muchas limitaciones, dado que los marcadores físicos siempre deben permanecer en el campo de visión de su sistema. También se pueden hacer referencias científicas a los artículos citados en el **Apéndice 1.**

El documento US2007248283A1 describe un método y un aparato para un sistema de vista previa de escena virtual de área amplia.

Breve compendio de la invención

De acuerdo con un primer aspecto de la invención, se da a conocer un sistema sin marcadores, incluyendo el sistema:

(i) una cámara de video;

(ii) sensores que incluyen un acelerómetro y un giroscopio que detectan más de seis grados de libertad;

(iii) dos cámaras testigo que forma un sistema estereoscópico; y

(iv) un procesador;

para mezclar o componer en tiempo real objetos 3D generados por ordenador y una señal de video de la cámara de video, para generar video de realidad aumentada para difusión de TV, cine o videojuegos, en el que:

(a) el cuerpo de la cámara de video se puede mover en 3D y los sensores en la cámara de video, o conectados directa o indirectamente a la misma, proporcionan datos de posicionamiento en tiempo real que definen la posición 3D y la orientación 3D de la cámara de video, o permiten calcular la posición 3D y la orientación 3D de la cámara de video;

(b) las dos cámaras testigo que forman el sistema estereoscópico están fijadas directa o indirectamente a la cámara de video;

(c) el sistema está configurado para usar esos datos de posicionamiento en tiempo real automáticamente para crear, recuperar, representar o modificar objetos 3D generados por ordenador;

(d) el sistema está configurado para mezclar o combinar los objetos 3D generados por ordenador resultantes con la señal de video de la cámara de video para proporcionar video de realidad aumentada para difusión de TV, cine o videojuegos;

y en el que:

(e) el sistema está configurado para determinar la posición 3D y la orientación de la cámara de video haciendo referencia a un mapa 3D del mundo real, donde el sistema está configurado para generar el mapa 3D del mundo real, al menos en parte, utilizando los datos de posicionamiento 3D en tiempo real de los sensores más un flujo óptico en el que las dos cámaras testigo que forman el sistema estereoscópico inspeccionan una escena, y en el que un software que se ejecuta en el procesador está configurado para detectar marcadores naturales en la escena que no se han añadido manual o artificialmente a esa escena;

(f) el sistema está configurado para usar un modelo de velocidad constante asociado con los datos de posicionamiento 3D en tiempo real de los sensores para predecir la siguiente posición de la cámara de video usando una posición previamente calculada o confirmada correctamente, y el sistema está configurado para usar esa predicción para proyectar una nube de puntos 3D en un cuadro de la cámara testigo actual, y para usar un algoritmo de puesta en correspondencia de puntos para poner en correspondencia puntos identificados en una señal de video en tiempo real del sistema estereoscópico y puntos proyectados en la nube de puntos 3D proyectada.

De acuerdo con un segundo aspecto de la invención, se da a conocer un método sin marcadores para mezclar o componer en tiempo real objetos 3D generados por ordenador y una señal de video de una cámara de video, para generar video de realidad aumentada para difusión de TV, cine o videojuegos, en el que:

(a) el cuerpo de la cámara de video se puede mover en 3D y los sensores, que incluyen un acelerómetro y un giroscopio que detectan más de seis grados de libertad, en la cámara de video o conectados directa o indirectamente a la misma, proporcionan datos de posicionamiento en tiempo real que definen la posición 3D y la orientación 3D de la cámara de video, o permiten calcular la posición 3D y la orientación 3D de la cámara de video;

(b) dos cámaras testigo que forman un sistema estereoscópico están fijadas directa o indirectamente a la cámara de video;

(c) dichos datos de posicionamiento en tiempo real se utilizan automáticamente para crear, recuperar, representar o modificar objetos 3D generados por ordenador;

(d) los objetos 3D generados por ordenador resultantes se mezclan o combinan, a continuación, con la señal de video de la cámara de video para proporcionar video de realidad aumentada para difusión de TV, cine o videojuegos;

y en el que:

(e) la posición 3D y la orientación de la cámara de video se determinan haciendo referencia a un mapa 3D del mundo real, donde el mapa 3D del mundo real se genera, al menos en parte, utilizando los datos de posicionamiento 3D en tiempo real de los sensores más un flujo óptico en el que las dos cámaras testigo que forman el sistema estereoscópico inspeccionan una escena, y se usa un software que se ejecuta en un procesador, para detectar marcadores naturales en la escena que no se han añadido manual o artificialmente a esa escena;

(f) se usa un modelo de velocidad constante asociado con los datos de posicionamiento 3D en tiempo real de los sensores para predecir la siguiente posición de la cámara de video usando una posición previamente calculada o confirmada correctamente, y esa predicción se usa para proyectar una nube de puntos 3D en un cuadro de la cámara testigo actual, y se usa un algoritmo de puesta en correspondencia de puntos para poner en correspondencia puntos identificados en una señal de video en tiempo real del sistema estereoscópico y puntos proyectados en la nube de puntos 3D proyectada.

Las características opcionales, de las que algunas o todas se pueden combinar entre sí, incluyen lo siguiente:

- los objetos 3D generados por ordenador se mezclan o componen en tiempo real con la señal de video en tiempo real de la cámara de video.

- Los ajustes de zoom, enfoque e iris en tiempo real de la cámara de video se miden y utilizan, junto con los datos de posicionamiento en tiempo real, para que los objetos 3D se representen correctamente en una ubicación y orientación deseadas en una escena 3D.

- Los sensores incluyen un acelerómetro de 3 ejes que mide la aceleración traslacional en 3D, un giroscopio de 3 ejes que mide la velocidad angular en 3D y un magnetómetro que mide el rumbo absoluto en 3D y, por lo tanto, constituyen un sensor 9DOF.

- Los sensores incluyen un sensor de distancia 3D, como una luz estructurada o una cámara de tiempo de vuelo.

- El sensor de distancia 3D captura la profundidad de cada píxel en una salida de video de la cámara.

- La profundidad de los bordes se refina al reproyectar las profundidades del sensor de distancia 3D en la señal de video de alta resolución de la cámara de video.

- Los sensores están conformados en una unidad que se puede fijar de forma segura a la cámara de video.

  ° La unidad incluye una o dos cámaras testigo.

  ° La unidad incluye un sensor de distancia 3D que captura la profundidad de cada píxel en una salida de video.

  ° La unidad incluye un sensor 6DOF o 9DOF.

  ° La unidad utiliza perchas intercambiables para permitir que se fije a cámaras de video y varillas de diferentes tamaños y diseños.

  ° La unidad es extensible para incluir otras formas de sensores.

  ° La unidad puede formar un dispositivo de inspección que se puede usar para inspeccionar una escena compleja y transmitir datos que definen la escena 3D que se está inspeccionando, de forma inalámbrica a un ordenador que, a continuación, sigue o recupera el seguimiento de la escena.

  ° La cámara de video incluye codificadores convencionales para leer el zoom, el enfoque y el iris de la cámara.

- El sistema incluye dos cámaras testigo (caja estereoscópica), equipadas con lentes que son lentes de ojo de pez de 180 grados.

- La cámara o cámaras testigo están desplazadas respecto de la cámara de video y esa compensación se obtiene utilizando un gráfico de calibración que incluye un primer y un segundo grupos de círculos, estando cada círculo en una ubicación conocida con respecto a uno o más de los otros círculos, y reconociéndose cada círculo utilizando un algoritmo de imágenes de regiones.

- Una lente de la cámara de video se calibra para la distorsión óptica utilizando un gráfico de calibración que incluye varios círculos, estando cada círculo en una ubicación conocida con respecto a uno o más de los otros círculos y reconociéndose cada círculo utilizando un algoritmo de imágenes de regiones.

- La cámara de video es cualquiera de las siguientes: cámara de grúa; steadicam; cámara portátil; cámara montada en plataforma rodante, cámara montada en trípode, teléfono inteligente, gafas de realidad aumentada.

- El sistema utiliza una o dos cámaras testigo de alta velocidad (tal como de, por lo menos, 100 fps) para permitir que el sistema se inicialice completamente sin una etapa independiente de inspección pura de la escena a seguir (denominada "inspección instantánea"), sino que la inspección se lleva a cabo continuamente mientras la cámara se usa para capturar video.

- El sistema estereoscópico permite que el software procese las imágenes e, incluso sin mover en absoluto el sistema de cámaras, genere una nube de puntos 3D instantánea (por ejemplo, asociar una gran cantidad de puntos en la escena a su posición en un espacio 3D utilizando el conocimiento de la separación entre las dos cámaras y la geometría epipolar).

- La profundidad de cada píxel en la nube de puntos 3D se obtiene utilizando correspondientes parches de textura 2D obtenidos de cada cámara testigo estereoscópica y un algoritmo de búsqueda de línea epipolar.

- El sistema ejecuta un algoritmo de fusión que combina datos de flujo óptico del sistema de cámaras testigo con los datos de posicionamiento en tiempo real de los sensores de hardware.

- El algoritmo de fusión se basa en una técnica de predicción/corrección de filtro de Kalman extendido, para integrar las salidas de todos los sensores y recalibrarlos, que pueden incluir un acelerómetro, un giroscopio, un magnetómetro, un sensor de distancia 3D, para determinar la posición y orientación de la cámara.

- El algoritmo de fusión EKF utiliza datos de nivel de confianza, asociados con la salida de cada sensor, al determinar cómo fusionar los datos de cada sensor.

- Los cuadros clave generados por el sistema de cámaras testigo son parte del proceso de seguimiento visual y son imágenes en tiempo real calculadas a 4 diferentes niveles de resolución de la señal de video de la cámara testigo.

- El sistema incluye (a) un ordenador generador de contenido que proporciona animación 3D generada por ordenador de figuras, objetos y lugares virtuales, y (b) un ordenador de representación (que puede o no ser independiente del ordenador generador de contenido), y en el que los datos de posicionamiento en tiempo real que definen la posición 3D de la cámara de video son utilizados por uno del ordenador generador de contenido y el ordenador de representación, o por ambos, para hacer que se generen objetos 3D generados por ordenador en

tiempo real que pueden insertarse y mezclarse en tiempo real con la señal de video de la cámara de video para formar una parte natural de la escena mostrada en esa señal de video.

- Los objetos 3D generados por ordenador son animaciones que pueden moverse a cualquier lugar dentro de la escena y pueden alterar su forma y apariencia de una manera determinada por el ordenador generador de contenido.

- Los objetos 3D generados por ordenador son figuras animadas de personas o criaturas que se mueven (por ejemplo, corren, bailan, caminan, luchan, vuelan, saltan, ...) de manera realista cuando se mezclan en la escena.

- Los datos de posicionamiento o seguimiento de la cámara también están disponibles para su uso en la posproducción para facilitar el CGI de la posproducción.

- El sensor de distancia 3D se usa para mejorar la precisión de una medición de profundidad asociada con un punto 3D reconstruido, o para rechazar ese punto 3D reconstruido.

- El sensor de distancia 3D se utiliza para modulación de profundidad en tiempo real con el fin de permitir oclusión dinámica y suprimir el uso eventual de un escenario verde.

- El sistema utiliza un pequeño objeto de registro de cámara, como un tablero de tamaño conocido y cubierto con un patrón conocido, colocado en la escena para que una esquina del patrón detectado se trate como el origen de la nube de puntos 3D (y, por lo tanto, del mundo).

- El objeto de registro de la cámara comprende al menos dos esferas de tamaño conocido dispuestas en una verdadera vertical y reconocidas mediante un algoritmo de reconocimiento de imágenes de regiones.

- El sistema incluye un magnetómetro para indicar el norte magnético, un acelerómetro para indicar la dirección de la gravedad (y, por lo tanto, dar la verdadera vertical), un giroscopio para indicar si el sistema está inclinado hacia arriba o hacia abajo, si se ha desplazado hacia la izquierda o hacia la derecha o si se ha girado sobre el eje óptico, y un acelerómetro de 3 ejes para permitir inferir la traslación en 3D desde una posición de inicio.

- El software intenta generar una nube de puntos uniformemente distribuida en el mapa 3D para reducir en gran medida las pérdidas de seguimiento y aumentar la precisión del seguimiento (se genera más paralaje, por lo que la posición estimada de la cámara es más precisa).

- El sistema de seguimiento de la cámara se puede conectar de forma inalámbrica al sistema de cámaras y, por lo tanto, se puede mover rápidamente por el plató al generar la nube de puntos 3D, sin necesidad de arrastrar cables por el plató, a diferencia de los anteriores sistemas de cámaras testigo.

- El sistema de seguimiento de la cámara combina la inspección instantánea (caso estereoscópico) con el seguimiento de la cámara de video a medida que el director/camarógrafo sigue, desplaza, inclina el sistema de seguimiento de la cámara conectado a la cámara de video.

- El sistema automatiza completamente todos los aspectos del seguimiento de la cámara de video, incluida la rotación, la traslación, el enfoque, el iris, la distancia focal; y automatiza el escalado, el posicionamiento y la orientación del contenido 3D generado por ordenador para mezclarlo con el video.

- El sistema permite la inspección continua en tiempo real de una escena para generar una nube de puntos más completa que define la escena.

- El sistema acopla descriptores invariantes bajo rotación, por ejemplo, mediante ORB, a los puntos característicos detectados en la escena para facilitar la recuperación del seguimiento.

- El sistema utiliza un modelo de velocidad constante asociado con la información proporcionada por los sensores para predecir la siguiente posición de la cámara de video utilizando la posición previamente calculada o confirmada correctamente. Utiliza esa predicción para reproyectar la nube de puntos 3D en el cuadro actual, para permitir que un algoritmo de puesta en correspondencia de puntos ponga en correspondencia puntos identificados en la señal de video en tiempo real del sistema de cámaras testigo y los puntos proyectados en la nueva nube de puntos 3D.

- El sistema utiliza un esquema de minimización de Levenberg-Marquardt para el seguimiento de la cámara a fin de minimizar el error entre los puntos identificados en la señal de video en tiempo real del sistema de cámaras testigo y los puntos proyectados en la nueva nube de puntos 3D.

- El usuario puede utilizar la nube de puntos 3D generada por el sistema de seguimiento de la cámara para definir máscaras 3D, tales como máscaras de mates de datos sobrantes 3D.

- Los objetos 3D incluyen objetos estáticos, animaciones dinámicas, mundos virtuales, personas virtuales, edificios virtuales, escenarios virtuales, platós virtuales y cualesquiera datos en una base de datos de animación.

- La cámara de video y la cámara testigo se calibran para el retardo de adquisición de cuadros utilizando una fuente de luz modulada, por ejemplo, comparando curvas de intensidad de luz asociadas, con un LED parpadeante.

Otros conceptos - cada uno puede combinarse con cualquiera de las características definidas anteriormente, o con cualquier otro concepto definido a continuación:

Un método para mezclar o componer en tiempo real objetos 3D generados por ordenador y una señal de video de una cámara cinematográfica, en el que el cuerpo de la cámara cinematográfica se puede mover en 3D y los sensores en la cámara o conectados a la misma proporcionan datos de posicionamiento en tiempo real que definen la posición 3D y la orientación 3D de la cámara, o permiten calcular la posición 3D.

Un método para mezclar o componer objetos 3D generados por ordenador en tiempo real y una señal de video de una cámara cinematográfica, como una cámara de video, para generar video de realidad aumentada para difusión de TV, cine o videojuegos, en el que:

(a) el cuerpo de la cámara cinematográfica se puede mover en 3D y los sensores en la cámara cinematográfica o conectados directa o indirectamente a la misma proporcionan datos de posicionamiento en tiempo real que definen la posición 3D y la orientación 3D de la cámara cinematográfica, o permiten calcular la posición 3D y la orientación 3D de la cámara cinematográfica y

(b) dichos datos de posicionamiento en tiempo real son, a continuación, automáticamente utilizados por el sistema para crear, recuperar, representar o modificar objetos 3D generados por ordenador y

(c) los objetos 3D generados por ordenador resultantes se mezclan o combinan a continuación con la señal de video de la cámara cinematográfica para proporcionar video de realidad aumentada para difusión de TV, cine o videojuegos.

Los métodos o sistemas definidos anteriormente y utilizados para permitir que un director (o un director de fotografía) monte los activos 3D generados por ordenador de preproducción en la cámara, generalmente activos de previsualización o efectos visuales, en tiempo real en la placa de la película de acción en vivo o en imágenes de video capturadas por la cámara, lo que permite al director explorar posibles ángulos y movimientos de la cámara en tiempo real, mezclándose automáticamente los activos 3D generados por ordenador en el video según lo ve el director.

Métodos o sistemas como los definidos anteriormente en los que los datos de posicionamiento en tiempo real se registran y se estampan con un código de tiempo para proporcionar datos de movimientos en correspondencia para procesos de postproducción.

Métodos o sistemas como se definen anteriormente y utilizados para permitir la inserción de imágenes u objetos virtuales en un flujo de video de difusión.

Un método o sistema como se define anteriormente, para permitir uno o más de los siguientes:

- Seguimiento en tiempo real para cámaras de estudio
- Seguimiento en tiempo real para steadicam
- Seguimiento en tiempo real para cámaras montadas en grúas
- Seguimiento en tiempo real para cámaras montadas en plataforma rodante
- Seguimiento en tiempo real para steadicam
- Seguimiento en tiempo real para difusión en exteriores (OB, Oustside Broadcast)
- Uso de datos en tiempo real (por ejemplo, datos de seguimiento) para posproducción 2D
- Uso de datos en tiempo real (por ejemplo datos de seguimiento) para posconversión de contenido estereoscópico 3D
- Uso de datos en tiempo real (por ejemplo, datos de seguimiento) para contenido estereoscópico 3D nativo
- inserción de gráficos 3D
- Inserción de gráficos 3D para la colocación de productos en estudio o en plató
- Inserción de gráficos 3D para OB
- Inserción de gráficos 3D para otras imágenes patrocinadas

- Inserción de gráficos 3D que es específica de la ubicación del espectador
- Inserción de gráficos 3D que es específica del espectador
- Inserción de gráficos 3D que es específica del tiempo
- Inserción de gráficos 3D para rellenar escenas de multitudes
- Inserción de gráficos 3D para reemplazo de pantalla verde
- Inserción de gráficos 3D de contenido educativo para ayudar al aprendizaje, en museos y centros de interpretación en sitios culturales, históricos o naturales.
- Medición del tamaño absoluto o relativo de los objetos en la escena.

Métodos o sistemas tal como se definen anteriormente, donde la cámara cinematográfica es una de las siguientes:

- Todas las cámaras con un anclaje estándar
- Cámaras que requieren conexión de fibra óptica táctica
- Cámara que requiere conexión RF/inalámbrica

Método o sistemas tal como se definen anteriormente, desplegados en uno de los siguientes mercados:

- Cine/TV (no en vivo)
- Anuncios (no en vivo)
- Anuncios en vivo
- Difusión (no deportes)
- OB de difusión
- Basado en estudio deportivo
- Basado en OB deportiva
- Publicidad indirecta de TV en vivo
- Uso de internet (no en vivo)
- internet en vivo
- internet en vivo basado en territorio
- Publicidad indirecta de internet en vivo
- Contenido museístico/patrimonial
- Anuncios de museos/patrimonio
- Arquitectónico
- Juegos

Métodos o sistemas como se definen anteriormente y que se utilizan para permitir que cualquier dispositivo de visualización muestre imágenes de realidad aumentada, incluyendo un teléfono inteligente y gafas de realidad aumentada, alterándose automáticamente el aspecto de las imágenes de realidad aumentada en función de los datos de posicionamiento en tiempo real.

Un film, película, programa de televisión o videojuego en el que se mezclan objetos 3D generados por ordenador en tiempo real con una señal de video de una cámara, en el que el cuerpo de la cámara se puede mover en 3D, y sensores de, o conectados a la cámara proporcionan datos de posicionamiento en tiempo real que definen la posición 3D de la cámara o permiten calcular la posición 3D.

Un dispositivo de inspección y seguimiento de escenas adaptado para acoplarse a una cámara convencional, en el que el cuerpo de la cámara se puede mover en 3D y los sensores de hardware en el dispositivo de inspección y seguimiento de escenas proporcionan datos de posicionamiento en tiempo real que definen la posición 3D de la cámara, o permiten calcular la posición 3D.

Una cámara de mano o portátil que incluye sensores en la cámara o conectados a la misma proporciona datos de posicionamiento en tiempo real que definen la posición 3D y la orientación 3D de la cámara en relación con un marco de referencia 3D, o permiten, completamente o como parte de un sistema que analiza otros datos, tales como datos de flujo óptico, que se calcula la posición 3D.

Una cámara cinematográfica que incluye o está conectada a un sistema de cámaras testigo estereoscópico, generando el sistema de cámaras testigo imágenes estereoscópicas de gran angular (por ejemplo, 180 grados), lo que permite que el software procese las imágenes y, sin que el sistema de cámaras sea seguido/movido en absoluto, generar una nube de puntos 3D instantánea.

Un sistema de seguimiento de cámara para acoplar a una cámara cinematográfica, con el sistema de cámaras testigo generando imágenes estereoscópicas, que permite que el software procese las imágenes y, sin que el sistema de cámaras se mueva en absoluto, genere una nube de puntos 3D instantánea y proporcione un seguimiento en tiempo real (posición, orientación, zoom, enfoque e iris) de la cámara cinematográfica.

Gafas de realidad aumentada que incluyen sensores en, o acoplados a las gafas, proporciona datos de posicionamiento en tiempo real que definen la posición 3D y la orientación 3D de las gafas en relación con un marco de referencia 3D, o permiten, completamente o como parte de un sistema que analiza otros datos, tales como datos de flujo óptico, la posición 3D a calcular.

El presente sistema se implementa en un sistema denominado Ncam. Varias aplicaciones de Ncam incluyen las siguientes:

Efectos visuales en cine, TV y anuncios

Efectos visuales en preproducción y producción: Ncam busca ayudar a llenar el vacío entre la previsualización (preproducción) y VFX (postproducción) durante el proceso de filmación (producción). Ncam ofrece una solución para aquellos que deseen montar en el VFX en la cámara, mientras disparan, en tiempo real. Usando contenido creado previamente, a menudo del proceso de previsualización, Ncam puede componer esos activos de previsualización, normalmente modelos 3D y animación, en la placa de la película de acción en vivo en tiempo real.

Llevar al plató la previsualización cuidadosamente elaborada y aprobada puede ahorrar una gran cantidad de tiempo. Los cineastas pueden montar la previsualización o los efectos visuales, ya sea filmando en pantalla verde y componiendo un fondo virtual o superponiendo una criatura u objeto en primer plano. Entonces, los cineastas pueden recuperar el control de la previsualización y los efectos visuales, explorando posibles ángulos de cámara y movimientos sobre la marcha y en tiempo real, sin las restricciones habituales.

A su vez, los datos se graban y se marcan con un código de tiempo, lo que proporciona al departamento de VFX datos de movimiento de correspondencia de cámara (la pista de 6 DOF) y, al mismo tiempo, proporciona al editor un 'provisional' del VFX final.

Al filmar el VFX 'provisional' en la cámara y diseñar el VFX previamente en el proceso, es posible eliminar muchas conjeturas del VFX como proceso de posproducción. Se puede disminuir el proceso de diseño de VFX, ahorrando así el derroche.

Publicidad virtual - difusión en vivo

La tecnología Ncam se presta bien a la publicidad virtual en tiempo real con el espacio de difusión. La publicidad por emplazamiento digital, donde el producto se inserta después del rodaje durante la posproducción, a diferencia de un producto físico real que está en cámara, se está volviendo más popular. Sin embargo, con la capacidad de aumentar la publicidad indirecta digital en vivo, se pueden abrir varias posibilidades. Los ejemplos pueden incluir logotipos de marca en campos deportivos o bebidas enlatadas en una telenovela. Aquí es donde la capacidad de representar imágenes CG fotorrealistas, en tiempo real, en función de la iluminación del entorno de rodaje, se vuelve fundamental.

Juegos holográficos

La combinación de la tecnología Ncam con pantallas holográficas y gafas (como Google Glass) podría proporcionar una experiencia de juego totalmente inmersiva. Mezclando el mundo real con mundos y personajes virtuales. Las posibilidades son prácticamente infinitas.

Museo y patrimonio

La tecnología Ncam ofrece una amplia gama de aplicaciones potenciales a los sectores de museos y patrimonio.

Como herramienta de exploración, podría ser útil para la reconstrucción de sitios patrimoniales como ruinas, mostrando cómo se veía el sitio en su antigua gloria.

Dentro de los museos, Ncam puede usarse como una herramienta educativa, tal vez para demostrar el tamaño y el movimiento de un dinosaurio, sus órganos internos o su estructura esquelética. Otro ejemplo puede ser explorar el funcionamiento del motor de combustión integral, viendo efectivamente un diagrama animado con las piezas desmontadas, pero en un espacio 3D totalmente inmersivo.

Cámara virtual - VCS

La tecnología Ncam se presta perfectamente a esta aplicación. Una cámara virtual es esencialmente una pantalla LCD que muestra un entorno 3D CG completo. A medida que el operador mueve la pantalla LCD, esta se comporta como una cámara física en términos de movimiento. La cámara LCD es rastreada en sus traslaciones y rotaciones XYZ y muestra el entorno CG completo desde un motor CG, en tiempo real. Actualmente hay varias soluciones VCS (sistema de cámaras virtual) disponibles, pero todas tienen limitaciones en términos de tecnología. Estas tienden a requerir mucho tiempo de instalación, están limitadas en el espacio y el entorno en el que operarán y son costosas. Un VCS basado en Ncam probablemente adoptaría la forma de una tableta, por lo que el procesamiento se calcularía localmente combinado con una pantalla LCD. Los cálculos de CG podrían tomarse de un PC remoto o posiblemente realizarse localmente, según los requisitos. La idea es permitir el acceso abierto a través de la usabilidad y los puntos de precio a muchos departamentos dentro de una película, que anteriormente no han podido tener acceso a tales herramientas. Esto es muy útil para diseñadores, arquitectos, VFX, empresas de juegos, CG y casas de animación, etc.

Este VCS también formaría la columna vertebral de la posible solución de museo y patrimonio.

Sistema de exploración virtual - VSS

Si imagina que está rodando una película y es predominantemente en pantalla azul/pantalla verde. ¿Cómo, como cineasta o director, decido qué ángulos o tomas serán las mejores? Después de todo, es posible que tenga una pequeña cantidad de escenario físico y algunos actores, pero tengo poca idea, aparte de mi imaginación, de qué ángulos de cámara son los mejores, y mucho menos explicarle a todo el equipo dónde debemos colocar las cámaras y hacer una elección válida de lentes. Actualmente, se mueve una cámara cinematográfica y toda su parafernalia a una ubicación en el plató, se configura Ncam y se echa un vistazo, solo para darnos cuenta de que sería mejor en otra posición. Si tan solo tuviéramos un sistema portátil liviano para tomar esas decisiones, de manera rápida y eficiente. Esto es Ncam VSS.

VSS es Ncam tal como es hoy, pero integrado en una cámara de factor de forma pequeño: piense en DSLR. Es esencialmente un visor digital con los beneficios de Ncam. Los inventores están combinando todos los elementos de Ncam tal como es hoy, en un plató, pero en un dispositivo portátil. La producción virtual para episodios de televisión también se beneficiaría enormemente de esta herramienta, al igual que los museos/sitios patrimoniales, así como los arquitectos que deseen mostrar a los inversores potenciales sus nuevas creaciones.

Breve descripción de las figuras

Figura 1

Descripción del hardware para configuración 2.3. Cabe señalar que la versión dibujada aquí es la que tiene la estación de trabajo de seguimiento y la estación de trabajo de representación fusionadas.

Figura 2

Algoritmo de seguimiento global para obtener la posición y rotación de una cámara cinematográfica en tiempo real.

Figura 3

El objeto de registro de nivelación automática 3D creado para el sistema de los inventores. Ese objeto se usa en la escena real para calcular automáticamente el origen, la orientación y el tamaño del mundo real en el software de los inventores. En la imagen de la derecha de la figura 5 se muestra una versión plana.

Figura 4

La pirámide de imágenes utilizada para la detección de marcadores naturales. Esta figura también muestra algunos parches de textura 2D asociados a los marcadores.

Figura 5

Gráficos de calibración desarrollados y utilizados en Ncam. Izquierda: tabla de calibración utilizada para calibrar todas las lentes. Centro: tabla de calibración utilizada o un cálculo de alineación entre una cámara cinematográfica y una cámara testigo. El gráfico de la izquierda también se puede utilizar para los casos en que la distancia focal de la cámara cinematográfica es lo suficientemente corta (decidida experimentalmente en el plató). Derecha: versión plana del objeto de registro. Más fácil de manejar y se puede colocar en cualquier lugar de la superficie del suelo.

Los círculos proporcionan una base ortonormal dando orientación, escala y origen (0,0,0) del mundo cuando se añaden objetos virtuales a la escena real. El centro del círculo superior izquierdo es el origen del mundo.

Figura 6

La barra de la cámara se monta en una cámara cinematográfica con los diferentes tamaños de percha (para varillas de 15 mm y 19 mm).

Figura 7

El dispositivo de inspección inalámbrica. Cuando se requiere una inspección (es decir, un caso monoscópico, por ejemplo (ver la sección 2.1, o platós muy complejos (ver el final de la sección 2.1.3)), todo el proceso de seguimiento se consigue utilizando ese dispositivo de inspección. Cuando se completa la inspección, todos los datos de seguimiento (nube de puntos, puntos clave, etc.) se transmiten de forma inalámbrica a la estación de trabajo de seguimiento/representación que recupera instantáneamente el seguimiento y comienza a transmitir en continuo tanto los datos de seguimiento como el resultado compuesto final.

Figura 8

La caja de conexiones de la cámara/ordenador. Cabe señalar que aquí se muestra la versión que utiliza varios cables RJ45. Todas estas cajas de conexiones también tienen una variante en la que esos cables se reemplazan por un solo cable táctico de fibra óptica.

Descripción detallada

Esta tecnología trata de mezclar objetos 3D generados por ordenador en tiempo real y una señal de video de la vida real procedentes de una cámara real (denominada cámara cinematográfica en el resto de este documento) en un plató, una difusión de TV o un videojuego. Los principales objetivos de este sistema son:

- Proporcionar una herramienta de encuadre en tiempo real de personajes y entornos virtuales para directores de cine.

- Mostrar en tiempo real el resultado fotorrealista de los objetos virtuales fusionados directamente con el video real.

- Proporcionar una estimación precisa de la posición de una cámara cinematográfica (denominado seguimiento de la cámara) tanto para rodaje en plató como para posproducción de películas. Se manejan todas las cámaras cinematográficas: cámaras de grúa, steadicam, cámara de mano, cámaras en trípodes, etc.

- Proporcionar una aproximación al modelado geométrico 3D de la escena real, manejando la oclusión entre los objetos reales (y/o actores) y el entorno virtual (objetos 3D, personajes, etc.).

La tecnología descrita aquí es, en primer lugar, un poderoso algoritmo de fusión de sensores, que fusiona marcadores naturales (es decir, no añadidos manualmente a la escena física) calculados a partir de 3 (caso estereoscópico) o 2 (caso monoscópico) diferentes flujos ópticos y la medición física de otros 4 sensores (giroscopio, acelerómetro, magnetómetro, sensor de distancia 3D). También es un sistema totalmente no intrusivo que no requiere añadir manualmente ningún marcador físico ni nada directamente visible en la escena real para hacer su trabajo. El caso de una sola cámara testigo monoscópica no es según la invención y se describe únicamente con fines ilustrativos.

1 Hardware

La tecnología tiene 2 posibles configuraciones de hardware diferentes según el tipo de datos que hay que transmitir en continuo.

Estas 2 configuraciones tienen en común el siguiente hardware montado en un dispositivo denominado barra de la cámara (ver la figura 6 para obtener esquemas detallados), así como una caja de conexiones separada que es esencialmente una caja de conexiones que fusiona o separa todos los diferentes cables en uno o varios haces independientes (ver la figura 8 para obtener esquemas detallados):

- Una o dos cámaras de alta velocidad (al menos 100 fps), denominadas cámaras testigo, según el enfoque seleccionado (monoscópico o estereoscópico).

- Cada cámara testigo tiene una lente de ojo de pez de 180 grados montada.

- Un sensor de 9 DOF, que incluye un giroscopio, un magnetómetro y un acelerómetro (o 6 DOF cuando no se puede usar el magnetómetro).

- Un sensor de distancia 3D que captura la profundidad de un píxel.

Además de la barra de la cámara, todas las configuraciones tienen codificadores de lente regulares montados en la cámara cinematográfica para leer los valores de zoom, enfoque e iris. También tienen en común un dispositivo de inspección inalámbrico (ver la figura 7 para obtener esquemas detallados) para inspeccionar la escena y aprenderla. Sin embargo, se debe observar que en el caso de un enfoque estereoscópico completo (ver la sección 2.5) y, más precisamente, una inspección instantánea (ver la sección 2.5.1), dicho dispositivo de inspección no es necesario.

Aparte de estas características compartidas, las 2 configuraciones se describen a continuación:

Configuración 1: transmisión en continuo con cámara

1. Ordenador inalámbrico integrado:

(a) Un ordenador de seguimiento integrado está montado en la cámara cinematográfica. Transmite en continuo de forma inalámbrica la traslación y rotación (RT) de la cámara cinematográfica a una estación de trabajo externa de representación, así como el zoom, el enfoque y el iris (ZFI) que proceden de los codificadores.

(b) Una estación de trabajo de generación de contenido externo. Esta máquina generalmente ejecuta un software de gráficos por ordenador para proporcionar animación 3D generada por ordenador y contenido CG relevante para la composición final. También transmite en continuo los datos animados a la estación de trabajo de representación externa (1-1c).

(c) Una estación de trabajo de representación externa que, por un lado, recibe los datos del ordenador integrado (1-1a) y, por otro lado, también maneja los datos animados procedentes de la estación de trabajo generadora de contenido (1-1b). Esta estación de trabajo de representación utiliza la cámara cinematográfica RT+ZFI procedente de 1-1a para mostrar el entorno virtual 3D procedente de 1-1b y mezcla el resultado con el flujo de video real procedente de la cámara cinematográfica. También transmite en continuo el resultado de la composición final, de vuelta al ocular de la cámara cinematográfica o a un monitor de control, sin ningún cable.

2. Ordenador cableado integrado

(a) Un ordenador de seguimiento integrado está montado en la cámara cinematográfica. Transmite en continuo la traslación, la rotación (RT) y los datos del codificador (ZFI) de la cámara cinematográfica a una estación de trabajo externa de representación, utilizando cables de fibra óptica tácticos.

(b) Similar a la configuración 1-1b.

(c) Similar a la configuración 1-1c, excepto que el resultado compuesto final se envía de vuelta a la cámara a través de un cable HD-SDI.

Configuración 2: transmisión total en continuo

1. Ordenador inalámbrico integrado:

(a) Un ordenador integrado de seguimiento está montado en la cámara cinematográfica. No se obtiene ningún procesamiento real en esa unidad. Esta solo transmite de forma inalámbrica todos los datos de sensor a una estación de trabajo externa de seguimiento. Se transmiten en continuo flujos de video 2x de las cámaras testigo, datos 1x9DOF, datos de sensor de distancia 1x. La cantidad total de datos requiere un ancho de banda mínimo superior a 142 MB/s (las cámaras testigo requieren 63 MB/s, el sensor de distancia 74 MB/s, el sensor 9DOF 4,4 MB/s).

(b) Una estación de trabajo externa de seguimiento que recibe los datos del ordenador de seguimiento integrado. Esta unidad procesa todos los datos de sensor para calcular las matrices de rotación y traslación de la cámara cinematográfica. Transmite en continuo la cámara RT+ZFI a la estación de trabajo externa de representación (2-1d).

(c) Una estación de trabajo externa de generación de contenido. Esta máquina generalmente ejecuta un software de gráficos por ordenador para proporcionar animación 3D generada por ordenador y contenido CG relevante para la composición final. También transmite en continuo los datos animados a la estación de trabajo externa de representación (2-1d).

(d) Una estación de trabajo externa de representación que, por un lado, recibe los datos del ordenador integrado (2-1b) y, por otro lado, también maneja

los datos animados procedentes de la estación de trabajo generadora de contenido (2-1c). Esta estación de trabajo de representación utiliza la cámara cinematográfica RT+ZFI procedente de 2-1b para mostrar el entorno virtual 3D procedente de 2-1c y mezcla el resultado con el flujo de video real procedente de la cámara cinematográfica. También transmite en continuo el resultado de la composición final al ocular de la cámara cinematográfica o a un monitor de control sin ningún cable.

2. Ordenador cableado integrado:

(a) Un ordenador de seguimiento integrado está montado en la cámara cinematográfica. Transmite en continuo la traslación y rotación (RT) de la cámara cinematográfica a una estación de trabajo de representación externa (2-2c) utilizando cables de fibra óptica tácticos o múltiples cables Cat-6 RJ45. De nuevo, se utilizan codificadores para zoom, enfoque e iris (ZFI).

(b) Similar a la configuración 2-1c.

(c) Similar a la configuración 2-1d, excepto que el resultado compuesto final se envía de vuelta a la cámara a través de un cable HD-SDI.

3. Sin ordenador integrado (ver la figura 1 para ver una ilustración gráfica: todas las demás versiones se pueden derivar fácilmente de esa):

(a) Un solo dispositivo de seguimiento de la cámara (barra de la cámara) está acoplado a la cámara cinematográfica y conectado a la caja de conexiones.

(b) Una estación de trabajo externa de seguimiento que recibe los datos del dispositivo de seguimiento integrado. Esta unidad procesa todos los datos de sensor para calcular las matrices de rotación y traslación de la cámara cinematográfica. De nuevo, se utilizan codificadores para zoom, enfoque e iris (ZFI). Transmite en continuo la cámara RT+ZFI a una estación de trabajo de representación 2-3d.

(c) Similar a la configuración 2-1c.

(d) Similar a la configuración 2-1d, excepto que el resultado compuesto final se envía de vuelta a la cámara a través de un cable HD-SDI.

Además de las configuraciones anteriores, cada versión también tiene otras dos variantes. Una tiene la estación de trabajo de generación de contenido fusionada con la estación de trabajo de representación externa, y la segunda variante tiene la estación de trabajo de representación externa fusionada con la estación de trabajo de seguimiento.

2 Software

Se han creado varios métodos y software científicos nuevos para conseguir resultados de seguimiento precisos y realidad aumentada en tiempo real [19]. Estos métodos se describen en las siguientes secciones.

2.1 Seguimiento monoscópico sin marcadores desde la señal de video

2.1.1 Visión general del proceso

La técnica monoscópica utilizada en la tecnología está construida sobre dos pases separados que en realidad son técnicamente casi iguales, excepto que uno requiere interacción del usuario mientras que el otro es completamente automático.

El primer pase se llama *inspección.* Un usuario utiliza un dispositivo integrado en el hardware descrito en el segundo párrafo de la sección 1 (ver también la figura 7) para escanear la escena real y el objeto de registro. Durante esa fase, el sistema aprende y calcula automáticamente los marcadores naturales en 3D a medida que el usuario escanea el entorno caminando. En cuanto el usuario ha comenzado a hacer la inspección, también tiene que capturar el **objeto de registro** (ver sección 2.1.3.2) para estimar automáticamente la escala, el origen y la orientación del mundo. Una vez se ha conseguido esto, el usuario puede continuar escaneando el resto de la escena para construir la mejor nube de puntos 3D posible de marcadores naturales (un marcador natural es un punto característico invisible al ojo humano y que los algoritmos de los inventores calculan). Cuanto más completa se consiga que sea esta inspección, mejor será el resultado final. La parte de inspección es la tarea más importante de todas y es muy crítica.

El segundo pase es el seguimiento de la cámara cinematográfica (ver la sección 2.1.4) que se realiza desde el dispositivo de seguimiento que se ha colocado en la cámara cinematográfica. Se aplican diferentes configuraciones, siguiendo la descripción del apartado anterior titulado "Hardware" y según diversas situaciones. Esta etapa es totalmente automática y, por lo general, no requiere ninguna interacción humana.

2.1.2 Visión general del algoritmo

La tecnología descrita aquí es un marco de realidad aumentada basado en una técnica de fusión de múltiples sensores (ver sección 2.3).

De hecho, no se basa solo en los datos de flujo óptico habituales para seguir una cámara e insertar un objeto virtual. Tales técnicas han demostrado sus limitaciones científicas y prácticas en muchos casos. Por ejemplo, si un usuario oculta la cámara utilizada para el seguimiento, el seguimiento falla y se pierde. Es exactamente el mismo problema

si la cámara apunta a una región que no ha sido inspeccionada (ver sección 2.1.3). Además, incluso si la cámara cinematográfica no se mueve en absoluto, todavía hay algunos problemas inevitables de leves oscilaciones debido a la precisión de los marcadores naturales detectados calculados por el algoritmo de seguimiento visual puro. Filtrar los datos resuelve parcialmente el problema, pero nunca se obtiene un objeto virtual perfectamente estable, especialmente cuando se usa una lente focal muy larga. Como la tecnología de los inventores utiliza otros sensores para saber si la cámara se está moviendo o no, el algoritmo de seguimiento de los inventores no tiene ese problema.

Técnicas regulares de flujo óptico como SLAM[9], PTAM[8], RSLAM[13], etc. utilizan información contenida en el video capturado por la cámara para aproximar su posición y orientación. La tecnología de los inventores tiene un enfoque similar para su parte de seguimiento visual (denominada *seguimiento visual*), excepto que, por un lado, el núcleo del seguimiento no es una simple búsqueda de cuadro clave cuando este se pierde y, por otro lado, utiliza todos los datos de los 7 sensores disponibles (una cámara cinematográfica, dos cámaras testigo, un giroscopio, un acelerómetro, un magnetómetro y un sensor de distancia 3D) para calcular una posición y orientación precisas de la cámara.

Los inventores utilizan un *filtro Kalman extendido* (EKF, ver la sección 2.3.2) [7, 21] para integrar los siguientes sensores: giroscopio, acelerómetro, sensores de magnetómetro y cámaras testigo. La técnica EKF es el corazón del software de los inventores y todo el algoritmo de predicción/corrección se basa en un método de fusión de datos que permite utilizar lo mejor de cada sensor. Esto proporciona una solidez inigualable en casos simples y críticos en los que fallan todas las demás técnicas. De hecho, cuando un sensor ha perdido el seguimiento (lo que en realidad significa que sus mediciones ya no son fiables), el EKF aún puede obtener una estimación de la posición/orientación fusionando los datos de los otros sensores fiables y restantes. Además de eso, como cada sensor tiene un nivel de confianza, esto incide en el algoritmo de fusión para evitar que el sistema use información inexacta. Por ejemplo, un giroscopio normal tiene un defecto natural denominado *deriva* que tiende a modificar su orientación incluso cuando no se mueve. Cuanto más tiempo pasa, más error genera el giroscopio. La corrección de deriva del giroscopio se realiza utilizando seguimiento visual, y el giroscopio corrige el seguimiento visual cuando su error se vuelve demasiado grande. Por lo tanto, el EKF puede encontrar la mejor posición y rotación de la cámara cinematográfica prediciendo los movimientos de la cámara cinematográfica usando los valores previamente estimados correctamente y, a continuación, corrigiendo su predicción de acuerdo con las nuevas mediciones proporcionadas por todos los sensores.

Además, la mayoría de las técnicas de seguimiento ópticas puras utilizan cuadros clave (es decir, instantáneas) para localizar la cámara cuando esta se pierde. Si no se tiene un cuadro clave de la vista actual a la que se apunta con la cámara, la recuperación falla. Por supuesto, se pueden utilizar técnicas más sólidas, como las técnicas SIFT[11] o SURF[1], para resolver casos en los que, por ejemplo, simplemente se ha rotado verticalmente la cámara. De hecho, como todos estos sistemas comparan el cuadro actual con el cuadro clave más similar, es bastante interesante utilizar descriptores invariantes bajo rotación y escala para obtener una correspondencia mejor y más rápida. Sin embargo, si no se tiene ningún cuadro clave de la posición donde está la cámara, no hay forma de que el sistema pueda recuperar su posición. La técnica de los inventores toma lo mejor de cada técnica (se debe tener en cuenta que se utilizan ORB[18] como descriptores de puntos clave) y se ejecutan simultáneamente tres subprocesos diferentes para recuperar el seguimiento perdido (ver la sección 2.1.4.4 para obtener más detalles). El proceso de recuperación se puede ver como un algoritmo global basado en un enfoque probabilístico y estocástico. Cuando se pierde el seguimiento, el espacio 3D se muestrea instantáneamente alrededor de las últimas posiciones conocidas utilizando un enfoque elipsoidal y todos los cuadros clave dentro de ese elipsoide se comparan con el cuadro clave actual. Además, todos los demás datos procedentes de los sensores restantes (giroscopio, acelerómetro, magnetómetro, sensor de distancia 3D) siguen fusionándose en la búsqueda, lo que permite que el software de los inventores desambigüe todos los posibles buenos candidatos. Por supuesto, si ningún cuadro clave es un candidato lo suficientemente bueno, el sistema utilizará todos los demás sensores además del de seguimiento visual para calcular una aproximación de la posición actual. La consecuencia obvia de esa técnica es que la tecnología de los inventores permite al usuario apuntar la cámara a un lugar que no ha sido inspeccionado sin perder el seguimiento. Sin embargo, los algoritmos de los inventores nunca dejan de muestrear el espacio 3D para encontrar un cuadro clave (es decir, una mejor aproximación de RT) que se corresponda con la instantánea actual. Si se recupera un seguimiento visual, todos los demás datos de sensor se utilizan y actualizan para corregir su propia información además de la de la cámara (ver filtro de Kalman extendido, sección 2.3.2).

La versión más reciente (es decir, estable) de la tecnología de los inventores es estereoscópica (ver la sección 2.5), lo que significa que utiliza dos cámaras testigo separadas por 21 centímetros (ver la figura 6). Esta tecnología no requiere ningún cálculo de una homografía ni ninguna traslación manual de una sola cámara testigo (ver la sección 2.1.3) para calcular la información de profundidad de los marcadores naturales detectados en la señal de video. Este método simplifica el algoritmo principal de los inventores haciendo obsoleta la fase de inspección en la mayoría de los casos (ver sección 2.5). Al igual que en la técnica monoscópica, el seguimiento visual en sí tiene un enfoque completo de subpíxeles que permite que el sistema siga la posición en la cámara a 4 niveles diferentes de una pirámide de imágenes (ver la figura 4), mientras que todas las demás tecnologías basadas en flujo óptico utilizan incorrectamente un enfoque de píxeles en solo dos resoluciones diferentes de imágenes cuando buscan puntos característicos.

2.1.3 1ª fase: la inspección

2.1.3.1 Etapa de inicialización

La primera etapa es que el usuario haga una traslación horizontal física/manual con el dispositivo para capturar dos cuadros. El algoritmo detecta automáticamente marcadores naturales (es decir, puntos característicos) en las imágenes mediante el uso de un método de detección de esquinas denominado FASTER[17]. Se aplica un emparejamiento para encontrar la correspondencia entre un par de puntos. Para una imagen, todos los puntos detectados tienen que permanecer en el mismo plano 3D para poderse calcular una correcta *homografía* de ese plano a la foto. Esto da una rotación y traslación de la cámara con respecto al plano 3D. La segunda imagen sigue el mismo principio y se obtiene una segunda posición y traslación de la cámara para la imagen actual. Ahora es posible calcular la transformación de una cámara a otra y obtener una primera nube de puntos 3D.

Considérense dos vistas (izquierda y derecha) capturadas por una cámara. Para cada vista, la cámara testigo apuntaba a un punto *M* en un plano. M^L y M^R son las proyecciones en perspectiva de *M* en las vistas izquierda y derecha respectivamente.

Se puede escribir:

$$M^L = H \cdot M^R \quad (1)$$

$$H = \left(D \cdot R + \vec{T} \cdot \vec{N}^t\right) \quad (2)$$

dónde:

- *R* es la matriz de rotación mediante la que M^L se rota con respecto a M^R.
- $\vec{N}$ (*A B C*) es el vector normal al plano al que pertenece el punto *M*.
- *D* es la distancia al plano, tal como $Ax + By + Cz + D = 0$.

Resolviendo para H en la ecuación 1 se obtiene:

$$H = \begin{pmatrix} x_{M^L} & 0 \\ y_{M^L} & 0 \\ 1 & 0 \\ 0 & x_{M^L} \\ 0 & y_{M^L} \\ 0 & 1 \\ -x_{M^L} * x_{M^R} & -x_{M^L} * y_{M^R} \\ -y_{M^L} * x_{M^R} & -y_{M^L} * y_{M^R} \\ -x_{M^R} & -y_{M^R} \end{pmatrix}^t$$

El algoritmo de los inventores requiere un mínimo de 4 puntos coplanares para realizar una inicialización correcta. Es bastante habitual que en realidad se tengan muchos más puntos que eso. Por eso se calculan n homografías posibles utilizando 4 puntos entre todo el conjunto de puntos para cada homografía. A continuación, se usa un método RANSAC[4] para encontrar realmente la mejor homografía posible y construir un conjunto de candidatos a valores atípicos e internos. Los valores atípicos se rechazarán, mientras que los valores internos se refinarán mediante un método de Gauss-Newton que minimice los errores de reproyección de los puntos en las dos vistas. Habiendo calculado la mejor homografía a partir del conjunto filtrado de puntos, ahora es necesario estimar las matrices de rotación y traslación entre las vistas izquierda y derecha. Esto se consigue fácilmente usando la descomposición en valores singulares en dos matrices ortogonales U, *V* y una matriz diagonal Y. La matriz *H* ahora se puede escribir como: $H = UYV^t$.

$$\Upsilon = \left(D' \cdot R' + T' \cdot N'^t\right) \quad (3)$$

con:

$$R = det(U) \cdot det(V) \cdot U \cdot R' \cdot V^t$$

$$T = U' \cdot T'$$

$$N = V \cdot N'$$

$$D = det(U) \cdot det(V) \cdot D'$$

Como no se manejan casos de homografías indeterminadas, la ecuación 3 tiene dos posibles soluciones:

- Solución 1:

$$T' = (\lambda_1 - \lambda_3) \begin{pmatrix} \varepsilon_1 \sqrt{\frac{\lambda_1^2 - \lambda_2^2}{\lambda_1^2 - \lambda_3^2}} \\ 0 \\ -\varepsilon_3 \sqrt{\frac{\lambda_2^2 - \lambda_3^2}{\lambda_1^2 - \lambda_3^2}} \end{pmatrix} \quad (4)$$

$$R' = \begin{pmatrix} \cos\theta & 0 & -\operatorname{sen}\theta \\ 0 & 1 & 0 \\ \operatorname{sen}\theta & 0 & \cos\theta \end{pmatrix} \quad (5)$$

con:

$$\begin{cases} \cos\varphi = \frac{(\lambda_1\lambda_3 - \lambda_2^2)}{\Lambda} \\ \operatorname{sen}\varphi = \varepsilon_1\varepsilon_3 \frac{\sqrt{(\lambda_1^2 - \lambda_2^2)(\lambda_2^2 - \lambda_3^2)}}{\Lambda} \\ \Lambda = \lambda_2(\lambda_1 - \lambda 3) \end{cases}$$

λ_1, λ_2 y λ_3 son los valores propios de la matriz H de la ecuación 2, y ordenados como λ_1 λ_2 λ_3 y $\lambda_1 \neq \lambda_2 \neq \lambda_3$.

- Solución 2:

$$T' = (\lambda_1 + \lambda_3) \begin{pmatrix} \varepsilon_1 \sqrt{\frac{\lambda_1^2 - \lambda_2^2}{\lambda_1^2 - \lambda_3^2}} \\ 0 \\ \varepsilon_3 \sqrt{\frac{\lambda_2^2 - \lambda_3^2}{\lambda_1^2 - \lambda_3^2}} \end{pmatrix} \quad (6)$$

$$R' = \begin{pmatrix} \cos\varphi & 0 & \operatorname{sen}\varphi \\ 0 & -1 & 0 \\ \operatorname{sen}\varphi & 0 & -\cos\varphi \end{pmatrix} \quad (7)$$

con:

$$\begin{cases} \cos\varphi = \frac{(\lambda_1\lambda_3 - \lambda_2^2)}{\Lambda} \\ \operatorname{sen}\varphi = \varepsilon_1\varepsilon_3 \frac{\sqrt{(\lambda_1^2 - \lambda_2^2)(\lambda_2^2 - \lambda_3^2)}}{\Lambda} \\ \Lambda = \lambda_2(\lambda_1 - \lambda 3) \end{cases}$$

λ_1, λ_2 y λ_3 son los valores propios de la matriz H de la ecuación 2, y ordenados como λ_1 λ_2 λ_3 y $\lambda_1 \neq \lambda_2 \neq \lambda_3$.

Además del conjunto de puntos característicos, el algoritmo también captura dos posiciones clave como instantáneas de lo que ven las cámaras testigo desde cada punto de vista. El aprendizaje de puntos adicionales es siempre una suma de tres componentes: nuevos puntos detectados, posiciones clave y parches, que se describen en la sección 2.1.3.3.

Finalmente, se debe tener en cuenta que en el caso de un enfoque estereoscópico (ver sección 2.5), la fase de inspección se vuelve automática e instantánea. Una inspección manual aún podría usarse para casos extremadamente complejos (decididos experimentalmente en el plató), pero esto sigue siendo anecdótico.

2.1.3.2 Escala, origen y orientación del mundo 3D automáticas

Durante una inspección monoscópica, no es posible calcular un origen, una escala y una orientación precisas del mundo, dado que esto requeriría el conocimiento previo de un objeto real, incluyendo sus forma y dimensiones exactas, o la distancia utilizada entre las dos vistas para calcular la homografía. Cabe señalar que en ese último caso, esto no daría el origen o la orientación del mundo de todos modos. La mayoría de los sistemas no tienen en

cuenta esa parte. A menudo solicitan al usuario que haga la escala manualmente ingresando las dimensiones de un objeto 3D visible en los videos. Otras opciones son que el usuario escale los personajes virtuales 3D dentro de un software de modelado/animación específico durante el rodaje y también alineándolos correctamente con el mundo real. Sin embargo, cualquier error de escala, incluso pequeño, tiene un impacto dramático en la calidad de la composición final y es por eso que la tecnología de los inventores lo consigue con tanto cuidado (ver más abajo). Una de las consecuencias más comunes observadas de un mal escalado es la sensación por parte del usuario de que los objetos virtuales se deslizan sobre el suelo, en lugar de permanecer perfectamente fijados al mismo.

Los inventores proponen un nuevo método que requiere poner un **objeto de registro** propietario en la escena cuando se realiza la inspección. Este objeto se detecta automáticamente porque tiene una forma muy conocida, y también se conocen todas sus dimensiones. No se requiere interacción del usuario en absoluto.

El objeto de registro de los inventores se compone de varias partes que son esencialmente esferas y patas, como se muestra en la figura 3. Las esferas son partes cruciales de ese objeto en el sistema de los inventores, dado que permiten usar algoritmos de detección de regiones para reconocer el patrón a través del video. A continuación se calculan los centros de gravedad de las esferas proyectadas (es decir, círculos) y, como se conocen con precisión las dimensiones y la distancia entre cada esfera, el sistema puede dar una escala muy precisa para el mundo. También se tiene una versión plana del objeto de registro (ver la imagen de la derecha de la figura 5), que a menudo es más cómodo mover de un lugar a otro en un escenario real.

Una vez se ha conseguido esto, un usuario puede, por ejemplo, colocar un objeto virtual de 1,80 m de altura en el video real y asegurarse de que realmente se verá así de alto y correcto. Otros enfoques no logran obtener tal cualidad debido al aspecto manual de la tarea.

El escalado automático es en realidad algo que se realiza durante la inspección en sí, pero como sucede al comienzo de la inspección, es importante considerarlo también como una etapa de inicialización. De hecho, la escala calculada se usa en las siguientes etapas de la inspección para construir un conjunto preciso de marcadores naturales 3D. Veamos ahora en detalle la inspección en sí.

2.1.3.3 Proceso de inspección y construcción de nubes de puntos

La inspección requiere que el usuario se mueva a través de la escena usando el *dispositivo de inspección*.

Como el sistema siempre calcula nuevos puntos de acuerdo con la confianza que tiene en los aprendidos previamente, la inspección siempre se construye de manera que la cámara testigo vea suficientes puntos del conjunto anterior para añadir nuevos candidatos.

Durante los movimientos de la cámara testigo, el algoritmo crea un **mapa** compuesto de tres conjuntos de datos aprendidos en los 4 niveles diferentes de una **pirámide de imágenes** (ver figura 4):

- Una posición clave. Una posición clave contiene una instantánea en 4 resoluciones diferentes de lo que ve la cámara testigo. El nivel inferior de la pirámide es la resolución original de la cámara testigo (640 × 480 en nuestro caso). El algoritmo también utiliza esa posición clave para recuperar la posición de la cámara cuando el seguimiento no va bien.

- Un conjunto de puntos característicos 2D estimados por *FASTER*[17] en todas las imágenes de la posición clave. FASTER es un algoritmo muy conocido cuya característica principal es ser un algoritmo detector de esquinas. Cada punto clave también tiene unido un descriptor ORB para garantizar una recuperación mucho más rápida cuando se pierde el seguimiento (ver la sección 2.1.4.4).

- Un conjunto de parches (16 × 16 texturas 2D) centrados en cada punto característico detectado por la etapa anterior. Durante la detección de nuevos puntos, no hay forma de comenzar desde una sola vista para calcular su profundidad. Para eso sirven los parches. Una búsqueda epipolar (ver figura 2, rectángulo *generador de nubes de puntos*) se puede aplicar a través de las 4 imágenes de las posiciones clave encontrando una correspondencia entre dos parches en dos posiciones clave lo más cercanas posible. Una vez se ha detectado un punto (es decir, un parche) en ambas vistas, es posible calcular un punto característico 3D. El conjunto de puntos característicos 3D se denomina *mapa*. También es importante comprender que esta búsqueda entre la posición clave A y B se consigue atravesando niveles iguales de la pirámide A y B, pero también el subnivel de la pirámide B (ver figura 4).

Durante la construcción del mapa, la cámara se mueve según el desplazamiento del operador de inspección. En este momento, el software de los inventores solo conoce los parámetros de la cámara que fueron calculados previamente (es decir, seguidos). Para calcular la nueva posición de la cámara, se necesita el cuadro actual y la nube de puntos 3D de marcadores naturales. Por un lado, FASTER calcula un conjunto de marcadores 2D en varios niveles de la pirámide de imágenes (nivel actual y nivel actual+1) y, por otro lado, la nube de puntos 3D se reproyecta en el cuadro actual. Esta última etapa solo se puede conseguir si se conoce la posición de la cámara a medida que se reproyectan los puntos desde su punto de vista. Pero eso es precisamente lo que estamos tratando de calcular. Por lo tanto, el software de los inventores utiliza un modelo de velocidad constante asociado a la información proporcionada por el sensor de los inventores de 9 DOF (ver la sección 2.3) para predecir la siguiente

posición de la cámara utilizando la posición previamente calculada correctamente. Usando esa predicción, la nube de puntos 3D se puede reproyectar en el cuadro actual y se aplica un algoritmo de puesta en correspondencia de puntos para encontrar una correspondencia entre los puntos 2D que fueron detectados por FASTER y los puntos proyectados de la nube de puntos 3D. El error entre los dos conjuntos de marcadores se minimiza utilizando un algoritmo de Levenberg-Marquardt [10, 12, 15], dado que se sabe que es el mejor algoritmo de optimización para ese tipo de problemas. Si el número de puntos en correspondencia dividido por el número total de puntos proyectados es mayor que un determinado umbral, el sistema puede seguir la cámara con éxito (el seguimiento es *bueno*) y se le permite añadir nuevas posiciones clave. Los puntos 2D detectados por FASTER que no encontraron una correspondencia en la nube de puntos 3D se almacenan en la memoria para usuarios posteriores, así como sus parches de textura 2D 16 × 16 relacionados. Serán necesarios para generar nuevos puntos característicos 3D (ver el siguiente párrafo).

Se añaden nuevas posiciones clave (y nuevos puntos característicos) si se cumplen 3 condiciones. En primer lugar, como se dijo en el párrafo anterior, el seguimiento tiene que ser *bueno,* lo que significa que es lo suficientemente preciso o no se pierde. En segundo lugar, la posición clave se añade cada 30 cuadros (1 por segundo) para evitar la creación de un conjunto de datos demasiado grande. En tercer lugar, la nueva posición clave tiene que estar a una distancia mínima de 10*cm* desde la posición clave más cercana. Esto evita que el sistema aprenda puntos adicionales cuando está parado.

Cuando todas estas pruebas se han pasado con éxito, el software puede añadir una nueva instantánea y nuevos puntos característicos 3D. La primera etapa es usar FASTER nuevamente para detectar nuevos puntos 2D relevantes en el nuevo cuadro (instantánea). Como se tiene un conjunto de puntos característicos 2D sin correspondencia procedentes del seguimiento, ahora se intenta poner correspondencia el conjunto de puntos característicos 2D "antiguo" y el recién calculado. Esto se hace mediante una búsqueda epipolar clásica utilizando los parches de textura 2D del nuevo conjunto de puntos. Los parches se mueven a lo largo de las líneas epipolares (ver figura 2) tratando de corresponderse con el conjunto de parches 2D del conjunto "antiguo" de puntos. Gracias a la restricción epipolar, es posible calcular la profundidad del punto 3D procedente de la correspondencia de dos puntos 2D. En realidad, esto es similar a la etapa de puesta en correspondencia de la fase de homografía (ver la sección 2). Si se han añadido nuevos puntos al mapa, se aplica un ajuste de paquetes local. Después de haber refinado la posición y la rotación de la cámara usando estos nuevos puntos característicos, se usa un algoritmo de minimización de Levenberg-Marquardt para refinar la posición 3D de todos los puntos característicos. Esto se hace sobre una ventana de $k + 1$ cuadros, lo que significa que la minimización tiene en cuenta la instantánea actual más las k más cercanas para estimar correctamente la posición actual y la rotación de la cámara (ajuste de paquetes local). El valor de k puede determinarse empíricamente o calcularse adaptativamente de acuerdo con un umbral de error dado para el algoritmo de minimización.

Además del ajuste de paquetes local, también se aplica un ajuste de paquetes global en todas las posiciones clave existentes. A medida que el conjunto de posiciones clave se hace cada vez mayor con el tiempo, el sistema tarda cada vez más en realizar el ajuste global. Si se añaden nuevas posiciones clave y el software de los inventores no tuvo tiempo suficiente para realizar los ajustes de paquetes locales y/o globales, los ajustes se cancelan para dejar la prioridad a la adición de nuevas posiciones clave. Estos se aplicarán nuevamente en cuanto se consiga la incorporación de nuevas posiciones clave.

Una vez existe una nube de puntos, se puede seguir la cámara. Esto es descrito en la siguiente sección.

2.1.4 2ª Fase: seguimiento de la cámara cinematográfica

2.1.4.1 Calibración de cámara geométrica

La calibración de la cámara tiene varios objetivos. Por un lado da una relación matemática para describir cómo se crea la imagen en el sensor. Por otro lado, calcula una matriz de transformación entre los puntos 3D y su proyección sobre el plano de imagen de la cámara.

Este importante requisito previo del proceso de inspección y seguimiento significa que se tienen que calcular dos tipos diferentes de propiedades: los parámetros intrínsecos y extrínsecos. Los parámetros intrínsecos están relacionados con las propiedades de la cámara y la lente y se mueven de acuerdo con la forma en que estas se han construido físicamente. Abarcan el formato de la imagen, la distancia entre el centro óptico de la cámara y el plano de la imagen, y el punto principal. El formato de la imagen se compone de dos factores de escala que relacionan los píxeles con la distancia. El punto principal es la proyección del centro óptico sobre el plano de la imagen (lo ideal es que esté en el centro de la imagen). Además de eso, algunas lentes requieren que se calcule su distorsión óptica y se tenga en cuenta para el seguimiento. Durante un proceso de calibración, la cámara captura sucesivamente una superficie plana que tiene varios patrones circulares (ver la figura 5). Para todas las posiciones, los parámetros extrínsecos (es decir, la posición y la rotación de la cámara) y los parámetros intrínsecos (incluida la distorsión óptica) se calculan mediante un algoritmo de minimización de Levenberg-Marquardt que minimiza el error de reproyección entre todas las instantáneas.

Este proceso es una técnica muy famosa basada en los trabajos de Devernay y Faugeras[3].

2.1.4.2 Compensación de cámara cinematográfica

El sistema global desarrollado aquí requiere colocar un dispositivo de seguimiento conectado directamente a la cámara cinematográfica objetivo. La posición y orientación de la cámara testigo se estiman utilizando la técnica de seguimiento descrita anteriormente. Sin embargo, existe un problema de escala al usar el mismo patrón para las cámaras cinematográficas y testigo. De hecho, como las cámaras testigo de los inventores utilizan lentes de 180 grados, las formas circulares en el patrón de calibración tienden a aparecer muy pequeñas en la vista de la cámara testigo, lo que dificulta su detección.

Los inventores han creado su propia tabla de calibración que en realidad se compone de dos patrones de diferente tamaño y contraste. Una parte del gráfico tiene círculos negros puros sobre un fondo blanco puro (para las cámaras testigo). Una segunda parte del gráfico está formada por círculos de color blanco puro sobre un fondo negro puro (cámara cinematográfica). Entonces, se utiliza un detector de patrones muy simple para encontrar ambos patrones para las cámaras testigo y cinematográficas. Como se conoce el posicionamiento relativo exacto de ambos patrones, la posición y la rotación de la cámara cinematográfica se pueden encontrar "propagando" la posición y la rotación calculadas de la cámara testigo, y viceversa. De hecho, dado que ambas cámaras están conectadas directamente entre sí, podemos utilizar un algoritmo de alineación para calcular con precisión el desplazamiento 3D (*R, T*) entre ambas. Los inventores han desarrollado un algoritmo de alineación basado en los patrones circulares descritos anteriormente. Los patrones cuyos tamaños, formas y posiciones exactos se conocen son reconocidos por separado por ambas cámaras, surgiendo así una relación matemática trivial (rotación y traslación) entre las cámaras testigo y cinematográficas.

2.1.4.3 Fase de seguimiento

Separar la inspección del proceso de seguimiento es solo una forma pragmática de describir cómo se usa el sistema en el plató. De hecho, el proceso de seguimiento no es diferente del seguimiento descrito en la figura 2 y utilizado durante una inspección. Sin embargo, este utiliza un dispositivo más simple (ver figura 6) que es muy parecido al dispositivo de inspección, excepto que no incluye una pantalla táctil por razones obvias.

El rectángulo de *seguimiento visual* de la figura 2 muestra cómo se encuentra una correspondencia entre los marcadores naturales 2D detectados por FASTER en un nuevo cuadro y la nube de puntos 3D existente. Los parches de textura 2D asociados con cada marcador natural y que proceden de ambos conjuntos de datos se comparan linealmente. Una vez se han procesado todos los puntos característicos, se aplica un algoritmo de Gauss Newton para encontrar la mejor correspondencia posible y calcular la posición y orientación de la cámara. Tal algoritmo de minimización generalmente requiere alrededor de 10 iteraciones para converger a la solución. Con el fin de aumentar la solidez de los criterios para minimizar, se utiliza un estimador estadístico M de Tukey (estimador de tipo de máxima verosimilitud) [6]. Esto también asegura que la convergencia no se vea interrumpida por valores atípicos.

Cada seguimiento se realiza dos veces y en dos niveles diferentes de la pirámide antes de obtener los parámetros finales de la cámara. En primer lugar, la convergencia del algoritmo aumenta considerablemente al calcular una aproximación de la posición y la orientación de la cámara a través de un nivel aproximado de la pirámide de imágenes y al usar un subconjunto de marcadores naturales. En segundo lugar, se consigue el mismo cálculo en un conjunto mucho mayor (alrededor de 20 veces mayor) y se fusiona con la aproximación anterior para obtener el resultado final preciso.

2.1.4.4 Fallos del seguimiento y recuperación

Durante un seguimiento de cámara en un plató real, pueden ocurrir muchas situaciones en las que se pierde el seguimiento visual. Esto ocurre a menudo, por ejemplo, cuando las personas se quedan frente a la cámara testigo o la barra simplemente está oculta por objetos. Para todos los casos en los que se pierde el seguimiento visual, se inician tres hilos de recuperación diferentes al mismo tiempo y se acepta el que obtiene una recuperación primero.

Usando la función de predicción del filtro de Kalman extendido (ver sección 2.3.2), es posible intentar primero una recuperación de la posición de la cámara comenzando una búsqueda desde los parámetros predichos de EKF. De hecho, como el EKF es un algoritmo de predicción y corrección, siempre trata de adivinar cuál será la siguiente posición de la cámara antes de que esta se mueva de hecho a esa posición. Cuando la predicción es correcta, es posible obtener una recuperación instantánea. En la práctica, esa predicción funciona muy bien cuando la cámara testigo está oculta y cuando el sistema está rastreando usando la IMU. Sin embargo, a menudo falla si ha habido también una traslación significativa de la cámara cinematográfica mientras están ocultas las cámaras testigo.

El segundo nivel de recuperación es muy diferente. Cada punto característico tiene asociado un descriptor invariante bajo rotación, siguiendo la técnica ORB (Oriented FAST and Rotated BRIEF, FAST orientado y BRIEF rotado) [18]. En primer lugar, se basa en la famosa técnica FAST[16] para calcular características multiescala (es decir, en los diferentes niveles de la pirámide de imágenes, véase la figura 4) asociadas a una puntuación de Harris [5]. La orientación del descriptor se calcula utilizando un centroide de intensidad. El segundo descriptor es rBRIEF, que es un descriptor BRIEF [2] girado siguiendo la orientación del punto clave. Usando estos dos descriptores, la técnica de los inventores puede recuperarse de posiciones donde fallan las técnicas regulares de seguimiento visual. Por

ejemplo, supongamos que perdemos el seguimiento visual por cualquier motivo. Supongamos que ahora trasladamos y giramos 90 grados alrededor del eje Z de la cámara mientras sigue perdida. La imagen actual vista por la cámara nunca se ha aprendido antes, pero sigue apuntando a una dirección en la que se añaden posiciones clave y puntos clave aprendidos antes. Sin embargo, como un proceso de recuperación normal no es invariante bajo rotación, esos sistemas no pueden recuperar el seguimiento. En este caso, el uso de descriptores invariantes bajo rotación asociados a cada punto clave permite una recuperación rápida (usando una técnica de puesta en correspondencia por fuerza bruta) en cuanto la cámara cinematográfica haya estado apuntando a algún lugar en esa posición antes (la rotación no importa).

Finalmente, el tercer hilo de recuperación es más brutal y computacionalmente intensivo. Se construye un elipsoide (elipse 3D) alrededor de la última posición conocida de la cámara. Ese elipsoide tiene un tamaño hecho de la covarianza del parámetro calculado pero el filtro de Kalman extendido. A continuación se generan múltiples muestras dentro del elipsoide siguiendo la covarianza de error del EKF. Se aplica un algoritmo RANSAC a ese conjunto para encontrar la mejor correspondencia posible para la posición y orientación de la cámara buscada.

2.2 Zoom dinámico, enfoque e iris

Durante el proceso de seguimiento global, un director o un director de fotografía, por ejemplo, puede decidir acercar o enfocar un personaje/objeto en la escena. La tecnología de los inventores integra varios codificadores regulares como los propuestos por Arri, pero también es capaz de manejar cualquier otro codificador para capturar un valor de zoom, enfoque e iris en tiempo real. Esto permite el zoom dinámico y la profundidad de campo en el plató.

2.3 Técnica de fusión de sensores

Todas las técnicas puramente ópticas sin marcadores pueden fallar en el seguimiento en muchas situaciones. El caso más común es que no queden suficientes marcadores naturales en la vista actual de la cámara testigo. En situaciones ideales, esto rara vez sucede, ya que el usuario presta atención a no ocultar la lente con la mano, por ejemplo. En platós reales, esa situación sucede mucho simplemente porque el equipo de cámara a menudo necesita cambiar lentes, modificar el equipo de la cámara, limpiar la lente, trasladarse a otro lugar de rodaje, etc. En una técnica normal basada en marcadores y sin marcadores (basada en el flujo óptico), esto evita que el sistema de seguimiento encuentre una buena estimación de la posición y la rotación de la cámara.

Los inventores han creado un método de fusión de sensores que corrige automáticamente la posición y la rotación de una cámara testigo utilizando múltiples sensores: cámaras testigo, giroscopio, acelerómetro y magnetómetro. Esa técnica es el corazón de la tecnología de los inventores.

La primera etapa es la calibración del sistema, lo que significa calibrar la IMU de 9 DOF en relación con las cámaras testigo. Es una etapa crítica para garantizar que todos los sensores realmente funcionen de la misma manera. La relación entre el sensor de 9 DOF y las cámaras testigo es cercana a una técnica mano-ojo[20], pero los inventores han añadido varias contribuciones científicas interesantes para que encaje con el algoritmo de seguimiento de los inventores.

2.3.1 IMU-calibración de cámaras testigo

Consideremos la transformación $R_{IMU \to C}$ que calcula la rotación del sensor 9-DOF (IMU) a una de las dos cámaras testigo (C). R_{IMU} es conocido en su base local y R_C también se conoce en su propia base local. El objetivo es resolver $R_{IMU \to C}$ para $R_C = R_{IMU \to C} \cdot R_{IMU}$.

$R_{IMU \to C}$ se asegura que sea siempre constante ya que la relación entre la IMU y las cámaras no cambia con el tiempo (ambos están acoplados a la barra de la cámara estática).

Por lo tanto, es posible escribir para t_i, $i \in [0, n]$:

$$\begin{aligned} R_C^{t_0} &= R_{IMU \to C} \cdot R_{IMU}^{t_0} \\ &\vdots \\ R_C^{t_{n-1}} &= R_{IMU \to C} \cdot R_{IMU}^{t_{n-1}} \\ R_C^{t_n} &= R_{IMU \to C} \cdot R_{IMU}^{t_n} \end{aligned}$$

De las ecuaciones anteriores se puede deducir:

$$\begin{aligned} R_C^{t_1 - t_0} &= R_{IMU \to C} \cdot R_{IMU}^{t_1 - t_0} \\ &\vdots \\ R_C^{t_{n-1} - t_{n-2}} &= R_{IMU \to C} \cdot R_{IMU}^{t_{n-1} - t_{n-2}} \\ R_C^{t_n - t_{n-1}} &= R_{IMU \to C} \cdot R_{IMU}^{t_n - t_{n-1}} \end{aligned}$$

con las relaciones:

$$R_C^{t_i - t_{i-1}} = R_C^{t_i} \cdot (R_C^{t_{i-1}})^{-1}$$

$$R_{IMU}^{t_i - t_{i-1}} = R_{IMU}^{t_i} \cdot (R_{IMU}^{t_{i-1}})^{-1}$$

Se supone que la medición de rotación proporcionada por el giroscopio y el seguimiento visual son iguales. Por lo tanto, es lógico considerar que R_{IMU} y R_C describen los mismos ángulos de rotación pero en diferentes bases. Usando el mapa logarítmico de la matriz de rotación en algebra de Lie definida como *registro* : $SO(3) \rightarrow so(3)$, es posible convertir la matriz R_C a una representación eje-ángulo ($\vec{r}_c$,α):

$$\alpha = \arccos\left(\frac{tr(R_C) - 1}{2}\right)$$

$$\vec{r_c} = \frac{1}{2\,\mathrm{sen}\,\alpha} \begin{bmatrix} R_c(3,2) & R_c(2,3) \\ R_c(1,3) & R_c(3,1) \\ R_c(2,1) & R_c(1,2) \end{bmatrix}$$

siendo $tr(R_C)$ la *traza* de la matriz, como $tr(R_c) = \sum_{j=1}^{3} R_c(j,j)$. Ahora se puede escribir el siguiente sistema de ecuaciones sobredeterminado:

$$S = \begin{cases} \vec{r}_c^{\,t_1 - t_0} = R_{IMU \rightarrow C} \cdot \overrightarrow{r_{IMU}}^{t_1 - t_0} \\ \vdots \\ \vec{r}_c^{\,t_{n-1} - t_{n-2}} = R_{IMU \rightarrow C} \cdot \overrightarrow{r_{IMU}}^{t_{n-1} - t_{n-2}} \\ \vec{r}_c^{\,t_n - t_{n-1}} = R_{IMU \rightarrow C} \cdot \overrightarrow{r_{IMU}}^{t_n - t_{n-1}} \end{cases} \quad (8)$$

Como para cualquier sistema sobredeterminado, es posible resolver la matriz rectangular anterior usando una descomposición en valores singulares (SVD) como $S = U\sigma V^t$, siendo U la matriz $m \times n$ de los vectores propios de $S \cdot S^t$, V la matriz $n \times n$ de los vectores propios $S^t \cdot S$ y D la matriz $n \times n$ diagonal de los valores singulares de S ordenados.

La solución anterior de la ecuación está sujeta a ruido de las mediciones. Por lo tanto, se aplica un esquema de minimización de Levenberg-Marquardt para minimizar el error:

$$E = \left|\left|\left(R_{IMU \rightarrow C} \times \overrightarrow{r_{IMU}}^{t_i - t_{i-1}}\right) \cdot \vec{r}_C^{\,t_i - t_{i-1}} - 1\right|\right| \quad (9)$$

Ahora que se tiene una relación matemática entre la IMU y las cámaras testigo, es posible inyectar todas las mediciones en el filtro de Kalman extendido para extraer lo mejor de cada sensor y calcular la rotación y la traslación de la cámara cinematográfica.

2.3.2 Filtro de Kalman extendido

El filtro de Kalman extendido (EKF) es una técnica muy conocida en las áreas de investigación aeroespacial y robótica para fusionar diferentes datos procedentes de múltiples sensores diferentes. Se utiliza esencialmente para estimar los parámetros de un sistema no lineal a partir de mediciones ruidosas calculando un error de estimación y aplicando dos fases; predicción y corrección. En la fase de predicción, el filtro de Kalman calcula los nuevos parámetros, así como un error relacionado en un intervalo de tiempo utilizando los parámetros y el error estimado en el intervalo de tiempo anterior. La fase de corrección actualiza la predicción utilizando nuevas mediciones ruidosas. Esas nuevas mediciones permiten calcular nuevos parámetros y errores estimados a partir de los parámetros y errores previstos. Si el error es mayor que un determinado umbral, los parámetros se corrigen.

En nuestro caso aquí, el EKF se desarrolla de una manera muy específica ya que se tienen múltiples mediciones relacionadas con rotaciones 3D (giroscopio + cámaras testigo) y estas no se pueden interpolar fácilmente.

La idea de un EKF para la tecnología de los inventores es poner algún tipo de valor de confianza a cada medición (cámara testigo, giroscopio, magnetómetro, acelerómetro) y corregir de forma iterativa la posición y la rotación estimadas actuales utilizando el sensor que obtiene la mejor puntuación de confianza. En la práctica, el EKF es un poco más complicado que eso y puede describirse matemáticamente como un algoritmo de corrección de predicción para sistemas no lineales. Considerando un vector de estado $\hat{X}(t)$en un intervalo de tiempo *yo,* el algoritmo aplica el

modelo físico actual de nuestras ecuaciones de movimiento para predecir un nuevo estado del vector $\hat{X}(t)$, que pasa a $\hat{X}(t)^{\Delta t}$. La fase de corrección genera entonces un nuevo vector de estado $X(t + \Delta t)$ en un intervalo de tiempo $t + \Delta t$.

El modelo EKF se puede escribir de la siguiente manera en nuestro caso:

$$\begin{cases} Y(t) & = & h(X(t), B) \\ \hat{X}(t)^{\Delta t} & = & f(\hat{X}(t), \Delta t) \end{cases} \quad (10)$$

siendo h las ecuaciones de estado para nuestros observables y siendo X_t el vector de traslación, rotación, velocidad de traslación y velocidad angular en la base global, y f la función de actualización del vector de estado. B es el ruido global en el sistema.

Definición de h

$$Y(t) = h(X(t), B) \quad (11)$$

Si, por un lado, se consideran las mediciones procedentes del seguimiento visual sin marcadores (es decir, cámaras testigo aquí) como $h_{testigo}$, y por otro lado las mediciones procedentes de la IMU como h_{IMU}, podemos decir que h se compone de dos tipos de observables, $h_{testigo}$ y h_{IMU}. Por definición, h es una función trivial y su forma general viene dada por identificación: $Y_{testigo}(t) = X(t)$ e $Y_{IMU}(t) = X(t)$. Dependiendo del tipo de sensores que envíen las mediciones, ahora se pueden escribir dos versiones diferentes de h y por lo tanto reescribiendo la ecuación 11 para el caso de seguimiento visual como:

$$Y(t) = \begin{cases} Y_{testigo}\ (t).T & = & X(t).T \\ Y_{testigo}\ (t).R & = & X(t).R \\ Y_{testigo}\ (t).\vec{V} & = & X(t).\vec{V} \\ Y_{testigo}\ (t).\vec{\Omega} & = & X(t).\vec{\Omega} \end{cases} \quad (12)$$

$Y_{Testigo}(t).\{\alpha \in T, R, V, \Omega\}$ siendo los observables $y^{(y \in \Re^{12})}$ del seguimiento visual en un intervalo de tiempo t para la traslación T, en *metros*), rotación (R, en *radianes*), velocidad (V, en $m \cdot s^{-1}$) y velocidad angular ($\vec{\Omega}$, en $rad \cdot s^{-1}$). En caso de que el sensor sea un giroscopio, la ecuación 11 se convierte en:

$$Y(t) = \begin{cases} Y_{IMU}(t).R & = & X(t).R \\ Y_{IMU}(t).\vec{\Omega} & = & X(t).\vec{\Omega} \end{cases} \quad (13)$$

$Y_{IMU}(t).\{\alpha \in R,\Omega\}$ siendo los observables de la IMU en un intervalo de tiempo t para la rotación (R, en *radianes*) y velocidad angular ($\vec{\Omega}$, en $rad \cdot s^{-1}$). De hecho, no hay traslación o velocidad computable usando un sensor de 9 DOF.

Definición de f

$$\hat{X}(t)^{\Delta t} = f(\hat{X}(t), \Delta t) \quad (14)$$

con $\hat{X}(t)^{\Delta t}$ siendo el vector de estado **predicho** en el intervalo de tiempo $t + \Delta t$, y $X(t)$ el vector de estado **estimado** en el intervalo de tiempo t.

La función f es la función de actualización del vector de estado y se define como:

$$\hat{X}(t)^{\Delta t} = \begin{cases} \hat{X}(t).T & + & \hat{X}(t).V \cdot \Delta_t \\ \hat{X}(t).R & \times & \hat{X}(t).\Omega \cdot \Delta_t \\ \hat{X}(t).\vec{V} & & \\ \hat{X}(t).\vec{\Omega} & & \end{cases} \quad (15)$$

Cabe señalar que la función de actualización regular del filtro de Kalman no se puede utilizar en nuestro caso, ya que las matrices de rotación no se pueden interpolar y se escriben de forma lineal $A \cdot X + B$.

Etapa de predicción

La etapa de predicción se escribe utilizando la definición del filtro de Kalman:

$$\hat{X}(t)^{\Delta t} = f(\hat{X}(t), \Delta t) \quad (16)$$

$$P_t^{\Delta t} = A_t \cdot P_t \cdot A_t^T + Q \cdot \Delta t \qquad (17)$$

con:

- $P_t^{\Delta t}$ la matriz de covarianza del error predicho en el intervalo de tiempo $t + \Delta t$,
- P_t la matriz de covarianza del error en el intervalo de tiempo t,
- A la matriz de transición como $A_t = \frac{\partial f}{\partial X_t} \mid \hat{X}_{t-1}$, A_t es la matriz jacobiana de f y se calcula como una suma de diferencias finitas: $\frac{f(x+\Delta)-f(x)}{\Delta}$,
- q es la matriz de covarianza 12 × 12 del ruido del vector de estado. El ruido del vector de estado de traslación se pone heurísticamente en $1e^{-2}$, así como el ruido del vector de estado de rotación a $1 \cdot e^{-3}$.

$$\mathbf{Q} = \begin{bmatrix} \boxed{0.01\ \mathrm{T}} & \boxed{0.0} & \boxed{0.0} & \boxed{0.0} \\ \boxed{0.0} & \boxed{0.001\ \mathrm{R}} & \boxed{0.0} & \boxed{0.0} \\ \boxed{0.0} & \boxed{0.0} & \boxed{0.0\ \mathrm{V}} & \boxed{0.0} \\ \boxed{0.0} & \boxed{0.0} & \boxed{0.0} & \boxed{0.0\ \Omega} \end{bmatrix}$$

Etapa de corrección

Siguiendo la definición de Kalman, la corrección en nuestro caso se escribe como:

$$K(t+\Delta t) = P_t^{\Delta t} \cdot H^T(t+\Delta t) \cdot \left(U(t+\Delta t) + H(t+\Delta t) \cdot P_t^{\Delta t} \cdot H^T(t+\Delta t)\right)^{-1} \qquad (18)$$

$$P(t+\Delta t) = (I - K(t+\Delta t) \cdot H(t+\Delta t)) \cdot P_t^{\Delta t} \qquad (19)$$

$$\hat{X}(t+\Delta t) = \begin{cases} \hat{X}(t)^{\Delta t}.T + K(t+\Delta t) \times J_{testigo} \mid_{IMU}.T \\ \hat{X}(t)^{\Delta t}.R \times K(t+\Delta t) \times J_{testigo} \mid_{IMU}.R \\ \hat{X}(t)^{\Delta t}.V + K(t+\Delta t) \times J_{testigo} \mid_{IMU}.V \\ \hat{X}(t)^{\Delta t}.\Omega \times K(t+\Delta t) \times J_{testigo} \mid_{IMU}.\Omega \end{cases} \qquad (20)$$

con:

- I la matriz de identidad,
- $k(t + \Delta t)$ la ganancia de Kalman en el intervalo de tiempo $t + \Delta t$,
- siendo H la matriz jacobiana de la función h, que, en sí misma, es $h_{testigo}$ o h_{IMU} dependiendo del sensor seleccionado actualmente.

$$H(t+\Delta t) = \frac{\partial h}{\partial X} \mid \hat{X}_t^{\Delta t} \qquad (21)$$

En nuestro caso, el jacobiano se calcula como una suma de diferencias finitas:

$$\frac{h_{testigo} \mid_{IMU}(x+\Delta) - h_{testigo} \mid_{IMU}(x)}{\Delta}$$

- $U(t + \Delta t)$ es la matriz de covarianza de las mediciones de ruido del sensor actualmente seleccionado (cámara testigo o IMD). Ese valor se modifica en función de la calidad de la medición (valor de confianza). En el caso de un seguimiento visual, depende del error de seguimiento ε:

$$U(t+\Delta t) = \begin{cases} 0.0001 \text{ if } \varepsilon < 1.0\, pixel \\ 10.0 \text{ if } \varepsilon \geq 1.0\, pixel \end{cases} \quad (22)$$

Es aún más fácil determinar $U(t + \Delta t)$ para la IMU ya que los errores de ruido son proporcionados directamente el fabricante para todos los sensores. La ecuación anterior es reemplazada directamente por esos errores.

- $P(t + \Delta t)$ es la matriz de covarianza del error en el intervalo de tiempo $t + \Delta t$.
- V y Ω son la representación matricial del vector velocidad V y de la velocidad angular Ω, respectivamente.
- $j_{testigo|IMU}$ de la ecuación 20 también se escribe siguiendo el tipo de sensores:

$$\begin{aligned} J_{testigo}\ .T &= y(t+\Delta t).T - \hat{Y}_t^{\Delta t}.T \\ J_{testigo}\ .R &= y(t+\Delta t).R \times \hat{Y}_t^{\Delta t}.R^{-1} \\ J_{testigo}\ .V &= y(t+\Delta t).V - \hat{Y}_t^{\Delta t}.V \\ J_{testigo}\ .\Omega &= y(t+\Delta t).\Omega \times \hat{Y}_t^{\Delta t}.\Omega \\ J_{IMU}.R &= y(t+\Delta t).R \times \hat{Y}_t^{\Delta t}.R^{T} \\ J_{IMU}.\Omega &= y(t+\Delta t).\Omega \times \hat{Y}_t^{\Delta t}.\Omega^{-1} \end{aligned}$$

- $y(t + \Delta t)$ es la medición actual dada por el sensor seleccionado. $y(t + \Delta t).(\alpha \in \{ R, T, V, \Omega\})$ es la medición de la traslación, rotación, velocidad y velocidad angular respectivamente. $j.(\alpha \in \{R, T, V, \Omega\})$ se conoce como la innovación de Kalman.

2.4 Reconstrucción 3D y modulación de profundidad

La reconstrucción 3D en tiempo real se consigue utilizando un sensor de distancia 3D integrado en nuestra barra de la cámara. Ese sensor de distancia puede ser un dispositivo de luz estructurada (Microsoft Kinect, Asus Xtion, por ejemplo) o una cámara de tiempo de vuelo (como TigerEye de Advanced Scientific Concept, por ejemplo), también conocida como Flash Lidar. Ambas técnicas tienen limitaciones y la tecnología de los inventores solo usa un sensor de distancias cortas Asus Xtion por ahora. Como es habitual, se requiere una etapa de calibración para conocer la posición del Asus Xtion con respecto a las cámaras testigo. Simplemente usamos exactamente la misma técnica que para la alineación de testigos a película (ver la sección 2.1.4.2). Como el sensor de distancia tiene una cámara en escala de grises, se usa el patrón de calibración habitual de los inventores para que sea detectado en ambas cámaras: la cámara del sensor de distancia y una de las cámaras testigo de los inventores. Ese proceso es bastante sencillo y similar a 2.1.4.2.

El sensor de distancia se utiliza para dos propósitos en la tecnología de los inventores. Por un lado, ayuda a eliminar puntos característicos malos creados lejos detrás de los objetos, por ejemplo. Una simple reproyección de los puntos característicos en el mapa de profundidad del sensor de distancia puede indicar si el punto característico es correcto o no. Durante el seguimiento estereoscópico (ver sección 2.5), cada punto creado por la geometría epipolar puede seguir estando más cerca o más lejos que la estimación proporcionada por las líneas epipolares. Comparamos el punto 3D reconstruido con el valor actual de su reproyección en el mapa de profundidad del sensor de distancia para mejorar su precisión o simplemente rechazarlo.

El segundo uso del sensor de distancia se basa en el propio mapa de profundidad. Como proporciona la profundidad a una distancia máxima de 12 m, se puede ocultar a todas las personas (actores) en el plató dentro de esa distancia mediante los personajes/objetos virtuales, y viceversa. Esto se denomina *oclusiones dinámicas* y plantea otro nivel de realismo para los efectos visuales tanto en películas como en difusión en vivo. La tecnología de los inventores utiliza la señal de la cámara cinematográfica principal (esto también se puede conseguir añadiendo una cámara testigo de alta resolución adicional) para mejorar la resolución del sensor de distancia reproyectando la profundidad en la cámara cinematográfica y refinando los resultados (mejora de bordes). Dicha tecnología es capaz de proporcionar modulación de profundidad en tiempo real y, por lo tanto, elimina el uso de una pantalla verde vinculante, especialmente cuando se usa un sensor de distancia más potente (*Flash Lidar*).

2.5 Técnica estereoscópica sin marcadores

La técnica descrita anteriormente que utiliza un enfoque monoscópico en realidad no es lo suficientemente eficiente en situaciones reales. Como requiere que un objeto real conocido se use como registro para calcular la escala, la orientación y la posición del mundo, a menudo es difícil usarlo en películas donde muchas personas siempre están

en medio cuando se intenta registrar (detectar) el objeto. Pero la principal limitación de ese enfoque monoscópico es el hecho de que requiere inspeccionar la escena. Como se tiene que utilizar una homografía para la inicialización del sistema, esto implica que un técnico tiene que ir al plató para empezar a inspeccionarlo. En la práctica, esa operación rara vez se consigue con éxito ya que las personas que trabajan en el plató nunca abandonan el lugar antes de los últimos segundos. Los inventores han construido un enfoque estereoscópico para deshacerse de esa limitación y, de ese modo, traer la noción completamente original de inspección y seguimiento instantáneos.

2.5.1 Inspección instantánea

La técnica estereoscópica integrada en el seguimiento de la cámara es un gran paso adelante comparada con cualquier otra técnica, incluido el enfoque monoscópico de los inventores descrito anteriormente. De hecho, el uso de 2 cámaras testigo precalibradas aporta varias ventajas. El primero es la *inspección instantánea.* A diferencia de la técnica descrita en el enfoque monoscópico, no se requiere una inspección previa de la escena. Como se conoce perfectamente la distancia exacta entre las dos cámaras testigo, y como estas están sincronizadas mediante sus capacidades internas de hardware de bloque generador (“genlock”), se puede obtener una nube de puntos al instante sin mover la barra. La homografía descrita en el párrafo ya no se usa, dado que la reconstrucción 3D de un punto es traída por las dos cámaras testigo a través de la teoría de la estereovisión. El segundo punto importante es que no se requiere una escala manual. Durante la captura de los dos flujos de video, ejecutamos un esquema de minimización de Levenberg-Marquardt para minimizar la suma del error de píxel entre la proyección del punto 3D estimado por geometría epipolar y el punto característico 2D detectado en el video actual.

2.5.2 Registro automático

Sin embargo, el uso de una inspección instantánea plantea un nuevo problema que es la imposibilidad de utilizar un objeto de registro como el que se detectó en el enfoque monoscópico. La razón principal de esa limitación procede del hecho de que la barra de la cámara está conectada directamente a la cámara cinematográfica y no es posible pedirle al equipo de cámara que mueva su cámara para descubrir el objeto de registro. Es por eso que los inventores han añadido un segundo registro automático que no requiere que se añada ningún objeto de registro físico a la escena. Sin embargo, esto requiere que el usuario tenga dos interacciones diferentes con la nube de puntos: una selección de puntos para indicarle al sistema cuál es el punto que se pretende utilizar como referencia terrestre y una rotación manual alrededor del eje Y (vertical) para orientar adecuadamente el contenido CG en relación con la escena.

Para poder tener solo estas dos interacciones de usuario muy rápidas y muy simples, es necesario calcular todos los demás grados de libertad restantes, lo que significa calcular una base ortonormal que tenga un eje Y perfectamente vertical y una escala correcta. La tecnología de los inventores simplemente aprovecha el acelerómetro integrado en la barra para calcular la orientación de la gravedad, es decir, el eje Y. El magnetómetro da la dirección norte y, por lo tanto, proporciona un vector aproximado que permite calcular una base ortonormal utilizando productos cruzados regulares.

La escala es aún mucho más fácil de calcular. Esto es completamente instantáneo y automático y no requiere ninguna interacción del usuario. Como la distancia entre las cámaras testigo se conoce con precisión, todos los puntos reconstruidos en 3D tienen directamente la escala correcta, por lo que no hay ningún objeto adicional que detectar como en los algoritmos regulares de estructura pura a partir del movimiento (SFM).

2.5.3 Seguimiento instantáneo

El proceso de seguimiento es ligeramente diferente del descrito para el enfoque monoscópico en la sección 2.1.4.3. Cada nuevo punto se crea mediante el uso de geometría epipolar que garantiza que la escala y la profundidad de un punto característico 3D sean siempre correctas. Esto significa que en cuanto se inicia nuestro sistema, este está ya rastreando gracias a esa inspección instantánea y precisa a la velocidad de la luz.

Se calcula la unión de todos los puntos 3D reproyectados que se ven desde las cámaras izquierda y derecha, para garantizar la máxima estabilidad del proceso de seguimiento. La nube de puntos se genera simplemente por la intersección 3D de las líneas epipolares combinadas con el descriptor ORB para la puesta en correspondencia de características (ver la sección 2.1.4.4). El esquema de puesta en correspondencia de características y minimización global sigue siendo similar al utilizado en el enfoque monoscópico.

Para poder confiar en la inspección instantánea, los inventores han modificado mucho, asimismo, la forma en que se construyen los nuevos puntos en la nube de puntos. Además de la reconstrucción estereoscópica del punto, los inventores han añadido la noción de dispersión de cuadrícula y uniformidad de puntos característicos. Durante un algoritmo de seguimiento regular, se inicia un algoritmo de detección de características 2D y los puntos se reconstruyen en 3D, tal como se describe en las secciones 2.1.3.3 y 2.1.4.4. Sin embargo, muchos de estos puntos no son fiables (mala profundidad, puntos de borde, puntos en movimiento, puntos de sombra, etc.) y el algoritmo tiende a aprender mucho del punto en áreas de alto contraste y nada en áreas de bajo contraste. Esto genera mucho ruido en los datos de seguimiento dando la sensación de que la cámara está temblando. La técnica de dispersión en cuadrícula que los inventores han creado intenta distribuir los marcadores naturales aprendidos por el sistema de forma casi uniforme. Se empieza primero usando un umbral muy alto para generar un máximo de puntos

característicos. A continuación, se reduce el número de puntos característicos de acuerdo con un umbral proporcionado por el usuario (cuadrícula de 16 × 16, 2 muestras por celda, por ejemplo) usando una puntuación de Harris (se eligen los 2 mejores puntos, por ejemplo). Cada punto, por lo tanto, tiene un criterio de calidad que se utiliza para decir cuán fiable es. Si la uniformidad no es lo suficientemente buena, el sistema también cambia localmente el umbral de contraste para garantizar que puede capturar puntos incluso en áreas de bajo contraste. El hecho de que el algoritmo de los inventores distribuya los puntos característicos a lo largo de la imagen también evita que el sistema pierda el seguimiento fácilmente. De hecho, los enfoques monoscópicos regulares sin una técnica de dispersión de cuadrícula como la de los inventores pueden tener muchos puntos para el seguimiento en la parte inferior izquierda de la imagen, por ejemplo. Si un actor entra en lo que están ocultando todos estos marcadores naturales, el seguimiento se perderá instantáneamente, algo que no puede suceder en nuestro caso, ya que distribuimos la función en toda la imagen.

2.6 Aumentando el mundo real

2.6.1 El problema del retardo de la cámara/testigo

Como la cámara cinematográfica y las cámaras testigo funcionan a diferentes velocidades, pueden generar diferentes retardos en el proceso de adquisición de cuadros. La cámara cinematográfica es en realidad la que crea el retardo más alto, lo que obliga a calcularlo para aumentar adecuadamente el mundo real.

Se usa de nuevo el propio dispositivo de calibración de los inventores para calcular automáticamente el retardo entre dos cámaras. Se utiliza un pequeño LED parpadeante como el sol que parpadea a 1 Hz frente a ambas cámaras. Se ha desarrollado un algoritmo detector de regiones mejorado, mediante una técnica de umbral de intensidad para detectar el LED en ambas cámaras. Las curvas de intensidad relacionadas que se generan a partir de la captura de la intensidad del LED en las cámaras se comparan mediante un algoritmo de minimización para encontrar la mejor coherencia temporal entre muestras. De este modo, se minimizan varios parámetros estadísticos y se extrae automáticamente un retardo en milisegundos, a partir de la calibración en unos 10 segundos.

2.6.2 Representación y realidad aumentada

Una vez se tiene una cámara seguimiento funcionando, el software de los inventores puede recibir datos de diferentes softwares generadores de contenido CG, tal como *Autodesk Motion Builder*, por ejemplo. Esos datos se integran y representan directamente dentro del software de los inventores (en el caso de que las estaciones de trabajo de seguimiento y de representación se fusionen, ver la figura 1) para combinarse con el fondo o el primer plano reales (en el caso de un escenario verde).

Otra característica interesante inherente a la tecnología de los inventores es el uso de mates de datos sobrantes 3D. El uso de mates de datos sobrantes ciertamente no es invención de los inventores pero se hace de manera innovadora en el caso de los inventores. De hecho, como el algoritmo de seguimiento de los inventores genera una nube de puntos 3D precisa (gracias al enfoque estereoscópico), todos estos puntos tienen una representación tanto en 3D como en 2D cuando se reproyectan en el video de la cámara cinematográfica. Los inventores han desarrollado una interfaz de usuario simple que permite al usuario seleccionar múltiples puntos dentro de la nube de puntos 3D. Por lo tanto, se crea un objeto 3D triangulando todos los puntos seleccionados utilizando, un algoritmo de triangulación de orejas [14]. Esa superficie recién creada ahora se puede activar como objeto de mates de datos sobrantes, lo que significa que cada píxel cubierto por la proyección 2D de ese objeto en el video de la cámara cinematográfica ahora se reemplaza por el entorno virtual 3D generado por la estación de trabajo de generación de contenido 3D (ver la sección 1). Cabe señalar que en el caso del uso de un sensor de profundidad de largo alcance (es decir, al menos 50 m), la técnica de mates de datos sobrantes se vuelve inútil ya que el método de modulación de profundidad de los inventores la reemplaza fácilmente (ver la sección 2.4).

Apéndice 1

Referencias

[1] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: speeded up robust features. En Proceedings de la 9th European conference on Computer Vision - Volumen Parte 7, ECCV'06, páginas 404 a 417. Berlin, Heidelberg, 2006. Springer-Verlag.

[2] Michael Calender, Vincent Lepetit, Christoph Strecha, and Pascal Fua. Brief: binary robust independent elementary features. In Proceedings of the 11th European conference on Computer vision: parte IV, ECCV'10, páginas 778 a 792, Berlin, Heidelberg, 2010. Springer-Verlag.

[3] Frédéric Devernay and Olivier Faugeras. Straight lines have to be straight. En SPIE, volumen 2567, 2001.

[4] Martin A. Fischler and Robert C. Bolles. Random sample consensus: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 24(6):381-395, 1981.

[5] Chris Harris and Mike Stephens. A combined corner and edge detector. In Proc. of Fourth Alvey Vision Conference, páginas 147-151, 1988.

[6] D.C. Hoaglin, F. Mosteller, and J.W. Tukey. Understanding Robust and Exploratory Data Analysis. Wiley series in probability and mathematical statistics: applied probability and statistics. John Wiley and Sons, 1983.

[7] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. Transactions of the ASME-Journal of Basic Engineering, 82(Series D):35-45, 1960.

[8] Georg Klein and David Murray. Parallel tracking and mapping for small ar workspaces. En Procedimientos de 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, ISMAR '07, páginas 1-10, Washington, DC, USA, 2007. IEEE Computer Society.

[9] J. J. Leonard and Durrant H. Whyte. Mobile robot localization by tracking geometric beacons. IEEE Transactions on Robotics and Automation, 7(3), 1991.

[10] K. Levenberg. A method for the solution of certain problems in least squares. Quart. Appl. Math, 2:164-168, 1944.

[11] David G. Lowe. Object recognition from local scale-invariant features. En Proceedings de la International Conference on Computer Vision- Volumen 2 - Volumen 2, ICCV '99, páginas 1150-, Washington, DC, USA, 1999. IEEE Computer Society.

[12] D. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. SIAM J. Appl. Math, 11:431-441, 1963.

[13] Christopher Mei, Gabe Sibley, Mark Cummins, Paul Newman, and Ian Reid. Rslam: A system for large-scale mapping in constant-time using stereo. Int. J. Comput. Vision, 94(2):198-214, septiembre de 2011.

[14] Gary Hosler Meisters. Polygons have ears. Amer. Math. Monthly, 82:648-651, 1975.

[15] J. J. More'. The Levenberg-Marquardt algorithm: implementation and theory, páginas 105-116. Lecture Notes in Mathematics 630. SPRINGER, 1977.

[16] Edward Rosten and Tom Drummond. Machine learning for high-speed corner detection. En Proceedings de la 9th European conference on Computer Vision - Volumen Parte 7, ECCV'06, páginas 430-443, Berlin, Heidelberg, 2006. Springer-Verlag.

[17] Edward Rosten, Reid Porter, and Tom Drummond. Faster and better: A machine learning approach to corner detection. IEEE Trans. Pattern Analysis and Machine Intelligence, 32:105-119, 2010.

[18] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. En Proceedings de la 2011 International Conference on Computer Vision, ICCV '11, páginas 2564-2571, Washington, DC, USA, 2011. IEEE Computer Society.

[19] Ivan E. Sutherland. A head-mounted three dimensional display. En Proceedings del 9-11 de diciembre de 1968, conferencia conjunta de informática de otoño, parte I, AFIPS '68 (Otoño, part I), páginas 757-764, New York, NY, USA, 1968. ACM.

[20] Roger Y. Tsai and Reimer K. Lenz. A new technique for fully autonomous and efficient 3d robotics hand-eye calibration. En Proceedings del 4th international symposium on Robotics Research, páginas 287-297, Cambridge, MA, USA, 1988. MIT Press.

[21] J.K. Ulhmann. Algorithms for multiple target tracking. American Scientist, 80(2):128-141, 1992.

REIVINDICACIONES

1. Un sistema sin marcadores, incluyendo el sistema:

(i) una cámara de video;

(ii) sensores que incluyen un acelerómetro y un giroscopio que detectan más de seis grados de libertad;

(iii) dos cámaras testigo formando un sistema estereoscópico; y

(iv) un procesador;

para mezclar o componer en tiempo real, objetos 3D generados por ordenador y una señal de video de la cámara de video, para generar video de realidad aumentada en tiempo real para difusión de TV, cine o videojuegos, en el que:

(a) el cuerpo de la cámara de video se puede mover en 3D y los sensores en la cámara de video o conectados directa o indirectamente a la misma proporcionan datos de posicionamiento en tiempo real que definen la posición 3D y la orientación 3D de la cámara de video, o permiten calcular la posición 3D y la orientación 3D de la cámara de video;

(b) las dos cámaras testigo que forman el sistema estereoscópico están fijadas directa o indirectamente a la cámara de video;

(c) el sistema está configurado para usar esos datos de posicionamiento en tiempo real automáticamente para crear, recuperar, representar o modificar objetos 3D generados por ordenador;

(d) el sistema está configurado para mezclar o combinar los objetos 3D generados por ordenador resultantes con la señal de video de la cámara de video para proporcionar video de realidad aumentada para difusión de TV, cine o videojuegos;

y en el que:

(e) el sistema está configurado para determinar la posición 3D y la orientación de la cámara de video haciendo referencia a un mapa 3D del mundo real, donde el sistema está configurado para generar el mapa 3D del mundo real, al menos en parte, utilizando los datos de posicionamiento 3D en tiempo real de los sensores más un flujo óptico en el que las dos cámaras testigo que forman el sistema estereoscópico inspeccionan una escena, y en el que el software que se ejecuta en el procesador está configurado para detectar marcadores naturales en la escena que no se han añadido manual o artificialmente a esa escena;

(f) el sistema está configurado para usar un modelo de velocidad constante asociado con los datos de posicionamiento 3D en tiempo real de los sensores para predecir la siguiente posición de la cámara de video usando una posición previamente calculada o confirmada correctamente, y el sistema está configurado para usar esa predicción para proyectar una nube de puntos 3D en un cuadro de la cámara testigo actual, y para usar un algoritmo de puesta en correspondencia de puntos para poner en correspondencia puntos identificados en una señal de video en tiempo real del sistema estereoscópico y puntos proyectados en la nube de puntos 3D proyectada.

2. El sistema según la reivindicación 1, en el que los sensores incluyen un sensor de distancia 3D, que captura la profundidad de cada píxel en una salida de video de la cámara.

3. El sistema según la reivindicación 2, en el que el sensor de distancia 3D es una luz estructurada o una cámara de tiempo de vuelo.

4. El sistema según cualquier reivindicación anterior, en el que las dos cámaras testigo que forman el sistema estereoscópico funcionan al menos a 100 fps para permitir que el sistema se inicialice completamente sin una etapa independiente de inspección pura de la escena que se va a seguir, sino que la inspección se lleva a cabo continuamente mientras la cámara se utiliza para capturar video.

5. El sistema según la reivindicación 4, en el que las dos cámaras testigo estereoscópicas forman el sistema estereoscópico que permite que el software procese las imágenes e, incluso sin mover el sistema de cámaras, genere una nube de puntos 3D instantánea.

6. El sistema según la reivindicación 4, en el que la profundidad de cada punto en la nube de puntos 3D se obtiene utilizando correspondientes parches de textura 2D obtenidos de cada cámara testigo estereoscópica y un algoritmo de búsqueda de línea epipolar.

7. El sistema según cualquier reivindicación anterior, que ejecuta un algoritmo de fusión que combina datos de flujo óptico del sistema de cámaras testigo con los datos de posicionamiento en tiempo real de los sensores.

8. El sistema según la reivindicación 7, en el que el algoritmo de fusión se basa en una técnica de predicción/corrección de filtro de Kalman extendido para integrar las salidas de todos los sensores y recalibrarlos, que pueden incluir un acelerómetro, un giroscopio, un magnetómetro, un sensor de distancia 3D, para determinar la posición y orientación de la cámara.

9. El sistema según la reivindicación 8, en el que el algoritmo de fusión del filtro Kalman extendido usa datos de nivel de confianza, asociados con la salida de cada sensor, al determinar cómo fusionar los datos de cada sensor.

10. El sistema según cualquier reivindicación anterior, en el que los cuadros clave generados por las cámaras testigo son parte de un proceso de seguimiento visual y son imágenes en tiempo real calculadas en cuatro diferentes niveles de resolución de la señal de video de la cámara testigo.

11. El sistema según cualquier reivindicación anterior, incluyendo el sistema un sensor de distancia 3D, en el que el sensor de distancia 3D se usa para mejorar la precisión de una medición de profundidad asociada con un punto 3D reconstruido obtenido usando las cámaras testigo que forman el sistema estereoscópico, o para rechazar ese punto 3D reconstruido obtenido usando las cámaras testigo que forman el sistema estereoscópico.

12. El sistema según la reivindicación 11, en el que el sensor de distancia 3D se usa para modulación de profundidad en tiempo real para permitir oclusión dinámica y suprimir el uso eventual de un escenario verde.

13. El sistema según cualquier reivindicación anterior, en el que el sistema está configurado para cambiar localmente un umbral de contraste para incluir puntos incluso en áreas de bajo contraste, en la nube de puntos.

14. El sistema según cualquiera de las reivindicaciones 4 a 6, que incluye un sistema de seguimiento de la cámara que combina que el sistema se inicializa completamente sin una etapa independiente de inspección pura de la escena que se va a seguir, con el seguimiento de la cámara de video cuando un director/camarógrafo sigue, desplaza, inclina el sistema de seguimiento de la cámara acoplado a la cámara de video.

15. El sistema según la reivindicación 4, en el que las dos cámaras testigo que forman el sistema estereoscópico permiten la inspección continua en tiempo real de una escena para generar una nube de puntos que define la escena.

16. El sistema según cualquier reivindicación anterior, que acopla descriptores invariantes bajo rotación, por ejemplo utilizando ORB, a puntos característicos detectados en la escena para facilitar la recuperación del seguimiento.

17. El sistema según cualquier reivindicación anterior, que utiliza un esquema de minimización de Levenberg-Marquardt para el seguimiento de la cámara para minimizar el error entre los puntos identificados en la señal de video en tiempo real del sistema estereoscópico y los puntos proyectados en la nube de puntos 3D proyectada.

18. El sistema según cualquier reivindicación anterior, en el que un usuario puede usar una nube de puntos 3D generada por el sistema de seguimiento de la cámara para definir máscaras 3D, tales como máscaras de mates de datos sobrantes 3D.

19. El sistema según cualquier reivindicación anterior, en el que la cámara de video y una cámara testigo se calibran para retardo de adquisición de cuadros utilizando una fuente de luz modulada, por ejemplo comparando las curvas de intensidad de luz asociadas con un LED parpadeante.

20. El sistema según cualquier reivindicación anterior, en el que la cámara de video es cualquiera de las siguientes: cámara de grúa; steadicam; cámara portátil; cámara montada en plataforma rodante, cámara montada en trípode, teléfono inteligente, gafas de realidad aumentada.

21. Un método sin marcadores para mezclar o componer en tiempo real objetos 3D generados por ordenador y una señal de video de una cámara de video, para generar video de realidad aumentada para difusión de TV, cine o videojuegos, en el que:

(a) el cuerpo de la cámara de video se puede mover en 3D y los sensores, que incluyen un acelerómetro y un giroscopio que detectan más de seis grados de libertad, en la cámara de video o conectados directa o indirectamente a la misma, proporcionan datos de posicionamiento en tiempo real que definen la posición 3D y la orientación 3D de la cámara de video, o permiten calcular la posición 3D y la orientación 3D de la cámara de video;

(b) dos cámaras testigo que forman un sistema estereoscópico están fijadas directa o indirectamente a la cámara de video;

(c) dichos datos de posicionamiento en tiempo real se utilizan automáticamente, a continuación, para crear, recuperar, representar o modificar objetos 3D generados por ordenador;

(d) los objetos 3D generados por ordenador resultantes se mezclan o combinan, a continuación, con la señal de video de la cámara de video para proporcionar video de realidad aumentada para difusión de TV, cine o videojuegos;

y en el que:

(e) la posición 3D y la orientación de la cámara de video se determinan haciendo referencia a un mapa 3D del mundo real, donde el mapa 3D del mundo real se genera, al menos en parte, utilizando los datos de posicionamiento 3D en tiempo real de los sensores más un flujo óptico, en el que las dos cámaras testigo que forman el sistema estereoscópico inspeccionan una escena, y se usa un software que se ejecuta en un procesador, para detectar marcadores naturales en la escena que no se han añadido manual o artificialmente a esa escena;

(f) se usa un modelo de velocidad constante asociado con los datos de posicionamiento 3D en tiempo real de los sensores para predecir la siguiente posición de la cámara de video usando una posición previamente calculada o confirmada correctamente, y esa predicción se usa para proyectar una nube de puntos 3D en un cuadro de la cámara testigo actual, y se usa un algoritmo de puesta en correspondencia de puntos para poner en correspondencia puntos identificados en una señal de video en tiempo real del sistema estereoscópico y puntos proyectados en la nube de puntos 3D proyectada.

22. El método según se define en la reivindicación 21, en el que los datos de posicionamiento en tiempo real se registran y se marcan con un código de tiempo para proporcionar datos de movimientos en correspondencia, para procesos de postproducción.

23. El método según se define en la reivindicación 21 o 22, utilizado para:

(i) seguimiento en tiempo real para cámaras de estudio; o

(ii) seguimiento en tiempo real para Steadicam; o

(iii) seguimiento en tiempo real para cámaras montadas en grúas; o

(iv) seguimiento en tiempo real para cámaras montadas en plataforma rodante; o

(v) seguimiento en tiempo real para difusión externa; o

(vi) proporcionar datos de seguimiento en tiempo real para posproducción 2D; o

(vii) proporcionar datos de seguimiento en tiempo real para la conversión posterior de contenido estereoscópico 3D; o

(viii) proporcionar datos de seguimiento en tiempo real para contenido estereoscópico 3D nativo; o

(ix) inserción de gráficos 3D; o

(x) inserción de gráficos 3D para publicidad por emplazamiento en estudio o en plató; o

(xi) inserción de gráficos 3D para difusión exterior; o

(xii) inserción de gráficos 3D para imágenes patrocinadas; o

(xiii) inserción de gráficos 3D que es específica de la ubicación del espectador; o

(xiv) inserción de gráficos 3D que es específica del espectador; o

(xv) inserción de gráficos 3D en un tiempo específico; o

(xvi) inserción de gráficos 3D para rellenar escenas de multitudes; o

(xvii) inserción de gráficos 3D para reemplazo de pantalla verde; o

(xviii) inserción de gráficos 3D de contenido educativo para ayudar al aprendizaje, en museos y centros de interpretación en sitios culturales, históricos o naturales; o

(xix) medición del tamaño absoluto o relativo de los objetos en la escena.

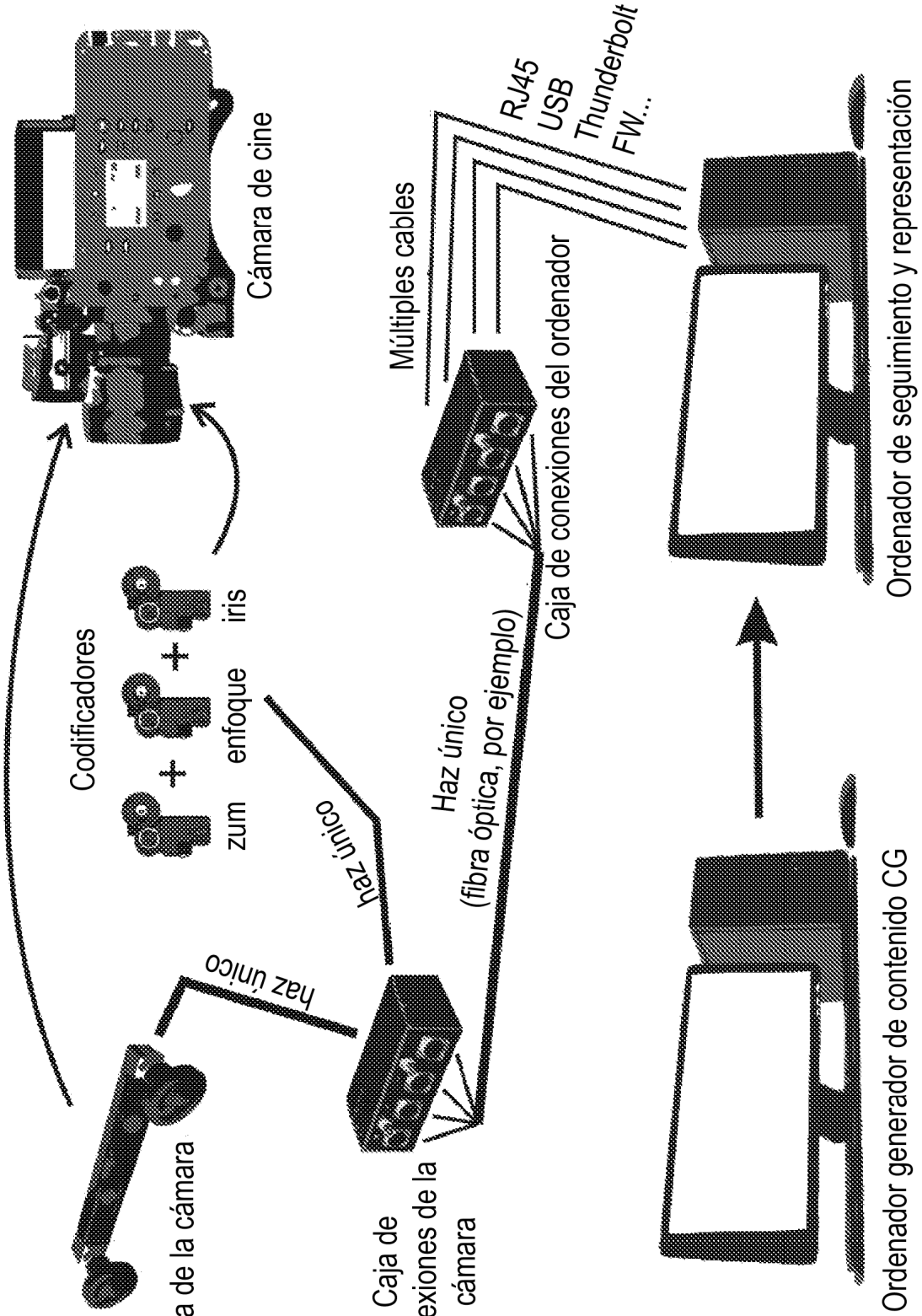
Cámara de cine
Codificadores
zum
enfoque
iris
Múltiples cables
RJ45
USB
Thunderbolt
FW...
Caja de conexiones del ordenador
Haz único
(fibra óptica, por ejemplo)
haz único
haz único
Barra de la cámara
Caja de conexiones de la cámara
Ordenador de seguimiento y representación
Ordenador generador de contenido CG

Figura 1

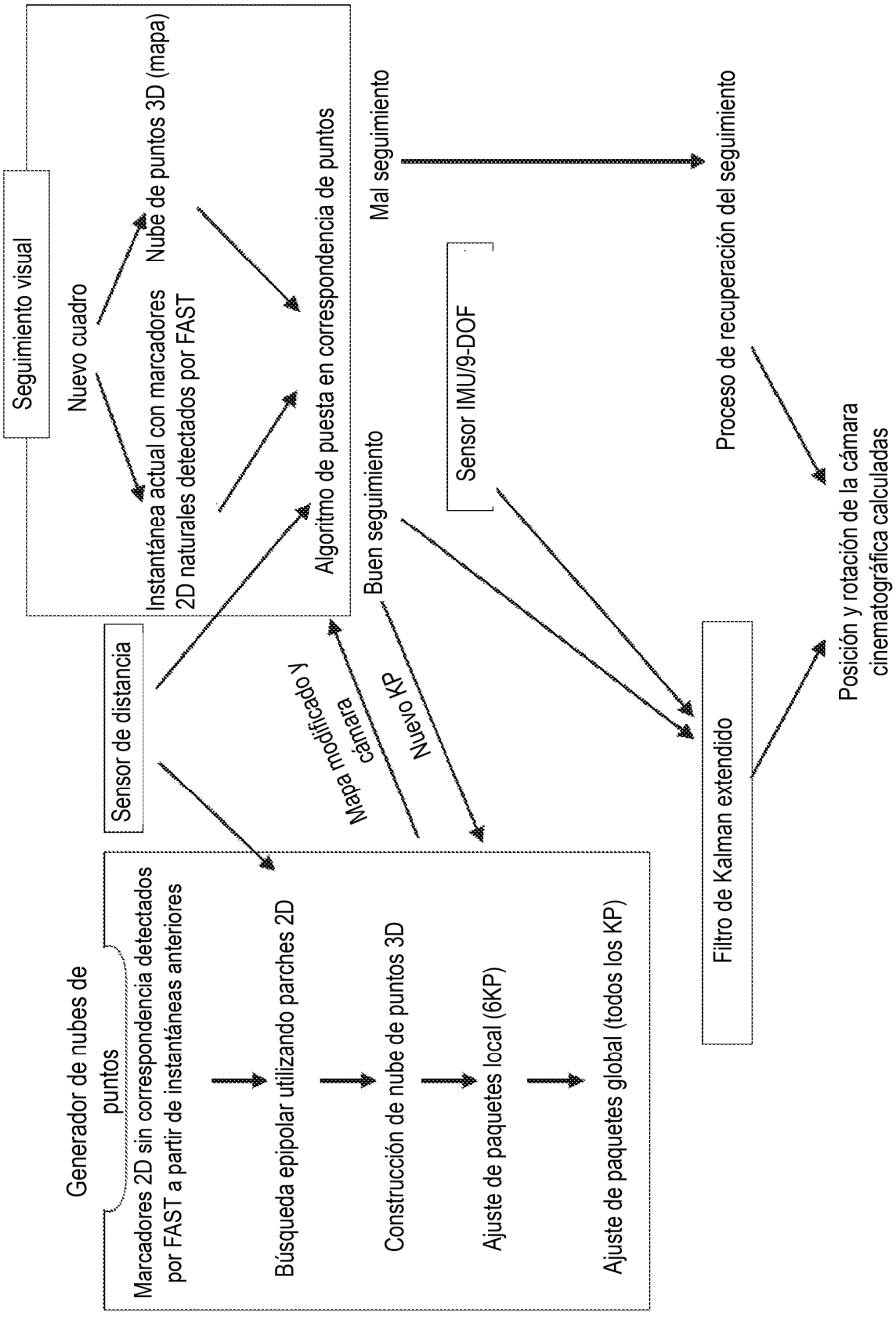
Seguimiento visual
Nuevo cuadro
Instantánea actual con marcadores 2D naturales detectados por FAST
Nube de puntos 3D (mapa)
Algoritmo de puesta en correspondencia de puntos
Mal seguimiento
Proceso de recuperación del seguimiento
Sensor IMU/9-DOF
Buen seguimiento
Sensor de distancia
Mapa modificado y cámara
Nuevo KP
Filtro de Kalman extendido
Posición y rotación de la cámara cinematográfica calculadas
Generador de nubes de puntos
Marcadores 2D sin correspondencia detectados por FAST a partir de instantáneas anteriores
Búsqueda epipolar utilizando parches 2D
Construcción de nube de puntos 3D
Ajuste de paquetes local (6KP)
Ajuste de paquetes global (todos los KP)

Figura 2

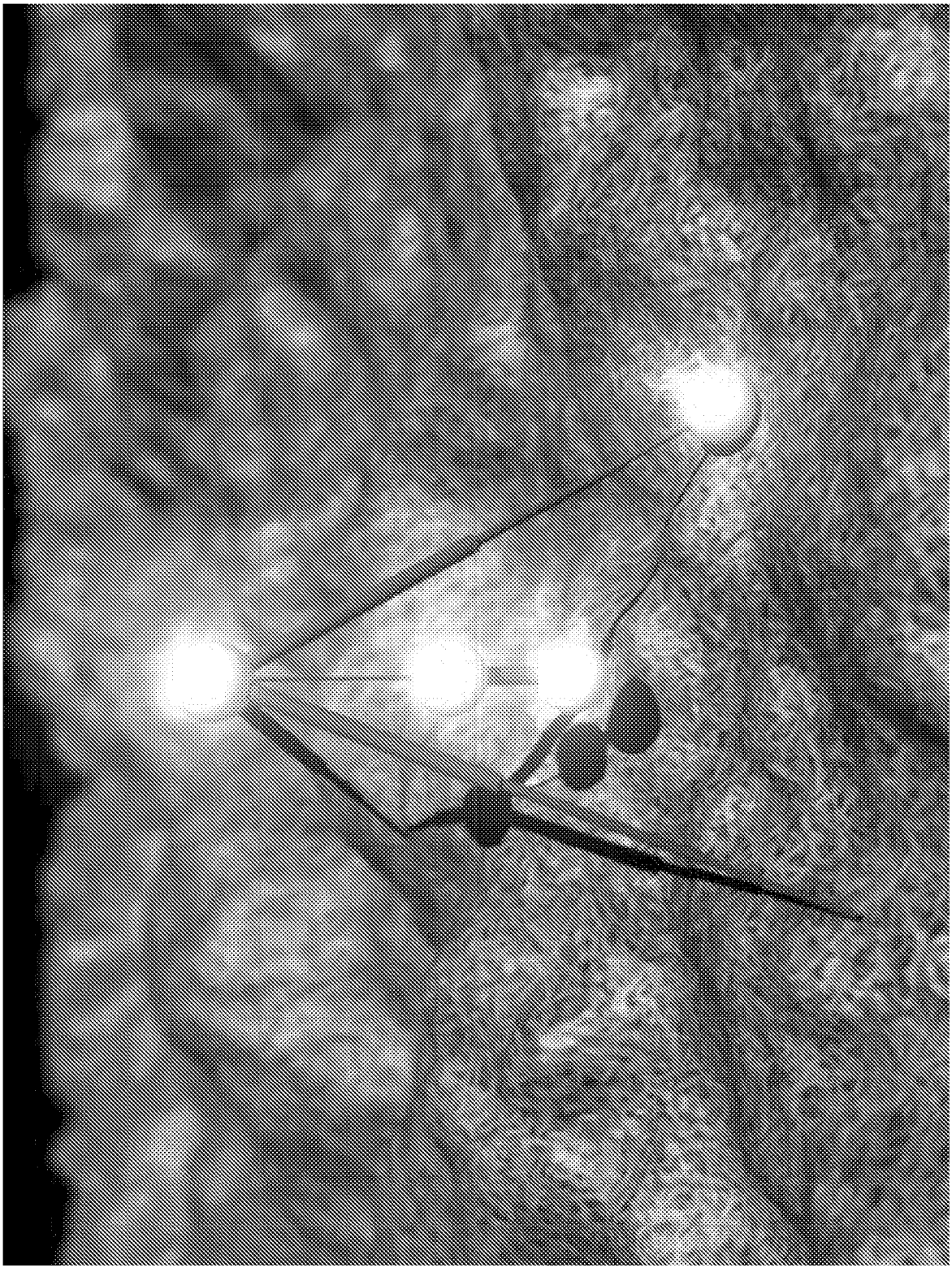

Figura 3

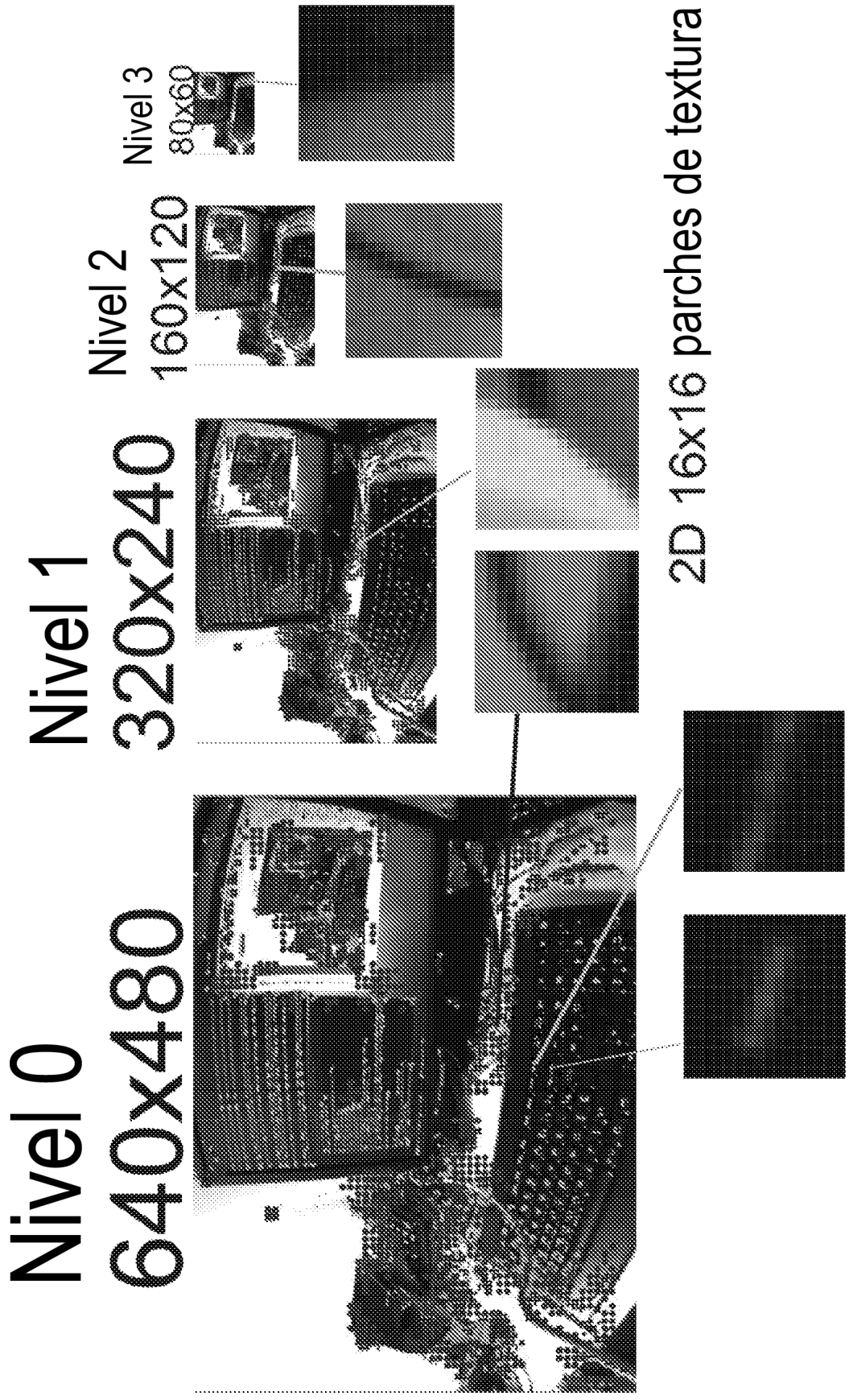
Nivel 0
640x480
Nivel 1
320x240
Nivel 2
160x120
Nivel 3
80x60
2D 16x16 parches de textura

Figura 4

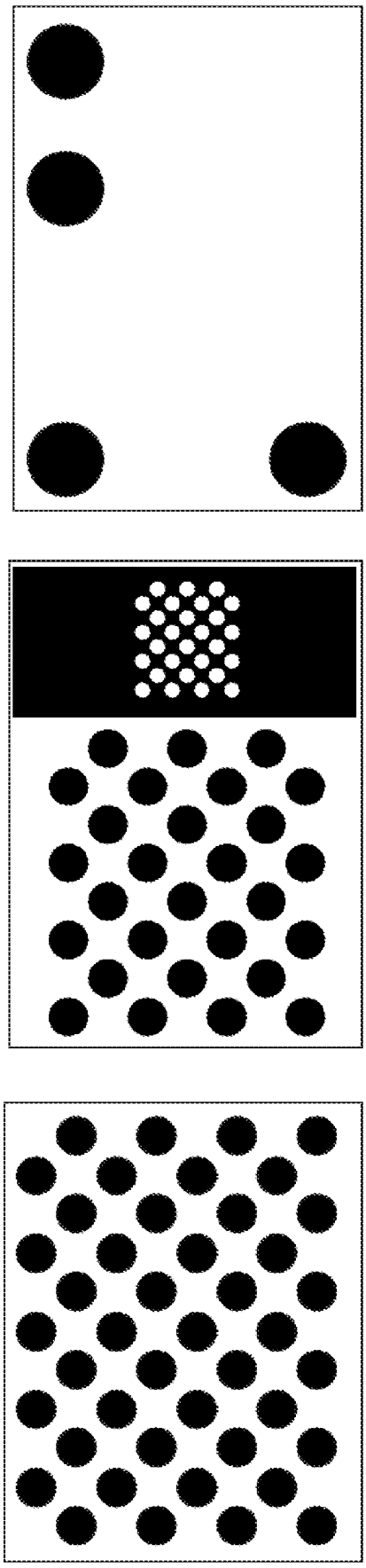

Figura 5

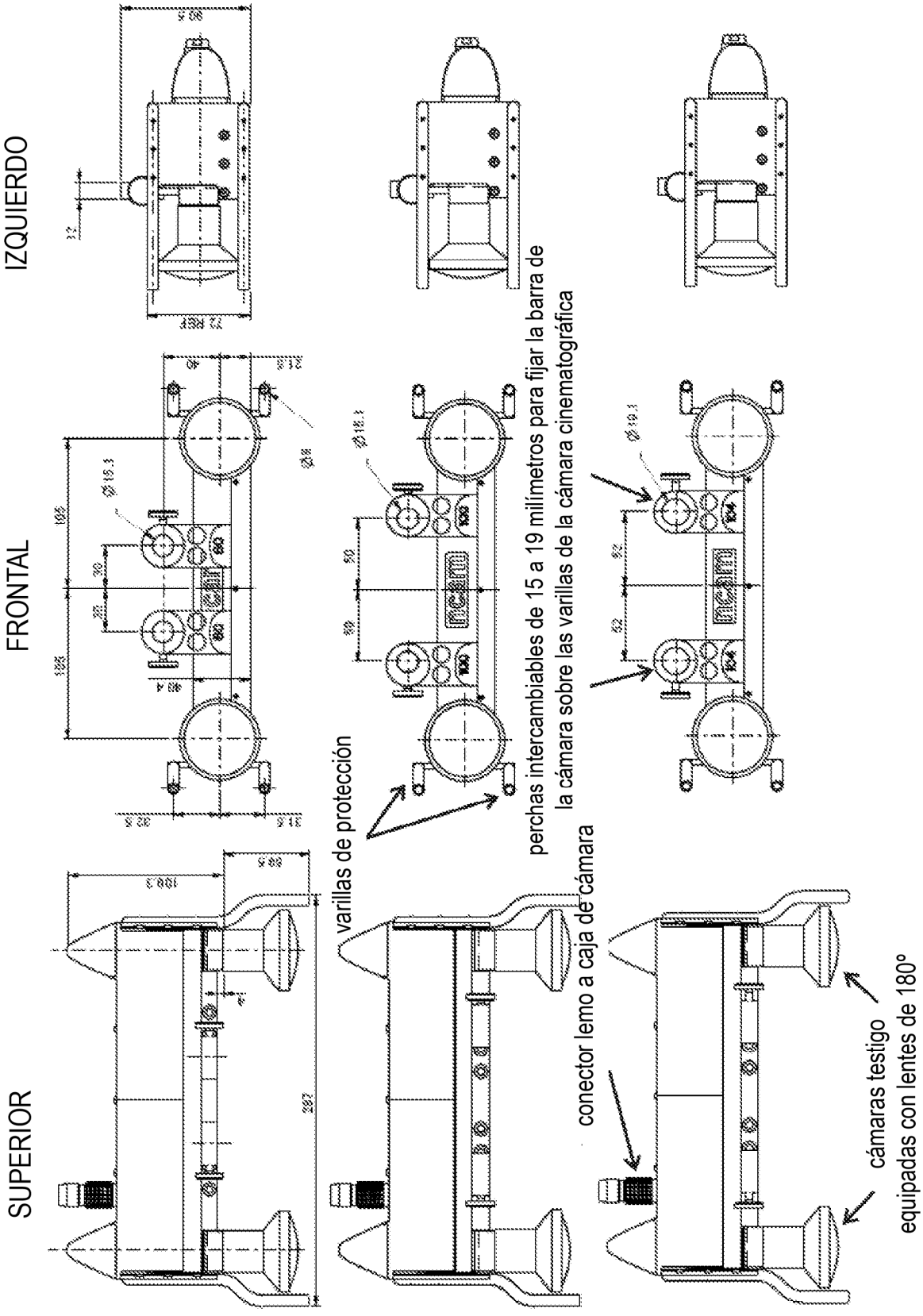
IZQUIERDO
FRONTAL
SUPERIOR
varillas de protección
perchas intercambiables de 15 a 19 milímetros para fijar la barra de la cámara sobre las varillas de la cámara cinematográfica
conector lemo a caja de cámara
cámaras testigo equipadas con lentes de 180º

Figura 6

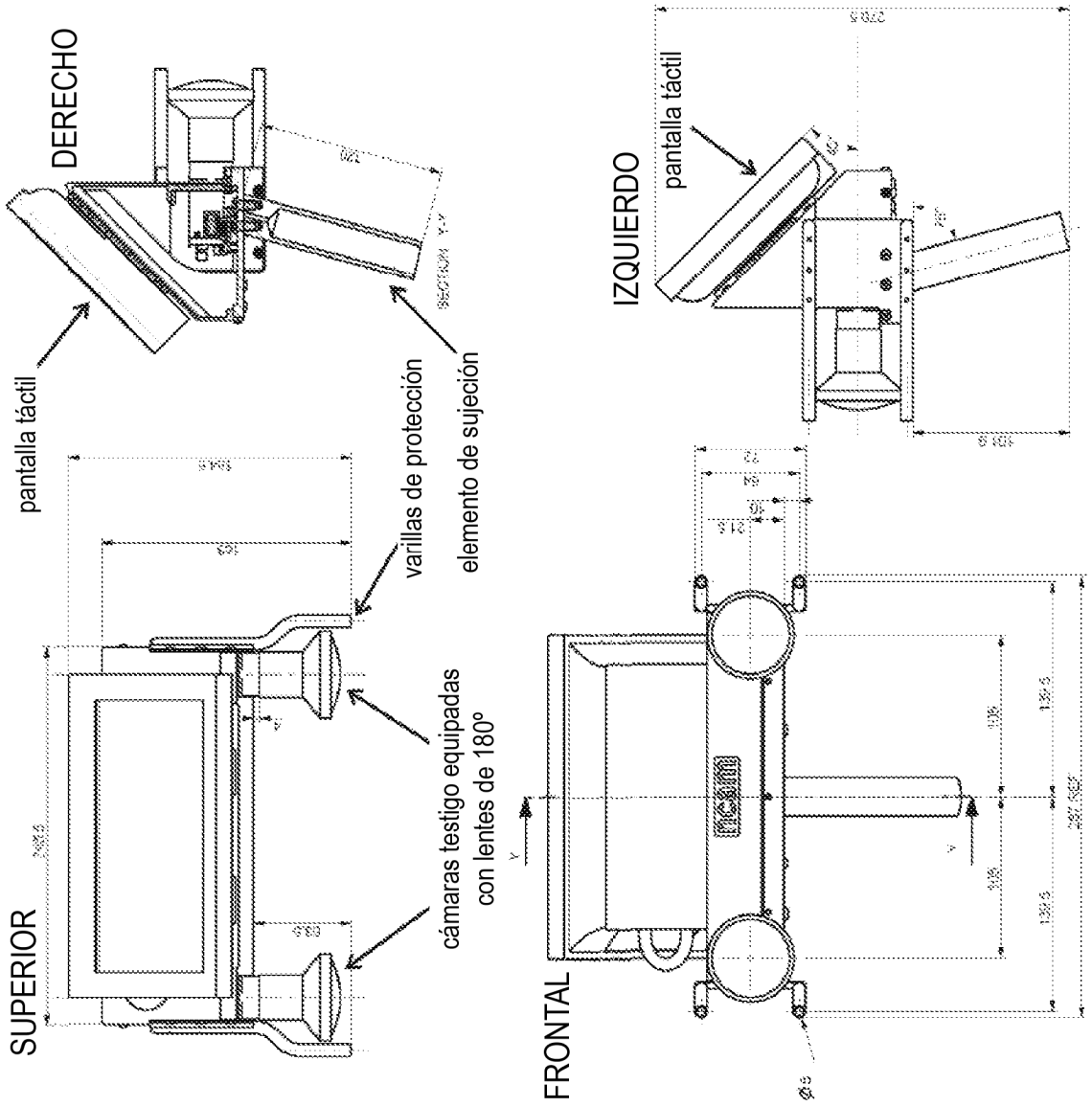
DERECHO
pantalla táctil
varillas de protección
elemento de sujeción
SUPERIOR
cámaras testigo equipadas
con lentes de 180º
IZQUIERDO
pantalla táctil
FRONTAL

Figura 7

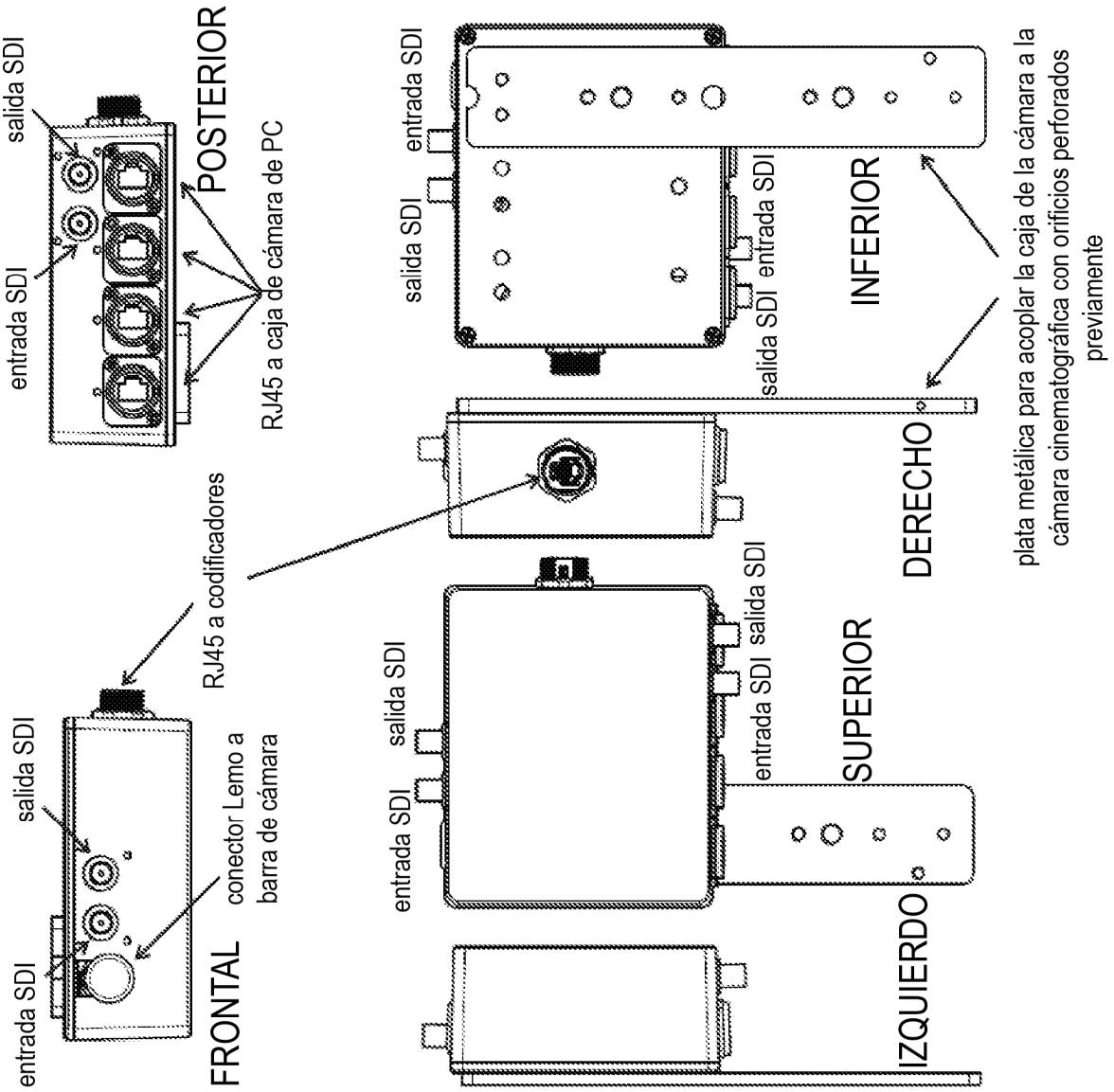
salida SDI
entrada SDI
POSTERIOR
RJ45 a caja de cámara de PC
RJ45 a codificadores
salida SDI
entrada SDI
conector Lemo a barra de cámara
FRONTAL
salida SDI
entrada SDI
INFERIOR
salida SDI
entrada SDI
DERECHO
salida SDI
entrada SDI
SUPERIOR
IZQUIERDO
plata metálica para acoplar la caja de la cámara a la cámara cinematográfica con orificios perforados previamente

Figura 8