(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 553 258**

(51) Int. CI.:

G06T 7/00 (2006.01)
G06T 15/20 (2011.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **29.04.2011 E 11405250 (9)**

(97) Fecha y número de publicación de la concesión europea: **19.08.2015 EP 2383699**

(54) Título: **Método para la estimación de una pose de un modelo de objeto articulado**

(30) Prioridad:

30.04.2010 EP 10405091

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
07.12.2015

(73) Titular/es:

LIBEROVISION AG (100.0%)
Technoparkstrasse 1
8005 Zürich, CH

(72) Inventor/es:

GERMANN, MARCEL;
WÜRMLIN STADLER, STEPHAN;
KEISER, RICHARD;
ZIEGLER, REMO;
NIEDERBERGER, CHRISTOPH;
HORNUNG, ALEXANDER y
GROSS, MARKUS

(74) Agente/Representante:

DE ELZABURU MÁRQUEZ, Alberto

DESCRIPCIÓN

Método para la estimación de una pose de un modelo de objeto articulado

Campo de la invención

La invención se refiere al campo del procesamiento de imágenes de vídeo. Se refiere a un método para la estimación de una pose de un modelo de objeto articulado.

Antecedentes de la invención

La representación basada en imágenes (IBR, por sus siglas en inglés) se introdujo en el trabajo pionero de Levoy et al. [LH96] y Gortler et al. [GGSC96]. El objetivo básico es simple: IBR se esfuerza por crear una sensación de una escena del mundo real en 3D basada en datos de imágenes capturadas. Muchos trabajos posteriores han explorado los fundamentos teóricos, por ej., la dependencia de la geometría y las imágenes relativas al requerimiento de muestreo mínimo [CCST00], o han desarrollado implementaciones más eficientes y menos restrictivas [BBM*01]. Una importante visión general desde estas obras es que un proxy geométrico suficientemente preciso de la escena reduce considerablemente el número de imágenes de entrada requeridas.

Un pequeño número de vistas de entrada es un requisito previo importante para la aplicación de IBR en entornos y aplicaciones del mundo real. Un ejemplo destacado es la transmisión de deportes, donde se observa una creciente demanda de repetición sin perspectiva de análisis de la escena. Sin embargo, para estos y la mayoría de las otras aplicaciones que no son de estudio, IBR debería trabajar sobre la base de la infraestructura existente, tales como cámaras de televisión operadas manualmente. Esto plantea la cuestión fundamental de cómo se puede generar robustamente un proxy geométrico suficientemente preciso, a pesar de las cámaras de gran línea de base, las condiciones de adquisición no controladas, la baja calidad de la textura y resolución, y la calibración de cámara inexacta. Estos problemas son aún más graves para el procesamiento de secuencias de vídeo en lugar de las imágenes fijas. En estas difíciles condiciones del mundo real, las técnicas de reconstrucción 3D clásicas tales como cascos visuales [MBR*00] o estéreo de múltiples vistas [Mid09] por lo general son inaplicables. Debido a las dificultades involucradas, uno de los enfoques actualmente más populares en este campo sigue siendo el uso de carteleras planas simples [HS06], a pesar de los artefactos visuales inevitables tales como imágenes fantasmas.

Existe una variedad de diferentes representaciones en 3D y existen métodos de representación que utilizan imágenes o vídeos como origen. La mayoría de ellos están estrechamente relacionados con determinadas configuraciones de adquisición:

Si muchas cámaras con diferentes puntos de vista están disponibles, se puede calcular el campo de luz [LH96] de la escena, el cual representa la radiación como una función del espacio. Buehler et al. [BBM*01] generaliza este enfoque para incluir proxys geométricos. El sistema EyeVision utilizado para el Super Bowl [Eye09] utiliza más de 30 cámaras controladas para las repeticiones de los eventos deportivos. El método de Reche et al. [RMD04] para los árboles requiere de 20 a 30 imágenes por objeto. Un enfoque reciente de Mahajan et al. [MHM*09] utiliza interpolación de vista basada en gradientes. En contraste con estos métodos, este método no requiere una densa colocación de la cámara.

Muchos métodos utilizan, además, datos de intervalo o estimación de profundidad en su representación. Shade et al. [SGwHS98] utiliza la información de profundidad estimada para la representación de imágenes en capas de profundidad. Waschbusch et al. [WWG07] utiliza el color y la profundidad para calcular nubes de cartelera de vídeo en 3D, que permite representaciones de alta calidad desde los puntos de vista arbitrarios. Pekelny y Gotsman [PG08] utilizan un único sensor de profundidad para la reconstrucción de la geometría de un carácter articulado. Si bien estos métodos requieren ya sea datos de profundidad o siluetas precisas y densas, esto no está disponible en las escenas no controladas con sólo unas pocas cámaras de vídeo y calibraciones débiles.

Se propusieron varios métodos para la coincidencia de siluetas de basada en plantillas para configuraciones de estudio controladas [CTMS03, VBMP08, dAST*08]. Para la representación de punto de vista libre, las imágenes de la cámara se mezclan sobre la superficie de un modelo de plantilla coincidente o deformado. Sin embargo, estos métodos requieren imágenes de origen exactas de configuraciones de estudio, mientras que las carteleras articuladas se pueden utilizar con cámaras escasamente colocadas y calibradas incorrectamente. En estas situaciones, la geometría de carteleras articuladas es mucho más robusta frente a errores que, por ej., un modelo de cuerpo de la plantilla completo en el que la textura tiene que ser proyectada con precisión sobre partes curvadas y con frecuencia delgadas (por ej., un brazo). Por otra parte, los modelos de plantilla 3D altamente teselados, que por lo general se requieren, no son eficientes para la representación de los sujetos a menudo pequeños con baja calidad y resolución de la textura. Debevec et al. [DTM96] propuso un método que utiliza la correspondencia estéreo con un modelo en 3D simple. Sin embargo, se aplica a la arquitectura y no es directamente ampliable a figuras articuladas sin líneas rectas.

Recientemente, los métodos mejorados para cascos visuales, el casco visual conservador y el casco visual dependiente de la vista, mostraron resultados prometedores [GTH*07, KSHG07]. Sin embargo, estos métodos están basados en el tallado de volumen que requiere posiciones de cámara seleccionadas para eliminar partes no del

cuerpo en todos los lados del sujeto. Este método no requiere una configuración especial de la cámara y se puede utilizar con sólo dos cámaras de origen para exhibir, por ejemplo, la perspectiva de vista de pájaro desde un punto de vista por encima de las posiciones de todas las cámaras. Un trabajo reciente de Guillemaut et al. [GKH09] aborda muchos retos para el vídeo sin el punto de vista libre en la transmisión de deportes por medio de la optimización conjunta de la segmentación de escenas y la reconstrucción de múltiples vistas. Su enfoque está dando lugar a una geometría más precisa que el casco visual, pero aún requiere un número bastante grande de cámaras colocadas en forma bastante densa (6 a 12). Se compara nuestro método con sus resultados de reconstrucción en la Sección 7.

Un método simple para configuraciones no controladas es la mezcla entre carteleras [HS06] por sujeto y cámara. Sin embargo, este tipo de carteleras estándar sufre de artefactos de efecto fantasma y no conservan la pose del cuerpo en 3D de una persona debido a su representación plana. La idea de subdividir el cuerpo en partes representadas por las carteleras es similar en espíritu a la representación de nubes de cartelera [DDS03, BCF*05], microfacetas [YSK*02, GM03] o subdivisión en impostores [ABB*07, ABT99]. Sin embargo, estos métodos no son adecuados para la aplicación diana, ya que dependen de escenas controladas, datos de profundidad o incluso modelos dados. Lee et al. [LBDGG05] propuso un método para extraer las carteleras del flujo óptico. Sin embargo, utilizaron imágenes de entrada generadas de los modelos sintéticos de alta calidad.

También se relacionado con este enfoque el gran cuerpo de trabajo en la estimación de pose de humanos y la segmentación del cuerpo a partir de imágenes. Aquí, sólo se pueden discutir las obras más relevantes. Efros et al. [EBMM03] ha presentado un enfoque interesante para el reconocimiento de la acción humana a una distancia con aplicaciones para la estimación de poses. Su método requiere una estimación del flujo de la escena óptica que a menudo es difícil de estimar en entornos dinámicos y no controlados. Agarwal y Triggs [AT06], Jaeggli et al. [JKMG07], y Gammeter et al. [GEJ*08] presentan métodos basados en el aprendizaje para la estimación y el seguimiento de poses humanas en 3D. Sin embargo, las poses calculadas a menudo son sólo aproximaciones, mientras que se requieren estimaciones precisas de posiciones de la articulación del sujeto. Por otra parte, por lo general hay que lidiar con una calidad y resolución de imagen mucho más baja en este ajuste. Por lo tanto, se presenta un enfoque semiautomático, basado en datos, dado que una cantidad limitada de la interacción del usuario es aceptable en muchos escenarios de aplicación si conduce a una mejora considerable en la calidad.

Y. Kameda et al. "Tri Dimensional Pose Estimation of an Articulated Object from its Silhouette Image", Conferencia Asiática sobre la Visión por Computador, 23 de noviembre de 1993, páginas 612 a 615, describe un método para la estimación de la pose de un modelo de objetos en 3D que corresponde a un objeto del mundo real articulado y que tiene varias partes. Con este fin, el modelo de la silueta coincide parte por parte con la silueta del objeto en la imagen de origen, en el que una lista de siluetas candidatas se determina para cada parte del modelo articulado, y para cada parte se determina la silueta que mejor se ajusta, basado en un error de coincidencia.

Descripción de la invención

Por lo tanto, un objetivo de la invención es crear un método para la estimación de una pose de un modelo de objeto articulado del tipo mencionado inicialmente, que supere las desventajas mencionadas con anterioridad.

El método para la estimación de una pose de un **modelo de objeto articulado**, en el que el modelo de objeto articulado es un modelo en 3D basado en ordenador de un **objeto del mundo real** observado por una o más **cámaras de origen,** y el modelo de objeto articulado representa una pluralidad de **articulaciones** y de **enlaces** que enlazan las articulaciones, y en el que la **pose** del modelo de objeto articulado está definida por la localización espacial de las articulaciones, comprende los pasos de

- obtener por lo menos una **imagen de origen** a partir de una corriente de vídeo que comprende una vista del objeto del mundo real grabado por una cámara de origen (en el que la corriente de vídeo es en vivo o de una grabación);
- procesar por lo menos una imagen de origen para extraer un **segmento de imagen de origen** correspondiente que comprende la vista del objeto del mundo real separado del fondo de la imagen;
- mantener, en una base de datos en forma legible por ordenador, un conjunto de **siluetas de referencia**, cada silueta de referencia está asociada con un modelo de objeto articulado y con una **pose de referencia** particular de este modelo de objeto articulado;
- comparar el por lo menos un segmento de imagen de origen de las siluetas de referencia y seleccionar un número predeterminado de siluetas de referencia, teniendo en cuenta, para cada silueta de referencia,
    - un **error de coincidencia** que indica cuán estrechamente la silueta de referencia coincide con el segmento de imagen de origen y
    - un **error de consistencia** que indica cuánto la pose de referencia es consistente con la pose del mismo objeto del mundo real de acuerdo con lo estimado a partir de por lo menos una de imágenes de origen precedentes y siguientes de la corriente de vídeo;
- recuperar (de la base de datos) las poses de referencia de los modelos de objeto articulado asociados con las siluetas de referencia seleccionadas; y
- calcular una estimación de la pose del modelo de objeto articulado de las poses de referencia de las siluetas de

referencia seleccionadas. Esta pose es la pose en 2D del modelo de objeto articulado en la imagen de origen, por lo que para cada una de las imágenes de origen, se estima una pose separada en 2D.

Un enlace que enlaza dos articulaciones puede ser representado por una sección de línea recta entre las dos articulaciones, es decir, el enlace puede ser denominado un enlace lineal, sin subdivisiones o articulaciones intermedias. Un enlace puede estar asociado con una longitud de enlace, que posiblemente limita o define una distancia entre las articulaciones. Dependiendo del método utilizado para identificar la pose del modelo, una longitud de enlace se supone como constante, lo que limita el ajuste de las ubicaciones de las articulaciones, y/o la longitud del enlace se puede ajustar de acuerdo con una estimación de las posiciones de la articulación. Un enlace puede tener una relación geométrica por lo menos parcialmente restringida con una superficie de proyección asociada.

El problema de la estimación de una pose sobre la base de unas pocas o sólo una imagen de entrada está mal planteado debido a las ambigüedades y la información faltante. En la presente invención, se hace uso de una base de datos como una previa para superar este problema mal planteado.

El paso de procesamiento de por lo menos una imagen de origen para extraer un segmento de imagen de origen correspondiente con preferencia comprende por lo menos el paso de segmentar la imagen de origen. Los métodos de segmentación de imágenes, tales como son muy conocidos y pueden ser adaptados para su uso en la presente invención. En una realización preferida de la invención, el paso de procesamiento comprende un paso adicional tal como un paso de **compensación de movimiento**. Este paso de compensación de movimiento puede ser ejecutado antes del paso de segmentación (en la imagen no segmentada) o después del paso de segmentación (en los segmentos de la imagen). El paso de compensación de movimiento compensa, de una manera conocida, por ej., el movimiento de la cámara de origen y/o el objeto del mundo real.

La compensación de movimiento en esta etapa se puede utilizar para dar una estimación inicial de parte del cuerpo u otros segmentos en un marco particular de interés, teniendo en cuenta una parte del cuerpo o de otra segmentación de un marco anterior y/o posterior. La segmentación dada del último marco (es decir, anterior o posterior) de este último se mueve de acuerdo con la compensación de movimiento o flujo óptico entre el último marco y el marco de interés, y se utiliza como la estimación inicial para la segmentación del marco de interés.

Los segmentos de imagen de origen y las siluetas de referencia representan ambas las imágenes parciales; que se denominan "segmentos" y "siluetas", ya que se generan en diferentes contextos. Ellos se pueden representar conceptualmente y/o en una forma legible por ordenador en una variedad de diferentes maneras, tales como, por ejemplo

- una mancha de píxeles; o
- un esquema, por ej., en una representación basada en píxeles o vectores, en forma opcional con un color de relleno o modelo de color que caracteriza el área interior.

En una realización preferida de la invención, se toman en cuenta el error de coincidencia y el error de consistencia como una suma ponderada de ambos. El error de coincidencia puede estar basado en la información de imagen (píxel) de la silueta de referencia y el segmento de imagen de origen, y/o en su respectivo flujo óptico. En este último caso, un flujo óptico también se almacena como parte de la silueta de referencia. Esto hace que sea posible diferenciar entre las imágenes de objetos que tienen la misma pose pero que se mueven de una manera diferente. El error de consistencia, también, de acuerdo con una realización adicional preferida de la invención, tiene en cuenta el flujo óptico y/o la compensación de movimiento.

Por ejemplo, dado un marco de imagen actual y una imagen anterior, se calculan los parámetros de compensación de movimiento que describen la diferencia entre estas imágenes (por lo menos en el área relevante de las imágenes). A continuación, la pose que se estimó en el marco de imagen anterior se mueve de acuerdo con los parámetros de compensación de movimiento. Esta es una estimación de la pose en el marco actual. Ahora bien, esta estimación de la pose (en lugar de la pose del marco de imagen precedente) se compara con la pose de referencia cuando se determina el error de consistencia.

En otra realización preferida de la invención, la estimación de la pose se determina moviendo la pose del marco de imagen precedente de acuerdo con un flujo óptico. Este flujo óptico puede ser el de la imagen anterior, o de la imagen actual, o de la imagen de referencia, o de una combinación de los mismos, tal como un promedio (ponderado).

Los ejemplos anteriores y otros se refieren sobre todo a un marco de imagen anterior. Sin embargo, los mismos métodos se pueden aplicar teniendo en cuenta más de un marco de imagen, y también luego de marcos de imagen (o posteriores, o futuros). La información de todos estos marcos, ya sea relacionada con compensación de movimiento o flujo óptico, se puede combinar por medio de, por ej., promedio ponderado, en particular, por medio del pesaje de los marcos cercanos más que los marcos que están más lejos en el tiempo desde el marco actual.

En una realización preferida de la invención, el paso de comparar el por lo menos un segmento de imagen de origen con las siluetas de referencia comprende los pasos de, para cada silueta de referencia con la que se compara el segmento de imagen de origen:

- determinar una **transformación proyectiva** que mapea el segmento de imagen de origen en la silueta de referencia; y
- calcular el **error de coincidencia** ya sea como proporcional al tamaño relativo del área de la imagen en la que el segmento de imagen de origen mapeado y la silueta de referencia no se superponen, o como una medida de la distancia entre las líneas del segmento de imagen de origen escalado y de la silueta de referencia, con el error de coincidencia, en forma opcional, también dependiente de parámetros de la transformación proyectiva;
- y utilizar este error de coincidencia como una medida de cuán estrechamente coinciden el segmento de imagen de origen y la silueta de referencia.

Los parámetros de la transformación proyectiva en principio comprenden información acerca de hasta qué punto el segmento de imagen de origen y la silueta de referencia deben estar distorsionadas con el fin de que coincida con, con la condición de que una coincidencia de este tipo se puede conseguir por medio de la transformación proyectiva. Por esta razón, uno o más parámetros de la transformación proyectiva, o una combinación de los mismos, se incorpora con preferencia en el cálculo del error de coincidencia.

En una realización preferida de la invención, la determinación de la **transformación proyectiva** se efectúa por medio del **escalamiento** del segmento de imagen de origen para que sea del mismo tamaño que la silueta de referencia. Los parámetros de escalamiento, que no necesitan mantener la proporción del segmento de imagen, corresponden a los de una transformación proyectiva.

La distancia entre los contornos del segmento de imagen de origen escalado y de la silueta de referencia se puede calcular por medio de una métrica basada en línea, tal como la distancia de Hausdorff.

En una realización preferida de la invención,

- el paso de escalamiento mencionado con anterioridad se logra por medio del re-muestreo de ya sea el segmento de imagen de origen o las siluetas de referencia o ambas para tener cuadros delimitados del mismo tamaño de píxel, y
- tanto el segmento de imagen de origen como la silueta de referencia están representadas por imágenes binarias que tienen la misma dimensión de píxel, y calcular el valor de error por medio del recuento del número de píxeles correspondientes desde el segmento de imagen de origen y la silueta de referencia que difieren en valor.

En una realización preferida de la invención, el paso de comparar el por lo menos un segmento de imagen de origen con las siluetas de referencia comprende los pasos de, para cada silueta de referencia con la que se compara el segmento de imagen de origen:

- recuperar la pose del mismo objeto del mundo real estimado a partir de una imagen de origen anterior de la corriente de vídeo;
- calcular el **error de consistencia** como proporcional a la diferencia entre esta estimación de la pose precedente y la pose de referencia de la silueta de referencia, y utilizar este error de consistencia como una medida de la consistencia con la imagen de origen anterior

En una realización preferida de la invención, el paso de calcular una estimación de la pose del modelo de objeto articulado de las poses de referencia de las siluetas de referencia seleccionadas comprende

- repetir los pasos anteriores para una o más imágenes de origen adicionales de una o más corrientes de vídeo de cámaras de origen adicionales, cada imagen de origen además comprende una vista del mismo objeto del mundo real que ha sido grabado en el mismo tiempo, pero desde un punto de vista diferente, de ese modo se obtiene para cada imagen de origen adicional, un número predeterminado de siluetas de referencia seleccionadas y poses de referencia seleccionadas asociadas;
- llevar a cabo una optimización para seleccionar para cada imagen de origen una **pose de referencia más plausible**, por medio del cálculo para cada combinación de poses de referencia seleccionadas para las diferentes imágenes de origen de una **medida de consistencia de las articulaciones totales** por medio de,
    - la proyección de las articulaciones de las poses de referencia seleccionadas de esta combinación en el espacio 3D, la estimación de una posición en 3D de las articulaciones y el cálculo, para cada articulación, de una **medida de consistencia de las articulaciones** que expresa cuán estrechamente la posición de la articulación en 3D estimada coincide con la proyección de la articulación de las poses de referencia seleccionadas;
    - la combinación de las medidas de consistencia de las articulaciones de todas las articulaciones para obtener la **medida de consistencia de las articulaciones totales**;

- seleccionar la combinación de las poses de referencia para las diferentes imágenes de origen que optimiza la medida de consistencia de las articulaciones totales, lo que de ese modo determina una pose de referencia óptima para cada imagen de origen.

Además de determinar una pose de referencia óptima para cada imagen de origen, el procedimiento anterior

también proporciona una estimación de la posición en 3D de cada articulación.

En una realización preferida de la invención, el paso de llevar a cabo una optimización además comprende el paso de variar y optimizar un desplazamiento en 2D de cada silueta en el plano de su imagen de origen asociada con el fin de corregir los errores de calibración de la cámara de origen.

En una realización preferida de la invención, el paso adicional de exhibir, en un dispositivo de visualización, por lo menos una imagen de origen con posiciones de la articulación estimadas superpuestas sobre la imagen de origen y aceptar una entrada de usuario para modificar en forma interactiva una o más posiciones de la articulación.

Por lo general, los modelos de objeto articulado asociados con las poses de referencia tienen la misma topología de enlace que el modelo de objeto articulado del objeto del mundo real. Cada articulación de uno de estos modelos de objeto articulado coincide únicamente con una articulación de los otros modelos de objeto articulado.

De acuerdo con otro aspecto de la invención, se proporciona un método para la estimación de una pose de un **modelo de objeto articulado**, en el que, con el fin de determinar una pose en 3D que coincide con una pose en 2D dada asociada con una imagen de origen, se llevan a cabo los siguientes pasos:

- calcular, a partir de la pose en 2D dada, una pose en 3D aproximada que comprende posiciones de la articulación aproximadas que coinciden aproximadamente con las posiciones de las articulaciones correspondientes de la pose en 2D cuando se proyecta en el plano de la imagen de la imagen de origen asociada con la pose en 2D;
- modificar la pose en 3D aproximada para coincidir exactamente con la pose en 2D por medio de, para cada articulación, el movimiento de la posición de la articulación desde la posición de la articulación aproximada a una posición definida por la intersección de un rayo que pasa desde la cámara a través de la posición de la articulación en la imagen de origen con un plano paralelo al plano de imagen de la imagen de origen y que pasa a través de la posición de la articulación aproximada.

Estos dos pasos se pueden llevar a cabo solos, con el fin de mejorar una pose en 2D dada, a través de una pose en 3D estimada, en la que la pose en 2D dada se determina por medio de un método de estimación anterior arbitrario, o en combinación con las realizaciones preferidas de la invención que se han descrito con anterioridad.

De acuerdo con otro aspecto de la invención, se proporciona un método para la estimación de una pose de un **modelo de objeto articulado**, en el que el modelo de objeto articulado es un modelo en 3D basado en ordenador de un **objeto del mundo real** observado por dos o más **cámaras de origen**, y el modelo de objeto articulado representa una pluralidad de articulaciones y de **enlaces** que enlazan las articulaciones, y en donde la **pose** del modelo de objeto articulado está definida por la localización espacial de las articulaciones, denominada posiciones de la articulación en 3D, el método comprende los pasos de

- determinar una estimación inicial de la pose en 3D, es decir, las posiciones de la articulación en 3D del modelo de objeto articulado;
- asociar cada enlace con una o más **superficies de proyección**, en el que las superficies de proyección son superficies definidas en el modelo en 3D, y la posición y la orientación de cada superficie de proyección están determinadas por la posición y la orientación del enlace asociado;
- adaptar en forma iterativa las posiciones de la articulación en 3D por medio de, para cada articulación,
    - el cálculo de un **puntaje de posición** asignado a su posición de la articulación en 3D, el puntaje de posición es una medida del grado en el que los segmentos de imagen desde las diferentes cámaras de origen, cuando se proyectan sobre las superficies de proyección de enlaces adyacentes a la articulación, que son consistentes entre sí;
    - la variación de la posición de la articulación en 3D de la articulación hasta que se logre un puntaje de posición óptimo;

- repetir el paso de adaptación en forma iterativa de las posiciones de la articulación en 3D para todas las articulaciones durante un número predeterminado de veces o hasta que los puntajes de posición converjan.

Al repetir la adaptación iterativa para todas las articulaciones, las posiciones convergen después de algunos pasajes por todas las articulaciones. Se establecen los puntajes de posición que son convergentes, por ejemplo, cuando la mejora de los puntajes de posición cae por debajo de un límite predeterminado.

En una realización preferida de la invención, la estimación inicial de la pose en 3D se determina de acuerdo con uno o más de los aspectos anteriores de la invención, por ejemplo, cuando se determina una pose de referencia óptima para cada imagen de origen por la estimación de la pose en 2D, o después de mejorar por medio de patrón de montaje a una pose en 3D estimada.

En una realización preferida de la invención, el paso de variar la posición de la articulación en 3D de las articulaciones se lleva a cabo por medio de la variación de las posiciones de la articulación en 3D sujetas a limitaciones antropométricas, por lo menos una de las limitaciones antropométricas es:

- la articulación está en o por encima del suelo;
- las longitudes de los enlaces topológicamente simétricos no difieren más de 10%;
- las longitudes de los enlaces están dentro de los estándares antropométricos;
- las distancias entre las articulaciones que no están conectadas por un enlace están dentro de los estándares antropométricos.

En una realización preferida de la invención, las superficies de proyección, para cada enlace, comprenden un **ventilador de carteleras**, cada **cartelera** está asociada con una cámara de origen, y cada cartelera es una superficie plana atravesada por su enlace asociado y un vector que es normal tanto a este enlace como a una línea que conecta un punto del enlace a la cámara de origen. En otras palabras, cada cartelera es una superficie de proyección y está asociada con un enlace y con una cámara de origen.

En una realización preferida de la invención, el **puntaje de posición** de una posición de la articulación en 3D de una articulación se calcula por medio de los pasos de, para cada enlace adyacente a la articulación,

- proyectar las imágenes de las diferentes cámaras de origen sobre las superficies de proyección asociadas del enlace y de allí a una imagen virtual, de acuerdo con lo observado por una cámara virtual;
- para un área (o para todos los píxeles) **en la imagen virtual** que corresponden a la proyección de estas superficies de proyección en la imagen virtual, calcular un **puntaje de posición parcial** para este enlace de acuerdo con el grado en el que los segmentos de imagen de las distintas cámaras de origen se solapan y tienen un color similar;
- combinar (por ej., por medio de adición) los puntajes de posición parciales para obtener el puntaje de posición.

En otras palabras, la consistencia de las proyecciones de los segmentos de imagen de las diferentes cámaras de origen se evalúa en una vista correspondiente a la de la cámara virtual. Dado que la optimización se basa en funciones diana que están definidas en la imagen virtual, no se determina ningún parámetro a la larga innecesario, y la eficiencia global es alta.

En una realización preferida de la invención, el cálculo y la combinación del puntaje de posición parcial comprende los pasos de

- calcular el puntaje de posición parcial para cada par de cámaras de origen que contribuyen a la imagen virtual;
- combinar estos puntajes de posición parciales por medio de la adición de los mismos, y por medio del pesaje de cada puntaje de posición parcial de acuerdo con el ángulo entre las direcciones de visión del par asociado de cámaras de origen.

De acuerdo con otro aspecto, se proporciona un método para la **representación** de una imagen virtual, de acuerdo con lo observado desde una cámara virtual, dado un modelo de objeto articulado, en el que el modelo de objeto articulado es un modelo en 3D basado en ordenador de un **objeto del mundo real** observado por **dos** o más **cámaras de origen**, y el modelo de objeto articulado representa una pluralidad de **articulaciones** y de **enlaces** que enlazan las articulaciones, y en donde la **pose** del modelo de objeto articulado está definida por la localización espacial de las articulaciones, el método comprende los pasos de

- determinar una estimación de la pose en 3D, es decir, las posiciones de la articulación en 3D del modelo de objeto articulado;
- asociar cada enlace con una o más **superficies de proyección**, en las que las superficies de proyección son superficies definidas en el modelo en 3D, y la posición y la orientación de cada superficie de proyección están determinadas por la posición y la orientación del enlace asociado;
- en el que las superficies de proyección, para cada enlace, comprenden un ventilador de carteleras, cada cartelera está asociada con una cámara de origen, y cada cartelera es una superficie plana atravesada por su enlace asociado y un vector que es normal tanto a este enlace como a una línea que conecta un punto del enlace a la cámara de origen;
- para cada cámara de origen, proyectar segmentos de la imagen de origen asociada a la cartelera asociada, lo que crea las imágenes de cartelera;
- para cada enlace, proyectar las imágenes de cartelera en la imagen virtual y mezclar las imágenes de cartelera para formar una parte correspondiente de la imagen virtual.

Las imágenes de cartelera se mezclan, es decir, múltiples carteleras por un enlace no se ocluyen el uno al otro). Sin embargo, la oclusión puede ocurrir entre enlaces separados, es decir, las partes del cuerpo separadas.

De acuerdo con otro aspecto de la invención, se proporciona un método para la determinación de una segmentación de un segmento de imagen de origen, el método comprende los pasos de

- obtener por lo menos una **imagen de origen** a partir de una **corriente de vídeo** que comprende una vista de un objeto del mundo real grabado por una cámara de origen;
- procesar la por lo menos una imagen de origen para extraer un **segmento de imagen de origen** correspondiente que comprende la vista del objeto del mundo real separada del fondo de la imagen;

- mantener, en una base de datos en forma legible por ordenador, un conjunto de **siluetas de referencia**, cada silueta de referencia está asociada con una **segmentación de referencia**, la segmentación de referencia define sub-segmentos de la silueta de referencia, cada sub-segmento se le asigna una etiqueta única;
- determinar una silueta de referencia coincidente que se asemeja más estrechamente al segmento de imagen de origen y recuperar la segmentación de referencia de la silueta de referencia;
- para cada sub-segmento, superponer, tanto una versión engrosada como afinada del sub-segmento sobre el segmento de imagen de origen y etiquetar los píxeles de la imagen de origen que se encuentran tanto en la versión engrosada como afinada con la etiqueta del sub-segmento;
- etiquetar todos los píxeles restantes del segmento de imagen de origen como inseguros;
- para cada sub-segmento, determinar un modelo de color que sea representativo del color de los píxeles etiquetados con la etiqueta del sub-segmento;
- etiquetar los píxeles inseguros de acuerdo con el modelo de color, por medio de la asignación de cada píxel inseguro a un sub-segmento cuyo modelo de color se ajusta más estrechamente al color del píxel inseguro.

Los pasos anteriores para la segmentación de un segmento de imagen de origen se pueden llevar a cabo solos, con el fin de mejorar una pose en 2D dada, a través de una pose en 3D estimada, en el que la pose en 2D es determinada por un método de estimación anterior arbitraria, o en combinación con las realizaciones preferidas de la invención que se ha descrito con anterioridad.

Un modelo de color es un modelo probabilístico para la distribución de color de los píxeles. Al tener un modelo de este tipo para cada parte del cuerpo, se permite calcular las probabilidades para un nuevo píxel para estimar a qué parte del cuerpo que pertenece. Por ejemplo, un modelo de color puede ser un Modelo de Mezcla Gaussiana.

En una realización preferida de la invención, la asignación de píxeles inseguros no tiene en cuenta si el sub-segmento al que se le asigna un píxel permanece cerca del píxel. Esto permite asignar píxeles en forma correcta, incluso si no hay píxeles confidentes del sub-segmento visibles en absoluto.

En una realización alternativa, la asignación tiene en cuenta la ubicación de un píxel inseguro, y, en caso de que el color del píxel coincida con el modelo de color de más de un sub-segmento, lo asigna al sub-segmento que se encuentra más cercano al píxel.

Una observación general con respecto a la presente invención es que la pose en 3D y la forma de un carácter puede ser bien capturada por carteleras articuladas, es decir, por una subdivisión articulada del cuerpo en primitivas geométricas simples. En lugar de depender en la información de la silueta exacta para el cálculo de los cascos visuales o correspondencias en estéreo, esta representación requiere una estimación de la pose en 2D de un sujeto en las vistas de entrada. Esto se puede lograr de una manera sencilla y muy eficiente por medio de un algoritmo semi-automático basado en datos. Desde esta pose entonces es posible construir un modelo articulado de cartelera 3D, que es una representación fiel de la geometría de los sujetos y que permite un vídeo de punto de vista libre fotorrealista. Los diferentes aspectos son

- Carteleras articuladas, una novela de representación de forma de vídeo de punto de vista libre de los caracteres humanos en condiciones de adquisición desafiantes.
- Estimación de pose en 2D semi-automática basada en datos y basada en siluetas aproximadas.
- Segmentación automática de partes del cuerpo por medio de ajuste de plantilla 3D y aprendizaje de modelos de color.
- Generación del modelo de cartelera articulado por optimización de pose en 3D y corrección de costura para la consistencia de textura óptima.
- Mezcla basada en GPU de píxeles precisos y representación para la síntesis de vista realista y eficiente.

Las solicitudes de las carteleras articuladas son vídeos de vista múltiple de escenas dinámicas con los humanos capturados en ambientes no controlados. Incluso desde tan sólo dos imágenes de cámara de TV convencionales, una escena se puede representar en una alta calidad a partir de puntos de vista virtuales donde ninguna cámara de origen estaba grabando.

Por medio de la combinación de los diferentes aspectos del método descritos con anterioridad, se puede implementar el siguiente flujo de trabajo: La idea básica es la aproximación de la forma articulada en 3D del cuerpo humano por el uso de una subdivisión en carteleras texturizadas a lo largo de la estructura de esqueleto. Las carteleras se agrupan a los ventiladores de manera tal que cada hueso del esqueleto contenga una cartelera por cámara de origen. En primer lugar, para cada vista de entrada, se utiliza una estimación de pose en 2D basada siluetas de imagen, datos de captura de movimiento, y la consistencia temporal de vídeo para crear una máscara de segmentación para cada parte del cuerpo. Luego, a partir de las poses en 2D y la segmentación, el modelo de cartelera articulado real se construye por medio de una optimización de articulación en 3D y la compensación por los errores de calibración de la cámara. El método de representación aplicado posteriormente combina las contribuciones de textura de cada cartelera y con preferencia cuenta con una corrección de costura adaptativa para eliminar discontinuidades visibles entre las texturas de carteleras adyacentes. El uso de carteleras articuladas no sólo minimiza los artefactos del efecto fantasma conocidos de la representación de cartelera convencional, sino que también alivia las restricciones a la configuración y la sensibilidad a los errores de representaciones en 3D más

complejas y técnicas de reconstrucción de vista múltiple. Los resultados demuestran la flexibilidad y la solidez del enfoque con vídeos de punto de vista libre de alta calidad generados a partir de imágenes de difusión de entornos no controlados desafiantes.

Observación general con respecto a la terminología: la expresión "A está asociada con B" significa que hay una asignación, o, en otras palabras, una relación entre A y B. La relación puede ser una relación de uno a uno, de uno a muchos o de muchos a muchos.

Un producto de programa de ordenador para la estimación de una pose de un modelo de objeto articulado de acuerdo con la invención se puede cargar en una memoria interna de un ordenador digital o un sistema de ordenador que comprende una memoria de ordenador y una unidad de procesamiento acoplada a la memoria de ordenador, y comprende un código de programa de ordenador significa, es decir, instrucciones legibles por ordenador, para hacer, cuando dicho medio de código de programa de ordenador es cargado en el ordenador, el ordenador ejecuta el método de acuerdo con la invención. En una realización preferida de la invención, el producto de programa de ordenador comprende un medio de almacenamiento legible por ordenador, que tiene el medio de código de programa de ordenador grabado en el mismo.

Otras realizaciones preferidas son evidentes a partir de las reivindicaciones dependientes de la patente.

Breve descripción de los dibujos

El objetivo de la invención se explicará con más detalle en el siguiente texto con referencia a realizaciones representativas preferidas que se ilustran en los dibujos adjuntos, en los cuales:

Figura 1 representa en forma esquemática una visión general sobre una escena del mundo real;
Figura 2 es un modelo de objeto articulado con superficies de proyección asociadas;
Figura 3a es una silueta típica de una imagen segmentada;
Figura 3b son tres mejores poses de coincidencia a partir de una base de datos;
Figura 3c una pose del esqueleto en 2D estimada a partir de la mejor coincidencia;
Figura 4 estimación de articulaciones en 3D a partir de dos imágenes de la cámara;
Figura 5a una imagen con posiciones de la articulación (manualmente) corregidas;
Figura 5b una adaptación inicial de una plantilla de forma 3D pre-segmentada para una imagen;
Figura 5c un ajuste corregido que coincide exactamente con las posiciones de la articulación en 5a;
Figura 6a una segmentación inicial de una imagen de un cuerpo, con píxeles seguros derivados de un modelo de plantilla, y con píxeles limítrofes inseguros;
Figura 6b una segmentación después del etiquetado de acuerdo con un modelo de color capacitado;
Figura 6c una segmentación final después de la eliminación morfológica de los valores atípicos;
Figura 7a carteleras mal alineadas en un ventilador de cartelera;
Figura 7b un ventilador de cartelera antes de la optimización de la articulación;
Figura 7c el mismo ventilador de cartelera después de la optimización de la articulación;
Figura 8a errores de muestreo que provocan grietas entre las carteleras;
Figura 8b un artefacto de representación correspondiente;
Figura 8c corrección del artefacto de representación;
Figura 9a ejemplo de peso de mezcla para dos cámaras de origen;
Figura 9b imagen representada sin suavizado;
Figura 9c imagen representada con un suavizado de adaptación;
Figura 9d discontinuidades donde se ha aplicado el suavizado; y
Figura 10 un diagrama de flujo de un método de acuerdo con la invención.

Los símbolos de referencia utilizados en los dibujos, y sus significados, se enumeran en forma sintetizada en la lista de símbolos de referencia. En principio, las partes idénticas se proporcionan con los mismos símbolos de referencia en los dibujos.

Descripción detallada de realizaciones preferidas

La Figura 1 muestra en forma esquemática una visión general sobre una escena del mundo real 8, la escena 8 comprende un objeto del mundo real 14 tal como un ser humano, que está siendo observado por dos o más cámaras de origen 9, 9', cada una de los cuales genera una corriente de vídeo de imágenes de origen 10, 10'. El sistema y el método de acuerdo con la invención genera una imagen virtual 12 que muestra la escena 8 desde un punto de vista de una cámara virtual 11 que es distinto de los puntos de vista de las cámaras de origen 9, 9'. En forma opcional, a partir de una secuencia de imágenes virtuales 12 se genera una corriente de vídeo virtual. Un aparato de acuerdo con la invención comprende una unidad de procesamiento 15 que lleva a cabo los cálculos de procesamiento de imágenes que aplican el método de la invención, dadas las imágenes de origen 10, 10' y generan una o más imágenes virtuales 12. La unidad de procesamiento 15 está configurada para interactuar con una unidad de almacenamiento 16 para el almacenamiento de imágenes de origen 10, imágenes virtuales 12 y los resultados intermedios. La unidad de procesamiento 15 se controla por medio de una estación de trabajo 19 que en forma típica comprende un dispositivo de visualización, un dispositivo de entrada de datos tales como un teclado y un dispositivo de señalización tal como un ratón. La unidad de procesamiento 15 puede estar configurada para suministrar una corriente de vídeo virtual a un

transmisor de radiodifusión de televisión 17 y/o a los dispositivos de visualización de vídeo 18.

La **Figura 2** muestra en forma esquemática un modelo en 3D 1 de la escena 8, que comprende un modelo de objeto articulado 4 del objeto del mundo real 14. El modelo en 3D 1 en forma típica además comprende otros modelos de objetos, por ej., otros seres humanos, que representan el suelo, edificios, etc. (no mostrados). El modelo de objeto articulado 4 comprende articulaciones 2 que están conectadas por enlaces 3, que corresponde aproximadamente a los huesos o las extremidades en el caso del modelo de un humano. Cada articulación 2 está definida como un punto en el espacio 3D, y cada enlace 3 puede ser representado por una línea recta que conecta dos articulaciones 2 a través del espacio 3D. Además, se muestra una variedad de superficies de proyección 5 que puede estar asociada con los enlaces 3 del modelo de objeto articulado 4. Esta asociación comprende una relación por lo menos en parte geométrica fija entre las superficies de proyección 5 y el enlace, por consiguiente, las superficies de proyección 5 se mueven con el enlace. De acuerdo con diferentes realizaciones de la invención, las superficies de proyección 5 son (de izquierda a derecha)

- cuerpos elipsoidales;
- cuerpos cilíndricos; o
- un conjunto de carteleras 6 que forman un ventilador de cartelera 7.

La asociación entre un enlace y una superficie de proyección puede ser, por ejemplo, que el enlace (es decir, una línea recta que enlaza las dos articulaciones conectadas por el enlace) define un eje mayor de un cuerpo o cilindro elipsoidal tal, o se encuentra dentro del plano de una o más de tales carteleras. Las carteleras 6 per se, para la proyección de vistas virtuales, son conocidas en la técnica. En la presente invención, dos o más carteleras planas 6 están asociadas con un solo enlace 3 del modelo de objeto articulado 4. Cada cartelera 6 está asociada con una cámara de origen 9. El plano de la cartelera 6 comprende el enlace 3, con la orientación de la cartelera 6 alrededor del enlace 3 definido por la ubicación de la cámara de origen asociada 9. Con preferencia, la cartelera 6 es normal a la línea más corta entre la cámara de origen 9 y la dirección del enlace 3. Todas las carteleras 6 para un enlace 3 forman juntos un ventilador de cartelera 7. Las imágenes de las cámaras de origen 9 se proyectan sobre las carteleras asociadas 6 de cada enlace 3, y luego proyectado en la cámara virtual 11, y mezclado, desde las carteleras 6 del enlace 3, para formar la imagen virtual 12 del enlace 3. Por lo tanto, las carteleras 6 del enlace 3 no se ocluyen la una a la otra. Sin embargo, pueden ocluir las carteleras 6 de otro enlace 3.

1. Información general

Uno de los objetivos es permitir virtualmente sin restricciones la representación con libre punto de vista de los sujetos humanos de un pequeño conjunto de la línea de base en todo el metraje de vídeo. Utilizamos una representación basada en las carteleras articuladas 6. La base de este modelo es una estructura de esqueleto humano en 3D 4 (véase la **Figura 2**). Cada hueso o enlace 3, representado por un vector 3D $\mathbf{b}_i$ y la posición de su articulación extrema 2 $\mathbf{x}_i$, corresponde a un componente principal del cuerpo completo real 14, por ej., el torso o las extremidades. Con cada hueso que se asocia un ventilador 7 de carteleras 6, que contiene una cartelera 6 por cada imagen de entrada $\boldsymbol{I}_i$ de un sujeto (véase la **Figura 2**). En forma más específica, para cada $\boldsymbol{I}_i$ el plano de la cartelera correspondiente está definido por la articulación $\mathbf{x}_i$, la dirección de hueso $\boldsymbol{b}_i$, y el vector $b_i \times (c_j - x_i)$, donde $\mathbf{c}_i$ es la posición de la cámara de $\boldsymbol{I}_i$. Por lo tanto, las carteleras 6 están alineadas con los huesos de carácter y como ortogonal como sea posible a sus vistas de entrada asociadas 10, 10'.

La idea básica de este método es calcular una pose en 3D del modelo de cartelera articulado, es decir, una configuración de articulación espacial de la estructura del esqueleto subyacente 4, que trae su proyección 2D en correspondencia con la pose del sujeto en cada marco de entrada del vídeo. Después de esta alineación, un mapa de textura y máscara alfa se genera para cada cartelera 6 de su vista asociada 10, 10'. Sin embargo, un cálculo totalmente automático de una sola pose en 3D, que es perfectamente compatible con todas las vistas de entrada, puede no ser posible en presencia de cuestiones tales como la calibración de la cámara imperfecta o baja resolución de la textura. En tales casos, se aplica un enfoque semiautomático, basado en datos, que opera en tres fases consecutivas: la estimación de la pose en 2D y la segmentación de imágenes basada en plantillas, la construcción del modelo de cartelera 3D articulado, y la representación real.

En primer lugar, para la estimación de la pose en 2D en cada vista de entrada individual, se utiliza una base de datos de las siluetas, la consistencia temporal de movimiento de los sujetos en el vídeo, y los datos de captura de movimiento para ayudar al usuario en la colocación rápida y precisa de las articulaciones 2. Teniendo en cuenta estas posiciones de articulaciones 2D, una segmentación de la imagen en las diferentes partes del cuerpo, es decir, el torso o las extremidades, se calcula por el uso de un modelo de plantilla humana con el fin de mapear los píxeles de imágenes con las carteleras (véase la Sección 2 "Estimación de posos y Segmentación Basada en Plantillas").

La segunda fase del algoritmo integra la información de la pose y la textura de todas las vistas individuales y genera el modelo de cartelera articulado definitivo para la representación. Este paso de procesamiento incluye una optimización de las posiciones de la articulación en 3D y una compensación por los errores de calibración de la cámara, lo que optimiza la superposición de textura para cada segmento del modelo, es decir, para cada ventilador 7 de carteleras 6. Una optimización final de máscara alfa y textura elimina las costuras visibles y discontinuidades entre carteleras adyacentes (véase la Sección 3 "Construcción del Modelo de Cartelera Articulado 3D").

El último paso es la representación real en tiempo real de nuevos puntos de vista. La Sección 4 describe un algoritmo para un esquema de mezcla por píxel totalmente basada en el GPU, dependiente de la vista, que está optimizado para representar modelos de cartelera articulados de manera eficiente, mientras que preserva el fotorrealismo del vídeo de entrada original.

2. Estimación de la Pose y Segmentación Basada en Plantillas

En la primera fase del método se calcula una estimación inicial de 14 posiciones de la articulación del sujeto en el espacio de la imagen y una segmentación de los píxeles en las diferentes partes del cuerpo. Para la calibración de los parámetros intrínsecos y extrínsecos de la cámara actualmente se utiliza el método de Thomas [Tho06]. De acuerdo con lo mencionado con anterioridad, una estimación y segmentación de pose completamente automática es muy difícil debido a la baja resolución y la calidad. En consecuencia, se propone el siguiente enfoque semi-automático que minimiza la interacción del usuario necesaria para sólo unos pocos clics de ratón. Entonces, dadas las 2 posiciones de las articulaciones, la segmentación de las piezas del cuerpo del sujeto 14 se calcula por medio del ajuste de un modelo humano con una plantilla de segmentación conocida para los marcos de vídeo de entrada.

2.1. Estimación de la Pose en 2D

Si supone que una segmentación gruesa del sujeto 14 del fondo está disponible, por ej., por el uso de codificación de colores o sustracción de fondo. La **Figura 3a** muestra un ejemplo típico de una imagen segmentada 13 en este escenario de aplicación. La idea básica para calcular una estimación inicial de la pose de un sujeto, es decir, las posiciones en 2D de las articulaciones del esqueleto 2, es compararla con una base de datos de siluetas, para la que se conocen las respectivas poses de esqueleto (véase la **Figura 3b**). En primer lugar, para cada vista $\boldsymbol{I_i}$, se normaliza para los sujetos de diferentes tamaños por medio del re-muestreo de la silueta 13 en una cuadrícula de 32 x 40 y se apila de la información silueta binaria en cada punto de la cuadrícula en un vector $v_j \in [0,1]^n$, con $\boldsymbol{n}$ = 32 x 40. Entonces, para cada $\mathbf{v}_j$, este algoritmo encuentra las mejores entradas *k* coincidentes en la base de datos, que minimizan el error

$$E_S = (1-\lambda)\frac{1}{n}\sum_{i=0}^{n-1}|\mathbf{v}_j(i)-\mathbf{w}(i)| + \lambda\frac{1}{m}\sum_{r=0}^{m-1}|\mathbf{p}_j(r)-\mathbf{q}(r)|. \qquad (1)$$

donde **w** es una entrada en la base de datos **q,** sus correspondientes posiciones de la articulación en 2D, y $\boldsymbol{m}$ es el número de articulaciones de esqueleto. El vector $\mathbf{p}_i$ contiene las coordenadas de las articulaciones desde el marco de vídeo anterior. El primer término de la Ecuación (1) asegura una coincidencia adecuada de las siluetas mientras que el segundo término explota la consistencia de movimiento temporal del sujeto que está en el vídeo. En otras palabras, la reducción al mínimo de (1) devuelve la entrada de la base de datos que se parece más a la imagen actual y cuyas 2 posiciones de la articulación están más cerca de las posiciones de la articulación de la imagen anterior. Esto es de particular ayuda para resolver ambigüedades de izquierda a derecha en las siluetas. La influencia del segundo término puede ser ponderada por el valor λ. Para el primer marco de una secuencia, simplemente se ajusta λ = 0, para todos los demás marcos se utilizó un valor de λ = 0,5 para todos los ejemplos. Las posiciones de la articulación 2 también se procesan en coordenadas normalizadas con respecto al cuadro delimitador del objeto. El uso de este error $\boldsymbol{E_s}$, el *k* = 3 mejores siluetas coincidentes y sus correspondientes posiciones de la articulación en 2D para cada vista individual $\boldsymbol{I_i}$; se recuperan de la base de datos.

Con el fin de seleccionar la pose en 2D más plausible de cada uno de estos conjuntos se ejecuta una optimización de múltiples vistas para cada combinación de poses: se calculan los rayos 3D de cada centro de cámara $\boldsymbol{c_j}$ a través de las posiciones de la articulación recuperadas en $\boldsymbol{I_i}$. A continuación, se calcula el 3D representante para cada articulación 2 que está más próxima a los rayos correspondientes. La **Figura 4** muestra un ejemplo con dos cámaras 9, 9'.

La medida para la calidad de una combinación particular de poses es la suma acumulada de las distancias de cada articulación en 3D a partir de sus respectivos rayos. Con el fin de hacer este procedimiento más robusto para la calibración de la cámara a menudo inexacta, esta optimización de múltiples vistas también incluye un paso de corrección simple. Para cada silueta, un desplazamiento en 2D en el plano de la imagen se introduce como un parámetro adicional. Al minimizar la suma acumulada de las distancias, estos desplazamientos en 2D son variados también, por el uso del algoritmo de Levenberg-Marquardt. Esta corrección de calibración ha demostrado ser muy eficaz: para algunas imágenes de la silueta el desplazamiento en 2D necesario para minimizar la medida de error puede ser de hasta 8 píxeles.

En resumen, la optimización mencionada con anterioridad se lleva a cabo para cada combinación de las mejores siluetas coincidentes para cada vista. Por ejemplo, dadas dos cámaras, y habiendo encontrado para cada cámara (o vista) tres mejores siluetas coincidentes, a continuación, se lleva a cabo la optimización de múltiples vistas nueve veces. Para cada cámara, la pose en 2D es elegida que da la suma acumulada más pequeña de distancias a lo largo de todas las ejecuciones de optimización.

De acuerdo con lo demostrado en la Figura 3c, esta estimación de pose basada en siluetas y optimización de

articulaciones proporciona por lo general una buena suposición de posiciones de la articulación en 2D del sujeto en cada vista I_j. Con una interfaz sencilla el usuario puede corregir manualmente estas posiciones por medio del movimiento de las articulaciones (véase la **Figura 5a**). Después de este paso de refinamiento de articulaciones manual la silueta y posiciones de la articulación se añaden inmediatamente con preferencia a la base de datos. El aumento de poses en la base de datos ha demostrado llevar a significativamente mejores resultados para nuevas secuencias. En escenarios de aplicación, donde no se dispone de información de la silueta en absoluto, el usuario puede recurrir a la colocación de todas las articulaciones en forma manual.

2.2. Ajuste de Plantilla en 3D

Incluso con articulaciones en 2D precisas una segmentación robusta de la imagen en las partes del cuerpo del sujeto sigue siendo un problema difícil. El uso de una base de datos de siluetas segmentadas en lugar de la segmentación de siluetas binarias anterior no es una opción deseable, ya que la creación de una base de datos tal sería extremadamente compleja y requiere mucho tiempo, y todavía no podría esperar encontrar siempre coincidencias suficientemente precisas.

En su lugar, se ajusta un *modelo de plantilla genérico y pre-segmentado en 3D* a las imágenes. Esto tiene la ventaja considerable de que se obtiene una solución de partida buena para el proceso de segmentación y que se puede resolver las oclusiones con facilidad. Sin embargo, el ajuste de un modelo en 3D requiere, para cada vista de entrada particular, el cálculo de una pose en 3D cuya proyección está perfectamente alineada con las articulaciones en 2D. Una pose en 3D que lleva una coincidencia perfecta en *todas* las vistas a menudo puede no ser encontrada debido a imprecisiones de calibración o extravíos leves de las articulaciones. Por lo tanto, se ajusta un modelo en 3D por visión de entrada. Una solución para el cálculo de una pose en 3D aproximada para los modelos articulados a partir de una sola imagen ha sido presentada por Hornung et al. [HDK07]. Dadas las posiciones de la articulación en 2D $\mathbf{x}_i$ para una imagen I_j, su enfoque utiliza una base de datos de los datos de captura de movimiento 3D para encontrar un conjunto de posiciones de la articulación en 3D $\mathbf{x}$; cuya proyección coincide aproximadamente con las articulaciones de entrada en 2D (véase la **Figura 5b**). Se proporciona una modificación simple pero eficaz para su algoritmo para calcular el ajuste exacto requerido.

Esto se lleva a cabo de acuerdo con lo presentado a continuación: La coincidencia en 3D aproximada se deforma, tal como para alinearse con las articulaciones en 2D, de acuerdo con el siguiente algoritmo: A través de cada articulación en 3D $\mathbf{X}_i$, se crea un plano paralelo al plano de la imagen de I_j. A continuación, se echa un rayo desde el centro de la cámara c_j a través de la posición de la articulación diana correspondiente $\mathbf{x}_i$ en I_j y se calcula su intersección con el plano. La pose en 3D se actualiza a continuación, por medio del movimiento de cada $\mathbf{X}_i$ al punto de intersección respectivo y la actualización del sistema de coordenadas del hueso en 3D en consecuencia. En otras palabras: este procedimiento supone que la distancia desde la cámara a la articulación es correcta, y ajusta la posición en 3D de la articulación para que coincida con la imagen mientras mantiene la distancia de la constante de la cámara. El resultado es la pose en 3D requerida suponer que se proyecta exactamente en las articulaciones en 2D estimadas previamente. El modelo de plantilla en 3D ahora se puede ajustar a la imagen por medio de la deformación de acuerdo con esta pose en 3D calculada por el uso de técnicas estándar para la animación basada en esqueletos [LCF00] (véase la **Figura 5c**). Se debe tener en cuenta que este algoritmo por lo general no conserva las longitudes de las extremidades del esqueleto en 3D original y por lo tanto, permite una adaptación de la plantilla de malla en 3D para adaptarse a las dimensiones del sujeto con mayor precisión.

2.3. Segmentación de las Partes del Cuerpo

El modelo de plantilla pre-segmentado ajustado no segmenta perfectamente el marco de entrada I_j y podría no cubrir toda la silueta por completo. Por lo tanto, un refinamiento de la segmentación se lleva a cabo en tres sencillos pasos. En un primer paso, un modelo de color que se aprende por segmento corporal con base en los píxeles seguros seleccionados en forma automática de las partes del cuerpo pre-segmentadas (véase la **Figura 6a**). En un segundo paso, el modelo de color formado se utiliza para etiquetar los píxeles inseguros que conducen a una segmentación ajustada a las dimensiones del cuerpo y la silueta de los sujetos (véase la **Figura 6b**). En un tercer paso, una operación de cierre morfológica elimina valores atípicos de acuerdo con lo representado en la **Figura 6c**.

Para determinar los píxeles seguros, se proyecta una versión ligeramente adelgazada y engrosada del modelo de plantilla en la imagen y se etiquetan los píxeles de siluetas en consecuencia. Los píxeles que reciben la misma etiqueta en ambas proyecciones se marcan como píxeles seguros y se etiquetan con el segmento del cuerpo correspondiente. Todos los píxeles que quedan dentro de la silueta son etiquetados como inseguros de acuerdo con lo mostrado en la **Figura 6a**.

Al aprender el modelo de color en línea, se proporciona un algoritmo de segmentación robusto que es capaz de manejar la segmentación en entornos no controlados. Las condiciones cambiantes de iluminación, el aspecto específico del sujeto o el aspecto dependiente de la vista de este modo se pueden manejar en forma fiable.

El procedimiento de estimación de pose y segmentación se lleva a cabo para todas las vistas y marcos de entrada a partir de los cuales las representaciones de punto de vista libre se han de generar. Como resultado, el enfoque de segmentación por el uso de estimaciones de pose en 2D sucesivas y ajustes de plantillas en 3D maneja en forma

automática las partes del cuerpo ocluidas, es robusto incluso para una calidad y resolución de imagen baja, y requiere sólo una pequeña cantidad de interacción de usuario sencilla durante el refinamiento de las posiciones de la articulación.

3. Construcción del Modelo de Cartelera Articulado en 3D

Se utilizan las posiciones de articulaciones en 3D computadas de la Sección 2.1 como una pose inicial de la representación de cartelera articulada final. Si una articulación del modelo de cartelera en 3D articulado no está posicionada de manera óptima, la textura resultante de la representación de todas las carteleras de un ventilador de cartelera no se alineará (véase la **Figura 7a**). En esta sección, se describe cómo las posiciones de la articulación en 3D se pueden optimizar en base a una medida cuantitativa de la alineación de las texturas de la cartelera.

En lo siguiente, primero se define una función de puntaje para una posición de una articulación en una sola vista y para un par de cámara. Esta función de puntaje se extiende luego a varias vistas y cámaras. Por el uso de esta función de puntaje y las limitaciones antropométricas, se optimiza la pose en 3D del modelo de cartelera articulado. Por último, se describirá una corrección de costura que elimina las discontinuidades de textura entre las carteleras adyacentes.

3.1. Puntaje de Posición

Para calificar la calidad de una posición de la articulación de una vista de salida ***V***, se evalúan todas las carteleras adyacentes a esta articulación. Para cada ventilador de carteleras, la alineación de sus carteleras para un par de vistas de entrada (I_1,I_2) se puntúa por medio de una comparación de píxeles racional de las texturas proyectadas. Para cada píxel de salida ***p*** de ***V***, el puntaje por píxel $s_{I1, I2}(p)$ se define como

$$s_{I_1,I_2}(p) = \begin{cases} 1 - \varepsilon(V_{I_1}(p), V_{I_2}(p)), & p \text{ activo en } I_1 \text{ e } I_2 \\ 0, & \text{de lo contrario} \end{cases} \quad (2)$$

donde $V_{Ii}(p)$ es la contribución del color de una cartelera asociada con la vista I_j de píxeles ***p***. $\varepsilon(\cdot)$ es una medida de distancia de color en RGB. Los píxeles activos se definen como aquellos píxeles en la vista de salida ***V*** que reciben una contribución de color válida desde las vistas de entrada i_1 e i_2. La segmentación generada en la Sección 2.3 se utiliza para resolver la oclusión en forma fiable. El puntaje para una articulación en una vista ***V*** es la suma normalizada de todos los píxeles

$$s_{I_1,I_2}(V) = \frac{\sum_{p \in V} s_{I_1,I_2}(p) n(p)}{\sum_{p \in P_v} n(p)}. \quad (3)$$

El factor de normalización *n(p)* es 1, si por lo menos uno de los dos píxeles está activo y 0, de lo contrario. Por lo tanto, la función de puntaje mide la coincidencia de los valores de textura, mientras que *n(p)* penaliza las partes no alineadas como en la **Figura 7a**. Estas operaciones por píxeles se implementan de manera eficiente en la GPU por el uso de sombreadores de fragmentos. En resumen, el procedimiento de acuerdo con (1) y (2) determina a qué grado coinciden las contribuciones de imagen de diferentes cámaras, de acuerdo con lo observado desde el punto de vista virtual y en la imagen de salida virtual, y sólo para aquellos píxeles para los que la imagen de salida recibe una contribución de ambas cámaras de origen.

Durante más de dos vistas de entrada, se define el puntaje como un promedio ponderado de todos los pares de la cámara, donde el peso para cada par de la cámara depende del ángulo $\beta_{i1,i2}$ entre las direcciones de visión respectivas, con ángulos estrechos que reciben un mayor peso:

$$s(V) = \frac{\sum_{(I_1,I_2) \in \mathcal{I}} s_{I_1,I_2}(V) \omega(\beta_{I_1,I_2})}{\sum_{(I_1,I_2) \in \mathcal{I}} \omega(\beta_{I_1,I_2})}; \quad (4)$$

donde I es el conjunto de todos los pares de vistas de entrada y $\omega(\beta)$ es, por ejemplo, un peso de Gauss:

$$\omega(\beta) = e^{-\frac{\beta^2}{2\sigma^2}} \quad (5)$$

Un valor adecuado para σ se determinó en forma empírica como 0,32. Por último, el puntaje de la posición de articulación es la suma normalizada de las puntuaciones en todas las vistas evaluadas:

$$S_v = \frac{1}{|v|}\sum_{v \in V} s(V) \quad (6)$$

dónde *V* es el conjunto de todos las vistas evaluadas.

3.2. Optimización de la Pose en 3D

Dado que la puntuación de la posición de la articulación depende de las vistas evaluadas, se necesita un conjunto adecuado *V*. Con el fin de cubrir un intervalo razonable de posiciones de visualización, se evalúa la función de puntaje en las posiciones de cámara de todas las vistas de entrada y las vistas virtuales en el centro entre cada par de la cámara. Para la optimización de posición de una articulación, se evalúa S_v en las posiciones espacialmente cercanas de los candidatos en una cuadrícula en 3D adaptativa y discreta. La cuadrícula se refina de manera codiciosa en torno a esas posiciones candidatas que consiguen un puntaje más alta S_v, hasta que se alcanza una resolución de la cuadrícula dada (ajustada en forma empírica a 1,2 cm).

Para evitar configuraciones degeneradas con ventiladores de cartelera de longitud cero, en forma adicional se considera la consistencia antropométrica [NAS09] durante la evaluación de cada pose. Una posición de la articulación recibe un puntaje de cero si no se mantiene una de las siguientes limitaciones:

- La articulación está en o por encima del suelo.
- Las longitudes de los huesos del esqueleto topológicamente simétricos (por ej., el brazo izquierdo/derecho) no difieren más del 10%.
- Las longitudes de los huesos adyacentes están dentro de los estándares antropométricos.
- Las distancias a las articulaciones inconexas están dentro de los estándares antropométricos.

Para las dos últimas restricciones, se utiliza el quinto percentil de sujetos femeninos, que se redondea hacia abajo como longitudes mínimas, y el percentil número 95 de los sujetos masculinos se redondea como longitudes máximas.

Este proceso de optimización de búsqueda de cuadrícula se repite en forma iterativa a lo largo del esqueleto. Es decir, en cada iteración, la posición se optimiza por separado, de acuerdo con lo descrito, para cada articulación del conjunto de todas las articulaciones. En estos experimentos, se ha hallado que por lo general converge después de 4 iteraciones. Dado que la optimización está basada en funciones diana que están definidas en la imagen virtual, no se determinan parámetros innecesarios en última instancia, y la eficiencia global es alta. Véase la Figura 7 para un modelo de cartelera articulado antes (**7b**) y después (**7c**) de la optimización.

3.3. Corrección de la Costura de la Textura

Debido a la toma de muestras de las máscaras de segmentación de las carteleras durante la representación con la texturización proyectiva (véase la **Figura 8a**), pueden aparecer pequeñas discontinuidades (grietas visibles) entre las carteleras adyacentes en la vista de salida de acuerdo con lo mostrado en la **Figura 8b**: En la imagen virtual 12, un píxel de salida a partir de una primera cartelera 6 puede caer, cuando se proyecta en la imagen de origen segmentada 10, dentro de un segundo segmento 13b que se asigna a una segunda cartelera 6' adyacente, en lugar de en un primer segmento 13a asignado a la primera cartelera 6. En consecuencia, el píxel de salida no recibe ninguna contribución de color en absoluto. Para superar este problema, estos píxeles de costura tienen que ser representados para ambas carteleras adyacentes. Por lo tanto, se marcan los píxeles como píxeles de costura en las vistas de entrada si cubren las carteleras sobre dos huesos del esqueleto adyacentes o enlaces 3 (por ej., un píxel encerrado por líneas discontinuas en la Figura 8a).

Para detectar los píxeles de costura, la máscara de segmentación es atravesada para cada vista de entrada. Un píxel ***p*** está marcado como píxel de costura, si cumple ambas de las siguientes condiciones:

- Por lo menos un píxel ***p'*** en su barrio de 4 tiene una etiqueta diferente, pero viene de la misma materia
- |profundidad(*p*) - profundidad(*p*')| < φ

donde la profundidad(·) es el valor de profundidad en este píxel. El umbral φ distingue entre las partes de oclusión y las partes conectadas. Se ajustó en forma empírica a φ= 3 cm. Un ejemplo para la máscara de segmentación

corregida en su costura y la mejora de representación resultante se muestra en la **Figura 8c.**

4. Representación

En lo que sigue se describe un procedimiento de representación para las carteleras articuladas. Se ha diseñado este algoritmo de acuerdo con los criterios generales definidos por Buehler et al. [BBM*01]. Debido a este entorno desafiante con errores de calibración y un posicionamiento de la cámara muy escaso, este enfoque particular está en:

- Aspecto Consistente: Las carteleras adyacentes deben intersectar sin grietas o artefactos perturbadores y mezclarse de manera realista con el medio ambiente.
- Continuidad Visual: Las carteleras no deben cambiar repentinamente o aparecer al mover el punto de vista.
- Interpolación de Vistas: Al ver la escena desde un ángulo y posición originales de la cámara, la vista representada debe reproducir la de la cámara de entrada.

Las entradas al procedimiento de representación son el modelo de cartelera articulado, las vistas de entrada segmentadas *I* (Sección 2.3) y las costuras calculadas en la Sección 3.3. Para cada marco de salida representado, las carteleras articuladas están ordenadas de atrás hacia adelante para un manejo adecuado de las oclusiones. Con el fin de cumplir con los objetivos anteriores, se lleva a cabo un procedimiento de mezcla por píxel. Se separa entre los pesos por cámara que se calculan una vez por cartelera y los últimos pesos por píxel.

4.1. Pesos de la Mezcla de la Cámara

Para una mezcla suave de las carteleras 6 asociadas con un ventilador 7 de carteleras 6, se utiliza el mismo peso de Gauss que en la Ecuación (5). Para lograr una interpolación en una vista de cámara original 10, se introduce una función de atenuación que asegura que todas las vistas de una perspectiva de cámara original 9 son idénticas a las imágenes de origen de la cámara correspondientes 10 mientras que se sigue asumiendo una transición suave entre las diferentes vistas. La función de atenuación se define como $f(I_{\omega Máx}) = 1$ para la vista de origen $I_{\omega Máx}$ con el valor más alto de $\alpha(\cdot)$ (es decir, la cámara de origen 9 más cercana) y

$$f(I_{\omega_{Max}}) = 1 - e^{-\frac{d(V, I_{\omega_{Max}})^2}{2\sigma^2}} \quad (7)$$

para todas las otras cámaras I_j. $d(V, I_{\omega Máx})$ es la distancia euclídea desde la posición de la cámara virtual del espectador 11 hasta la posición de la cámara de origen 9 de la vista $I_{\omega Máx}$. La constante σ se determina en forma empírica como de 1 metro, que es inferior a la distancia mínima entre dos cámaras de origen 9 y por lo tanto no da lugar a ningún tipo de discontinuidad.

4.2. Procesamiento por Píxel

Las carteleras de un ventilador de cartelera se mezclan por píxel. De acuerdo con lo mostrado en la **Figura 8a**, se lleva a cabo una búsqueda de cámaras en la máscara de segmentación correspondiente de cada cartelera. Esto determina si el píxel de salida de corriente *p* está en la parte del cuerpo que pertenece a esta cartelera. Si es así, entonces la contribución de color correspondiente $V_{Ij}(p) = 0$ desde la vista de origen *Ij* y su valor alfa $\alpha_{Ij}(p)$ se puede añadir a la vista de salida *V*. De lo contrario, se ajusta $\alpha_{Ij}(p) = 0$, es decir, transparente. El último caso también se produce cuando la parte del cuerpo correspondiente se ocluye en I_j y la información de color se debe tomar de otras cámaras. El valor de color resultante *V(p)* del píxel de pantalla es entonces

$$V(p) = \frac{\sum_{I_j \in \mathcal{I}} V_{I_j}(p) w(I_j, p)}{\sum_{I_j \in \mathcal{I}} w(I_j, p)} \quad (8)$$

con el conjunto de todas las vistas de entrada *I* como en la Ecuación (2) y los pesos por píxel

$$w(I_j, p) = \alpha_{I_j}(p) \omega(\beta_{I_j}) f(I_{\omega_{Max}}). \quad (9)$$

Esto se hace para todos los canales de color por separado. El valor alfa resultante es

$$\alpha_V(p) = \begin{cases} \alpha_{I_{\omega_{Max}}}(p), & si \quad (I_{\omega_{Max}} \cdot p) \neq 0 \\ \dfrac{\sum_{I_j \in \mathcal{I}} \alpha_{I_j}(p) w(I_j, p)}{\sum_{I_j \in \mathcal{I}} \alpha_{I_j}(p) \omega(\beta_{I_j})}, & \text{de lo contrario} \end{cases} \qquad (10)$$

donde se aplica el primer caso, si la cámara más cercana se utiliza para este píxel. La Ecuación (8) y la Ecuación (10) se aseguran de que los valores de color se mezclen de manera tal que los factores sumen 1. Sin embargo, los valores de alfa no tienen sumar 1, por ej., si los valores de alfa continuos están disponibles en lugar de las máscaras de segmentación binaria.

Además de esto, las carteleras vistas en un ángulo oblicuo o desde la parte trasera, es decir, que tienen una normal en un ángulo próximo a o más de 90 grados lejos de la dirección de visión, simplemente se desvanecen. Para simplificar, estos factores no se muestran en las ecuaciones.

Un ejemplo para la mezcla de intensidades (es decir, un canal de color) de dos cámaras se muestra en la **Figura 9a** donde los ángulos de azimut y altitud son de coordenadas esféricas de la posición de vista alrededor del ventilador de carteleras. Los dos puntos pico en (0,0, 0,0) y (0,5, 0,5) corresponden a las posiciones de las cámaras de origen. De acuerdo con lo que se puede observar en el gráfico, al acercarse a estos puntos, el peso de la cámara correspondiente aumenta al modelo en 3D 1,0 y todos los otros pesos cámara disminuyen a 0,0. Por lo tanto, en este caso sólo se utiliza la cámara de origen, lo que da lugar a la reproducción exacta de la imagen de origen.

Por último, para evitar los bordes no suaves en los límites de un ventilador de carteleras con respecto al fondo, otros ventiladores de la cartelera, y en lugares donde otras vistas de entrada reciben el peso más alto (por ej., debido a las oclusiones en una cartelera), se aplica un paso de suavizado gaussiano adicional. Esto se lleva a cabo de forma adaptativa como un proceso posterior sólo en las discontinuidades detectadas y almacenadas mientras que se representan las carteleras. Las **Figuras 9b, c** y **d** muestran un ejemplo: **9b** imagen sin suavizado, **9c** con suavizado adaptativo, **9d** ubicaciones donde las discontinuidades se han eliminado a través de suavizado. La **Figura 10** muestra un diagrama de flujo de un método. En un primer paso 21, se adquiere por lo menos una imagen por cámara de origen 9, ya sea desde una corriente de vídeo en vivo, o a partir de imágenes o corrientes de vídeo almacenadas. En un segundo paso 22, se lleva a cabo la estimación de la pose en 2D. En un tercer paso opcional 23, se lleva a cabo la optimización de múltiples vistas. En un cuarto paso 24, se lleva a cabo el ajuste de plantilla en 3D. En un quinto paso 25, se lleva a cabo la segmentación de las partes del cuerpo. En un sexto paso 26, se lleva a cabo la optimización de la pose en 3D, con base en la puntuación de posición. En un séptimo paso 27, se lleva a cabo la corrección de la costura de la textura. En un octavo paso 28, se lleva a cabo la mezcla de la cámara de las carteleras 6 de cada ventilador de cartelera 7. En un noveno paso 29, la imagen final se almacena y/o se exhibe. Mientras que la explicación anterior se refiere a la representación y la presentación de un único objeto articulado, la imagen final puede comprender una pluralidad de objetos articulados e imágenes de un fondo y otros objetos.

Si bien la invención se ha descrito en las presentes realizaciones preferidas de la invención, se entiende claramente que la invención no está limitada a las mismas, pero se puede incorporar de otro modo de diversas maneras y se puede practicar dentro del alcance de las reivindicaciones.

Bibliografía

ABB*07
ANDÚJAR C., BOO J., BRUNET P., FAIRÉN M., NAVAZO I., VAZQUEZ P., VINACUA À.:

Omni-directional relief impostors.
Computer Graphics Forum 26, 3 (2007), 553 a 560.

ABT99
AUBEL A., BOULIC R., THALMANN D.:

Lowering the cost of virtual human rendering with structured animated impostors.
En WSCG'99 (1999).

AT06
AGARWAL A., TRIGGS B.:

Recovering 3d human pose from monocular images.
IEEE Trans. Pattern Anal. Mach. Intell. 28, 1 (2006), 44 a 58.

BBM*01
BUEHLER C., BOSSE M., MCMILLAN L., GORTLER S., COHEN M.:

Unstructured lumigraph rendering.
In SIGGRAPH '01 (2001), págs. 425 a 432.

BCF*05
BEHRENDT S., COLDITZ C., FRANZKE O., KOPF J., DEUSSEN O.:

Realistic real-time rendering of landscapes using billboard clouds.
Computer Graphics Forum 24, 3 (2005), 507 a 516.

CCST00
CHAI J.-X., CHAN S.-C., SHUM H.-Y., TONG X.:

Plenoptic sampling.
In SIGGRAPH '00 (New York, NY, EE.UU., 2000), ACM Press/Addison-Wesley Publishing Co., págs. 307 a 318.

CTMS03
CARRANZA J., THEOBALT C., MAGNOR M. A., SEIDEL H.-P.:

Free-viewpoint video of human actors.
En SIGGRAPH '03 (2003), págs. 569 a 577.

dAST*08
DE AGUIAR E., STOLL C., THEOBALT C., AHMED N., SEIDEL H.-P., THRUN S.:

Performance capture from sparse multi-view video.
En SIGGRAPH '08 (2008), págs. 1 a 10.

DDS03
DÉCORET X., DURAND F., SILLION F. X.:

Billboard clouds.
In SCG '03 (2003), págs. 376 a 376.

DTM96
DEBEVEC P. E., TAYLOR C. J., MALIK J.:

Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach.
SIGGRAPH'96 (1996), 11 a 20.

EBMM03
EFROS A. A., BERG A. C., MORI G., MALIK J.:

Recognizing action at a distance.
En ICCV (2003), págs. 726 a 733.

Eye09
EYEVISION:.

http://www.ri.cmu.edu/events/sb35/tksuperbowl.html (2009).

GEJ*08
GAMMETER S., ESS A., JAEGGLI T., SCHINDLER K., LEIBE B., GOOL L. J. V.:

Articulated multi-body tracking under egomotion.
En ECCV (2) (2008), págs. 816 a 830.

GGSC96
GORTLER S. J., GRZESZCZUK R., SZELISKI R., COHEN M. F.:

The lumigraph.
En SIGGRAPH '96 (1996), págs. 43 a 54.

GKH09
GUILLEMAUT J.-Y., KILNER J., HILTON A.:

Robust graph-cut scene segmentation and reconstruction for free-viewpoint video of complex dynamic scenes.
En ICCV (Kyoto, Japón, Septiembre de 2009).

GM03
GOLDLUECKE B., MAGNOR M.:

Real-time microfacet billboarding for free-viewpoint video rendering.
En ICIP'03 (2003), vol. 3, págs. 713 a 716.

GTH*07
GRAU O., THOMAS G. A., HILTON A., KILNER J., STARCK J.:

A robust free-viewpoint video system for sport scenes.
In Proceedings of the 3DTV Conference (Abril 2007).

HDK07
HORNUNG A., DEKKERS E., KOBBELT L.:

Character animation from 2D pictures and 3D motion data.
ACM Transactions on Graphics 26, 1 (2007).

HS06
HAYASHI K., SAITO H.:

Synthesizing free-viewpoint images from multiple view videos in soccer stadium.
In CGIV (2006), págs. 220 a 225.

JKMG07
JAEGGLI T., KOLLER-MEIER E., GOOL L. J. V.:

Learning generative models for monocular body pose estimation.
En ACCV (1) (2007), págs. 608 a 617.

KSHG07
KILNER J., STARCK J., HILTON A., GRAU O.:

Dual-mode deformable models for free-viewpoint video of sports events. 3dim (2007), 177 a 184.

LBDGG05
LEE O., BHUSHAN A., DIAZ-GUTIERREZ P., GOPI M.:

Capturing and view-dependent rendering of billboard models.
In ISVC (2005), págs. 601 a 606.

LCF00
LEWIS J. P., CORDNER M., FONG N.:

Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation.
En SIGGRAPH '00 (2000), págs. 165 a 172.

LH96
LEVOY M., HANRAHAN P.:

Light field rendering.
En SIGGRAPH ’96 (1996), págs. 31 a 42.

MBR*00
MATUSIK W., BUEHLER C., RASKAR R., GORTLER S. J., MCMILLAN L.:

Image-based visual hulls.
En SIGGRAPH ’00 (2000), págs. 369 a 374.

MHM*09
MAHAJAN D., HUANG F.-C., MATUSIK W., RAMAMOORTHI R., BELHUMEUR P.N.:

Moving gradients: a path-based method for plausible image interpolation.
ACM Trans. Graph. 28, 3 (2009).

Mid09
Middlebury multi-view stereo evaluation.
http://vision.middlebury.edu/mview/, Octubre de 2009.

NAS09
NASA:

Anthropometry and biomechanics.
http://msis.jsc.nasa.gov/sections/section03.htm (2009).

PG08
PEKELNY Y., GOTSMAN C.:

Articulated object reconstruction and markerless motion capture from depth video.
Comput. Graph. Forum 27, 2 (2008), 399 a 408.

RMD04
RECHE A., MARTIN I., DRETTAKIS G.:

Volumetric reconstruction and interactive rendering of trees from photographs.
SIGGRAPH’04 23, 3 (Julio de 2004).

SGwHS98
SHADE J., GORTLER S., WEI HE L., SZELISKI R.:

Layered depth images.
En SIGGRAPH’98 (1998), págs. 231 a 242.

Tho06
THOMAS G.:

Real-time camera pose estimation for augmenting sports scenes. Visual Media Production, 2006. CVMP 2006. 3rd European Conference on (2006), págs. 10 a 19.

VBMP08
VLASIC D., BARAN I., MATUSIK W., POPOVIC J.:

Articulated mesh animation from multi-view silhouettes.
En SIGGRAPH ’08 (2008), págs. 1 a 9.

WWG07
WASCHBÜSCH M., WÜRMLIN S., GROSS M.:

3d video billboard clouds.
Computer Graphics Forum 26, 3 (2007), 561 a 569.

YSK*02
YAMAZAKI S., SAGAWA R., KAWASAKI H., IKEUCHI K., SAKAUCHI M.:

Microfacet billboarding.
En EGRW ’02 (2002), págs. 169 a 180.

Lista de designaciones

1	modelo en 3D
2	articulación
3	enlace
4	modelo de objeto articulado
5	superficie de proyección
6	cartelera
7	ventilador de cartelera
8	escena
9, 9’	cámara de origen
10, 10’	imagen de origen
11	cámara virtual
12	imagen virtual
13, 13a, 13b	segmento de imagen de origen
14	objeto del mundo real
15	unidad de procesamiento
16	unidad de almacenamiento
17	transmisor
18	dispositivo de visualización de vídeo
19	estación de trabajo

REIVINDICACIONES

1. Un método implementado por ordenador para la estimación de una pose de un **modelo de objeto articulado** (4), en el que el modelo de objeto articulado (4) es un modelo en 3D basado en ordenador (1) de un **objeto del mundo real** (14) observado por una o más **cámaras de origen** (9), y el modelo de objeto articulado (4) representa una pluralidad de **articulaciones** (2) y de **enlaces** (3) que enlazan las articulaciones (2), y en el que la **pose** del modelo de objeto articulado (4) está definido por la localización espacial de las articulaciones (2), el método comprende los pasos de

- obtener por lo menos una **imagen de origen** (10) desde una **corriente de vídeo** que comprende una vista del objeto del mundo real (14) grabada por una cámara de origen (9);
- procesar la por lo menos una imagen de origen (10) para extraer un **segmento de imagen de origen** correspondiente (13) que comprende la vista del objeto del mundo real (14) separado del fondo de la imagen;
- mantener, en una base de datos en forma legible por ordenador, un conjunto de **siluetas de referencia**, cada silueta de referencia está asociada con un modelo de objeto articulado (4) y con una **pose de referencia** particular de este modelo de objeto articulado (4);
- comparar el por lo menos un segmento de imagen de origen (13) con las siluetas de referencia y seleccionar un número predeterminado de siluetas de referencia, teniendo en cuenta, para cada silueta de referencia,
    - un **error de coincidencia** que indica cuán estrechamente la silueta de referencia coincide con el segmento de imagen de origen (13) y
    - un **error de consistencia** que indica cuánto la pose de referencia es consistente con la pose del mismo objeto del mundo real (14) de acuerdo con lo estimado a partir de por lo menos una de imágenes de origen precedentes y siguientes (10) de la corriente de vídeo;
- recuperar de las poses de referencia de los modelos de objeto articulado (4) asociadas con la seleccionada de las siluetas de referencia; y
- calcular una estimación de la pose del modelo de objeto articulado (4) a partir de las poses de referencia de las siluetas de referencia seleccionadas.

2. El método de acuerdo con la reivindicación 1, en el que el paso de comparar el por lo menos un segmento de imagen de origen (13) con las siluetas de referencia comprende los pasos de, para cada silueta de referencia con la que se compara el segmento de imagen de origen (13):

- determinar una **transformación proyectiva** que mapea el segmento de imagen de origen (13) sobre la silueta de referencia por medio del escalamiento del segmento de imagen de origen (13) para ser del mismo tamaño que la silueta de referencia; y
- calcular el **error de coincidencia** ya sea como proporcional al tamaño relativo del área de la imagen en la que el segmento de imagen de origen mapeado (13) y la silueta de referencia no se superponen, o como una medida de la distancia entre los contornos del segmento de imagen de origen escalado (13) y de la silueta de referencia, el error de coincidencia, en forma opcional, también puede depender de parámetros de la transformación proyectiva;
- y utilizar este error de coincidencia como una medida de cuán estrechamente coinciden el segmento de imagen de origen (13) y la silueta de referencia.

3. El método de acuerdo con la reivindicación 2, en el que

- el paso de escalamiento se logra por medio del re-muestreo de ya sea el segmento de imagen de origen (13) o las siluetas de referencia o ambos para tener cuadros delimitadores del mismo tamaño de píxel, y
- tanto el segmento de imagen de origen (13) como la silueta de referencia están representados por imágenes binarias que tienen la misma dimensión de píxel, y el cálculo del valor de error se lleva a cabo por medio del recuento del número de píxeles correspondientes desde el segmento de imagen de origen (13) y la silueta de referencia que difieren en valor.

4. El método de acuerdo con una de las reivindicaciones precedentes, en el que el paso de comparar el por lo menos un segmento de imagen de origen (13) con las siluetas de referencia comprende los pasos de, para cada silueta de referencia con la que se compara el segmento de imagen de origen (13):

- recuperar la pose del mismo objeto del mundo real (14) estimada a partir de una imagen de origen anterior (10) de la corriente de vídeo;
- calcular el **error de consistencia** como proporcional a la diferencia entre esta estimación de pose anterior y la pose de referencia de la silueta de referencia, y utilizar este error de consistencia como una medida de la consistencia con la imagen de origen anterior (10).

5. El método de acuerdo con una de las reivindicaciones precedentes, en el que el paso de calcular una estimación de la pose del modelo de objeto articulado (4) a partir de las poses de referencia de las siluetas de referencia seleccionadas comprende

- repetir los pasos anteriores para las una o más imágenes de origen adicionales (10) desde uno o más corrientes de vídeo adicionales de otras cámaras de origen (9), cada imagen de origen adicional (10) que comprende una vista del mismo objeto del mundo real (14) que se ha grabado al mismo tiempo, pero desde un punto de vista diferente, para obtener de esta manera para cada imagen de origen adicional (10) un número predeterminado de siluetas de referencia seleccionadas y poses de referencia seleccionadas asociadas;
- llevar a cabo una optimización para seleccionar para cada imagen de origen (10) una **pose de referencia más plausible**, por medio del cálculo para cada combinación de poses de referencia seleccionadas para las diferentes imágenes de origen (10) una **medida de consistencia de las articulaciones totales** por,

    - proyectar las articulaciones (2) de las poses de referencia seleccionadas de esta combinación en el espacio 3D, la estimación de una posición en 3D de las articulaciones (2) y calcular, para cada articulación, una **medida de consistencia de las articulaciones** que expresa lo bien que la posición de la articulación en 3D estimada coincide con la proyección de la articulación (2) de las poses de referencia seleccionadas;
    - combinar las medidas de consistencia de las articulaciones de todas las articulaciones para obtener el **medida de consistencia de las articulaciones totales;**

- seleccionar la combinación de las poses de referencia para las diferentes imágenes de origen (10) que optimiza la medida de consistencia de las articulaciones totales.

6. El método de acuerdo con la reivindicación 5, en el que el paso de llevar a cabo una optimización además comprende el paso de variar y optimizar un desplazamiento en 2D de cada silueta en el plano de su imagen de origen asociada (10) con el fin de corregir los errores de calibración de la cámara de origen (9) .

7. El método de acuerdo con una de las reivindicaciones precedentes, que comprende el paso adicional de exhibir, en un dispositivo de visualización, por lo menos una imagen de origen (10) con posiciones de la articulación estimadas superpuestas sobre la imagen de origen (10) y aceptar una entrada de usuario para modificar en forma interactiva una o más posiciones de la articulación.

8. Un método implementado por ordenador para la estimación de una pose de un **modelo de objeto articulado** (4) de acuerdo con una de las reivindicaciones precedentes, en el que, con el fin de determinar una pose en 3D que coincide con una pose en 2D dada asociada con una imagen de origen (10), se llevan a cabo los siguientes pasos:

- calcular, desde la pose en 2D determinada una pose en 3D aproximada que comprende posiciones de la articulación (2) aproximadas que coinciden aproximadamente con la posición de las articulaciones correspondientes (2) de la pose en 2D cuando se proyecta en el plano de imagen de la imagen de origen (10) asociada con la pose en 2D;
- modificar la pose en 3D aproximada para coincidir exactamente con la pose en 2D por, para cada articulación (2) mover la posición de la articulación (2) desde la posición de la articulación aproximada (2), a una posición definida por la intersección de un rayo que pasa desde la cámara a través de la posición de la articulación (2) en la imagen de origen (10) con un plano paralelo al plano de imagen de la imagen de origen (10) y que pasa a través de la posición de la articulación aproximada (2) .

9. Un método implementado por ordenador para la estimación de una pose de un **modelo de objeto articulado** (4) de acuerdo con una de las reivindicaciones precedentes, en el que el modelo de objeto articulado (4) es un modelo en 3D basado en ordenador (1) de un **objeto del mundo real** (14) observado por **dos** o más **cámaras de origen** (9), y el modelo de objeto articulado (4) representa una pluralidad de **articulaciones** (2) y de **enlaces** (3) que enlazan las articulaciones (2), y en el que la **pose** del modelo de objeto articulado (4) está definido por la localización espacial de las articulaciones (2), denominada posiciones de la articulación en 3D, el método comprende los pasos de

- determinar una estimación inicial de la pose en 3D, es decir, las posiciones de la articulación en 3D del modelo de objeto articulado (4);
- asociar cada enlace (3) con una o más **superficies de proyección** (5), en las que las superficies de proyección (5) son superficies definidas en el modelo en 3D, y la posición y la orientación de cada superficie de proyección (5) se determinan por la posición y la orientación del enlace asociado (3);
- adaptar en forma iterativa las posiciones de la articulación en 3D por, para cada articulación (2),

    - calcular un **puntaje de posición** asignado a su posición de la articulación en 3D, el puntaje de posición es una medida del grado en el que los segmentos de imagen desde las diferentes cámaras de origen (9), cuando se proyectan sobre las superficies de proyección (5) de enlaces (3) adyacentes a la articulación (2), que son consistentes entre sí;
    - variar la posición de la articulación en 3D de la articulación (2) hasta que se logre un puntaje de posición óptimo;

- repetir el paso de adaptar en forma iterativa las posiciones de la articulación en 3D para todas las articulaciones (2) durante un número predeterminado de veces o hasta que los puntajes de posición converjan.

10. El método de acuerdo con la reivindicación 9, en el que el paso de variar la posición de la articulación en 3D de las articulaciones (2) varía las posiciones de la articulación en 3D sujetas a limitaciones antropométricas, por lo menos una de las limitaciones antropométricas es:

- la articulación está en o por encima del suelo;
- las longitudes de los enlaces topológicamente simétricos no difieren más de 10%;
- las longitudes de los enlaces están dentro de los estándares antropométricos;
- las distancias entre las articulaciones que no están conectadas por un enlace están dentro de los estándares antropométricos.

11. El método de acuerdo con la reivindicación 9 o 10, en el que las superficies de proyección (5), para cada enlace (3), comprenden un **ventilador (7) de carteleras** (6), cada **cartelera** (6) está asociada con una cámara de origen (9), y cada cartelera es una superficie plana atravesada por su enlace asociado (3) y un vector que es normal tanto a este enlace (3) como a una línea que conecta un punto de la enlace (3) con la cámara de origen (9).

12. El método de acuerdo con la reivindicación 9 o 10 o 11, en el que el **puntaje de posición** de una posición de la articulación en 3D de una articulación (2) se calcula por medio de los pasos de, para cada enlace (3) adyacente a la articulación (2),

- proyectar las imágenes de las diferentes cámaras de origen (9) sobre las superficies de proyección asociadas (5) del enlace (3) y desde allí hacia una imagen virtual (12) de acuerdo con lo observado por una cámara virtual (11);
- para un área (12) que corresponde a la proyección de estas superficies de proyección (5) en la imagen virtual (12), calcular una **puntaje de posición parcial** para este enlace de acuerdo con el grado en el que los segmentos de imagen de las distintas cámaras de origen (9) se solapan y tienen un color similar;
- la combinación de los puntajes de posición parciales para obtener el puntaje de posición.

13. El método de acuerdo con la reivindicación 12, en el que el cálculo y la combinación del puntaje de posición parcial comprende los pasos de

- calcular el puntaje de posición parcial para cada par de cámaras de origen (9) que contribuyen a la imagen virtual (12);
- combinar estos puntajes de posición parciales por medio de la adición de los mismos, por medio del pesaje de cada puntaje de posición parcial de acuerdo con el ángulo entre las direcciones de visión del par asociado de cámaras de origen (9).

14. Un método implementado por ordenador para la determinación de una segmentación de un segmento de imagen de origen (13) en combinación con el método de acuerdo con una de las reivindicaciones precedentes, el método comprende los pasos de

- obtener por lo menos una **imagen de origen** (10) a partir de una **corriente de vídeo** que comprende una vista de un objeto del mundo real (14) grabada por una cámara de origen (9);
- procesar la por lo menos una imagen de origen (10) para extraer un **segmento de imagen de origen** (13) correspondiente que comprende la vista del objeto del mundo real (14) separado del fondo de la imagen;
- mantener en una base de datos en una forma legible por ordenador, un conjunto de **siluetas de referencia**, cada silueta de referencia está asociada con una **segmentación de referencia,** la segmentación de referencia define sub-segmentos de la silueta de referencia, a cada sub-segmento se le asigna una etiqueta única;
- determinar una silueta de referencia coincidente que se asemeja más estrechamente al segmento de imagen de origen (13) y recuperar la segmentación de referencia de la silueta de referencia;
- para cada sub-segmento, superponer tanto una versión engrosada como afinada del sub-segmento sobre el segmento de imagen de origen (13) y etiquetar los píxeles de la imagen de origen que se encuentran tanto en la versión engrosada como afinada con la etiqueta de el sub-segmento;
- etiquetar todos los píxeles restantes del segmento de imagen de origen (13) como inseguros;
- para cada sub-segmento, determinar un modelo de color que sea representativo del color de los píxeles etiquetados con la etiqueta del sub-segmento;
- etiquetar los píxeles inseguros, de acuerdo con el modelo de color, por medio de la asignación de cada píxel inseguro a un sub-segmento cuyo modelo de color se ajusta más estrechamente al color del píxel inseguro.

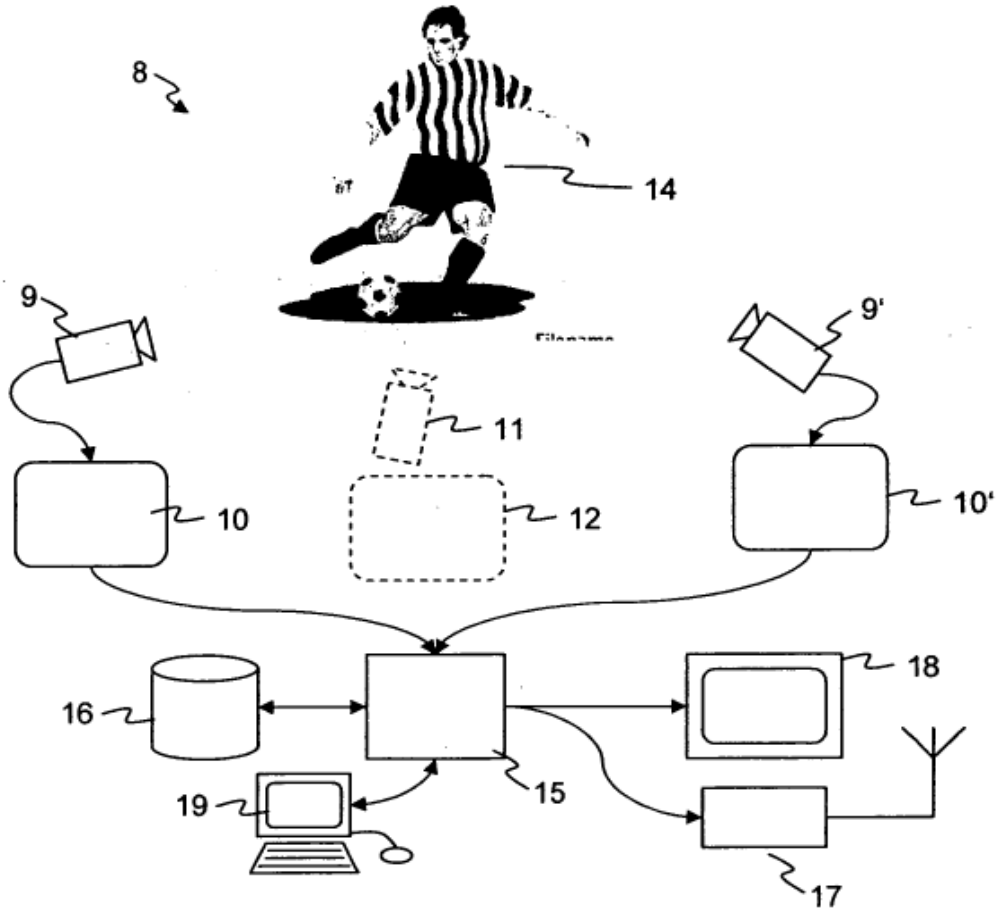

Fig. 1

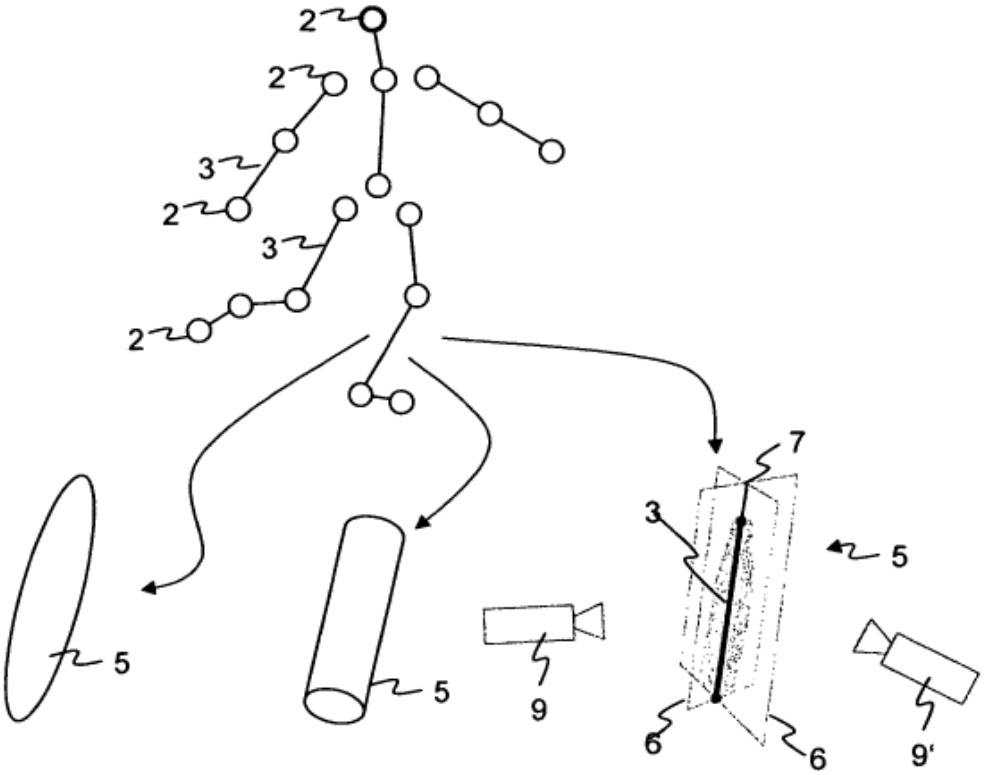

Fig. 2

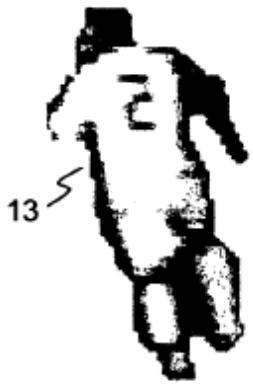

Fig. 3a

Fig. 3b

Fig. 3c

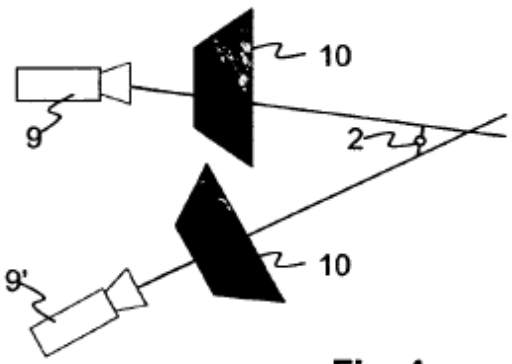

Fig. 4

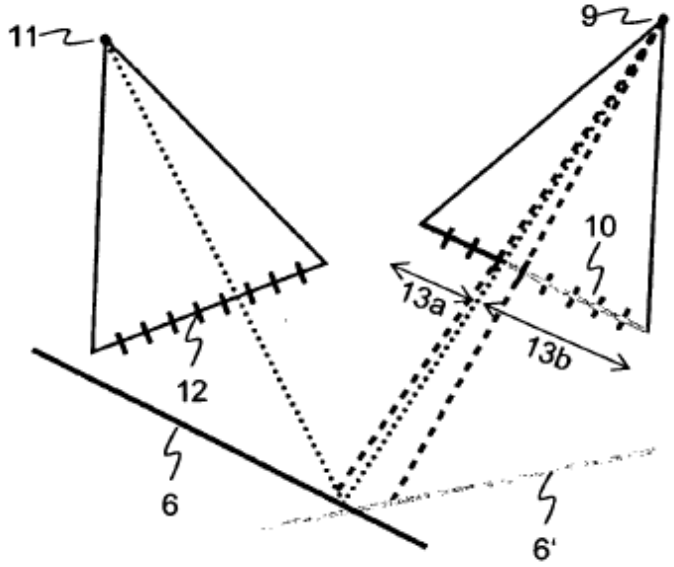

Fig. 8a

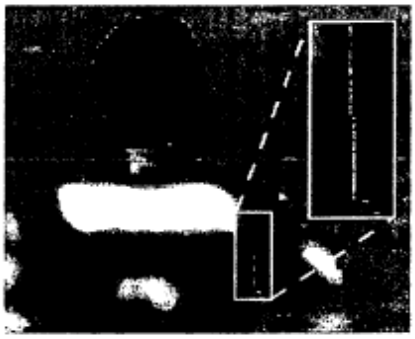

Fig. 8b

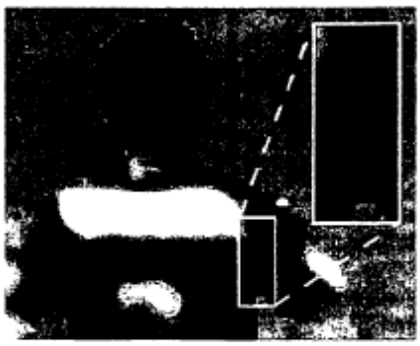

Fig. 8c

Fig. 5a

Fig. 5b

Fig. 5c

Fig. 6a

Fig. 6b

Fig. 6c

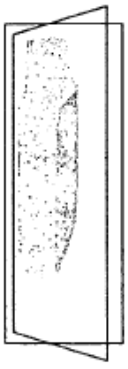

Fig. 7a

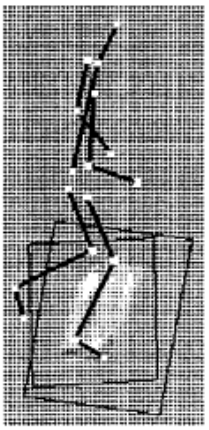

Fig. 7b

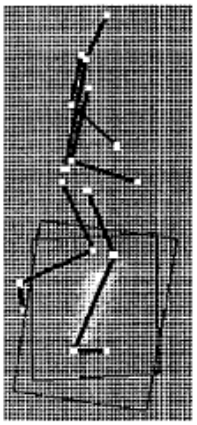

Fig. 7c

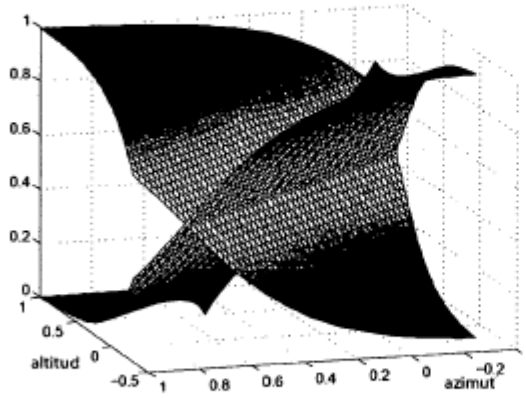

Fig. 9a

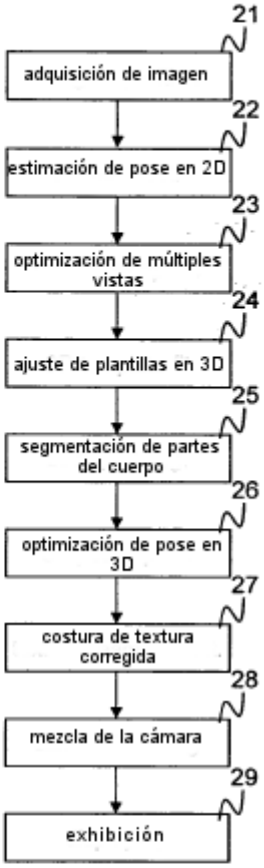

Fig. 10

Fig. 9b

Fig. 9c

Fig. 9d