(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 556 601**

(51) Int. CI.:

H04N 5/262 (2006.01)
H04N 5/268 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **07.05.2010 E 10737234 (4)**

(97) Fecha y número de publicación de la concesión europea: **16.09.2015 EP 2428036**

(54) Título: **Sistemas y métodos para la producción autónoma de vídeos a partir de múltiples datos detectados**

(30) Prioridad:

07.05.2009 GB 0907870

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
19.01.2016

(73) Titular/es:

KEEMOTION S.A. (100.0%)
Rue Louis de Geer 6
1348 Louvain-la-Neuve, BE

(72) Inventor/es:

DE VLEESCHOUWER, CHRISTOPHE y
CHEN, FAN

(74) Agente/Representante:

LINAGE GONZÁLEZ, Rafael

DESCRIPCIÓN

Sistemas y métodos para la producción autónoma de vídeos a partir de múltiples datos detectados

Campo de la invención

La presente invención se refiere a la integración de información a partir de múltiples cámaras en un sistema de vídeo, por ejemplo una producción de televisión o sistema de vigilancia inteligente, y a la producción automática de contenido de vídeo, por ejemplo para representar una acción que implica una o varias personas y/u objetos de interés.

Antecedentes técnicos

El proyecto APIDIS (Producción Autónoma de Imágenes basándose en la Detección Distribuida e Inteligente) intenta proporcionar una solución para generar contenidos personalizados para representación visual mejorada y de bajo coste de escenarios controlados tales como la televisión deportiva, donde la calidad de imagen y la comodidad perceptual son tan esenciales como la integración eficaz de información contextual [1].

En el contexto APIDIS, múltiples cámaras están distribuidas alrededor de la acción de interés, y la producción autónoma de contenido implica tres cuestiones técnicas principales con respecto a estas cámaras:

(i) cómo seleccionar puntos de vista óptimos, es decir parámetros de recorte en una cámara dada, de modo que puedan adaptarse a resolución de visualización limitada,

(ii) cómo seleccionar la cámara correcta para representar la acción en un momento dado, y

(iii) cómo suavizar secuencias de cámara/punto de vista para eliminar artefactos de producción.

Los artefactos de producción consisten tanto en artefactos visuales, que principalmente significan efectos de parpadeo debido a temblores o acercamiento/alejamiento rápido de puntos de vista, como en artefactos de narración tales como la discontinuidad de la historia producida por cambio de cámara rápido y movimientos de punto de vista drásticos.

La fusión de datos de múltiples cámaras se ha analizado ampliamente en la bibliografía. Estos trabajos anteriores podrían clasificarse de manera aproximada en tres categorías principales de acuerdo con sus diversos fines. Los métodos en la primera categoría tratan de la calibración de la cámara y control de cámara inteligente integrando información contextual del entorno de múltiples cámaras [4]. Reconstrucción de escena en 3D [5] o la síntesis de vídeo de punto de vista arbitrario [2] desde múltiples cámaras es también un tema candente. La tercera categoría usa múltiples cámaras para resolver ciertos problemas tales como oclusión en diversas aplicaciones, por ejemplo, seguimiento de personas [6]. Todos estos trabajos se centran mucho en la extracción de información contextual en 3D importante, pero consideran poco las cuestiones técnicas anteriormente mencionadas acerca de producción de vídeo.

Con respecto a producción de vídeo autónoma, existen algunos métodos propuestos en la bibliografía para seleccionar el área más representativa desde una imagen independiente. Suh y otros [7] definen la región de recorte óptima como el rectángulo mínimo que contiene notabilidad sobre un umbral dado, donde la notabilidad se calculó mediante el modelo de atención visual [8]. En la Ref. [9], se propuso otro método basado en modelo de atención, donde analizaron más la trayectoria de desplazamiento óptima de atención que la decisión del punto de vista. Se conoce también cómo aprovechar una red distribuida de cámaras para aproximar las imágenes que se capturarían mediante un sensor virtual localizado en una posición arbitraria, con cobertura de punto de vista arbitraria. Para pocas cámaras con lentes bastantes heterogéneas y cobertura de escena, la mayoría de los métodos de síntesis de punto de vista libres del estado de la técnica producen resultados borrosos [2][3].

En la referencia [10] se propone un sistema de producción automática para vídeos deportivos de fútbol y se analizó también la selección de punto de vista basándose en la comprensión de escena. Sin embargo, este sistema únicamente cambia puntos de vista entre tres tamaños de toma fijadas de acuerdo con varias reglas fijadas, que conduce a artefactos visuales molestos debido al cambio drástico de los tamaños de toma. Adicionalmente, únicamente analizaban el caso de la cámara única.

Además del estudio anterior de la bibliografía, varias solicitudes de patente han considerado sistemas de múltiples cámaras (omnidireccionales) para producir y editar contenido de vídeo de una manera semi-automática. Pueden identificarse tres categorías principales de sistemas.

La primera categoría selecciona una vista (es decir, un vídeo) entre los cubiertos mediante un conjunto de cámaras predefinido, basándose en algún mecanismo de detección de actividad. En [15], cada cámara se activa basándose en algún dispositivo externo, que acciona la adquisición de vídeo cada vez que se detecta un evento particular (por

ejemplo, un objeto que entra en el campo de visión). En [16], se usan sensores de audio para identificar la dirección en la que el vídeo debería capturarse.

La segunda categoría captura una señal visual rica, basándose en cámaras omnidireccionales o en ajuste de múltiples cámaras de gran angular, para ofrecer cierta flexibilidad en la manera en la que se representa la escena en el extremo de receptor. Por ejemplo, los sistemas en [17] y [18] respectivamente consideran sistemas de visualización de múltiples cámaras y omnidireccionales para capturar y difundir corrientes de vídeo de gran angular. En [17], una interfaz permite al observador monitorizar la o las corriente de vídeo de gran angular para seleccionar qué porción del vídeo sacar en tiempo real. Además, el operador puede detener la reproducción y controlar efectos de panorámica-inclinación-zoom en un fotograma particular. En [18], la interfaz se mejora basándose en la detección automática de las áreas de vídeo en las que está presente un participante de evento. Por lo tanto, el observador tiene la oportunidad de elegir de manera interactiva a qué participante(s) de evento desearía mirar.

De manera similar, [19-21] detecta personas de interés en una escena (típicamente un conferencista o un participante de videoconferencia). Sin embargo, la mejora sobre [18] es doble. En primer lugar, en [19-21] se proponen métodos para definir automáticamente un conjunto de tomas candidatas basándose en análisis automático de la escena. En segundo lugar, se definen mecanismos para seleccionar automáticamente una toma entre las tomas candidatas. En [19], la definición de toma se basa en la detección y seguimiento del conferencista, y se usan reglas probabilísticas para cambiar pseudo-aleatoriamente del público a la cámara del conferencista durante una charla. En [20] y [21], se define también una lista de tomas candidatas basándose en la detección de algún objeto de interés particular (típicamente una cara), pero se consideran efectos de edición más sofisticados para crear una representación dinámica (videoconferencia). Por ejemplo, una toma puede hacer panorámica desde una persona a otra, o varias caras pueden pegarse una junto a la otra en una única toma. El vídeo de salida editado se construye a continuación seleccionando una mejor toma entre las tomas candidatas para cada escena (en [20] y [21], una escena corresponde a un periodo de tiempo particular). La mejor toma se selecciona basándose en un conjunto pre-definido de reglas cinemáticas, por ejemplo, para evitar demasiado de la misma toma en una fila.

Vale la pena destacar que los parámetros de toma (es decir, los parámetros de recorte en la vista disponible) permanecen fijos hasta que se cambia la cámara. Además, en [19-21] una toma está asociada directamente a un objeto, de modo que, al final, la selección de toma finaliza al seleccionar el objeto u objetos a representar, que puede ser difícil e irrelevante en contextos que son más complejos que una videoconferencia o una conferencia. Específicamente, [19-21] no seleccionan la toma basándose en el procesamiento conjunto de las posiciones de los múltiples objetos.

La tercera y última categoría de sistemas de producción de vídeo semi-automáticos diferencia las cámaras que están dedicadas a análisis de escena de las que se usan para capturar las secuencias de vídeo. En [22], se usa una rejilla de cámaras para fines de análisis de escena deportiva. Las salidas del módulo de análisis se aprovechan a continuación para calcular estadísticas acerca del juego, pero también para controlar cámaras de panorámica-inclinación-zoom (PTZ) que recopilan vídeos de jugadores de interés (típicamente el que sujeta el disco o la pelota). [22] debe implementar todos los algoritmos de análisis de escena en tiempo real, puesto que tiene por objeto controlar los parámetros de PTZ de la cámara instantáneamente, como una función de la acción observada en la escena. Más importante y fundamentalmente, [22] selecciona los parámetros de PTZ para capturar un objeto detectado específico y no ofrecer la representación apropiada de una acción de equipo, compuesta potencialmente de múltiples objetos de interés. En esto es similar a [19-21]. También, cuando se recopilan múltiples vídeos, [22] no proporciona ninguna solución para seleccionar uno de ellos. Solamente remite todos los vídeos a una interfaz que los presenta de una manera integrada a un operador humano. Este es el origen de un cuello de botella cuando se consideran muchas cámaras de origen.

El documento US 2008/0129825 desvela el control de cámara motorizada para capturar imágenes de un objeto seguido individual, por ejemplo para deportes individuales como competiciones de atletismo. El usuario selecciona la cámara a través de una interfaz de usuario. Las unidades de localización se unen al objeto. Por lo tanto son intrusivas.

El documento GB 2402011 desvela un control de cámara automatizado que usa parámetros de eventos. Basándose en el seguimiento de jugador y un conjunto de reglas de accionamiento, el campo de visión de las cámaras se adapta y cambia entre las vistas cercana, media y lejana. Se selecciona una cámara basándose en eventos de accionamiento. Un evento de accionamiento típicamente corresponde a movimientos específicos o acciones de deportistas, por ejemplo el servicio de un jugador de tenis, o actualizaciones de información del marcador.

El documento US 2004/0105004 A1 se refiere a representar ponencias o reuniones. Las cámaras de seguimiento se aprovechan para representar al presentador o a un miembro del público que pregunta una cuestión. El presentador y los miembros del público se siguen basándose en una localización de fuente de sonido, usando un conjunto de micrófonos. Dada la posición de la cámara de seguimiento objetivo, los parámetros de PTZ de la cámara motorizada se controlan para proporcionar un vídeo editado suave del objetivo. El método y sistema descritos son únicamente adecuados para seguir a una única persona individual. Con respecto a la selección de la cámara, se desvela el cambio entre un conjunto de vistas muy distintas (una vista general de la sala, una vista de las diapositivas, una vista

cercana del presentador y una vista cercana de un miembro del público que habla). El proceso de selección de cámara se controla basándose en la detección de evento (por ejemplo, una nueva aparición de diapositiva, o un miembro del público que habla) y reglas de videografía definidas por profesionales, para emular un equipo humano de producción de vídeo.

El documento US 5745126 desvela la selección dinámica automatizada de una cámara de vídeo/imagen desde múltiples cámaras de vídeo/imágenes reales (o virtuales) de acuerdo con una perspectiva particular, un objeto en la escena o un evento en la escena de vídeo.

Referencias

[1] Homepage of the APIDIS project, *http://www.apidis.org/*
Vídeos de demostración relacionados con este artículo: *http://www.apidis.org/Initial Results/APIDIS%20Initial%20Results.htm*

[2] S. Yaguchi y H. Saito, Arbitrary viewpoint video synthesis from multiple uncalibrated cameras, IEEE Trans. Syst. Man. Cybern. B, 34(2004) 430-439.

[3] N. Inamoto, y H. Saito, Free viewpoint video synthesis and presentation from multiple sporting videos, Electronics and Communications in Japan (Part III: Fundamental Electronic Science), 90(2006) 40-49.

[4] I.H. Chen, y S.J. Wang, An efficient approach for the calibration of multiple PTZ cameras, IEEE Trans. Automation Science and Engineering, 4(2007) 286-293.

[5] P. Eisert, E. Steinbach y B. Girod, Automatic reconstruction of stationary 3-D objects from multiple uncalibrated camera views, IEEE Trans. Circuits and Systems for Video Technology, Special Issue on 3D Video Technology, 10(1999) 261-277.

[6] A. Tyagi, G. Potamianos, J.W. Davis y S.M. Chu, Fusion of Multiple camera views for kernel-based 3D tracking, WMVC'07, 1(2007) 1-1.

[7] B. Suh, H. Ling, B.B. Bederson, y D.W. Jacobs, Automatic thumbnail cropping and its effectiveness, Proc. ACM UIST 2003, 1(2003) 95-104.

[8] L. Itti, C. Koch, y E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Analysis and Machine Intelligence, 20(1998) 1254-1259.

[9] X. Xie, H. Liu, W.Y. Ma, H.J. Zhang, "Browsing large pictures under limited display sizes, IEEE Trans. Multimedia, 8(2006) 707-715.

[10] Y. Ariki, S. Kubota, y M. Kumano, Automatic production system of soccor sports video by digital camera work based on situation recognition, ISM'06,1(2006) 851-860.

[11] J. Owens, Television sports production, 4ª edición, Focal Press, 2007.

[12] J.W. Gibbs, Elementary principles in statistical mechanics, Ox Bow Press, 1981.

[13] D. Chandler, Introduction to modem statistical mechanics, Oxford University Press, 1987.

[14] C. De Vleeschouwer, F. Chen, D. Delannay, C- Parisot, C. Chaudy, E. Martrou, y A. Cavallaro, Distributed video acquisition and annotation for sport-event summarization, NEM summit, (2008).

[15] Documento EP 1289282 (AI) Video sequence automatic production method and system Inventor: AYER SERGE [CH]; MOREAUX MICHEL [CH] (+1); Solicitante: DARTFISH SA [CH]; EC: H04N5/232 IPC: *H04N5/232*; H04N5/232; (IPC1-7): H04N5/232

[16] Documento US 20020105598, documento EP 1352521 AUTOMATIC MULTI-CAMERA VIDEO COMPOSITION; INTEL CORP

[17] Documento US 6741250 Method and system for generation of multiple viewpoints into a scene viewed by motionless cameras and for presentation of a view path; BE HERE CORP

[18] Documento US 20020191071 Automated online broadcasting system and method using an omni-directional camera system for viewing meetings over a computer network; MICROSOFT CORP

[19] Documento US 20020196327 Automated video production system and method using expert video production

rules for online publishing of lectures; MICROSOFT CORP; Microsofit Corporation

[20] Documento US 20060251382 A1 System and method for automatic video editing using object recognition MICROSOFT CORP

[21] Documento US 20060251384 Automatic video editing for real-time multi-point video conferencing; MICROSOFT CORP

[22] Documento WO 200599423 AUTOMATIC EVENT VIDEOING, TRACKING AND CONTENT GENERATION SYSTEM; AMAN JAMES A; BENNETT PAUL MICHAEL

[23] Documento US 5745126 Machine synthesis of a virtual video camera/image of a scene from multiple video cameras/images of the scene in accordance with a particular perspective on the scene, an object in the scene, or an event in the scene.

Aspectos de la presente invención

Un objeto de la presente invención es proporcionar métodos y sistemas basados en ordenador para la producción autónoma de un vídeo editado, compuesto basándose en las múltiples corrientes de vídeo capturadas mediante una red de cámaras, distribuidas alrededor de una escena de interés.

La presente invención proporciona un método y un sistema autónomos basados en ordenador para producción personalizada de vídeos tales como vídeos de deporte en equipo tales como vídeos de baloncesto a partir de múltiples datos detectados bajo resolución de visualización limitada. Sin embargo la invención tiene un alcance de aplicación más amplio y no está limitada a solamente este ejemplo. Las realizaciones de la presente invención se refieren a la selección de una vista para presentar de entre las múltiples corrientes de vídeo capturadas mediante la red de cámaras. Las soluciones técnicas se proporcionan para proporcionar comodidad perceptual así como una integración eficaz de información contextual, que se implementa, por ejemplo, suavizando secuencias de punto de vista/cámara generadas para mitigar los artefactos visuales de parpadeo y artefactos de narrativa discontinuos. Se desvela un diseño e implementación del proceso de selección de punto de vista que se ha verificado mediante experimentos, que muestra que el método y sistema de la presente invención distribuyen de manera eficaz la carga de procesamiento a través de las cámaras, y selecciona de manera eficaz puntos de vista que cubren la acción del equipo disponible mientras evita artefactos perceptuales principales.

Por consiguiente la presente invención proporciona un método basado en ordenador que comprende las etapas de la reivindicación 1.

La selección de parámetros de representación puede ser para todos los objetos u objetos de interés simultáneamente. El conocimiento acerca de la posición de los objetos en las imágenes puede aprovecharse para decidir cómo representar la acción capturada. El método puede incluir seleccionar parámetros de campo de visión para la cámara que representa la acción como una función de tiempo basándose en un equilibrio óptimo entre métricas de cercanía y completitud. Por ejemplo, los parámetros de campo de visión se refieren al corte en la vista de cámara de cámaras estáticas y/o a la panorámica-inclinación-zoom o parámetros de desplazamiento para cámaras dinámicas y potencialmente en movimiento.

Las métricas de cercanía y completitud pueden adaptarse de acuerdo con preferencias de usuario y/o recursos. Por ejemplo, un recurso de usuario puede ser resolución de codificación. Una preferencia de usuario puede ser al menos una de objeto preferido o cámara preferida. Las imágenes desde todas las vistas de todas las cámaras pueden mapearse a las mismas coordenadas temporales absolutas basándose en una referencia temporal única común para todas las vistas de cámara. En cada instante de tiempo, y para cada vista de cámara, se seleccionan parámetros de campo de visión que optimizan el equilibrio entre completitud y cercanía. El punto de vista seleccionado en cada vista de cámara puede puntuarse de acuerdo con la calidad de su equilibrio de completitud/cercanía, y a su grado de oclusiones. Para el segmento temporal disponible, los parámetros de una cámara virtual óptima que realiza panorámica, zoom y cambia a través de las vistas pueden calcularse para conservar altas puntuaciones de puntos de vista seleccionados mientras se minimiza la cantidad de movimientos de cámara virtual.

El método puede incluir seleccionar el campo de visión óptimo en cada cámara, en un instante de tiempo dado.

Un campo de visión v_k en la $k^{ésima}$ vista de cámara se define mediante el tamaño S_k y el centro c_k de la ventana que se recorta en la $k^{ésima}$ vista para visualización real. Se selecciona para incluir los objetos de interés y para proporcionar una descripción de alta resolución de los objetos, y se selecciona un campo de visión óptimo v_k^* para maximizar una suma ponderada de objetos interesantes como sigue

$$\mathbf{v}_k^* = \arg\max_{\{S_k, \mathbf{c}_k\}} \sum_{n=1}^{N} I_n \cdot \alpha(S_k, \mathbf{u}) \cdot m\left(\frac{\left\| \mathbf{x}_{n,k} - \mathbf{c}_k \right\|}{S_k} \right)$$

donde, en la ecuación anterior:

• I_n indica el nivel de interés asignado al n^{ésimo} objeto detectado en la escena.

• $x_{n,k}$ indica la posición del n^{ésimo} objeto en la vista de cámara *k*.

• La función m(…..) modula los pesos del n^{ésimo} objeto de acuerdo con su distancia al centro de la ventana del punto de vista, en comparación con el tamaño de esta ventana.

• El vector *u* refleja las preferencias de usuario, en particular, su componente u_{res} define la resolución de la corriente de salida, que está generalmente restringida por el ancho de banda de transmisión o la resolución del dispositivo del usuario final.

• La función α(.) refleja la penalización inducida por el hecho de que la señal nativa capturada mediante la k^{ésima} cámara tiene que sub-muestrearse una vez que el tamaño del punto de vista se hace mayor que la resolución máxima u_{res} permitida mediante el usuario.

Preferentemente α(....) se reduce con Sk y la función α(....) es igual a uno, cuando $S_k < u_{res}$, y se reduce posteriormente. α(....) se define mediante:

$$\alpha(S, \mathbf{u}) = \left[\min\left(\frac{u_{res}}{S}, 1 \right) \right]^{u_{cercano}},$$

donde el exponente $u_{cercano}$ es mayor que 1, y aumenta a medida que el usuario prefiere representación a resolución completa de área de acercamiento, en comparación con puntos de vista grades pero sub-muestreados.

El método incluye puntuar el punto de vista asociado a cada cámara de acuerdo con la calidad de su equilibrio de completitud/cercanía, y a su grado de oclusiones. La puntuación más alta debería corresponder a una vista que (1) hace a la mayoría del objeto de interés visible, y (2) está cerca a la acción, que significa que presenta objetos importantes con muchos detalles, es decir a alta resolución. Formalmente, dado el interés I_n de cada jugador, la puntuación $I_k(v_k,u)$ asociada a la k^{ésima} vista de cámara se define como sigue:

$$I_k(\mathbf{v_k}, \mathbf{u}) = \sum_{n=1}^{N} I_n \cdot o_k\left(\mathbf{x}_n \middle| \bar{\mathbf{x}} \right) \cdot h_k(\mathbf{x}_n) \cdot \beta_k(S_k, \mathbf{u}) \cdot m\left(\frac{\left\| \mathbf{x}_{n,k} - \mathbf{c}_k \right\|}{S_k} \right)$$

donde, en la ecuación anterior:

■ I_n indica el nivel de interés asignado al n^{ésimo} objeto detectado en la escena.

■ x_n indica la posición del n^{ésimo} objeto en el espacio en 3D;

■ $o_k(x_{n|}\ x)$ mide la relación de oclusión del n^{ésimo} objeto en la vista de cámara *k*, conociendo la posición de todos los otros objetos, definiéndose la relación de oclusión de un objeto para que sea la fracción de píxeles del objeto que se ocultan por otros objetos cuando se proyectan en el sensor de cámara;

■ La altura $h_k(x_n)$ se define para que sea la altura en píxeles de la proyección en la vista k de una altura de referencia de un objeto de referencia localizado en x_n. El valor de $h_k(x_n)$ se calcula directamente basándose en la calibración de la cámara, o cuando la calibración no está disponible, puede estimarse basándose en la altura del objeto detectado en la vista k.

■ La función $\beta_k(.)$ refleja el impacto de las preferencias de usuario en términos de vista de cámara y resolución de

visualización. $\beta_k(.)$ se define como

$$\beta_k(S,\mathbf{u}) = u_k \cdot \alpha(S,\mathbf{u}),$$

donde u_k indica el peso asignado a la k^{ésima} cámara, y $\alpha(S,u)$ se ha definido anteriormente.

El método puede comprender suavizar la secuencia de índices de cámara y parámetros de punto de vista correspondientes, en el que el proceso de suavizado se implementa, por ejemplo, basándose en dos Campos Aleatorios de Markov, mecanismo de filtrado de paso bajo lineal o no lineal, o mediante un formalismo de modelo de grafos, resuelto basándose en el algoritmo Viterbi convencional.

La captura de las múltiples corrientes de vídeo puede ser mediante cámaras estáticas o dinámicas.

La presente invención incluye también un sistema basado en ordenador que comprende las características de la reivindicación 11.

El sistema basado en ordenador puede tener:

• Medios para detectar objetos/personas de interés en las imágenes de las corrientes de vídeo, por ejemplo conociendo sus coordenadas del mundo en 3D reales.

• Medios para seleccionar para cada cámara el campo de visión que representa la escena de interés de manera que (permite al observador) seguir la acción llevada a cabo mediante los múltiples objetos/personas que interactúan que se han detectado. Los parámetros del campo de visión se refieren, por ejemplo, a la ventana de recorte en una cámara estática y/o a los parámetros de panorámica-inclinación-zoom y de posición en una cámara motorizada y en movimiento. El concepto de *acción que sigue* puede cuantificarse midiendo la cantidad de píxeles asociados a cada objeto/personas de interés en la imagen presentada. El seguimiento preciso de la acción resulta de la representación completa y cercana, donde la completitud cuenta el número de objetos/personas en la imagen visualizada, mientras la cercanía mide la cantidad de píxeles disponibles para describir cada objeto.

• Medios para montar el vídeo editado seleccionando y concatenando segmentos de vídeo proporcionados mediante una o más cámaras individuales, de manera que maximizan las métricas de completitud y cercanía a lo largo del tiempo, mientras suavizan la secuencia de parámetros de representación asociados a los segmentos concatenados.

La presente invención proporciona también un producto de programa informático que comprende segmentos de código que cuando se ejecutan en un motor de procesamiento ejecutan cualquiera de los métodos de la invención o implementan cualquier sistema de acuerdo con la invención.

La presente invención incluye también un medio de almacenamiento de señal legible por máquina no transitorio que almacena el producto de programa informático.

La presente invención puede tratar con escenas que implican varias personas/objetos de interés en movimiento. A continuación, estas escenas se indican como acciones de equipo, y corresponden típicamente a las escenas encontradas en contexto de deportes de equipo.

Automatizar el proceso de producción permite:

• reducir los costes de producción, evitando procesos hechos a mano largos y tediosos, tanto para el control de cámara como la selección de cámara;

• aumentar el ancho de banda de producción y calidad, manejando potencialmente un número infinito de cámaras simultáneamente;

• crear contenido personalizado, repitiendo el proceso de producción varias veces, con parámetros distintos.

Un objeto de la presente invención está dirigido a la producción semánticamente significativa, es decir mostrar la acción de interés, y contenidos perceptualmente cómodos desde múltiples datos detectados en bruto. El sistema de acuerdo con la presente invención está basado en ordenador, incluyendo memoria y un motor de procesamiento y es un sistema de producción computacionalmente eficaz, por ejemplo, basándose en un paradigma divide y vencerás (véase la Figura 15).

En unas realizaciones, el mejor campo de visión se calcula en primer lugar para cada cámara individual, y a continuación se selecciona la mejor cámara para representar la escena. Juntos el índice de cámara y su campo de

visión definen el *punto de vista* para representar la acción. Cuando la cámara está fija, la definición del campo de visión está limitada a un recorte de la imagen capturada mediante la cámara. Cuando la cámara está motorizada, el campo de visión resulta directamente de los parámetros de panorámica-inclinación-zoom de la cámara, y puede por lo tanto capturar una porción rectangular arbitraria del campo de luz que alcanza el centro de la cámara.

Para definir de una manera cuantitativa la noción de mejor campo de visión o mejor índice de cámara, la presente invención introduce tres conceptos importantes, que son "completitud", "cercanía" y "suavidad". Completitud establece la integridad de representación de acción. En el contexto de representación de acción de equipo, la completitud mide cómo de bien se incluyen los objetos/personas de interés en la escena (típicamente los jugadores que participan en un deporte de equipo) en la imagen visualizada. La cercanía define la precisión de descripción de detalle (típicamente la cantidad media de píxeles que están disponibles para representar las personas/objetos de interés), y la suavidad es un término que hace referencia a la continuidad de la selección del punto de vista. Equilibrando entre estos factores, se proporcionan métodos para seleccionar (como una función del tiempo) puntos de vista óptimos para ajustar la resolución de visualización y otras preferencias de usuario, y para suavizar estas secuencias para una narrativa continua y elegante. La presente invención es completamente autónoma y auto-regida, en el sentido de que puede seleccionar los píxeles para presentar sin ninguna intervención humana, basándose en un conjunto por defecto de parámetros de producción y en los resultados de sistemas de detección de personas. Pero la invención puede tratar también con preferencias de usuario, tal como el perfil de narrativa de usuario, y capacidades de dispositivo. Las preferencias de narrativa pueden resumirse en cuatro descriptores, es decir, grupo de objetos preferidos por el usuario o "equipo", objeto preferido por el usuario o "jugador", 'tipo de vista' preferida por el usuario (por ejemplo vistas de acercamiento cercano o alejamiento lejano), y "cámara" preferida por el usuario. Todas las restricciones de dispositivo, tal como resolución de visualización, velocidad de red, rendimiento de decodificador, se resumen como el parámetro de resolución de salida, que indica la resolución a la que se codifica el vídeo de salida a transportar y presentar en el anfitrión final.

La capacidad para tener en cuenta estas preferencias depende del conocimiento capturado acerca de la escena, por ejemplo, a través de herramientas de análisis de vídeo. Por ejemplo, se ha implementado una realización de la presente invención en "Detection and Recognition of Sports(wo)men from Multiple Views", D. Delannay, N. Danhier, y C. De Vleeschouwer, *Third ACM/IEEE International Conference on Distributed Smart Cameras*, Como, Italia, septiembre de 2009 para seguir y reconocer automáticamente los jugadores en movimiento en la escena de interés. Este documento se incluye como el Apéndice 2.

En primer lugar, en las realizaciones de la presente invención se considera un conjunto de cámaras que (parcialmente) cubren la misma área, que es probable que se activen simultáneamente basándose en cualquier mecanismo de detección de actividad que es otra ventaja importante de la presente invención sobre la técnica anterior. El fin de la invención, por lo tanto, no es seleccionar una vista de cámara basándose en el hecho de que se detectó alguna actividad en la vista. En su lugar, el objetivo es seleccionar a lo largo del tiempo la vista de la cámara y sus variaciones correspondientes en parámetros tales como parámetros de recorte o de PTZ, para representar mejor la acción que tiene lugar en el área cubierta. En este punto calidad de representación se refiere a la optimización de un equilibrio entre medidas de cercanía, completitud y suavidad.

En segundo lugar, la presente invención tiene una ventaja de adaptar dinámicamente y suavizar parámetros de puntos de vista con el tiempo, que es una mejora sobre sistemas de la técnica anterior en los que los parámetros de toma (por ejemplo, los parámetros de recorte en la vista disponible) permanecen fijos hasta que se cambia la cámara.

En tercer lugar, en las realizaciones de la presente invención no se hace una elección entre un objeto u otro, sino en su lugar se realiza una selección del punto de vista basándose en el procesamiento conjunto de las posiciones de los múltiples objetos que se han detectado. De acuerdo con las realizaciones de la presente invención se realiza una selección de la secuencia de puntos de vista que es óptima en la manera en que maximiza y suaviza las métricas de cercanía y completitud, por ejemplo, para todos los objetos simultáneamente.

Esas diferencias en comparación con la técnica anterior proporcionan beneficios significativos cuando tratan el problema de producción de contenido, por ejemplo, en un contexto de deporte de equipo. Permite principalmente seguir la acción de jugadores en movimiento y que interactúan, que no era posible basándose en métodos de la técnica anterior.

Preferentemente, los métodos y sistemas de la presente invención capturan y producen contenido automáticamente, sin la necesidad de procesos costosamente hechos a mano (no es necesario equipo técnico u operador de cámara).

Como una consecuencia de su rentabilidad, la presente invención tiene por objeto mantener la producción de contenido lucrativo incluso para públicos dirigidos de pequeño o medio tamaño. De esta manera, promueve la emergencia de mercados novedosos, ofreciendo una gran elección de contenidos que son de interés para un número relativamente pequeño de usuarios (por ejemplo, el resumen de un evento deportivo regional, una charla universitaria, o un día en la guardería).

Además, automatizar la producción posibilita personalización de acceso de contenido. Generar un vídeo personalizado simplemente consiste en (re-)ejecutar el proceso de producción con parámetros de entrada que corresponden a las preferencias específicas o restricciones expresadas por el usuario.

Un objeto de la presente invención es producir un informe de vídeo de un evento basándose en la concatenación de segmentos de vídeo (y opcionalmente audio correspondiente) capturados mediante un conjunto de cámaras. En la práctica, tanto las cámaras estáticas como dinámicas pueden manipularse mediante la presente invención:

○ Usar sensores estáticos se añade a la rentabilidad puesto que permite almacenar todo el contenido relevante y procesarlo fuera de línea, para seleccionar los fragmentos de corrientes que merece la pena presentar al observador.

Los principios de producción autónoma descritos a continuación podrían también usarse para controlar una (conjunto de) cámara o cámaras de PTZ dinámicas. En ese caso, la información acerca de la localización de objetos de interés tiene que proporcionarse en tiempo real, por ejemplo, basándose en el análisis en tiempo real de la señal capturada mediante algún sensor audio-visual (como se hace en [ref]), o basándose en información recopilada desde transmisores embebidos. Además, el espacio de campos de visión candidatos se define mediante los parámetros de posición y de control de la cámara de PTZ, y no mediante la imagen recortada en el ángulo de visión cubierto mediante la cámara estática.

La principal suposición que subyace el ajuste de adquisición en red es la existencia de una referencia temporal única común para todas las vistas de cámara, de modo que las imágenes desde todas las cámaras pueden mapearse a las mismas coordenadas temporales absolutas de la escena disponible. Las cámaras se supone por lo tanto que están ligeramente, pero no necesariamente estrechamente, sincronizadas. En este punto, la sincronización ligera se refiere a un conjunto de cámaras que capturan imágenes independientemente, y se basa en indicaciones de tiempo para asociar las imágenes que se han capturado a instantes de tiempo similares, pero no necesariamente idénticos. En contraste, una sincronización estrecha se referiría a captura sincronizada de las imágenes mediante las cámaras, como se hace cuando se controla la adquisición mediante una señal de accionamiento común.

Para decidir acerca de cómo representar la acción de equipo disponible, la invención tiene que conocer la posición de objetos de interés en la escena. Este conocimiento puede ser una estimación (propensa a errores), y puede referirse a la posición de objetos en la escena en 3D, o a la posición de objetos en cada una de las vistas de cámara.

Esta información puede proporcionarse basándose en transmisores que son llevados por los objetos a seguir en la escena de interés. Este conocimiento puede proporcionarse también mediante una alternativa no intrusiva, por ejemplo mediante el aprovechamiento de un conjunto de señales de vídeo capturadas mediante una red de cámaras estáticas, por ejemplo las usadas para producción de reportaje de vídeo, para detectar y seguir los objetos de interés. El método se describe en “Detection and Recognition of Sports(wo)men from Multiple Views, D. Delannay, N. Danhier, y C. De Vleeschouwer, *Third ACM/IEEE International Conference on Distributed Smart Cameras*, Como, Italia, septiembre de 2009” que se incorpora en el presente documento por referencia en su totalidad. Se fundamenta en un modelo de referencia de segundo plano para identificar los píxeles que cambian en cada vista. Cuando se calibran las múltiples vistas, por ejemplo a través de un proceso fuera de línea, las máscaras de detección de cambio que se recopilan en cada vista pueden unirse, por ejemplo en una máscara de ocupación de superficie, para identificar la posición de objetos de interés en el espacio en 3D (véase por ejemplo el enfoque representado en la Figura 16). Pueden usarse a continuación filtros de partículas o técnicas basadas en grafos para enlazar apariciones del mismo objeto a lo largo de la línea de tiempo. Obsérvese que tal detección y técnicas de seguimiento son bien conocidas para los expertos en la materia, y no se describirán en detalle en el presente documento. La realización de estos algoritmos que se han implementado se describe en la referencia anterior, y ofrece la ventaja de manejar oclusiones de una manera computacionalmente eficaz.

Una vez que se conocen las posiciones de los objetos de interés, la invención soporta producción autónoma (= selección de puntos de vista a lo largo del tiempo) del contenido capturado mediante la red de cámaras estáticas[1]. El enfoque es genérico en el sentido de que puede integrar una gran variedad de preferencias de usuario incluyendo recursos de transmisión o de visualización, interés semántico (como jugador preferido), o preferencias de narrativa (que tratan de la manera preferida para visualizar la historia, por ejemplo, cámara preferida o factor de acercamiento).

A través de un periodo de tiempo dado, la presente invención tiene por objeto seleccionar la secuencia de puntos de vista que optimiza la representación de escena a lo largo del tiempo, con respecto a las personas/objetos de interés detectados. En este punto, un punto de vista se refiere a un índice de cámara y a la ventana que se recorta en esa vista de cámara particular, para visualización real.

La optimización de la secuencia de puntos de vista fundamenta un número de nociones y principios que pueden describirse como sigue.

En cada instante de tiempo, la optimización de la representación tiene que:

○ Maximizar la noción de *completitud*, que mide hasta qué punto los (píxeles de los) objetos de interés están incluidos y visibles en el punto de vista presentado. Opcionalmente esto implica minimizar el grado de oclusión de objeto, que mide la fracción de un objeto que está presente en la escena, pero está (por ejemplo, al menos parcialmente) oculto por otros objetos;

○ Maximizar la noción de *cercanía*, que se refiere a la precisión de detalles, es decir, la densidad de píxeles o resolución, cuando se representan los objetos de interés.

Estos dos objetivos son en ocasiones antagonistas. Por esta razón, los métodos y sistemas de acuerdo con las realizaciones de la presente invención proponen equilibrar la completitud y cercanía, opcionalmente como una función de preferencias de usuario individuales (por ejemplo, en términos de resolución de punto de vista, o cámara o jugadores preferidos).

Finalmente, la suavidad de las transiciones entre los parámetros de representación de fotogramas consecutivos del vídeo editado se tiene que tener en cuenta cuando se considera la producción de un segmento temporal. En otras palabras, es importante conservar la uniformidad entre la cámara y, por ejemplo, parámetros de recorte que se seleccionan a lo largo de la línea de tiempo, para evitar distraer al observador de la historia mediante cambios abruptos o parpadeo constante.

Basándose en estos principios de guiado, se ha desarrollado el proceso de tres etapas representado en la Figura 14. Puede describirse como sigue:

Etapa 1: en cada instante de tiempo, y para cada vista de cámara, seleccionar las variaciones en parámetros tales como parámetros de recorte que optimizan el equilibrio entre completitud y cercanía. Opcionalmente, el equilibrio de completitud/cercanía se mide como una función de las preferencias de usuario. Por ejemplo, dependiendo de la resolución a la que accede al contenido producido, un usuario puede preferir un punto de vista pequeño (acercamiento) o uno grande (alejamiento).

Etapa 2: puntuar el campo de visión seleccionado en cada vista de cámara de acuerdo con la calidad (en términos de preferencias de usuario) de su equilibrio de completitud/cercanía, y a su grado de oclusiones.

Etapa 3: Para el segmento temporal disponible, calcular los parámetros de una cámara virtual óptima que realiza panorámica, zoom y cambia a través de las cámaras para conservar altas puntuaciones de puntos de vista seleccionados mientras se minimiza la cantidad de movimientos de cámara virtuales.

La primera etapa consiste en seleccionar el campo de visión óptimo para cada cámara, en un instante de tiempo dado. Para simplificar las notaciones, a continuación, omitimos el índice de tiempo *t*.

Un campo de visión v_k en la $k^{ésima}$ cámara estática se define mediante el tamaño S_k y el centro c_k de la ventana que se recorta en la $k^{ésima}$ vista para visualización real.

Se ha de seleccionar para:

○ Incluir los objetos de interés;

○ Proporcionar una descripción precisa, es decir, a alta resolución, de estos objetos.

El campo de visión óptimo v_k^* se selecciona preferentemente de acuerdo con preferencias de usuario, para maximizar una suma ponderada de los intereses del objeto como sigue

$$\mathbf{v}_k^* = \underset{\{S_k, \mathbf{c}_k\}}{\arg\max} \sum_{n=1}^{N} I_n \cdot \alpha(S_k, \mathbf{u}) \cdot m\left(\frac{\left\|\mathbf{x}_{n,k} - \mathbf{c}_k\right\|}{S_k}\right)$$

En la ecuación anterior:

○ I_n indica el nivel de interés asignado al $n^{ésimo}$ objeto reconocido en la escena.

Esta asignación puede hacerse mediante cualquier método adecuado y la presente invención supone que esta asignación se ha completado y los resultados pueden usarse mediante la presente invención. Estos niveles de

interés pueden definirse mediante el usuario, por ejemplo una vez para todo el evento, y ponerse a disposición de la presente invención. En escenarios de aplicación para los que se detectan objetos pero no se etiquetan, el peso se omite, es decir, se sustituye mediante un valor unitario constante.

○ $x_{n,k}$ indica la posición del n^ésimo objeto en la vista de cámara *k*.

○ La función m(.) modula los pesos del n^ésimo objeto de acuerdo con su distancia al centro de la ventana de visualización, en comparación con el tamaño de esta ventana. De manera intuitiva, el peso debería ser alto y positivo cuando el objeto de interés está localizado en el centro de la ventana de visualización, y debería ser negativo o cero cuando el objeto radica fuera del área de visualización. Por lo tanto, m(.) debería ser positivo entre 0 y 0,5, e inferior o igual a cero por encima de 0,5. Son apropiadas muchas funciones, y la elección de un caso particular podría controlarse, por ejemplo, basándose en asuntos computacionales. Ejemplos de funciones son las funciones de sombrero mexicano o gaussiana bien conocidas. Se proporciona otro ejemplo en detalle en una realización particular de la invención descrito en el apéndice 1 de esta solicitud.

○ El vector *u* refleja las restricciones o preferencias de usuario en términos de resolución de ventana de visualización e índice de cámara. En particular, su componente u_{res} define la resolución de la corriente de salida, que está restringida generalmente por el ancho de banda de transmisión o resolución del dispositivo de usuario final. Su componente $u_{cercano}$ se establece a un valor mayor de 1 que aumenta para favorecer puntos de vista cercanos en comparación con vistas de alejamiento grandes. Los otros componentes de *u* tratan de preferencias de cámara, y se definen a continuación, mientras se describe la segunda etapa de la invención.

○ La función α(.) refleja la penalización inducida por el hecho de que la señal nativa capturada mediante la k^ésima cámara tiene que sub-muestrearse una vez que el tamaño del punto de vista se hace mayor que la resolución máxima u_{res} permitida mediante el usuario. Esta función típicamente se reduce con S_k. Una elección apropiada consiste en ajustar la función igual a uno cuando $Sk<u_{res}$, y hacerla reducir posteriormente. Un ejemplo de α(.) se define mediante

$$\alpha(S,\mathbf{u})=\left[\min\left(\frac{u_{res}}{S},1\right)\right]^{u_{cercano}},$$

donde el exponente $u_{cercano}$ es mayor de 1, y aumenta para favorecer puntos de vista cercanos en comparación con campos de visión de alejamiento grandes.

Vale la pena señalar que los equilibrios reflejados en la ecuación anterior pueden formularse en muchas maneras diferentes pero equivalentes. Un ejemplo de formulación alternativa, pero equivalente, se ha implementado en la realización de la invención definida en el apéndice 1. En esta formulación la suma del producto se ha sustituido por un producto de sumas, sin afectar fundamentalmente a la idea clave de la invención, que consiste en equilibrar cercanía y completitud de acuerdo con las restricciones del usuario (con respecto a resolución de salida) y preferencias (con respecto a puntos de vista de alejamiento o acercamiento).

La segunda etapa puntúa el punto de vista asociado a cada cámara de acuerdo con la calidad de su equilibrio de completitud/cercanía, y a su grado de oclusiones. La puntuación más alta debería corresponder a una vista que (1) hace a la mayoría del objeto de interés visible, y (2) está cerca de la acción, que significa que presenta objetos importantes con muchos detalles, es decir a alta resolución.

Formalmente, dado el interés I_n de cada jugador, la puntuación $I_k(vk, u)$ asociada a la k^ésima vista de cámara se define como sigue:

$$I_k(\mathbf{v_k},\mathbf{u})=\sum_{n=1}^{N} I_n\cdot o_k\left(\mathbf{x}_n\middle|\bar{\mathbf{x}}\right)\cdot h_k(\mathbf{x}_n)\cdot \beta_k(S_k,\mathbf{u}).m\left(\frac{\left\|\mathbf{x}_{n,k}-\mathbf{c}_k\right\|}{S_k}\right)$$

En la ecuación anterior:

■ I_n indica el nivel de interés asignado al n^ésimo objeto detectado en la escena.

■ x_n indica la posición del n^ésimo objeto en el espacio en 3D;

■ $o_k(x_n| x)$ mide la relación de oclusión del $n^{ésimo}$ objeto en la vista de la cámara *k*; conociendo la posición de todos los otros objetos. La relación de oclusión de un objeto se define para que sea la fracción de píxeles del objeto que se ocultan por otros objetos cuando se proyectan en el sensor de la cámara;

■ La altura $h_k(x_n)$ se define para que sea la altura en píxeles de la proyección en la vista k de un objeto vertical de 182,88 centímetros (seis pies) de alto localizado en x_n. 182,88 centímetros (seis pies) es la altura media de los jugadores.

El valor de $h_k(x_n)$ se calcula directamente basándose en la calibración de la cámara. Cuando la calibración no está disponible, puede estimarse basándose en la altura del objeto detectado en la vista k.

■ La función $\beta_k(.)$ refleja el impacto de las preferencias de usuario en términos de vista de cámara y resolución de visualización. Formalmente, $\beta_k(.)$ puede definirse como

$$\beta_k(S,\mathbf{u}) = u_k \cdot \alpha(S,\mathbf{u}),$$

donde u_k indica el peso asignado a la $k^{ésima}$ cámara, y $\alpha(S,u)$ se define como anteriormente.

Similar a lo que se ha contado acerca de la primera etapa, vale la pena mencionar que puede imaginarse la formulación alternativa de la misma idea básica. Por ejemplo, la realización de la invención que se describe en el apéndice 1 define la función para maximizar basándose en el producto de un factor de cercanía con un factor de completitud, midiendo cada factor una suma ponderada de resolución y visibilidad de objeto individual. Por lo tanto, sustituye la suma del producto por un producto de sumas, pero sigue aún la misma idea básica de tener en cuenta las preferencias de usuario mientras equilibra dos términos antagonistas, que refleja el concepto de cercanía y completitud, respectivamente.

De manera similar, una formulación basándose en la suma ponderada de dos términos que reflejan los conceptos de la cercanía y la completitud anteriormente descritos es también una realización de la presente invención.

La tercera y última etapa consiste en suavizar la secuencia de los índices de cámara y parámetros de puntos de vista correspondientes.

En la realización propuesta de la invención, el proceso de suavizado se implementa basándose en la definición de dos Campos Aleatorios de Markov (véase la Figura 5, y la descripción de la realización a continuación). Otras realizaciones se pueden fundamentar asimismo en cualquier mecanismo de filtrado de paso bajo lineal o no lineal para suavizar la secuencia de índices de cámara y parámetros de punto de vista. El suavizado podría hacerse también a través de un formalismo de modelo de grafo, resuelto basándose en el algoritmo Viterbi convencional. En ese caso, los vértices del grafo corresponderían a parámetros de representación candidatos para un fotograma dado, mientras que los bordes conectarían estados de representación candidatos a lo largo del tiempo. El coste asignado a cada borde reflejaría la perturbación inducida por un cambio de parámetros de representación entre dos fotogramas consecutivos.

El sistema y método de producción de vídeo automatizado incluye también un director virtual, por ejemplo un módulo de vector virtual para seleccionar y determinar cuál de las múltiples corrientes de vídeo de cámara es una corriente de cámara actual para visualizarse. El director virtual, en cada instante de tiempo, y para cada vista de cámara, selecciona las variaciones en parámetros, por ejemplo en parámetros de recorte que optimizan el equilibrio entre completitud y cercanía. El equilibrio completitud/cercanía se mide como una función de preferencias de usuario. Por ejemplo, dependiendo de la resolución en la que un usuario accede al contenido producido, un usuario puede preferir un punto de vista pequeño (acercamiento) o uno grande (alejamiento). El módulo de director virtual también puntúa el punto de vista seleccionado en cada vista de cámara de acuerdo con la calidad (en términos de preferencias de usuario) de su equilibrio de completitud/cercanía, y a su grado de oclusiones. Finalmente el módulo de director virtual calcula los parámetros de una cámara virtual óptima que realiza panorámica, zoom y cambia a través de las vistas para el segmento temporal disponible, para conservar altas puntuaciones de puntos de vista seleccionados mientras se minimiza la cantidad de movimientos de cámara virtual.

Se experimenta que los puntos de vista seleccionados por el director virtual, de acuerdo con las realizaciones de la presente invención, basándose en las funciones anteriores, coinciden con las expectativas del usuario final. Incluso más, ensayos subjetivos revelan que los observadores prefieren en general los puntos de vista seleccionados basándose en el sistema automático que los seleccionados por un productor humano. Esto se explica parcialmente mediante la fuerte carga impuesta al operador humano cuando el número de cámaras aumenta. Por lo tanto, la presente invención también mitiga el cuello de botella experimentado por un operador huando, cuando procesa conjuntamente y de manera simultánea un gran número de cámaras de origen.

Breve descripción de las figuras

Figura 1: flujo de trabajo jerárquico

Figura 2: estructura jerárquica

Figura 3: función de ponderación

Figura 4: comportamiento de selección de punto de vista

Figura 5: modelo de estimación de dos etapas de movimiento de punto de vista

Figura 6: planos de cámara

Figura 7: vistas de muestra desde cámaras

Figura 8: clip de vídeo corto

Figura 9: secuencias de punto de vista

Figura 10: comportamiento de secuencia de cámara/punto de vista

Figura 11: comparación de secuencias de cámara y de punto de vista

Figura 12: fotogramas en secuencias generadas

Figura 13: comparación de secuencias de cámara generadas

Figura 14: realización de 3 etapas, de la presente invención

Figura 15: realización de divide y vencerás de la presente invención

Figura 16: uso de máscaras para detección

Se muestran dibujos adicionales en el apéndice 2. Estos dibujos hacen referencia al apéndice 2 y el texto del apéndice 2 debería leerse junto con estos dibujos y las referencias específicas a este apéndice.

Descripción detallada de la presente invención

La presente invención proporciona métodos y sistemas basados en ordenador para la generación rentable y autónoma de contenidos de vídeo desde múltiples datos detectados incluyendo la extracción automática de contenidos inteligentes desde una red de sensores distribuidos alrededor de la escena disponible. En este punto, contenidos inteligentes se refiere a la identificación de segmentos sobresalientes en el contenido audiovisual, usando algoritmos de análisis de escena distribuidos. Este conocimiento puede aprovecharse para automatizar la producción y personalizar el resumen de contenidos de vídeo.

Sin pérdida de generalidad y sin limitar la presente invención, únicamente se describirán principalmente cámaras estáticas como una realización ilustrativa.

Una entrada son las posiciones de los objetos de interés. Para identificar segmentos sobresalientes en el contenido de vídeo en bruto, se considera análisis de múltiples cámaras, en el cual la detección de objetos relevantes tales como métodos de detección de personas que se basan en que puede usarse la fusión de la información de probabilidad de primer plano calculada en cada vista. El análisis multi-vista puede superar obstáculos tradicionales tales como oclusiones, sombras e iluminación cambiante. Esto es en contraste con el análisis de señal de sensor único, que se somete en ocasiones a ambigüedades de interpretación, debido a la ausencia de modelo preciso de la escena, y a configuraciones de escena adversas coincidentes.

De acuerdo con algunas realizaciones de la presente invención, las posiciones de los objetos de interés se suponen que se conocen (al menos parcialmente) como una función del tiempo. Por ejemplo, las realizaciones de la presente invención infieren este conocimiento desde el análisis de los campos de luz capturados mediante un conjunto distribuido de cámaras estáticas. En una realización de este tipo puede calcularse una máscara de ocupación de superficie uniendo la probabilidad de primer plano medida en cada vista. Las posiciones de jugador reales pueden deducirse a continuación a través de un proceso voraz iterativo y sensible a oclusión. El análisis multi-vista puede usarse para proporcionar las entradas requeridas al método y sistema de producción de deporte de equipo autónomo de la presente invención y se describe en el artículo “Detection and Recognition of Sports(wo)men from Multiple Views”, D. Delannay, N. Danhier, y C. De Vleeschouwer, *Third ACM/IEEE International Conference on*

Distributed Smart Cameras, Como, Italia, septiembre de 2009 se incorpora en el presente documento por referencia en su totalidad como el apéndice 2.

Las realizaciones de la presente invención continúan a continuación en dos etapas.

En una primera etapa, dadas las posiciones de cada objeto de interés con el tiempo, la invención selecciona un conjunto de denominados parámetros relevantes para representar la escena de interés como una función del tiempo, usando una cámara localizada en un punto que puede ser cualquier punto en 3D arbitrario alrededor de la acción.

En este punto, los parámetros de representación definen un *campo de visión* para la cámara, y dependen de la infraestructura de la cámara que se ha desplegado para capturar las imágenes de la escena. Por ejemplo, las realizaciones de la presente invención hacen uso de una cámara fija, y los parámetros de representación definen cómo recortar sub-imágenes en la vista de la cámara. En otras realizaciones puede usarse una cámara articulada y motorizada, y los parámetros de representación puede a continuación hacer referencia a los parámetros de panorámica, inclinación y zoom de la cámara. La noción de parámetros *relevantes* tiene que hacerse con la definición de informativo, es decir visualizar las personas y objetos de interés, e imágenes perceptualmente agradables.

En una segunda etapa, las realizaciones de la presente invención suponen que múltiples cámaras (PTZ) están distribuidas alrededor de la escena, y se determina a continuación cómo seleccionar la cámara correcta para representar la acción en un tiempo dado. Esto se hace seleccionando o promoviendo cámaras informativas, y evitando el cambio perceptualmente inoportuno entre cámaras.

Juntos el índice de cámara y su campo de visión definen el *punto de vista* para representar la acción.

Para producir resúmenes de vídeo semánticamente significativos y perceptualmente cómodos basándose en la extracción o interpolación de imágenes desde el contenido en bruto, la presente invención introduces tres conceptos fundamentales, es decir “completitud”, “suavidad” y cercanía (o “precisión”), para resumir el requisito de semántica y narrativa de los contenidos de vídeo. Basándose en estos conceptos, puede determinarse la selección de puntos de vista de cámara y la de los segmentos temporales en el resumen, siendo estos dos problemas de optimización independientes.

• Completitud establece tanto la integridad de la vista que representa en la selección de cámara/punto de vista, como la de la narrativa al resumir. Un punto de vista de alta completitud incluye más objetos sobresalientes, mientras que una historia de alta completitud consiste en más acciones clave.

• Suavidad se refiere al desplazamiento elegante del punto de vista de la cámara virtual, y a la narrativa continua resultante de la selección de segmentos temporales contiguos. Conservar la suavidad es importante para evitar distraer al observador de la historia por cambios abruptos de puntos de vista o saltos temporales constantes (Owen, 2007).

• Cercanía o precisión se refiere a la cantidad de detalles proporcionados acerca de la acción representada. Espacialmente, favorece vistas cercanas. Temporalmente, implica narrativa redundante, que incluye repeticiones. Aumentar la precisión de un vídeo no mejora únicamente la experiencia de visualización, sino que también es esencial en guiar la implicación emocional de los observadores mediante tomas de cerca.

De acuerdo con las realizaciones de la presente invención estos tres conceptos se optimizan, por ejemplo se maximizan para producir un contenido significativo y visualmente agradable. En la práctica, la maximización de los tres conceptos puede dar como resultado decisiones en conflicto, bajo algunas restricciones de recursos limitados, típicamente expresado en términos de la resolución espacial y duración temporal del contenido producido. Por ejemplo, una resolución de vídeo de salida fija, que aumenta la completitud generalmente induce puntos de vista más grandes, que a su vez reduce la precisión de objetos sobresalientes. De manera similar, la suavidad aumentada de movimiento de punto de vista evita la búsqueda precisa de acciones de interés a lo largo del tiempo. Las mismas observaciones se mantienen con respecto a la selección de segmentos y a la organización de historias a lo largo del tiempo, bajo algunas restricciones de duración globales.

Por consiguiente, las realizaciones de la presente invención relacionadas con métodos y sistemas basados en ordenador proporcionan un buen equilibrio entre los tres factores principales. Por ejemplo, se definen métricas cuantitativas para reflejar completitud, precisión/cercanía. La optimización restringida puede usarse a continuación para equilibrar estos conceptos.

Además, para eficacia computacional mejorada, se prevén tanto la producción como el resumen en el paradigma de divide y vencerás (véase la figura 15). Esto tiene sentido especialmente puesto que los contenidos de vídeo tienen intrínsecamente una estructura jerárquica, que empieza desde cada fotograma, tomas (conjunto de fotogramas consecutivos creados por un trabajo de cámara similar), a segmentos semánticos (tomas consecutivas relacionadas lógicamente a la acción idéntica), y finalizar con la secuencia global.

Por ejemplo, un fotograma de tiempo de evento puede cortarse en primer lugar en segmentos temporales semánticamente significativos, tales como una ronda de ataque/defensa de deportes de equipo, o una entrada en noticias. Para cada segmento, se consideran varias opciones narrativas. Cada opción define una historia local, que consiste en múltiples tomas con diferente cobertura de cámara. Una historia local no incluye únicamente tomas para representar la acción global disponible, sino también tomas para fines explicativos y decorativos, por ejemplo, repeticiones y vistas de cerca en deportes o datos de gráficos en noticias. Dadas las indicaciones de tiempo y la estrategia de producción (vista de cerca; repetición, etc.) de las tomas que componen una opción narrativa, el trabajo de cámara asociado a cada toma se planea automáticamente, teniendo en cuenta el conocimiento inferido acerca de la escena mediante los módulos de análisis de vídeo.

Los beneficios y costes se asignan a continuación a cada historia local. Por ejemplo, el coste puede corresponder simplemente a la duración del resumen. El beneficio refleja la satisfacción del usuario (bajo algunas preferencias individuales), y mide cómo se satisfacen algunos requisitos generales, por ejemplo, la continuidad y completitud de la historia. Estos pares de beneficios y costes se alimentan a continuación en un motor de resumen, que resuelve un problema de asignación de recursos para encontrar la organización de historias locales que consiguen el mejor beneficio bajo la duración de resumen restringida.

La planificación del trabajo de cámara se describirá con referencia a un ejemplo, por ejemplo producción de vídeo de baloncesto de vídeos de deportes de equipo. Aunque es extensible a otros contextos (por ejemplo, control de cámara de PTZ), el proceso se ha diseñado para seleccionar qué fracción de la cámara debería recortarse en un conjunto distribuido de cámaras fijas para representar la escena disponible de una manera semánticamente significativa y visualmente agradable suponiendo el conocimiento de las posiciones de los jugadores.

Etapa 1: selección de punto de vista a nivel de cámara

En cada instante de tiempo y en cada vista, se supone que se conocen los apoyos de los jugadores, y seleccionan los parámetros de recorte que optimizan el equilibrio entre completitud y precisión.

Formalmente, un punto de vista v_{ki}, en la k^{ésima} vista de cámara del *i-ésimo* fotograma se define mediante el tamaño S_{ki} y el centro c_{ki} de la ventana que se recorta en la k^{ésima} vista para visualización real. Se ha de seleccionar para que incluya los objetos de interés, y proporcione una descripción precisa, es decir a alta resolución, de estos objetos. Si hay N objetos sobresalientes en este fotograma, y la localización del n^{ésimo} objeto en la k^{ésima} vista se indica mediante x_{nki}, seleccionamos el punto de vista óptimo v_{ki}^*, maximizando una suma ponderada de los objetos de interés como sigue:

$$\mathbf{v}_{ki}^{*} = \arg\max_{\{S_{ki},\mathbf{c}_{ki}\}} \sum_{n=1}^{N} I_n \cdot \beta(S_{ki},\mathbf{u}) \cdot \alpha\left(\frac{\|\mathbf{x}_{nki} - \mathbf{c}_{ki}\|}{S_{ki}}\right)$$

En la ecuación anterior:

○ I_n indica el nivel de interés asignado al n^{ésimo} objeto detectado en la escena. Obsérvese que asignar pesos distintos a jugadores de deporte de equipo permite centrarse en un jugador preferido, pero también implica el reconocimiento de cada jugador. Un peso de unidad puede asignarse a todos los jugadores, produciendo de esta manera un vídeo que representa la acción de deporte de equipo global.

○ El vector u refleja las restricciones y preferencias de usuario en términos de resolución de punto de vista y vista de cámara, u=[$u^{cercano}$ u^{res} {u_k}]. En particular, su componente u^{res} define la resolución de la corriente de salida, que está generalmente restringida por el ancho de banda de transmisión o la resolución del dispositivo del usuario final. Su componente $u^{cercano}$ se establece a un valor mayor de 1, y aumenta para favorecer puntos de vista cercanos en comparación con vistas de alejamiento grandes. Los otros componentes de u se tratan de preferencias de cámara, y se definen en la segunda etapa a continuación.

○ La función $\alpha(.)$ modula los pesos de los objetos de acuerdo con su distancia al centro del punto de vista, en comparación con el tamaño de esta ventana. De manera intuitiva, el peso debería ser alto y positivo cuando el objeto de interés está localizado en el centro de la ventana de visualización, y debería ser negativo o cero cuando el objeto radica fuera del área de visualización. Muchos casos son apropiados, por ejemplo, la función de sombrero mexicano bien conocida.

○ La función $\beta(.)$ refleja la penalización inducida por el hecho de que la señal nativa capturada mediante la k^{ésima} cámara tiene que sub-muestrearse una vez que el tamaño del punto de vista se hace mayor que la resolución máxima u^{res} permitida mediante el usuario. Esta función típicamente se reduce con S_{ki}. Una elección apropiada consiste en establecer la función igual a uno cuando $S_{ki} < u^{res}$, y en hacerla reducir posteriormente. Un ejemplo de $\beta(.)$ se define mediante:

$$\beta(S_{ki},\mathbf{u}) = \left[\min\left(\frac{u^{res}}{S_{ki}},1\right)\right]^{u^{cercano}},$$

donde $u^{cercano} > 1$ aumenta para favorecer puntos de vista cercanos en comparación con vistas de alejamiento grandes.

Etapa 2: selección de cámara a nivel de fotograma

El punto de vista seleccionado en cada vista se puntúa de acuerdo con la calidad de su equilibrio de completitud/cercanía, y a su grado de oclusiones. La puntuación más alta debería corresponder a una vista que (1) hace a la mayoría del objeto de interés visible, y (2) está cerca de la acción, que significa que presenta objetos importantes con muchos detalles, es decir a alta resolución.

Formalmente, dado el interés I_n de cada jugador, la puntuación $I_{ki}(v_{ki}, u)$ asociada a cada vista de cámara se define como sigue:

$$I_{ki}(\mathbf{v}_{ki},\mathbf{u}) = u_k \cdot \sum_{n=1}^{N} I_n \cdot o_k\left(\mathbf{x}_{nki}\middle|\overline{\mathbf{x}}\right) \cdot h_k\left(\mathbf{x}_{nki}\right) \cdot \beta(S_{ki},\mathbf{u}) \cdot \alpha\left(\frac{\left\|\mathbf{x}_{nki}-\mathbf{c}_{ki}\right\|}{S_{ki}}\right)$$

En la ecuación anterior:

- u_k indica el peso asignado a la k^{ésima} cámara, mientras α y β se definen como en la primera etapa anterior.

- $o_k(\mathbf{x}_{nki}|\overline{\mathbf{x}})$ mide la relación de oclusión del n^{ésimo} objeto en la vista de cámara *k*, conociendo la posición de todos los otros objetos. La relación de oclusión de un objeto se define para que sea la fracción de píxeles del objeto que están ocultos por otros objetos cuando se proyectan en el sensor de la cámara.

- La altura $h_k(x_{nki})$ se define para que sea la altura en píxeles de la proyección en la vista *k* de un objeto vertical de 182,88 centímetros (seis pies) de alto localizado en x_{nki}. 182,88 centímetros (seis pies) es la altura media de los jugadores. El valor de $h_k(x_{nki})$ se calcula directamente basándose en la calibración de la cámara. Cuando la calibración no está disponible, puede estimarse basándose en la altura del objeto detectado en la vista *k*.

Etapa 3: suavizado de secuencias de cámara/punto de vista.

Para el segmento temporal disponible, se calculan los parámetros de una cámara virtual óptima que realiza panorámica, zoom y cambia a través de las vistas para conservar altas puntuaciones de puntos de vista seleccionados mientras se minimiza la cantidad de movimientos de cámara virtual.

El proceso de suavizado puede implementarse basándose en la definición de dos Campos Aleatorios de Markov. En primer lugar, se toma $\mathbf{S}_{ki}$ como datos observados en la *i-ésima* imagen y suponiendo que son salidas de ruido deformado de algún resultado suave subyacente v_{ki}. Dada la secuencia de punto de vista suave recuperada para cada cámara, se calculan las ganancias de cámara $I_{ki}(v_{ki}, u)$ de estos puntos de vista deducidos, y se infiere una secuencia de cámara suave desde el segundo campo de Markov realizando las probabilidades $P(k|v_{ki}, u)$ de cada cámara proporcionales a las ganancias $I_{ki}(v_{ki}, u)$.

En comparación con filtros de suavizado gaussianos sencillos, esto posibilita suavizado adaptativo estableciendo diferente intensidad de suavizado en cada fotograma individual. Adicionalmente, el suavizado ligero iterativo en nuestro método puede conseguir resultados más suaves que el suavizado intenso de una pasada.

La detección y reconocimiento de jugador multi-vista se obtiene en una producción autónoma de contenido visual basándose en la detección (y reconocimiento) del objeto de interés en la escena.

La probabilidad de primer plano se calcula independientemente en cada vista, usando técnicas de modelado de segundo plano convencionales. Estas probabilidades se fusionan a continuación proyectándolas en el plano de superficie, definiendo de esta manera un conjunto de denominadas máscaras de ocupación de superficie. El cálculo de la máscara de ocupación de superficie asociada a cada vista es eficaz, y estas máscaras se combinan y procesan para inferir la posición real de los jugadores.

Formalmente, el cálculo de la máscara de ocupación de superficie G_k asociada a la k^{ésima} vista se describe como sigue. En un momento dado, la k^{ésima} vista es la fuente de una imagen de probabilidad de primer plano $F_k \in [0,1]^{Mk}$,

donde M_k es el número de píxeles de la cámara *k, 0 < k< C*.

Debido a la suposición de la verticalidad del jugador, los segmentos de línea vertical anclados en posiciones ocupadas en el plano de superficie soportan una parte del objeto detectado, y por lo tanto proyectan hacia atrás en siluetas de primer plano en cada vista de cámara. Por lo tanto, para reflejar la ocupación de superficie en *x*, el valor de G_k en x se define para que sea la integración de la proyección (hacia delante) de F_k en un segmento vertical anclado en *x*. Evidentemente, esta integración puede calcularse de manera equivalente en F_k, a lo largo de la proyección hacia atrás del segmento vertical anclado en *x*. Esto es en contraste a métodos que calculan la máscara agregando las proyecciones de la probabilidad de primer plano en un conjunto de planos que son paralelos a la superficie.

Para acelerar los cálculos asociados a nuestra formulación, se observa que, a través de una transformación apropiada de F_k, es posible conformar el dominio de integración proyectado hacia atrás de modo que corresponde también a un segmento vertical en la vista transformada, haciendo de esta manera el cálculo de integrales particularmente eficaz a través del principio de imágenes integrales. La transformación se ha diseñado para tratar un doble objetivo. En primer lugar, los puntos del espacio en 3D localizado en la misma línea vertical tienen que proyectarse en la misma columna en la vista transformada (punto de fuga vertical en el infinito). En segundo lugar, los objetos verticales que permanecen en la superficie y cuyos pies se proyectan en la misma línea horizontal de la vista transformada tienen que mantener las mismas relaciones de alturas proyectadas. Una vez que se cumple la primera propiedad, los puntos en 3D que pertenecen a la línea vertical que permanecen por encima de un punto dado desde el plano de superficie proyectan simplemente en la columna de la vista transformada que permanece por encima de la proyección del punto de plano de superficie en 3D. Por lo tanto, $G_k(x)$ se calcula simplemente como la integral de la vista transformada sobre este segmento proyectado hacia atrás vertical. La conservación de la altura a lo largo de las líneas de la vista transformada incluso simplifica adicionalmente los cálculos.

Para las vistas laterales, estas dos propiedades pueden conseguirse moviendo virtualmente (a través de transformaciones de homografía) la dirección de visualización de la cámara (eje principal) para proporcionar el punto de fuga vertical en el infinito y asegurar que la línea del horizonte es horizontal. Para las vistas superiores, el eje principal se establece perpendicular a la superficie y se realiza un mapeo polar para conseguir las mismas propiedades. Obsérvese que en algunas configuraciones geométricas, estas transformaciones pueden inducir sesgado fuerte de las vistas.

Dadas las máscaras de ocupación de superficie G_k para todas las vistas, explicamos ahora cómo inferir la posición de las personas que permanecen en la superficie. A priori, en un contexto de deporte de equipo, sabemos que (i) cada jugador induce un grupo denso en la suma de máscaras de ocupación de superficie, y (ii) el número de personas a detectar es igual a un valor conocido N, por ejemplo N = 12 para baloncesto (10 jugadores + 2 árbitros).

Por esta razón, en cada localización de superficie *x*, consideramos la suma de todas las proyecciones -normalizadas por el número de vistas que realmente cubren *x*-, y averiguamos los puntos de intensidad superiores en esta máscara de ocupación de superficie agregada. Para localizar estos puntos, consideramos en primer lugar un enfoque voraz inicial que es equivalente a un procedimiento de búsqueda de coincidencia iterativa. En cada etapa, el proceso de búsqueda de coincidencia maximiza el producto interno entre un núcleo gaussiano traducido, y la máscara de ocupación de superficie agregada. La posición del núcleo que induce el producto interno mayor define la posición del jugador. Antes de ejecutar la siguiente iteración, la contribución del núcleo gaussiano se resta de la máscara agregada para producir una máscara residual. El proceso se itera hasta que se han localizado suficientes jugadores.

Este enfoque es sencillo, pero sufre de muchas falsas detecciones en la intersección de las proyecciones de distintas siluetas de jugadores desde diferentes vistas. Esto es debido al hecho de que las oclusiones no inducen linealidades en la definición de la máscara de ocupación de superficie. En otras palabras, la máscara de ocupación de superficie de un grupo de jugadores no es igual a la suma de máscaras de ocupación de superficie proyectadas por cada jugador individual. El conocimiento acerca de la presencia de algunas personas en el campo de superficie afecta al valor informativo de las máscaras de primer plano en estas localizaciones. En particular, si la línea vertical asociada a una posición *x* es ocluida por/ocluye otro jugador cuya presencia es muy probable, esta vista particular no debería aprovecharse para decidir si hay un jugador en *x* o no.

Un refinamiento implica inicializar el proceso definiendo $G_k^1(x) = G_k(x)$ para que sea la máscara de ocupación de superficie asociada a la $k^{ésima}$ vista, y establecer $w_k^1(x)$ a 1 cuando *x* se cubre mediante la $k^{ésima}$ vista, y a 0 de otra manera.

Cada iteración se ejecuta a continuación en dos etapas. En la iteración *n*, la primera etapa busca la posición más probable del $n^{ésimo}$ jugador, conociendo la posición de los *(n-1)* jugadores localizados en interacciones anteriores. La segunda etapa actualiza las máscaras de ocupación de superficie de todas las vistas para eliminar la contribución del jugador recién localizado.

Formalmente, la primera etapa de la iteración *n* agrega la máscara de ocupación de superficie desde todas las

vistas, y a continuación busca el grupo más denso en esta máscara. Por lo tanto, calcula la máscara agregada como:

$$G^n(x) = \frac{\sum_{k=1}^{C} w_k^n(x).G_k^n(x)}{\sum_{k=1}^{C} w_k^n(x)},$$

y a continuación define la posición más probable x_n para el n^{ésimo} jugador mediante

$$x_n = \underset{y}{\text{argmax}} < G^n, \phi(y) >$$

donde $\varphi(y)$ indica un núcleo gaussiano centrado en y, y cuyo apoyo espacial corresponde a la anchura típica de un jugador.

En la segunda etapa, la máscara de ocupación de superficie de cada vista se actualiza para tener en cuenta la presencia del n^{ésimo} jugador. En la posición de superficie x, consideramos que el apoyo típico de una silueta de jugador en la vista k es un cuadro rectangular de anchura W y altura H, y observamos que la parte de la silueta que ocluye o es ocluida por el jugador recién detectado no proporciona ninguna información acerca de la presencia potencial de un jugador en la posición x. Se estima la fracción $\varphi_k(x, x_n)$ de la silueta en la posición de superficie x que se hace no informativa en la k^{ésima} vista, como consecuencia de la presencia de un jugador en x_n. Se propone a continuación actualizar la máscara de ocupación de superficie y el peso de agregación de la k^{ésima} cámara en la posición x como sigue:

$$G_k^{n+1}(x) = \max\left(0, G_k^n(x) - \varphi_k(x, x_n).G_k^1(x_n)\right),$$

$$w_k^{n+1}(x) = \max\left(0, w_k^n(x) - \varphi_k(x, x_n)\right).$$

Para eficacia computacional mejorada, las posiciones x investigadas en el enfoque refinado se limitan al máximo local 30 que se han detectado por el enfoque inicial.

Por completitud, se observa que el procedimiento de actualización anteriormente descrito omite la interferencia potencial entre oclusiones producidas por distintos jugadores en la misma vista. Sin embargo, la consecuencia de esta aproximación está lejos de ser drástica, puesto que finaliza omitiendo parte de la información que fue significativa para evaluar la ocupación en posiciones ocluidas, sin afectar a la información que se aprovecha realmente. Tener en cuenta estas interferencias requeriría proyectar hacia atrás las siluetas del jugador en cada vista, tendiendo de esta manera hacia un enfoque caro computacionalmente y en memoria. El método y sistema de la presente invención no sufre de la debilidad habitual de los algoritmos voraces, tal como una tendencia engancharse en malos mínimos locales.

Los beneficios técnicos principales de la presente invención incluyen al menos uno o una combinación de:

• La capacidad de recortar píxeles apropiados en la memoria de imagen y/o controlar una PTZ motorizada, para representar una acción de equipo, es decir una acción que implica múltiples objetos/personas de interés en movimiento, desde un punto en 3D arbitrario.

• La capacidad para (i) controlar la selección del campo de visión mediante la cámara individual, y (ii) seleccionar una mejor cámara en un conjunto de cámaras. Tal capacidad hace posible manejar un número potencialmente muy grande de cámaras simultáneamente. Esto es especialmente cierto puesto que la selección de parámetros de representación para una cámara particular puede calcularse independientemente de otras cámaras.

• La posibilidad de reproducir y por lo tanto personalizar técnicamente el proceso de selección de punto de vista de acuerdo con preferencias de usuario individuales. Por ejemplo, en el contexto de un evento deportivo, los entrenadores (que prefieren puntos de vista grandes que muestran todo el juego) tienen diferentes expectativas con respecto a la selección de punto de vista que el espectador común (que prefiere imágenes más cercanas y emocionalmente más ricas). Por lo tanto estas preferencias están directamente relacionadas con parámetros técnicos de cómo se controlan las cámaras. Automatizar el proceso de producción proporciona una solución técnica a lo que equivale contestar a solicitudes individuales.

La presente invención incluye en su alcance mejoras adicionales. La presente invención incluye otros criterios para selección de puntos de vista computacionalmente eficaces y/o que pueden resolverse analíticamente. Incluye

también mejor representación de objetos sobresalientes tal como usar partículas en movimiento o modelos de cuerpo flexible en lugar de simples cuadros delimitadores. Adicionalmente, la selección y suavizado de puntos de vista y cámaras en cuatro sub-etapas en la versión actual simplifica la formulación. Sin embargo, pueden resolverse en una estimación unificada puesto que sus resultados afectan unos a los otros. La presente invención incluye también otros criterios de selección de puntos de vista y cámaras independientes de evaluaciones subjetivas.

El aprovechamiento de una red distribuida de cámaras para aproximar las imágenes que se capturarían mediante un sensor virtual localizado en una posición arbitraria, con cobertura de punto de vista arbitraria puede usarse con cualquiera de las realizaciones de la presente invención. La presente invención puede usarse con estos trabajos, puesto que de acuerdo con la presente invención se realiza una selección del punto de vista más apropiado en un conjunto/espacio de puntos de vista candidatos. Por lo tanto, la adición de algoritmos de representación de punto de vista libre a las realizaciones de la presente invención simplemente contribuye a ampliar el conjunto de candidatos potenciales.

Los métodos y el sistema de la presente invención pueden implementarse en un sistema informático que puede utilizarse con los métodos y en un sistema de acuerdo con la presente invención que incluye programas informáticos. Un ordenador puede incluir un terminal de visualización de vídeo, unos medios de entrada de datos tales como un teclado, y una interfaz de usuario gráfica que indica medios tales como un ratón. El ordenador puede implementarse como un ordenador de fin general, por ejemplo una estación de trabajo UNIX o un ordenador personal.

Típicamente, el ordenador incluye una Unidad de Procesamiento Central (“CPU”), tal como un microprocesador convencional del cual un procesador Pentium suministrado por Intel Corp. Estados Unidos es únicamente un ejemplo, y un número de otras unidades interconectadas mediante un sistema de bus. El sistema de bus puede ser cualquier sistema de bus adecuado. El ordenador incluye al menos una memoria. La memoria puede incluir cualquiera de una diversidad de dispositivos de almacenamiento de datos conocidos para el experto en la materia tal como memoria de acceso aleatorio (“RAM”), memoria de solo lectura (“ROM”), memoria de lectura/escritura no volátil tal como un disco duro como se conoce por el experto en la materia. Por ejemplo, el ordenador puede incluir adicionalmente memoria de acceso aleatorio (“RAM”), memoria de solo lectura (“ROM”), así como un adaptador de visualización para conectar el bus de sistema a un terminal de visualización de vídeo, y un adaptador de entrada/salida (I/O) opcional para conectar dispositivos periféricos (por ejemplo, unidades de disco y cinta) al bus de sistema. El terminal de visualización de vídeo puede ser la salida visual del ordenador, que puede ser cualquier dispositivo adecuado tal como una pantalla de vídeo basada en CRT bien conocida en la técnica de hardware informático. Sin embargo, con un ordenador de sobremesa, un ordenador portable o basado en portátil, el terminal de visualización de vídeo puede sustituirse por una pantalla de panel plano basada en LCD o basada en un plasma de gas. El ordenador incluye adicionalmente un adaptador de interfaz de usuario para conectar un teclado, ratón, altavoz opcional. El vídeo relevante requerido puede introducirse directamente en el ordenador mediante una interfaz de gráficos de vídeo o desde dispositivos de almacenamiento, después de lo cual un procesador lleva a cabo un método de acuerdo con la presente invención. Los datos de vídeo relevantes pueden proporcionarse en un medio de almacenamiento de señal adecuado tal como un disquete, un disco duro reemplazable, un dispositivo de almacenamiento óptico tal como un CD-ROM o DVD-ROM, una cinta magnética o similar. Los resultados del método pueden transmitirse a una localización cercana o remota adicional. Un adaptador de comunicaciones puede conectar el ordenador a una red de datos tal como internet, una intranet, una red de área local o amplia (LAN o WAN) o una CAN.

El ordenador incluye también una interfaz de usuario gráfica que reside en el medio legible por máquina para dirigir la operación del ordenador. Cualquier medio legible por máquina adecuado puede retener la interfaz de usuario gráfica, tal como una memoria de acceso aleatorio (RAM), una memoria de solo lectura (ROM), un disquete magnético, cinta magnética, o disco óptico (estando localizados los últimos tres en unidades de disco y de cinta). Cualquier sistema operativo adecuado e interfaz de usuario gráfica asociada (por ejemplo, Microsoft Windows, Linux) pueden dirigir la CPU. Además, el ordenador incluye un programa de control que reside en el almacenamiento de memoria informática. El programa de control contiene instrucciones que cuando se ejecutan en la CPU permiten al ordenador llevar a cabo las operaciones descritas con respecto a cualquiera de los métodos de la presente invención.

La presente invención proporciona también un producto de programa informático para llevar a cabo el método de la presente invención y este puede residir en cualquier memoria adecuada. Sin embargo, es importante que mientras que la presente invención haya sido, y continuará siendo, que los expertos en la materia apreciarán que los mecanismos de la presente invención pueden distribuirse como un producto de programa informático en una diversidad de formas, y que la presente invención se aplica igualmente independientemente del tipo particular de señal que lleve el medio usado para llevar a cabo realmente la distribución. Ejemplos de medios portadores de señal legible por ordenador incluyen: medio de tipo grabable tal como discos flexibles y CD ROM y medio de tipo de transmisión tal como enlaces de comunicación digitales y analógicos. Por consiguiente, la presente invención incluye también un producto de software que cuando se ejecuta en un dispositivo informático adecuado lleva a cabo cualquiera de los métodos de la presente invención. El software adecuado puede obtenerse programando en un lenguaje de alto nivel adecuado tal como C y compilando en un compilador adecuado para el procesador informático

objetivo o en un lenguaje interpretado tal como Java y a continuación compilarse en un compilador adecuado para implementación con la Máquina Virtual Java.

La presente invención proporciona software, por ejemplo un programa informático que tiene segmentos de código que proporcionan un programa que, cuando se ejecuta en un motor de procesamiento, proporciona un módulo de director virtual. El software puede incluir segmentos de código que proporcionan, cuando se ejecutan en el motor de procesamiento: cualquiera de los métodos de la presente invención o implementar cualquiera de los medios de sistema de la presente invención.

Otros aspectos y ventajas de la presente invención así como un entendimiento más completo de la misma serán evidentes a partir de la siguiente descripción tomada junto con las figuras embebidas y adjuntas, que ilustran a modo de ejemplo los principios de la invención. Además, se pretende que el alcance de la invención esté determinado mediante las reivindicaciones adjuntas y no mediante el resumen anterior o la siguiente descripción detallada.

Apéndice 1

1. Introducción

Dirigir la producción de contenidos semánticamente significativos y perceptualmente cómodos desde múltiples datos detectados en bruto, proponemos un sistema de producción computacionalmente eficaz, basándose en el paradigma de divide y vencerás. Resumimos factores principales de nuestro objetivo mediante tres palabras clave, que son “completitud”, “cercanía” y “suavidad”. La completitud establece la integridad de representación de vista. La cercanía define la precisión de descripción de detalle, y la suavidad es un término que hace referencia a la continuidad de tanto el movimiento de punto de vista como la narrativa. Equilibrando entre estos factores, desarrollamos métodos para seleccionar puntos de vista óptimos y cámaras para ajustar la resolución de visualización y otras preferencias de usuario, y para suavizar estas secuencias para una narrativa continua y elegante. Hay una lista larga de posibles preferencias de usuario, tales como perfil de usuario, historial de exploración del usuario, y capacidades del dispositivo. Resumimos las preferencias de narrativa en cuatro descriptores, es decir, equipo preferido del usuario, jugador preferido del usuario, evento preferido del usuario y cámara preferida del usuario. Todas las restricciones del dispositivo, tales como resolución de visualización, velocidad de red, rendimiento del decodificador, se resumen como la resolución de visualización preferida. Analizamos por lo tanto principalmente las preferencias de usuario con estos cinco elementos en el presente trabajo.

La capacidad para tener en cuenta estas preferencias depende evidentemente del conocimiento capturado acerca de la escena a través de las herramientas de análisis de vídeo, por ejemplo, detectando qué equipo está atacando o defendiendo. Sin embargo y más importantemente, vale la pena mencionar que nuestra estructura es genérica en que puede incluir cualquier tipo de preferencias de usuario.

En la sección 2, explicamos la estructura de estimación de tanto selección como suavizado de puntos de vista y vistas de cámara, y proporcionamos su formulación e implementación detalladas. En la sección 3, se proporcionan más detalles técnicos y experimentos realizados para verificar la eficacia de nuestro sistema. Finalmente, concluimos este trabajo y enumeramos un número de posibles rutas para investigación futura.

2. Producción autónoma de vídeos de baloncesto personalizados a partir de múltiples datos detectados

Aunque es difícil definir una regla absoluta para evaluar el rendimiento de historias organizadas y puntos de vista determinados al presentar un escenario genérico, la producción de vídeos deportivos tiene algunos principios generales. [11] Para juegos de baloncesto, resumimos estas reglas en tres equilibrios principales.

El primer equilibrio surge de la personalización de la producción. Específicamente, se origina desde el conflicto entre conservar reglas de producción generales de vídeos deportivos y maximizar la satisfacción de las preferencias de usuario. Algunas reglas básicas de producción de vídeo para juegos de baloncesto no podrían sacrificarse para mejor satisfacción de las preferencias de usuario, por ejemplo, la escena debe incluir siempre la pelota, y debería tomarse la ponderación bien equilibrada entre el jugador dominante y el jugador preferido del usuario cuando se representa un evento.

El segundo equilibrio es el balance entre completitud y cercanía de la escena representada. El interés intrínseco de los juegos de baloncesto proviene parcialmente de la complejidad del trabajo en equipo, cuya descripción evidente requiere completitud espacial en la cobertura de la cámara. Sin embargo, muchas actividades destacadas normalmente ocurren en un área de juego específica y delimitada. Una vista cercana que acentúa estas áreas aumenta la implicación emocional del público con el juego, moviendo al público más ceca de la escena. La cercanía se requiere también para generar una vista del juego con suficiente resolución espacial bajo una situación con recursos limitados, tales como tamaño de visualización pequeño o recursos de ancho de banda limitados de dispositivos portátiles.

El equilibrio final equilibra la búsqueda precisa de acciones de interés a lo largo del tiempo, y la suavidad del

movimiento del punto de vista. La necesidad para que el público conozca la situación general con respecto al juego a lo largo de toda la competición es un requisito primario y el fin principal del cambio de punto de vista. Cuando mezclamos ángulos de diferentes cámaras para destacar u otros efectos especiales, la suavidad del cambio de cámara debería tenerse en mente para ayudar al público a re-orientar rápidamente la situación del juego después de los movimientos de punto de vista. [11]

Dados los meta-datos recopilados desde los datos de vídeo de múltiples sensores, planeamos cobertura de punto de vista y cambio de cámara considerando los tres equilibrios anteriores. Proporcionamos una vista general de nuestra estructura de producción en la sección 2.1, e introducimos algunas notaciones sobre meta-datos en la sección 2.2. En la sección 2.3, proponemos nuestro criterio para seleccionar punto de vista y cámara en un fotograma individual. El suavizado de las secuencias de punto de vista y cámara se explica en la sección 2.4.

2.1. Vista general de la estructura de producción

Es inevitable proporcionar discontinuidad a contenidos de narrativa cuando se cambian vistas de cámara. Para suprimir la influencia de esta discontinuidad, normalmente localizamos puntos de vista drásticos o cambios de cámara durante el hueco entre dos eventos destacados, para evitar la posible distracción del público de la historia. Por lo tanto, podemos prever nuestra producción personalizada en el paradigma de divide y vencerás, como se muestra en la Figura 1. La historia total se divide en primer lugar en varios segmentos. Los puntos de vista óptimos y las cámaras se determinan localmente en cada segmento equilibrando entre beneficios y costes bajo preferencias de usuario especificadas. Adicionalmente, la estimación de la cámara óptima o los puntos de vista se realiza en una estructura jerárquica. La fase de estimación toma etapas de abajo a arriba desde todos los fotogramas individuales para la historia total. Empezando desde un fotograma independiente, optimizamos el punto de vista en cada vista de cámara individual, determinamos la mejor vista de cámara desde múltiples cámaras candidatas bajo los puntos de vista seleccionados, y finalmente organizamos la historia total. Cuando necesitamos representar la historia al público, se toma un procesamiento de arriba a abajo, que en primer lugar divide el vídeo en segmentos no solapados. Los fotogramas correspondientes para cada segmento se recogen a continuación, y se presentan en el dispositivo objetivo con cámaras y puntos de vista especificados.

La estructura jerárquica intrínseca de los juegos de baloncesto proporciona superficies razonables para la visión anterior, y proporciona también indicios en separación de segmentos. Como se muestra en la Figura 2, un juego se divide en reglas en una secuencia de periodos de posesión de pelota no solapados. Un periodo de posesión de pelota es el periodo de juego cuando el mismo equipo sujeta la pelota y hace varios intentos de anotación. En cada periodo, pueden ocurrir varios eventos durante el proceso de ataque/defensa. De acuerdo con si el evento está relacionado con el reloj de tiro de 24 segundos, los eventos en un juego de baloncesto podrían clasificarse como eventos de reloj y eventos no de reloj. Los eventos de reloj no solapan entre sí, mientras que los eventos no de reloj pueden solapar con tanto eventos de reloj/no de reloj. En general, un periodo de posesión de pelota es un periodo bastante fluido y requiere la continuidad de nivel de periodo de movimiento de punto de vista.

En este artículo, definimos en primer lugar los criterios para evaluar puntos de vista y cámaras en cada fotograma individual. La suavidad de puntos de vista a nivel de cámara se aplica a continuación a todos los fotogramas en cada periodo de posesión de pelota. Basándose en puntos de vista determinados, se selecciona y suaviza una secuencia de cámara.

2.2. Meta-datos y preferencia de usuario

Los datos de entrada alimentados en nuestro sistema incluyen datos de vídeo, meta-datos asociados y preferencias de usuario. Suponiendo que hemos recogido una base de datos de secuencias de vídeo de baloncesto, que se capturan simultáneamente mediante K cámaras diferentes. Todas las cámaras están sincronizadas ligeramente y producen el mismo número de fotogramas, es decir, N fotogramas, para cada cámara. En el i-ésimo fotograma capturado y en el tiempo t_i, M_i diferentes objetos sobresalientes, indicados mediante $\{o_{im} | m = 1, \ldots, M_i\}$, se detectan en total desde todas las vistas de cámara. Tenemos dos tipos de objetos sobresalientes definidos. La primera clase incluye regiones para jugadores, árbitros y la pelota, que se usan para entendimiento de escena. La segunda clase incluye la canasta, el banquillo del entrador y algunas marcas del terreno de la cancha, que se usan en tanto entendimiento de escena como calibración de la cámara. Los objetos de la primera clase se extraen automáticamente de la escena basándose típicamente en el algoritmo de resta de segundo plano, mientras aquellos de la segunda clase se etiquetan manualmente puesto que sus posiciones son constantes en cámaras fijas. Definimos el m-ésimo objeto sobresaliente como $o_{im} = \{o_{kim} | k = 1 \ldots K\}$, donde o_{kim} es el m-ésimo objeto sobresaliente en la $k^{ésima}$ cámara.

Todos los objetos sobresalientes se representan mediante regiones de interés. Una región r es un conjunto de coordenadas de píxeles que pertenecen a esta región. Si o_{im} no aparece en la $k^{ésima}$ vista de cámara, establecemos o_{kim} al conjunto vacío ϕ. Siendo r_1 y r_2 dos regiones arbitrarias, definimos en primer lugar varias funciones elementales sobre una o dos regiones como

$$\text{Área}: \mathcal{A}(\mathbf{r}_1) = \sum_{\mathbf{x}\in\mathbf{r}_1} 1; \quad (1)$$

$$\text{Centro}: \mathcal{C}(\mathbf{r}_1) = \frac{1}{\mathcal{A}(\mathbf{r}_1)} \sum_{\mathbf{x}\in\mathbf{r}_1} \mathbf{x}; \quad (2)$$

$$\text{Visibilidad}: \mathcal{V}(\mathbf{r}_1|\mathbf{r}_2) = \begin{cases} 1, & \mathbf{r}_1 \subseteq \mathbf{r}_2 \\ -1, & \textit{de otra manera}; \end{cases}; \quad (3)$$

$$\text{Distancia}: \mathcal{D}(\mathbf{r}_1, \mathbf{r}_2) = \|\mathcal{C}(\mathbf{r}_1) - \mathcal{C}(\mathbf{r}_2)\|; \quad (4)$$

que se usarán en nuestras secciones posteriores.

Adicionalmente, definimos preferencia de usuario mediante un parámetro establecido u, que incluye tanto preferencias de narrativa como restrictivas, tales como favoritos y capacidades del dispositivo.

2.3. Selección de cámara y puntos de vista en fotogramas individuales

Por simplicidad, dejamos a un lado el problema de suavizado en la primera etapa, y empezamos considerando la selección de un punto de vista apropiado en cada fotograma independiente. Usamos las siguientes dos subsecciones para explicar nuestra solución a este problema desde dos aspectos, es decir, evaluación de diversos puntos de vista en la misma vista de cámara y evaluación de diferentes vistas de cámara.

2.3.1. Cálculo de puntos de vista óptimos en cada cámara individual

Aunque la evaluación del punto de vista es una tarea altamente subjetiva que carece aún de una regla objetiva, tenemos algunos requisitos básicos en nuestra selección de punto de vista. Debería ser computacionalmente eficaz, y debería ser adaptable bajo diferentes resoluciones de dispositivo. Para un dispositivo con alta resolución de visualización, normalmente preferimos una vista completa de toda la escena. Cuando la resolución está limitada debido al dispositivo o a restricciones del canal, tenemos que sacrificar parte de la escena para representación mejorada de detalles locales. Para un objeto justo cerca del borde del punto de vista, debería incluirse para mejorar la completitud global de la narrativa si muestra alta relevancia al evento actual en fotogramas posteriores, y debería excluirse para evitar que la secuencia de punto de vista oscile si siempre aparece alrededor del borde. Para mantener un área segura para tratar con este tipo de objeto, preferimos que los objetos sobresalientes visibles dentro del punto de vista determinado estén más cerca al centro mientras los objetos invisibles deberían conducirse lejos del borde del punto de vista, tan lejos como sea posible.

Dejamos que el punto de vista para la construcción de escena en el i-ésimo fotograma de la k$^{\text{ésima}}$ cámara sea v_{ki}. El punto de vista v_{ki} se define como una región rectangular. Para representación natural de la escena, limitamos la relación de aspecto de todos los puntos de vista para que sea la misma relación de aspecto del dispositivo de visualización. Por lo tanto, para cada v_{ki}, tenemos únicamente tres parámetros libres para ajustar, es decir, el centro horizontal v_{kir}, el centro horizontal v_{kiy} y la anchura v_{kiw}. Se obtiene el punto de vista óptimo individual maximizando la ganancia de interés aplicando el punto de vista v_{ki} al i-ésimo fotograma de la k$^{\text{ésima}}$ cámara, que se define como una suma ponderada de intereses de atención desde todos los objetos sobresalientes visibles en ese fotograma, es decir,

$$\mathcal{I}_{ki}(\mathbf{v}_{ki}|\mathbf{u}) = \sum_{m} w_{kim}(\mathbf{v}_{ki}, \mathbf{u}) \mathcal{I}(\mathbf{o}_{kim}|\mathbf{u}), \quad (5)$$

donde $\mathcal{I}(\mathbf{o}_{kim}|\mathbf{u})$ es el interés de un objeto sobresaliente o_{kim} bajo la preferencia de usuario u. En el presente artículo, la función de interés pre-definida $\mathcal{I}(\mathbf{o}_{kim}|\mathbf{u})$ proporcionará diferente ponderación de acuerdo con diferentes valores de u, que refleja las preferencias de usuario de narrativa. Por ejemplo, un jugador especificado por el público se le asigna un interés superior que un jugador no especificado, y la pelota se le proporciona el interés más alto de modo que siempre se incluye en la escena. Explicamos un ajuste práctico de $\mathcal{I}(\mathbf{o}_{kim}|\mathbf{u})$ con más detalle en la siguiente sección.

Definimos $w_{kim}(v_{ki}, u)$ para ponderar la significancia de atención de un único objeto en un punto de vista. Matemáticamente, tomamos $w_{kim}(v_{ki}, u)$ en una forma como sigue:

$$w_{kim}(\mathbf{v}_{ki}, \mathbf{u}) = \frac{\mathcal{V}(\mathbf{o}_{kim}|\mathbf{v}_{ki})}{\ln \mathcal{A}(\mathbf{v}_{ki})} \exp\left[-\frac{\mathcal{D}(\mathbf{o}_{kim}, \mathbf{v}_{ki})^2}{2[u^{\text{DEV}}]^2}\right], \quad (6)$$

donde usamos u^{DEV} para indicar la limitación de resolución de dispositivo actual en la preferencia de usuario u. Nuestra definición de $w_{kim}(v_{ki}, u)$ consiste en tres partes principales: la parte exponencial que controla la intensidad de concentración de objetos sobresalientes alrededor del centro de acuerdo con la resolución de píxeles de la pantalla del dispositivo; la parte de cruce en cero $\mathcal{V}(\mathbf{o}_{kim}|\mathbf{v}_{ki})$ que separa intereses positivos de intereses negativos en el borde del punto de vista; y la parte de fracción añadida ln $\mathcal{A}(\mathbf{v}_{ki})$ que calcula la densidad de intereses para evaluar la cercanía y se establece como una función logarítmica. Obsérvese que $\mathcal{V}(\mathbf{o}_{kim}|\mathbf{v}_{ki})$ es positivo únicamente cuando el objeto sobresaliente o_{kim} está completamente contenido dentro del punto de vista v_{ki}, que muestra la tendencia de mantener un objeto sobresaliente intacto en la selección de punto de vista. Como se muestra en la Figura 3, la idea básica de nuestra definición es cambiar la importancia relativa de completitud y cercanía ajustando la agudeza de pico central y modificando la longitud de las colas. Cuando u^{DEV} es pequeño, la parte exponencial se descompone bastante rápido, que tiene a acentuar objetos más cercanos al centro e ignorar objetos fuera del punto de vista. Cuando u^{DEV} se hace mayor, se aumentan las penalizaciones para objetos invisibles, que es el incentivo para completarse y presentar todos los objetos sobresalientes. Por lo tanto, $\mathcal{I}_{ki}(\mathbf{v}_{ki}|\mathbf{u})$ describe el equilibrio entre completitud (que presenta tantos objetos como sea posible) y precisión (que representa los objetos con una resolución superior) de descripción de escena en fotogramas individuales.

Un punto de vista que maximiza $\mathcal{I}_{ki}(\mathbf{v}_{ki}|\mathbf{u})$ conduce los objetos visibles más cercanos al centro y conduce a mayores separaciones de objetos invisibles desde el centro. Sea $\hat{\mathbf{v}}_{ki}$ el punto de vista óptimo calculado individualmente para cada fotograma, es decir,

$$\hat{\mathbf{v}}_{ki} = \arg\max_{\mathbf{v}_{ki}} \mathcal{I}_{ki}(\mathbf{v}_{ki}|\mathbf{u}). \qquad (7)$$

Algunos ejemplos de $\hat{\mathbf{v}}_{ki}$ óptimo bajo diferente resolución de visualización se proporcionan en la Figura 4.

2.3.2. Selección de vistas de cámara para un fotograma dado

Aunque usamos datos desde múltiples sensores, lo que realmente importa no es el número de sensores o de su situación, sino la manera en la que utilizamos estos puntos de vista para producir un punto de vista virtual unificado que tenga un buen equilibrio entre acentuación de detalles locales y vista general global de escenarios. Puesto que es difícil generar vídeos de punto de vista libre de alta calidad con los métodos del estado de la técnica, únicamente consideramos seleccionar una vista de cámara desde todas las cámaras presentadas en el presente trabajo para hacer nuestro sistema más genérico. Definimos c = {c_i} como una secuencia de cámara, donde c_i indica el índice de cámara para el i-ésimo fotograma. Un entendimiento insignificante al evaluar una vista de cámara es que los objetos sobresalientes deberían representarse evidentemente con pocas oclusiones y alta resolución. Para el i-ésimo fotograma en la k^{ésima} cámara, definimos la tasa de oclusión de objetos sobresalientes como la relación normalizada del área unida de objetos sobresalientes con respecto a la suma de su área individual, es decir,

$$\mathcal{R}_{ki}^{\mathrm{OC}}(\mathbf{v}_{ki}) = \frac{\mathcal{N}_{ki}(\mathbf{v}_{ki})}{\mathcal{N}_{ki}(\mathbf{v}_{ki}) - 1} \left\{ 1 - \frac{\mathcal{A}\left[\bigcup_m (\mathbf{o}_{kim} \cap \mathbf{v}_{ki})\right]}{\sum_m \mathcal{A}\left[\mathbf{o}_{kim} \cap \mathbf{v}_{ki}\right]} \right\}$$

donde $\bigcup_m x_m$ calcula la unión de todos los cuadros delimitadores $\{x_m\}$. Usamos $\mathcal{N}_{ki}(\mathbf{v}_{ki}) = \sum_{m, \mathbf{o}_{kim} \cap \mathbf{v}_{ki} \neq \phi} 1$ para representar el número de objetos visibles dentro del punto de vista v_{ci}. Para normalizar la relación de oclusión frente a diversos números de objetos sobresalientes en diferentes fotogramas, reescalamos $\mathcal{R}_{ki}^{\mathrm{OC}}(\mathbf{v}_{ki})$ en el intervalo de 0 a 1 aplicando $\mathcal{N}_{ki}(\mathbf{v}_{ki})/(\mathcal{N}_{ki}(\mathbf{v}_{ki}) - 1)$. Definimos la cercanía de los objetos sobresalientes como áreas de píxeles medias usadas para representar objetos, es decir,

$$\mathcal{R}_{ki}^{\mathrm{CL}}(\mathbf{v}_{ki}) = \log \frac{1}{\mathcal{N}_{ki}(\mathbf{v}_{ki})} \sum_m \mathcal{A}\left[\mathbf{o}_{kim} \cap \mathbf{v}_{ki}\right]. \qquad (8)$$

También definimos la completitud de esta vista de cámara como el porcentaje de objetos sobresalientes incluidos, es decir,

$$\mathcal{R}_{ki}^{\mathrm{CP}}(\mathbf{v}_{ki},\mathbf{u})=\frac{1}{\sum_{m}\mathcal{I}(\mathbf{o}_{kim}|\mathbf{u})}\sum_{\substack{m\\ \mathbf{o}_{cim}\cap\mathbf{v}_{ci}\neq\phi}}\mathcal{I}(\mathbf{o}_{kim}|\mathbf{u}). \qquad (9)$$

Por consiguiente, la ganancia de interés de elegir la k[ésima] cámara para el i-ésimo fotograma se evalúa mediante $\mathcal{I}_i(k|\mathbf{v}_{ki},\mathbf{u})$, que se lee,

$$\mathcal{I}_i(k|\mathbf{v}_{ki},\mathbf{u})=w_k(\mathbf{u})\mathcal{R}_{ki}^{\mathrm{CL}}(\mathbf{v}_{ki})\mathcal{R}_{ki}^{\mathrm{CP}}(\mathbf{v}_{ki},\mathbf{u})\exp[-\frac{\mathcal{R}_{ki}^{\mathrm{OC}}(\mathbf{v}_{ki})^2}{2}]. \qquad (10)$$

Ponderamos el apoyo de la preferencia de usuario actual a la cámara *k* mediante $w_k(\mathbf{u})$, que asigna un valor superior a la cámara *k* si se especifica por el usuario y asigna un valor inferior si no se especifica. Definimos a continuación la probabilidad de tomar la k[ésima] cámara para el i-ésimo fotograma bajo {v_{ki}} como

$$P(c_i=k|\mathbf{v}_{ki},\mathbf{u})\equiv\frac{\mathcal{I}_i(k|\mathbf{v}_{ki},\mathbf{u})}{\sum_j\mathcal{I}_i(j|\mathbf{v}_{ki},\mathbf{u})} \qquad (11)$$

2.4. Generación de secuencias de punto de vista/cámara suaves

Una secuencia de vídeo con puntos de vista individualmente optimizados tendrá fluctuaciones evidentes, que conducen a artefactos visuales molestos. Resolvemos este problema generando una secuencia en movimiento suave de tanto cámaras como puntos de vista basándose en su óptimo individual. Usamos un grafo en la Figura 5 para explicar este procedimiento de estimación, que cubre dos etapas de todo el sistema, es decir, suavizado de movimientos de punto de vista a nivel de cámara y generación de una secuencia de cámara suave basándose en puntos de vista determinados. En primer lugar, tomamos $\widehat{\mathbf{v}}_{ki}$ como datos observados y suponemos que son salidas de ruido deformado de algún resultado suave subyacente v_{ki}. Usamos inferencia estadística para recuperar una secuencia de punto de vista suave para cada cámara. Teniendo en consideración las ganancias de cámara de estos puntos obtenidos, generamos a continuación una secuencia de cámara suave.

2.4.1. Suavizado, a nivel de cámara, de movimiento de punto de vista

Empezamos desde la suavidad de movimiento de punto de vista en un vídeo desde la misma cámara. Existen dos intensidades contradictorias que controlan la optimización del movimiento de punto de vista: por un lado, los puntos de vista optimizados deberían estar más cercanos al punto de vista óptimo de cada fotograma individual; por otro lado, la suavidad de puntos de vista entre fotogramas evita que tenga lugar el cambio drástico. Por consiguiente, modelamos movimiento de punto de vista suave como un Campo Aleatorio de Markov (MRF) gaussiano, donde la suavidad a nivel de cámara se modela como la configuración de punto de vista a priori, es decir,

$$P(\{\mathbf{v}_{ki}\}|\mathbf{u})=\frac{\exp(-\frac{1}{2}\sum_i\sum_{j\in\mathcal{N}_i}\mathcal{H}^{\mathrm{PRI}})}{\sum_{\{\mathbf{v}_{ki}\}}\exp(-\frac{1}{2}\sum_i\sum_{j\in\mathcal{N}_i}\mathcal{H}^{\mathrm{PRI}})}, \qquad (12)$$

$$\mathcal{H}_{ij}^{\mathrm{PRI}}=\frac{(v_{kix}-v_{kjx})^2}{2\sigma_{1x}^2}+\frac{(v_{kiy}-v_{kjy})^2}{2\sigma_{1y}^2}+\frac{(v_{kiw}-v_{kjw})^2}{2\sigma_{1w}^2}, \qquad (13)$$

donde $\mathcal{N}_i$ es el vecino del i-ésimo fotograma, mientras una distribución condicional

$$P(\{\widehat{\mathbf{v}}_{ki}\}|\mathbf{u},\{\mathbf{v}_{ki}\})=\prod_i\frac{\exp(-\mathcal{H}_i^{\mathrm{LL}})}{\sum_{\mathbf{v}_{ki}}\exp(-\mathcal{H}_i^{\mathrm{LL}})} \qquad (14)$$

$$\mathcal{H}_i^{\mathrm{LL}}=\frac{(v_{kix}-\widehat{v}_{kix})^2}{2\beta_{ki}\sigma_{2x}^2}+\frac{(v_{kiy}-\widehat{v}_{kiy})^2}{2\beta_{ki}\sigma_{2y}^2}+\frac{(v_{kiw}-\widehat{v}_{kiw})^2}{2\beta_{ki}\sigma_{2w}^2} \qquad (15)$$

describe el ruido que producen los resultados finales. Añadimos un parámetro β_{ki} para controlar la flexibilidad del fotograma actual al suavizar. Una β_{ki} más pequeña puede establecerse para aumentar la tendencia del fotograma actual a acercarse a su punto de vista localmente óptimo. La estimación de puntos de vista óptimos {v_{ki}} se hace maximizando la probabilidad posterior de {v_{ki}} sobre la observada {$\widehat{\mathbf{v}}_{ki}$}, es decir, $P(\{v_{ki}\}|u,\{\widehat{\mathbf{v}}_{ki}\})$, que se expresa

mediante una distribución canónica de Gibbs [12], es decir,

$$P(\{\mathbf{v}_{ki}\}|\mathbf{u},\{\hat{\mathbf{v}}_{ki}\}) = \frac{\exp\{-\mathcal{H}^{\mathrm{V}}\}}{\sum_{\{\mathbf{v}_{ki}\}}\exp\{-\mathcal{H}^{\mathrm{V}}\}} \quad (16)$$

con

$$\mathcal{H}^{\mathrm{V}} = \frac{1}{2}\sum_{i}\sum_{j\in\mathcal{N}_i}\mathcal{H}^{\mathrm{PRI}} + \sum_{i}\mathcal{H}_i^{\mathrm{LL}}. \quad (17)$$

En física estadística, la configuración óptima de la probabilidad posterior mayor se determina minimizando la siguiente energía libre [13]:

$$\mathcal{F}^{\mathrm{V}} = \langle\mathcal{H}^{\mathrm{V}}\rangle - \langle\ln P(\{\mathbf{v}_{ki}\}|\mathbf{u},\{\hat{\mathbf{v}}_{ki}\})\rangle \quad (18)$$

donde $\langle x\rangle = \sum_{\{\mathbf{v}_{ki}\}} xP(\{\mathbf{v}_{ki}\}|\mathbf{u},\{\hat{\mathbf{v}}_{ki}\})$ es el valor esperado de una cantidad $\boldsymbol{x}$. Formamos a continuación el siguiente criterio considerando la restricción de normalización de $P(\{\mathrm{v}_{ki}\}|\mathrm{u},\{\hat{\mathbf{v}}_{ki}\})$, como

$$\mathcal{L}^{\mathrm{V}} = \mathcal{F}^{\mathrm{V}} + \eta\left(1 - \sum_{\{\mathbf{v}_{ki}\}} P(\{\mathbf{v}_{ki}\}|\mathbf{u},\{\hat{\mathbf{v}}_{ki}\})\right) \quad (19)$$

donde η es un multiplicador lagrangiano. Usamos la aproximación de campo medio [13] que supone que $P(\{\mathbf{v}_{ki}\}|\mathbf{u},\{\hat{\mathbf{v}}_{ki}\}) \approx \prod_i P(\mathbf{v}_{ki}|\mathbf{u},\{\hat{\mathbf{v}}_{ki}\})$ para desacoplar correlaciones de dos cuerpos. Tomando el diferencial de $\mathcal{L}^{\mathrm{VP}}$ con respecto a $P(v_{kix}|\mathbf{u},\{\hat{\mathbf{v}}_{ki}\})$ y ajustándolo a cero, obtenemos la estimación óptima para $P(v_{kix}|\mathbf{u},\{\hat{\mathbf{v}}_{ki}\})$ como:

$$\begin{aligned} 0 &= \frac{\partial\mathcal{L}^{\mathrm{VP}}}{\partial P(v_{kix}|\mathbf{u},\{\hat{\mathbf{v}}_{ki}\})} \\ &= \sum_{j\in\mathcal{N}_i}\frac{v_{kix}^2 - 2v_{kix}\langle v_{kjx}\rangle^T}{2\sigma_{1x}^2} + \frac{(v_{kix}-\hat{v}_{kix})^2}{2\beta_{ki}\sigma_{2x}^2} + \ln P(v_{kix}|\mathbf{u},\{\hat{\mathbf{v}}_{ki}\}) - \eta. \end{aligned} \quad (20)$$

Por lo tanto, tenemos la probabilidad posterior:

$$P(v_{kix}|\mathbf{u},\{\hat{\mathbf{v}}_{ki}\}) \propto \exp\left\{-\sum_{j\in\mathcal{N}_i}\frac{(v_{kix}-\langle v_{kjx}\rangle)^2}{2\sigma_{1x}^2} - \frac{(v_{kix}-\hat{v}_{kix})^2}{2\beta_{ki}\sigma_{2x}^2}\right\}. \quad (21)$$

Puesto que es una distribución gaussiana cuyo valor medio tiene la probabilidad maximizada, el punto de vista óptimo para v_{kix} se resuelve como:

$$v_{kix}^* = \langle v_{kix}\rangle = \frac{\sum_{j\in\mathcal{N}_i}\sigma_{2x}^2\beta_{ki}\langle v_{kjx}\rangle + \hat{v}_{kix}\sigma_{1x}^2}{\sum_{j\in\mathcal{N}_i}\sigma_{2x}^2\beta_{ki} + \sigma_{1x}^2}, \quad (22)$$

$$v_{kiy}^* = \langle v_{kiy}\rangle = \frac{\sum_{j\in\mathcal{N}_i}\sigma_{2y}^2\beta_{ki}\langle v_{kjy}\rangle + \hat{v}_{kiy}\sigma_{1y}^2}{\sum_{j\in\mathcal{N}_i}\sigma_{2y}^2\beta_{ki} + \sigma_{1y}^2}, \quad (23)$$

$$v_{kiw}^* = \langle v_{kiw}\rangle = \frac{\sum_{j\in\mathcal{N}_i}\sigma_{2w}^2\beta_{ki}\langle v_{kjw}\rangle + \hat{v}_{kiw}\sigma_{1w}^2}{\sum_{j\in\mathcal{N}_i}\sigma_{2w}^2\beta_{ki} + \sigma_{1w}^2} \quad (24)$$

con resultados óptimos para v_{kiy} y v_{kiw} también proporcionados mediante derivación similar. Usamos $\mathbf{v}_{ki}^*$ en

las siguientes secciones para indicar el punto de vista óptimo representado mediante v^*_{kix}, v^*_{kiy}, y v^*_{kiw}.

2.4.2. Suavizado de secuencia de cámara

Una secuencia de cámara suave se generará desde puntos de vista determinados. Por simplicidad, usamos $\rho_{ki} = \log P(c_i = k|\mathbf{v}^*_{ki}, \mathbf{u})$ para acortar la formulación, que se calcula usando la Ec. 11. Tenemos que equilibrar entre minimizar el cambio de cámara y maximizar la ganancia global de las cámaras. Usamos otro MRF para modelar estos dos tipos de intensidades. La suavidad de la secuencia de cámara se modela mediante una distribución canónica de Gibbs, que se lee,

$$P(\{c_i\}|\{\mathbf{v}^*_{ki}\}, \mathbf{u}) = \frac{\exp(-\mathcal{H}^{\mathrm{C}})}{\sum_{\{c_i\}} \exp(-\mathcal{H}^{\mathrm{C}})}, \tag{25}$$

con

$$\mathcal{H}^{\mathrm{C}} = -\gamma \sum_{i,k} \delta_{c_i,k} \rho_{ki} - \frac{(1-\gamma)}{2} \sum_i \sum_{j \in \mathcal{N}_i} \alpha_{ij} \delta_{c_i,c_j}. \tag{26}$$

donde α_{ij} es un parámetro para normalizar la intensidad relativa de suavizado con respecto al tamaño de la cercanía, que se lee

$$\alpha_{ij} = \frac{K}{|j-i| \sum_{l \in \mathcal{N}_i} \frac{1}{|l-i|}}. \tag{27}$$

γ es un híper-parámetro para controlar la intensidad de suavizado. Usamos la aproximación de campo medio que supone que $P(\{c_i\}|\{\mathbf{v}^*_{ki}\}, \mathbf{u}) \approx \prod_i P(c_i|\{\mathbf{v}^*_{ki}\}, \mathbf{u})$ de nuevo para conseguir la estimación óptima. Omitimos la derivación detallada y mostramos únicamente el resultado final, que deduce que la probabilidad marginal de tomar la cámara *k* para el i-ésimo fotograma es

$$P(c_i = k|\{\mathbf{v}^*_{ki}\}, \mathbf{u}) = \langle \delta_{c_i,k} \rangle^{\mathrm{C}} = \frac{\exp\left\{(1-\gamma) \sum_{j \in \mathcal{N}_i} \alpha_{ij} \langle \delta_{c_j,k} \rangle^{\mathrm{C}} + \gamma \rho_{ki}\right\}}{\sum_k \exp\left\{(1-\gamma) \sum_{j \in \mathcal{N}_i} \alpha_{ij} \langle \delta_{c_j,k} \rangle^{\mathrm{C}} + \gamma \rho_{ki}\right\}}, \tag{28}$$

donde $\langle x \rangle^{\mathrm{C}} = \sum_{\{c_i\}} x P(\{c_i\}|\{\mathbf{v}^*_{ki}\}, \mathbf{u})$ es el valor esperado de una cantidad x. El proceso de suavizado se realiza iterando la siguiente regla de punto fijo hasta alcanzar la convergencia,

$$\langle \delta_{c_i,k} \rangle^{\mathrm{C}} = \frac{\exp\left\{(1-\gamma) \sum_{j \in \mathcal{N}_i} \alpha_{ij} \langle \delta_{c_j,k} \rangle^{\mathrm{C}} + \gamma \rho_{ki}\right\}}{\sum_k \exp\left\{(1-\gamma) \sum_{j \in \mathcal{N}_i} \alpha_{ij} \langle \delta_{c_j,k} \rangle^{\mathrm{C}} + \gamma \rho_{ki}\right\}}. \tag{29}$$

Después de la convergencia, seleccionamos la cámara que maximiza $\langle \delta_{ci,k} \rangle^{\mathrm{C}}$, es decir,

$$c^*_i = \arg\max_{c_i} \langle \delta_{c_i,k} \rangle^{\mathrm{C}}. \tag{30}$$

3. Resultados experimentales y análisis

Organizamos una adquisición de datos en la ciudad de Namur, Bélgica, bajo entorno de juego real, donde se usaron siete cámaras para grabar cuatro juegos. Todos estos vídeos se distribuyen públicamente en el sitio web del

proyecto APIDIS [1] y podría encontrarse explicación más detallada acerca de los ajustes de adquisición en la Ref. [14]. Brevemente, estas cámaras todas eran cámaras IP Arecont Vision AV2100M, cuyas posiciones en la cancha de baloncesto se muestran en la Figura 6. Las lentes de ojo de pez usadas para las cámaras de la vista superior son lentes Fujinon FE185C086HA-1. Los fotogramas de las siete cámaras se enviaron todos a un servidor, donde se usó el tiempo de llegada de cada fotograma al sincronizar diferentes cámaras. En la Figura 7, se proporcionan imágenes de muestra desde todas las siete cámaras. Debido al número limitado de cámaras, establecemos la mayoría de las cámaras para cubrir la cancha izquierda. Como resultado, nos centraremos principalmente en la cancha izquierda para investigar el rendimiento de nuestro sistema en producción personalizada de vídeos deportivos.

Puesto que la producción de vídeo carece aún de una regla objetiva para evaluación de rendimiento. Muchos parámetros de determinan heurísticamente basándose en evaluación subjetiva. Definimos varios objetos sobresalientes y se proporcionan las relaciones entre tipo de objeto e interés en la Tabla 1. Si el usuario muestra intereses especiales en un objeto sobresaliente, el peso se multiplicará por un factor de 1,2. Para suavizado de punto de vista, establecemos todos los β_{ki} a 1 para suavizado de punto de vista a nivel de cámara en los siguientes experimentos. Siendo también $\sigma_{1x} = \sigma_{1y} = \sigma_{1w} = \sigma_1$ y $\sigma_{2x} = \sigma_{2y} = \sigma_{2w} = \sigma_2$.

Se usa un clip de vídeo corto con aproximadamente 1200 fotogramas para demostrar las características de comportamiento de nuestro sistema, especialmente su capacidad de adaptación bajo resolución de visualización limitada. Este clip cubre tres periodos de posesión de pelota e incluye cinco eventos en total. En la Figura 8, mostramos los intervalos de tiempo de todos los eventos, cuyos momentos más destacados se marcan también mediante líneas continuas rojas. En la versión final de este proyecto, deberían generarse los meta-datos mediante el entendimiento automático de la escena. En el presente artículo que se centra en la producción personalizada, evaluamos en primer lugar nuestros métodos sobre meta-datos recopilados manualmente. Exploraremos la eficacia de cada etapa de procesamiento individual de nuestro método, y a continuación hacer una evaluación global basándose en salidas finalmente generadas. Debido a la limitación de la página, se proporcionan los resultados numéricos y se representan mediante grafos en el presente artículo mientras sus vídeos correspondientes están únicamente disponibles en el sitio web del proyecto APIDIS. [1] Los revisores están invitados a descargar muestras de vídeo producidas basándose en diferentes preferencias de usuario para evaluar subjetivamente la eficacia y relevancia del enfoque propuesto.

Tabla 1: Ponderación de diferentes objetos sobresalientes

Tipo de objeto	Pelota	Jugador	Juez	Canasta	Banquillo de entrenador	Otros
$\mathcal{I}(\mathbf{o}_{kim}\|\mathbf{u})$	2	1	0,8	0,6	0,4	0,2

Empezamos investigando el rendimiento de nuestro método para selección individual de puntos de vista. Las secuencias a nivel de cámara de puntos de vista determinados automáticamente mediante nuestro método se ponen en una tabla en la Figura 9, donde se presentan las anchuras de los puntos de vista óptimos bajo tres resoluciones de visualización diferentes, es decir, 160x120, 320x240, y 640x480 para todas las siguiente cámaras. El suavizado de punto de vista débil se ha aplicado para mejorar la legibilidad de los vídeos generados, donde la intensidad de suavizado se establece a $\sigma_2/\sigma_1 = 4$. A partir de la comparación de los resultados bajo tres resoluciones de visualización diferentes, el hallazgo más evidente es que una resolución de visualización superior conduce a una anchura de punto de vista mayor mientras una resolución de visualización inferior prefiere un tamaño de punto de vista más pequeño, tal como hemos esperado a partir de nuestro criterio de selección. Puesto que la cámara 1, 6 y 7 únicamente cubren la mitad de la cancha, sus tamaños de punto de vista se fijarán cuando estén todos los jugadores en la otra mitad de la cancha, que explica los segmentos planos en sus sub-grafos correspondientes. A partir de los datos de vídeo, podríamos confirmar adicionalmente que incluso cuando la resolución de visualización sea muy baja, nuestro sistema extraerá un punto de vista de un tamaño razonable donde la pelota se escala a un tamaño visible. Aunque en algunos fotogramas únicamente se visualiza la pelota para la resolución de visualización más baja, no producirá un problema puesto que estos fotogramas se filtrarán mediante selección de cámara posterior.

Los tamaños de punto de vista de secuencias suavizadas bajo diferentes intensidades de suavizado se comparan en la Figura 10(a). Siendo todos los otros parámetros los mismos, la relación de σ_2 a σ_1 se ajusta para todos los cinco casos. Una relación superior de σ_2 a σ_1 corresponde a un proceso de suavizado más intenso mientras que una relación más pequeña significa suavizado más débil. Cuando $\sigma_2/\sigma_1 = 1$ donde se aplica suavizado muy débil, obtenemos una secuencia bastante accidentada, que da como resultado un vídeo que parpadea con muchos movimientos de punto de vista drásticos. Con el aumento de la relación σ_2/σ_1, la curva de movimiento de punto de vista se hace para que tenga menos picos agudos, que proporciona contenidos perceptualmente más cómodos. Otra observación importante es que las secuencias generadas serán bastante diferentes de nuestra selección inicial basándose en información de notabilidad, si se ha realizado suavizado demasiado intenso con una σ_2/σ_1 muy grande. Esto producirá problemas tales como que el jugador favorito o la pelota estén fuera del punto de vista suavizado. La relación σ_2/σ_1 debería determinarse considerando el equilibrio entre puntos de vista optimizados localmente y secuencias de punto de vista globalmente suavizadas. Comprobando visualmente los vídeos

generados, consideramos que los resultados con un suavizado débil tal como $\sigma_2/\sigma_1 = 4$ ya son perceptualmente aceptables observando el vídeo de demostración.

Verificamos a continuación nuestro algoritmo de suavizado para secuencia de cámara. Las secuencias de cámara suavizadas bajo diversa intensidad de suavizado γ se representan en la Figura 10(b). El proceso de suavizado toma la probabilidad definida en la Ec. 11 como valores iniciales, e itera la regla de actualización de punto fijo con una cercanía de tamaño de treinta hasta convergencia. Una secuencia de cámara sin suavizar corresponde al sub-grafo más superior en la Figura 10(b), mientras que la secuencia con el suavizado más intenso se representa en el sub-grafo inferior. Es evidente que existen muchos cambios de cámara drásticos en una secuencia no suavizada, que conduce a incluso artefactos visuales más molestos que la posición de punto de vista fluctuada, como podemos observar a partir de los vídeos generados. Por lo tanto, preferimos suavizado intenso en las secuencias de cámara y usaremos $\gamma = 0{,}8$ en los siguientes experimentos.

En la Figura 11 (a) y (b), comparamos los puntos de vista y cámaras en secuencias generadas con respecto a diferentes resoluciones de visualización, respectivamente. Desde la parte superior a la inferior, mostramos los resultados para resolución de visualización u^{DEV} = 160, 320 y 640 en tres sub-grafos. Cuando se selecciona la misma cámara, observamos que se prefiere un punto de vista mayor mediante una resolución de visualización superior. Cuando se seleccionan diferentes cámaras, necesitamos considerar tanto la posición de la cámara seleccionada como la posición del punto de vista determinado al evaluar la cobertura de la escena de salida. De nuevo, confirmamos que los tamaños de los puntos de vista aumentan cuando la resolución de visualización se hace mayor. Antes del 400-ésimo fotograma, el evento tiene lugar en la cancha derecha. Hallamos que la 3ª cámara, es decir, la vista superior con lente de gran angular, aparece más a menudo en la secuencia de u^{DEV} = 640 que la de u^{DEV} = 160 y sus puntos de vista son también más amplios, que prueba que una resolución mayor prefiere una vista más ancha. Aunque la 2ª cámara aparece bastante a menudo en u^{DEV} = 160, sus puntos de vista correspondientes son mucho más pequeños en anchura. Esta cámara se selecciona puesto que proporciona una vista lateral de la cancha derecha con objetos sobresalientes recogidos más cerca que otras vistas de cámara debido a la geometría proyectiva. Por la misma razón, la 3ª cámara aparece más a menudo en u^{DEV} = 160 cuando el juego se mueve a la cancha izquierda desde el 450-ésimo fotograma al 950-ésimo fotograma. Esta conclusión se confirma adicionalmente mediante las miniaturas en la Figura 12, donde los fotogramas desde el índice 100 al 900 están dispuestos en una tabla para las tres resoluciones de visualización anteriores.

Debido al hecho de que se seleccionaron diferentes cámaras, los puntos de vista determinados bajo u^{DEV} = 640 parecen estar más cerca que aquellos bajo u^{DEV} = 320 en las últimas cinco columnas de la Figura 12. Esto refleja la no uniformidad de importancia relativa entre completitud y cercanía en selección de punto de vista. Puesto que únicamente se calculan puntos centrales de objetos sobresalientes en el criterio para selección de punto de vista, los puntos de vista resultantes no son continuos bajo diferente resolución. Aunque la cámara 7 es similar a la cámara 1 con acercamiento lineal, sus puntos de vista óptimos pueden tener diferentes acentuaciones en completitud y cercanía. Esta uniformidad existe también en selección separada de cámaras y puntos de vista. Si la selección de punto de vista se centra más en cercanía y la selección de cámara se centra más en completitud, se seleccionará en primer lugar un área de recorte pequeña en la cámara 7 en selección de punto de vista para u^{DEV} = 320, y a continuación se rechazará en la siguiente selección de cámara debido a completitud insuficiente. El ensayo subjetivo nos ayudará a ajustar la ponderación relativa de completitud y cercanía. Es más importante implementar la selección simultánea de puntos de vista y cámaras, que requiere tanto la inclusión de información posicional de las cámaras tales como usar homografía, y un criterio que pueda resolverse analíticamente para selección de punto de vista. Estos asuntos son nuestro mayor trabajo en el futuro próximo.

En todos los experimentos anteriores, no se incluyen preferencias de usuario de narrativa. Si el usuario tiene interés especial en una cierta vista de cámara, podríamos asignar una ponderación superior $w_k(\mathbf{u})$ a la cámara especificada. En nuestro caso establecemos $w_k(\mathbf{u})$ = 1,0 para las cámaras no especificadas y $w_k(\mathbf{u})$ = 1,2 para una cámara especificada por el usuario. Comparamos las secuencias de cámara bajo diferentes preferencias en la Figura 13. Como podemos observar fácilmente a partir del grafo, una cámara aparece más veces cuando se especifica, que refleja la preferencia de usuario en las vistas de cámara. Como para la preferencia de usuario en equipos o jugadores, la diferencia entre los puntos de vista con y sin preferencias de usuario es difícil de indicar sin una regla de evaluación bien definida, puesto que todos los jugadores se amontonan siempre juntos durante el juego. De hecho, estamos más interesados en reflejar las preferencias de usuario en los jugadores o equipos extrayendo sus fotogramas relativos. Omitimos por lo tanto los resultados en la selección de jugador o de equipo, pero los exploramos más tarde junto con resultados a partir de nuestro trabajo futuro sobre resumen de vídeo.

4. Observaciones finales

Se ha propuesto un sistema autónomo para producir vídeos personalizados desde múltiples vistas de cámara. Analizamos la adaptación automática de puntos de vista con respecto a resolución de visualización y contenidos de escenario, la fusión de datos en múltiples vistas de cámara, y suavidad de secuencias de punto de vista y cámara para fluida. Existen cuatro ventajas principales de nuestros métodos: 1) Orientado a semántica. En lugar de usar características inferiores tales como bordes o apariencia de fotogramas, nuestra producción se basa en

entendimiento semántico del escenario, que podría tratarse con preferencia de usuario semántica más compleja. 2) Computacionalmente eficaz. Tomamos una estrategia de divide y vencerás y consideramos un procesamiento jerárquico, que es eficaz al tratar con contenidos de vídeo largos puesto que su tiempo global es casi linealmente proporcional al número de eventos incluidos. 3) Genericidad. Puesto que nuestros sub-métodos en cada etapa individual son todos independientes de la definición de objetos sobresalientes e intereses, esta estructura no está limitada a vídeos de baloncesto, sino que también puede aplicarse a otros escenarios controlados. 4) No supervisado. Aunque hay algún parámetro que se deja establecer por los usuarios, el sistema no es supervisado.

Apéndice 2

Los métodos presentados en este artículo tienen por objeto detectar y reconocer jugadores en un campo deportivo, basándose en un conjunto distribuido de cámaras ligeramente sincronizadas. La detección supone la verticalidad del jugador, y suma la proyección acumulativa de las múltiples máscaras de actividad de primer plano de las vistas en un conjunto de planos que son paralelos al plano de superficie. Después de la suma, los valores de proyección grandes indican la posición del jugador en el plano de superficie. Esta posición se usa como un ancla para el cuadro delimitador del jugador proyectado en cada una de las vistas. En este cuadro delimitador, las regiones proporcionadas mediante segmentación de desplazamiento medio se ordenan basándose en características contextuales, por ejemplo, tamaño y posición relativos, para seleccionar las que es probable que correspondan a un dígito. La normalización y clasificación de las regiones seleccionadas a continuación proporciona el número e identidad del jugador. Puesto que el número de jugador puede leerse únicamente cuando se enfrenta hacia la cámara, se considera el seguimiento basado en grafo para propagar la identidad de un jugador a lo largo de su trayectoria.

I. Introducción

En la sociedad de hoy en día, la producción de contenido y consumo de contenido se confrontan con una mutación fundamental. Se observan dos tendencias complementarias. Por una parte, los individuos se hacen más y más heterogéneos en la manera en la que acceden al contenido. Desean acceder a contenido especializado a través de un servicio personalizado, que puede proporcionar en lo que están interesados, cuando lo deseen y a través del canal de comunicación de su elección. Por otra parte, los individuos y organizaciones tienen acceso más fácil a las facilidades técnicas requeridas para implicarse en la creación de contenido y proceso de difusión.

En este artículo, describimos las herramientas de análisis de vídeo que participan en las evoluciones futuras de la industria de producción de contenido hacia infraestructuras automatizadas que permiten que se produzca, almacene y acceda al contenido a bajo coste y de una manera personalizada y especializada. Más específicamente, nuestra aplicación dirigida considera el resumen autónomo y personalizado de eventos deportivos, sin la necesidad de procesos hechos a mano de manera costosa. En el escenario de aplicación apoyado mediante el conjunto de datos proporcionado, los sensores de adquisición cubren una cancha de baloncesto. El análisis distribuido e interpretación de la escena se aprovecha a continuación para decidir qué mostrar acerca de un evento, y cómo mostrarlo, para producir un vídeo compuesto de un subconjunto valioso a partir de las corrientes proporcionadas por cada cámara individual. En particular, la posición de los jugadores proporciona la entrada requerida para controlar la selección autónoma de parámetros de punto de vista [5], mientras la identificación y seguimiento de los jugadores detectados apoya la personalización del resumen, por ejemplo a través de destacar y/o repetición de acciones de jugador preferido [4].

Parte de este trabajo se ha encontrado mediante el proyecto europeo APIDIS FP7 y mediante el NSF Belga.

II. Vista global de sistema

Para demostrar el concepto de producción autónoma y personalizada, el proyecto de investigación europeo APIDIS FP7 (www.apidis.org) ha desarrollado un sistema de adquisición de múltiples cámaras alrededor de una cancha de baloncesto. Los ajustes de adquisición en un conjunto de 7 cámaras IP calibradas, recopilando cada una fotogramas de 2 Megapíxeles a una tasa superior de 20 fotogramas/s. Después de una sincronización temporal aproximada de las corrientes de vídeo, este artículo investiga cómo aumentar el conjunto de datos de vídeo basándose en la detección, el seguimiento y el reconocimiento de jugadores.

La Figura 1 analiza nuestro enfoque propuesto para calcular y etiquetar seguimientos de jugadores. Después de la detección multivista conjunta de personas que permanecen en el campo de superficie en cada instante de tiempo, un algoritmo de seguimiento basado en grafo coincide posiciones que están suficientemente cercanas - en posición y apariencia - entre fotogramas sucesivos, definiendo de esta manera un conjunto de seguimientos disjuntos potencialmente interrumpidos, llamado también seguimientos parciales. En paralelo, como se representa en la Figura 5, se considera el análisis y clasificación de imagen para cada fotograma de cada vista, para reconocer los dígitos que aparecen potencialmente en las camisetas de los objetos detectados. Esta información se agrega a continuación a través del tiempo para etiquetar los seguimientos parciales.

Las mayores contribuciones de este artículo se han encontrado en la solución de detección de personas propuesta,

que se representa en la Figura 2. En resumen, el proceso de detección sigue un enfoque de abajo a arriba para extraer grupos más densos en un mapa de ocupación de plano de superficie que se calcula basándose en la proyección de máscaras de actividad de primer plano. Se proponen dos mejoras fundamentales en comparación con el estado de la técnica. En primer lugar, la máscara de actividad de primer plano no se proyecta únicamente en el plano de superficie, como se recomienda en [9], sino sobre un conjunto de planos que son paralelos a la superficie. En segundo lugar, se implementa una heurística original para manejar oclusiones, y mitigar las falsas detecciones que tienen lugar en la intersección de las máscaras proyectadas desde distintas siluetas de jugadores mediante distintas vistas. Nuestras simulaciones demuestran que estas dos contribuciones mejoran bastante significativamente el rendimiento de detección.

El resto del artículo se organiza como sigue. Las secciones III, V, y IV se centran respectivamente en los problemas de detección, seguimiento y reconocimiento. Los resultados experimentales se presentan en la Sección VI para validar nuestro enfoque. La sección VII concluye.

Figura 1, cálculo de seguimiento de jugadores y etiquetado en cascada. La flecha discontinua refleja la inclusión opcional de los resultados de reconocimiento de dígito en el modelo de apariencia considerado para seguimiento.

III. Detección de personas multi-vista

El rastreo de las personas que se ocluyen entre sí usando un conjunto de *C* cámaras ampliamente espaciadas, calibradas, fijas y (ligeramente) sincronizadas es una cuestión importante puesto que este tipo de configuración es común a aplicaciones que varían de información de eventos (deportivos) a vigilancia en espacios públicos. En esta sección, consideramos un enfoque de detección de cambio para inferir la posición de jugadores en el campo de superficie, en cada instante de tiempo.

A. Trabajo relacionado

La detección de personas desde las máscaras de actividad de primer plano calculadas en múltiples vistas se ha investigado en detalle en los últimos años. Diferenciamos dos clases de enfoques.

Por un lado, los autores en [9], [10] adoptan un enfoque de abajo a arriba, y proyectan los puntos de la probabilidad de primer plano (siluetas restadas del segundo plano) de cada vista al plano de superficie. Específicamente, los mapas de probabilidad de cambio calculados en cada vista se distorsionan al plano de superficie basándose en homografías que se han inferido fuera de línea. Los mapas proyectados se multiplican a continuación juntos y se realizan umbrales para definir las manchas del plano de superficie para las cuales ha cambiado la apariencia en comparación con el modelo de segundo plano y de acuerdo con el algoritmo de detección de cambio de vista única.

Por otro lado, los trabajos en [2], [7], [1] adoptan un enfoque de arriba a abajo. Consideran una rejilla de puntos en el plano de superficie, y estiman las probabilidades de ocupación de cada punto en la rejilla basándose en la proyección hacia atrás de algún tipo de modelo generativo en cada una de las múltiples vistas calibradas. Por lo tanto, empiezan todos desde el plano de superficie, y validan la hipótesis de ocupación basándose en el modelo de apariencia asociado en cada una de las vistas. Los enfoques propuestos en esta segunda categoría se diferencian principalmente basándose en el tipo de modelo generativo que consideran (rectángulo o diccionario aprendido), y en la manera en la que deciden acerca de la ocupación en cada punto de la rejilla (combinación de múltiples clasificadores basados en vista en [2], rejilla de ocupación probabilística inferida a partir de máscaras de resta de segundo plano en [7], y mapa de ocupación binaria de escasez restringida para [1]).

La primera categoría de métodos tiene la ventaja de ser computacionalmente eficaz, puesto que la decisión acerca de la ocupación del plano de superficie se toma directamente a partir de la observación de la proyección de las

máscaras de detección de cambio de las diferentes vistas. En contraste, la complejidad de la segunda categoría de algoritmos depende del número de puntos de plano de superficie a investigarse (elegidos para limitar el área a monitorizar), y en la carga computacional asociada a la validación de cada hipótesis de ocupación. Este proceso de validación implica generalmente proyección hacia atrás de una plantilla del mundo en 3D en cada una de las vistas. A este respecto, observamos que, debido a las distorsiones de lente y de proyección, incluso la distorsión de una plantilla rectangular en 3D sencilla generalmente da como resultado patrones no rectangulares en cada una de las vistas, evitando de esta manera el uso de técnicas de imágenes integrales computacionalmente eficaces. Por lo tanto, en la mayoría de los casos prácticos, el segundo tipo de enfoque es significativamente más complejo que el primero. A cambio, ofrece rendimiento aumentado puesto que no únicamente los pies, sino toda la silueta del objeto se considera para tomar una decisión.

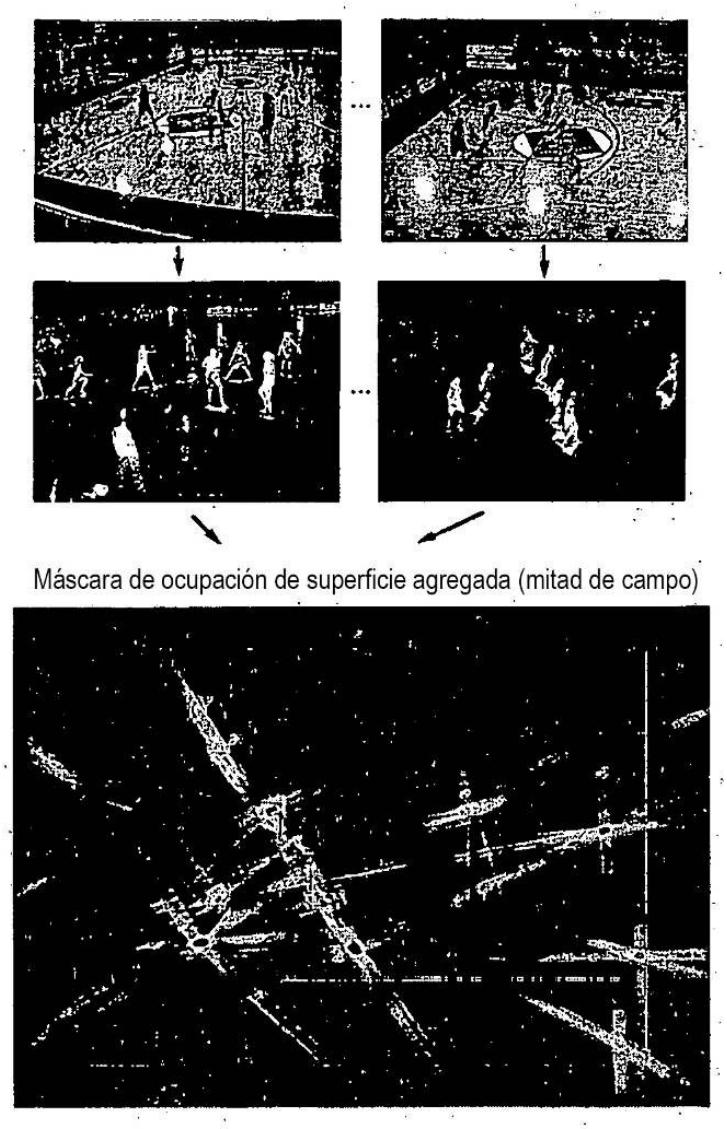


Figura 2. Detección de personas multi-vista. Las máscaras de primer plano se proyectan en un conjunto de planos que son paralelos al plano de superficie para definir un mapa de ocupación de plano de superficie, desde el que se infiere directamente la posición de los jugadores.

Nuestro enfoque es un intento para tomar lo mejor de ambas categorías. Propone un enfoque de abajo a arriba computacionalmente eficaz que puede aprovechar todo el conocimiento a priori que tenemos acerca de la silueta del objeto. Específicamente, el cálculo de abajo a arriba de la máscara de ocupación de superficie descrito en la sección III-B aprovecha el hecho de que la base de la silueta radica en el plano de superficie (de manera similar a soluciones de abajo a arriba anteriores), pero también que la silueta es una forma aproximadamente rectangular vertical (que se reservó anteriormente en enfoques de arriba a abajo). Como una segunda contribución, la sección III-C propone una heurística voraz sencilla para resolver la interferencia que tiene lugar entre las siluetas proyectadas desde distintas vistas por distintos objetos. Nuestros resultados experimentales revelan que esta interferencia fue el origen de muchas falsas detecciones mientras se infieren las posiciones de objetos reales desde la máscara de ocupación de

superficie. Hasta ahora, este fenómeno se tuvo únicamente en cuenta mediante el enfoque de arriba a abajo descrito en [7], a través de una aproximación iterativa compleja de las probabilidades de ocupación posteriores conjuntas. En contraste, aunque aproximado, nuestro enfoque parece que es tanto eficaz como efectivo.

B. Enfoque propuesto, cálculo de máscara de ocupación de plano de superficie

Similar a [9], [10], [7], [1], nuestro enfoque lleva a cabo la detección de cambio de vista única independientemente de cada vista para calcular un mapa de probabilidad de cambio. Para este fin, se implementa un algoritmo de resta de segundo plano convencional basándose en mezcla de modelado gaussiano. Para fusionar las siluetas de primer plano binarias resultantes, nuestro método las proyecta para montar una máscara de ocupación de superficie. Sin embargo, en contraste a enfoques de abajo a arriba anteriores [9], [10], no consideramos la proyección en el plano de superficie únicamente, sino en un conjunto de planos que son paralelos al plano de superficie, y cortan el objeto para detectar a diferentes alturas. Bajo la suposición de que el objeto de interés permanece aproximadamente de manera vertical, la proyección acumulativa de todas estas proyecciones en un plano de vista superior virtual refleja realmente la ocupación de plano de superficie. Esta sección explica cómo se calcula la máscara asociada a cada vista. La siguiente sección investiga cómo unir la información proporcionada mediante las múltiples vistas para detectar personas.

Formalmente, el cálculo de la máscara de ocupación de superficie G_i asociada a la i-ésima vista se describe como sigue. En un momento dado, la i-ésima vista es la fuente de una imagen de silueta restada de segundo plano binaria $B_i \in \{0,1\}^{M_i}$, donde M_i es el número de píxeles de la cámara i, $1 \leq i \leq C$. Como se ha explicado anteriormente, B_i se proyecta en un conjunto de L planos de referencia que se definen para que sean paralelos al plano de superficie, a intervalos de altura regulares, y hasta la altura típica de un jugador. Por lo tanto, para cada vista i, definimos G_i^j para que sea la proyección de la i-ésima máscara binaria en el j-ésimo plano. G_i^j se calcula aplicando la homografía que distorsiona cada píxel dese la cámara i a su posición correspondiente en el j-ésimo plano de referencia, con $0 \leq j < L$. Por construcción, los puntos desde B_i que se etiquetan a 1 debido a la presencia de un jugador en el j-ésimo plano de referencia se proyectan a la correspondiente posición de vista superior en G_i^j. Por lo tanto, se espera la suma G_i de las proyecciones obtenidas a diferentes alturas y desde diferentes vistas para destacar posiciones de vista superior de jugadores que permanecen verticalmente.

A medida que L aumenta, el cálculo de G_i en una posición de superficie x tiene hacia la integración de la proyección de B_i en un segmento vertical anclado en x. Esta integración puede calcularse de manera equivalente en B_i, a lo largo de la proyección hacia atrás del segmento vertical. Para acelerar adicionalmente los cálculos, observamos que, a través de la transformación apropiada de B_i, es posible conformar los dominios de integración proyectados hacia atrás de modo que corresponden a segmentos de líneas verticales en la vista transformada, haciendo de esta manera el cálculo de integrales particularmente eficaz a través del principio de imágenes integrales. La Figura 3 ilustra esa transformación específica para una vista particular. La transformación se ha diseñado para tratar un objetivo doble. En primer lugar, los puntos del espacio en 3D localizado en la misma línea vertical se han de proyectar en la misma columna en la vista transformada (punto de fuga vertical en el infinito). En segundo lugar, los objetos verticales que permanecen en la superficie y cuyos pies se proyectan en la misma línea horizontal de la vista transformada tienen que mantener mismas relaciones de alturas proyectadas.

Una vez que se cumple la primera propiedad, los puntos en 3D que pertenecen a la línea vertical que permanecen por encima de un punto dado desde el plano de superficie simplemente se proyectan en la columna de la vista transformada que permanece por encima de la proyección del punto de plano de superficie en 3D. Por lo tanto, $G_i(x)$ se calcula simplemente como la integral de la vista transformada a través de este segmento proyectado hacia atrás vertical. La conservación de la altura a lo largo de las líneas de la vista transformada, simplifica los cálculos incluso más.

Para vistas laterales, estas dos propiedades pueden conseguirse moviendo virtualmente - a través de transformaciones de homografía - la dirección de visualización de la cámara (eje principal) para proporcionar el punto de fuga vertical en el infinito y asegurar que la línea del horizonte es horizontal. Para las vistas superiores, el eje principal se establece perpendicular a la superficie y se realiza un mapeo polar para conseguir las mismas propiedades. Obsérvese que en algunas configuraciones geométricas, estas transformaciones pueden inducir sesgado fuerte de las vistas.

C. Enfoque propuesto: detección de personas a partir de ocupación de superficie

Dadas las máscaras de ocupación de superficie G_i para todas las vistas, explicamos ahora cómo inferir la posición de las personas que permanecen en la superficie. A priori, conocemos que (i) cada jugador induce un grupo denso en la suma de máscaras de ocupación de superficie, y (ii) el número de personas a detectar es igual a un valor conocido K, por ejemplo $K = 12$ para baloncesto (jugadores + árbitros).

Por esta razón, en cada localización de superficie x, consideramos la suma de todas las proyecciones -

normalizadas mediante el número de vistas que realmente cubren x -, y averiguar los puntos de intensidad superior en esta máscara de ocupación de superficie agregada (véase la Figura 2 para un ejemplo de máscara de ocupación de superficie agregada). Para localizar estos puntos, hemos considerado en primer lugar un enfoque voraz inicial que es equivalente a un procedimiento de búsqueda de coincidencia iterativa. En cada etapa el proceso de búsqueda de coincidencia maximiza el producto interno entre un núcleo gaussiano traducido, y la máscara de ocupación de superficie agregada. La posición del núcleo que induce el producto interno mayor define la posición del jugador. Antes de ejecutar la siguiente iteración, la contribución del núcleo gaussiano se resta de la máscara agregada para producir una máscara residual. El proceso se itera hasta que se hayan encontrado suficientes jugadores.

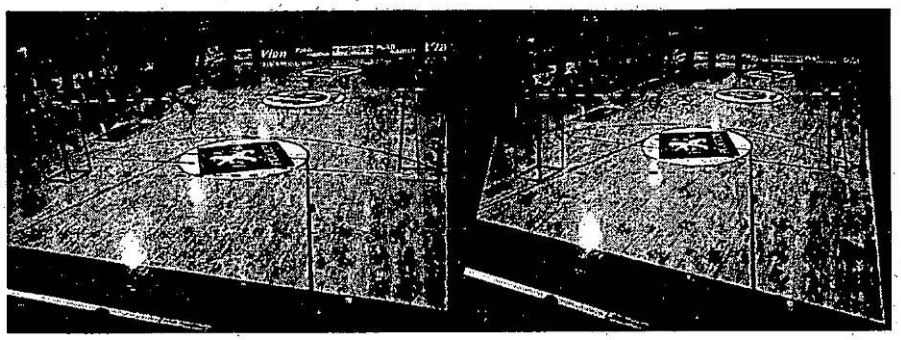


Figura 3. Cálculo eficaz de la máscara de ocupación de superficie: la vista original (a la izquierda) se mapea a un plano a través de una combinación de homografías que se eligen de modo que (1) se conserva la verticalidad durante la proyección desde la escena en 3D a la vista transformada, y (2) se conserva la relación de las alturas entre la escena en 3D y la vista proyectada para los objetos que radican en la misma línea en la vista transformada.

Este enfoque es sencillo, pero sufre de muchas falsas detecciones en la intersección de las proyecciones de distintas siluetas de jugadores desde diferentes vistas. Esto es debido al hecho de que las oclusiones no inducen linealidades[1] en la definición de la máscara de ocupación de superficie. Por lo tanto, una vez que se sabe que algunas personas están presentes en el campo de superficie afecta a la información que puede recuperarse desde las máscaras de cambio binario en cada vista. En particular, si la línea vertical asociada a una posición x es ocluida por/ocluye otro jugador cuya presencia es muy probable, esta vista particular no debería aprovecharse para decidir si hay o no un jugador en x.

Por esta razón, proponemos refinar nuestro enfoque inicial como sigue.

Para inicializar el proceso, definimos $G_i^0(\mathbf{x})$ para que sea la máscara de ocupación de superficie G_i asociada a la i-ésima vista (véase la sección III-B), y establecer $w_i^0(\mathbf{x})$ a 1 cuando x se cubre mediante la i-ésima vista, y a 0 de otra manera. Cada iteración se ejecuta a continuación en dos etapas. En la iteración n, la primera etapa averigua la posición más probable del n^{ésimo} jugador, conociendo la posición de los (n - 1) jugadores localizados en iteraciones anteriores. La segunda etapa actualiza las máscaras de ocupación de superficie de todas las vistas para eliminar la contribución del jugador recién localizado.

Formalmente, la primera etapa de la iteración n agrega la máscara de ocupación de superficie desde todas las vistas, y a continuación averigua el grupo más denso en esta máscara. Por lo tanto, calcula la máscara agregada G^n en la iteración n como

$$G^n(\mathbf{x}) = \frac{\sum_{i=1}^{C} w_i^n(\mathbf{x}) G_i^n(\mathbf{x})}{\sum_{i=1}^{C} w_i^n(\mathbf{x})}, \qquad (1)$$

y a continuación define la posición más probable x_n para el n^{ésimo} jugador mediante

$$\mathbf{x_n} = \underset{\mathbf{y}}{\operatorname{argmax}} < G^n(\mathbf{x}), k(\mathbf{y}) >, \qquad (2)$$

donde $k(y)$ indica un núcleo gaussiano centrado en y cuyo apoyo espacial corresponde a la anchura típica de un jugador.

[1]En otras palabras, la máscara de ocupación de superficie de un grupo de jugadores no es igual a la suma de máscaras de ocupación de superficie proyectadas por cada jugador individual.

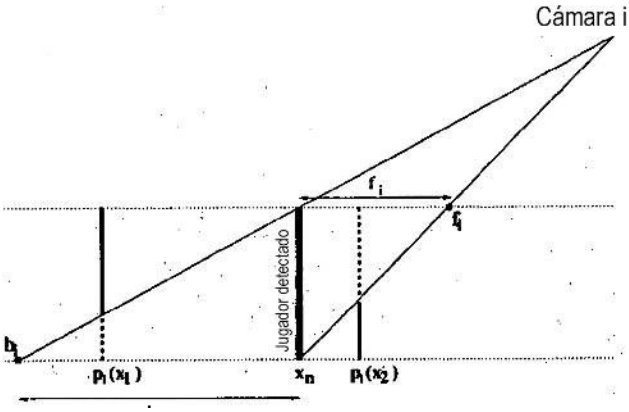


Figura 4. Impacto de oclusiones en la actualización de máscara de ocupación de superficie asociada a la cámara i. La parte discontinua de la silueta vertical que permanece en $p_i(x_1)$ y $p_i(x_2)$ son conocidas para etiquetarse como primer plano puesto que se sabe que un jugador permanece en x_n. Por lo tanto se hacen inútiles para inferir si un jugador está localizado en x_1 y x_2, respectivamente.

En la segunda etapa, la máscara de ocupación de superficie de cada vista se actualiza para tener en cuenta la presencia del n^{ésimo} jugador. En la posición de superficie x, consideramos que el apoyo típico de una silueta de jugador en la vista i es un cuadro rectangular de anchura *W* y altura *H*, y observamos que la parte de la silueta que ocluye o es ocluida por el jugador recién detectado no proporciona ninguna información acerca de la presencia potencial de un jugador en la posición x. Indicando $\alpha_i(x, x_n)$ la fracción de la silueta en la posición de superficie x que se hace no informativa en la vista i como consecuencia de la presencia de un jugador en x_n. Para estimar esta relación, consideramos la geometría del problema. La Figura 4 representa un plano $\mathcal{P}_i$ que es ortogonal a la superficie, mientras pasa a través de la i-ésima cámara y la posición de jugador x_n. En $\mathcal{P}_i$, consideramos dos puntos de interés, en concreto b_i y f_i, que corresponden a los puntos en los que los rayos, originados en la i-ésima cámara y que pasan a través de la cabeza y los pies del jugador, intersectan el plano de superficie y el plano paralelo a la superficie en la altura *H*, respectivamente. Indicamos f_i (b_i) para que sea la distancia entre f_i (b_i) y la línea vertical que apoya al jugador n en $\mathcal{P}_i$. Consideramos también $p_i(x)$ para indicar la proyección ortogonal de x en $\mathcal{P}_i$, y midiendo $d_i(x)$ la distancia entre x y $\mathcal{P}_i$. Basándose en estas definiciones, la relación $\alpha_i(x, x_n)$ se estima mediante

$$\alpha_i(\mathbf{x}, \mathbf{x_n}) = [(\delta - \min(||\mathbf{p_i}(\mathbf{x}) - \mathbf{x_n}||, \delta))/\delta] \cdot [1 - \min(d_i(\mathbf{x})/W, 1)] \qquad (3)$$

siendo δ igual a f_i o b_i dependiendo de si $p_i(x)$ radica delante o detrás de x_n, con respecto a la cámara. En (3), el primer y segundo factores reflejan la desalineación de x y x_n en $\mathcal{P}_i$ y la ortogonalidad a $\mathcal{P}_i$, respectivamente.

Dada $\alpha_i(x, x_n)$, la máscara de ocupación de superficie y peso de agregación de la i-ésima cámara en posición x se actualiza como sigue:

$$G_i^{n+1}(\mathbf{x}) = \max(G_i^n(\mathbf{x}) - \alpha_i(\mathbf{x}, \mathbf{x_n}) G_i^0(\mathbf{x_n}), 0) \qquad (4)$$

$$w_i^{n+1}(\mathbf{x}) = \max(w_i^n(\mathbf{x}) - \alpha_i(\mathbf{x}, \mathbf{x_n}) \qquad (5)$$

Para eficacia computacional mejorada, limitamos las posiciones x investigadas en el enfoque refinado al máximo local 30 que se ha detectado mediante el enfoque inicial.

Para completar, se observa que el procedimiento de actualización anteriormente descrito omite la interferencia potencial entre oclusiones producidas por distintos jugadores en la misma vista. Sin embargo, la consecuencia de esta aproximación está lejos de ser drástica, puesto que finaliza, sin afectar a la información que se aprovecha realmente. Teniendo en cuenta estas interferencias requeriría proyectar hacia atrás las siluetas del jugador en cada vista, tendiendo de esta manera hacia un enfoque de arriba a abajo computacionalmente y en memoria caro tal como el presentado en [7].

Además, vale la pena mencionar que, en un contexto de arriba a abajo, los autores en [1] o en [7] proponen formulaciones que averiguan simultáneamente las K posiciones que explican mejor las múltiples observaciones de máscara de primer plano. Sin embargo, considerar conjuntamente todas las posiciones aumenta la dimensionalidad del problema, e impactan drásticamente la carga computacional. Puesto que nuestros resultados experimentales

muestran que nuestro método propuesto no sufre de la debilidad habitual de los algoritmos voraces, tal como una tendencia a engancharse en malos mínimos locales, creemos que compara muy favorablemente a cualquier formulación conjunta del problema, se resuelve típicamente basándose en técnicas de optimización proximales iterativas.

IV. Reconocimiento de dígito de jugadores

Esta sección considera el reconocimiento de los caracteres digitales impresos en las camisetas deportivas de los atletas. El enfoque propuesto se representa en la Figura 2. Para cada posición detectada en el plano de superficie, se proyecta un cuadro delimitador conservador de 0,8 m x 2 m en cada una de las vistas. Cada cuadro se procesa a continuación de acuerdo con un enfoque que es similar al método de aproximado a preciso introducido en [12]. En la etapa inicial, la imagen del cuadro delimitador se segmenta en regiones. Las regiones candidatas de dígitos se filtran a continuación basándose en atributos contextuales. Eventualmente, las regiones seleccionadas se clasifican en dígitos '0-9' o clases binarias, y la identidad del jugador se define mediante voto mayoritario, basándose en los resultados obtenidos en diferentes vistas. Nuestro enfoque de método propuesto se diferencia de [12] en la manera en que se implementa cada una de estas etapas.

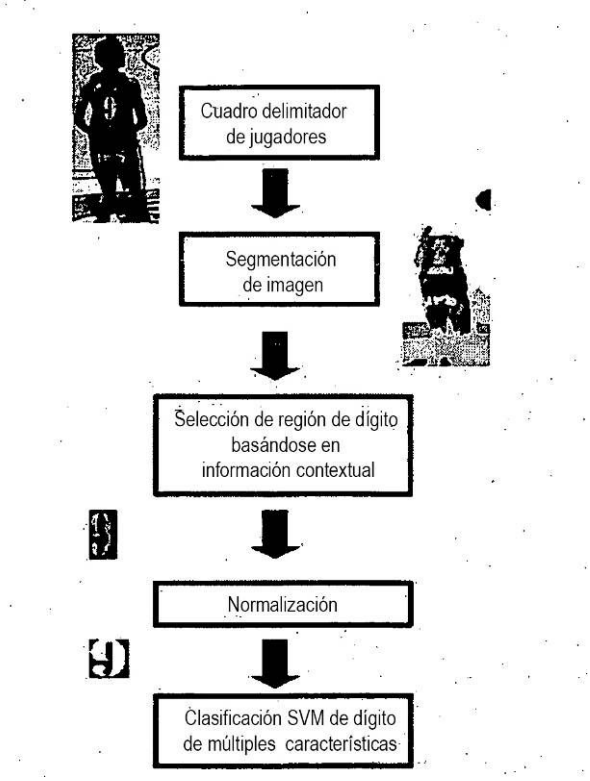


Figura 5. Reconocimiento de dígitos impresos en camisetas de jugadores a través de segmentación, selección y clasificación de regiones que es probable que representen dígitos.

Nuestra etapa de segmentación está basada en el algoritmo de desplazamiento de media [6], que es una técnica de reconocimiento de patrón que es particularmente bien adecuada para delinear regiones más densas en algún espacio característico arbitrariamente estructurado. En la segmentación de imagen de desplazamiento de media, la imagen se representa típicamente como una malla bidimensional de píxeles L*u*v de 3 dimensiones. El espacio de la malla se conoce como el dominio espacial, mientras la información de color corresponde al dominio de rango. Los vectores de localización y de rango se concatenan en un dominio espacial-rango conjunto, y se define un núcleo multivariado como el producto de dos núcleos radialmente simétricos en cada dominio, que permite la definición independiente de los parámetros de ancho de banda h_s y h_r para los dominios espacial y de rango, respectivamente [6]. Los máximos locales de la densidad de dominio conjunto se calculan a continuación, y los modos que están más cercanos al h_s en el dominio espacial y a h_r en el dominio de rango se cortan en modos significativos. Cada píxel se asocia a continuación con un modo significativo de la densidad de dominio conjunto localizado en su cercanía. Eventualmente, se eliminan las regiones espaciales que contienen menos de M píxeles. En nuestro caso, puesto que hay un fuerte contraste entre dígito y camiseta, podemos permitir un alto valor para h_r, que se establece a 8 en nuestras simulaciones. El parámetro h_s equilibra el tiempo de ejecución de la segmentación y filtrado posterior y las etapas de clasificación. De hecho, un valor de h_r pequeño define un núcleo menor, que hace la segmentación más rápida pero también da como resultado un número mayor de regiones a procesar en etapas posteriores. En nuestras simulaciones, h_r se ha establecido a 4, mientras M se ha fijado a 20.

Para filtrar regiones que evidentemente no corresponden a dígitos, nos basamos en las siguientes observaciones:

• Regiones de dígitos validas nunca tocadas por el borde del cuadro delimitador (conservativo).

• Regiones de dígitos válidas que son rodeadas mediante una única región coloreada homogéneamente. En la práctica, nuestro algoritmo selecciona las regiones para las que los vecinos de los 4 puntos extremos (superior/inferior, derecho/izquierdo) de la región pertenecen a la misma región.

• La altura y anchura de regiones válidas varía entre dos valores que se definen relativamente al tamaño del cuadro delimitador. Puesto que el tamaño de la delimitación se define de acuerdo con métricas del mundo real, el criterio de tamaño adapta implícitamente el rango de los valores de altura y anchura al efecto de la perspectiva resultante de la distancia entre el objeto detectado y la cámara.

Para completar, vale la pena mencionar que algunas fuentes particulares dividen algunos dígitos en dos regiones distintas. Por esta razón, las regiones de dígitos candidatas están compuestas de una región única o un par de regiones que satisfacen los criterios anteriores.

Las (parejas de) regiones que se han seleccionado como elegibles para procesamiento posterior a continuación se normalizan y clasifican. La normalización implica alineación horizontal del eje principal mayor, como se deduce a través del cálculo de momentos de inercia, y conversión a una máscara binaria de 24 x 24. La clasificación está basada en la estrategia SVM de múltiples clases ‘uno contra uno’ [8], como se recomienda e implementa mediante la biblioteca LIBSVM [3]. Se entrena una SVM de dos clases para cada par de clases, y se aprovecha una estrategia de voto mayoritario para inferir la clase (dígito de 0 a 9 o clase binaria) desde el conjunto de decisiones de clasificación binaria. En la práctica, alimentar el clasificador a cada muestra de región se describe mediante un vector característico de 30 dimensiones, en concreto:

• 1 valor para definir el número de huecos en la región;

• 3 valores que corresponden a momentos de segundo orden m02, m20, y m22;

• 2 valores para definir el centro de masa de la región;

• 2 x 12 valores para definir el histograma de la región a lo largo del eje vertical y horizontal.

Los números con dos dígitos se reconstruyen basándose en la detección de dos dígitos adyacentes. Para ejecutar nuestras simulaciones, hemos entrenado el clasificador de SVM basándose en más de 200 muestras segmentadas manualmente de cada dígito, y en 1200 muestras de la clase binaria. Las muestras de clase binaria corresponden a regiones no de dígitos que se segmentan automáticamente en una de las vistas, y cuyo tamaño es coherente con uno de un dígito.

V. Seguimiento de jugadores detectados

Para seguir jugadores detectados, hemos implementado un algoritmo rudimentario aunque eficaz. La propagación de seguimientos se hace actualmente a través de un horizonte de 1 fotograma, basándose en el algoritmo de asignación general de Munkres [11]. Se usa Gating para evitar coincidencias improbables, y se usa un módulo de análisis de alto nivel para enlazar juntos seguimientos parciales usando estimación de color de camiseta. En el futuro, deberían usarse técnicas de coincidencia de grafos para evaluar hipótesis de coincidencia de horizonte más largas. Deberían implementarse también análisis más sofisticados de alto nivel, por ejemplo para aprovechar la información de reconocimiento de jugador disponible o para duplicar los seguimientos parciales que siguen a dos jugadores que están muy cercanos entre sí.

VI. Validación experimental

A. Detección y seguimiento de jugador

Para evaluar nuestro algoritmo de detección de jugador, hemos medido las puntuaciones de detección perdida media y falsa detección a través de 180 instantes de tiempo diferentes y espaciados regularmente en el intervalo desde 18:47:00 a 18:50:00, que corresponde a un segmento temporal para el que está disponible una verdad sobre el terreno manual. Esta información de verdad sobre el terreno consiste en las posiciones de los jugadores y árbitros en el sistema de referencia de coordenadas de la cancha. Consideramos que dos objetos no puedan coincidirse si la distancia medida en la superficie es mayor de 30 cm. La Figura 6 presenta varias curvas ROC, obteniéndose cada curva variando el umbral de detección para un método de detección dado. Se comparan tres métodos, y para cada uno de ellos evaluamos nuestro algoritmo propuesto para mitigar falsas detecciones. Como un método de referencia y primero, consideramos el enfoque seguido por [9], [10], que proyectan las máscaras de primer plano de todas las vistas únicamente en el plano de superficie. El pobre rendimiento de este último enfoque se debe principalmente a

las sombras de los jugadores, y a la pequeña contribución de los pies de los jugadores a las máscaras de primer plano. Para validar esta interpretación, en el segundo método, hemos proyectado las máscaras de primer plano en un único plano localizado un metro por encima del plano de superficie. Haciendo esto, la influencia de las sombras se atenúa drásticamente, mientras que la contribución principal ahora se origina desde las partes centrales del cuerpo, que están normalmente bien representadas en las máscaras de primer plano. Observamos mejoras significativas en comparación con [9], [10]. El tercer y último método de detección presentado en la Figura 6 es nuestro método propuesto. Observamos que el beneficio obtenido desde nuestra integración de ocupación de superficie es sorprendente. La mejora conseguida por nuestro detector de falsas alarmas es también bastante evidente. Además, los cruces en la Figura 6 presentan un punto de operación conseguido después de seguimiento rudimentario de posiciones detectadas. Observamos que tener en cuenta la uniformidad temporal puede mejorar aún los resultados de detección.

En la configuración de APIDIS, todas las áreas de la cancha de baloncesto no se cubren mediante el mismo número de cámaras. La Figura 7 muestra la influencia de la cobertura de la cámara en las tasas de detecciones perdidas y falsas. Muestra también que en las áreas con alta cobertura, la mayoría de las detecciones perdidas son debido a jugadores que permanecen muy cercanos entre sí.

B. Reconocimiento de jugador

Para validar la cascada de reconocimiento de jugador, hemos seleccionado 190 cuadros delimitadores de jugadores desde cámaras de vistas laterales. En cada cuadro delimitador seleccionado, el dígito fue visible y podía leerse por un observador humano, a pesar de distorsiones de apariencia posiblemente significativas. La Tabla I resume nuestros resultados de reconocimiento. La tasa de reconocimiento está por encima del 73 %. De manera más interesante, observamos que cuando el dígito no se reconoció, se asignó más a menudo a la clase binaria, o no pasó el análisis contextual debido a error de segmentación. Además, el restante 4 % de falsos positivos no incluyeron ningún desemparejamiento real entre dos dígitos. Realmente, el 75 % de falsos positivos se debieron a la pérdida de un dígito en un número de dos dígitos. En otros casos, se han reconocido dos dígitos, el correcto y uno falso detectado.

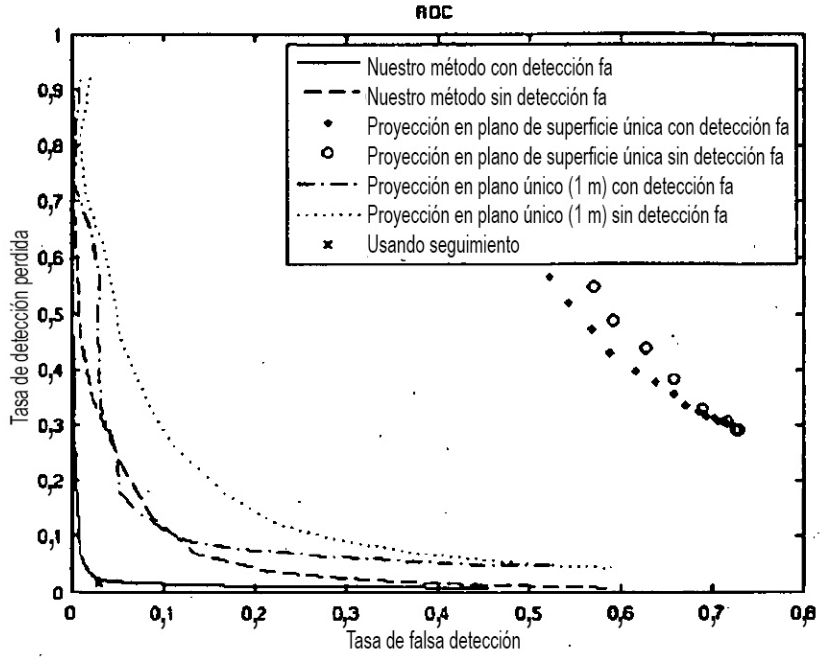


Fig. 6. Análisis ROC de rendimiento de detección de jugador

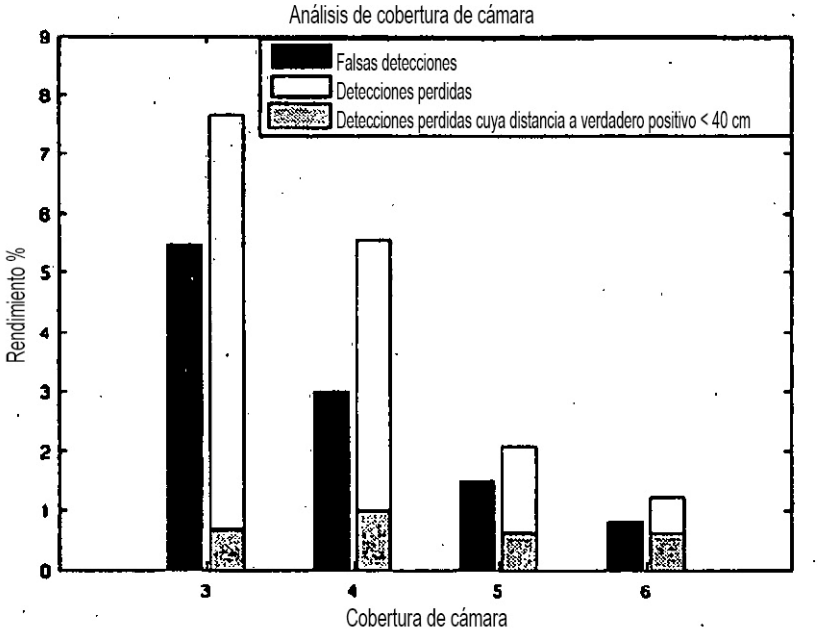


Fig. 7. Detección de jugador con respecto a cobertura de cámara

Además, un análisis más detallado ha revelado que la mayoría de los jugadores no reconocidos permanecieron en el lado opuesto del campo, en comparación con la vista de la cámara desde la que se extrajo el cuadro delimitador. En este caso, la altura del dígito se reduce a menos de 15 píxeles, que explica el pobre rendimiento de reconocimiento, por debajo del 50 %. En contraste, una cámara localizada en el mismo lado del campo que el jugador consigue cerca del 90 % de tasa de reconocimiento correcta.

Basándose en estas observaciones, estamos razonablemente convencidos que el rendimiento de reconocimiento de nuestro sistema será suficientemente bueno para asignar una etiqueta correcta a segmentos cortos de trayectorias de jugador, proporcionando de esta manera una herramienta valiosa tanto para localizar ambigüedades de seguimiento o para favorecer un jugador preferido durante producción de resumen de vídeo.

Reconocimiento	73 %
Error de segmentación	11 %
Falso negativo	12 %
Falso positivo	4 %

TABLA I
RENDIMIENTO DE RECONOCIMIENTO DE JUGADOR.

VII. Conclusión

Hemos presentado algoritmos de procesamiento de vídeo para definir la posición e identidad de atletas que juegan en un campo deportivo, rodeados por un conjunto de cámaras ligeramente sincronizadas. La detección se basa en la definición de un mapa de ocupación de superficie, mientras que el reconocimiento del jugador se fundamenta en pre-filtrado de regiones segmentadas y en clasificación de SVM de múltiples clases. Los experimentos sobre el conjunto de datos de la vida real del APIDIS demuestran la relevancia de los enfoques propuestos.

Referencias

[1] A. Alahi, Y. Boursier, L. Jacques, y P. Vandergheynst, “A sparsity constrained inverse problem to locate people in a network of cameras”, en *Proceedings of the 16th International Conference on Digital Signal Processing (DSP)*, Santorini, Grecia, julio de 2006.

[2] J. Berclaz, F. Fleuret, y P. Fua, “Principled detection-by-classification from multiple views”, en *Proceedings of the*

International Conference on Computer Vision Theory and Application (VISAPP), vol. 2, Funchal, Madeira, Portugal, enero de 2008, págs. 375-382.

[3] C.-C. Chang y C.-J. Lin, "LIBSVM: A library for support vector machines", en http://www.csie.ntu.edu.tw/cjlin/papers/libsvm.pdf.

[4] F. Chen y C. De Vleeschouwer, "A resource allocation framework for summarizing team sport videos", en *IEEE International Conference on Image Processing*. El Cairo, Egipto, noviembre de 2009.

[5] —, "Autonomous production of basket-ball videos from multi-sensored data with personalized viewpoints", en *Proceedings of the 10th International Workshop on Image Analysis for Multimedia Interactive Services*, Londres. RU, mayo de 2009.

[6] D. Comaniciu y P. Meer, "Mean shift: a robust approach toward feature space analysis". *IEEE Transactions on Pattern Analysis y Machine Intelligence*, vol. 24, nº 5, págs. 603-619, mayo de 2002.

[7] F. Fleuret. J. Berclaz, R. Lengagne, y P. Fua, "Multi-camera people tracking with a probabilistic occupancy map", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, nº 2, págs. 267-282, febrero de 2008.

[8] C.-W. Hsu y C.-J. Lin, "A comparison of methods for multiclass support vector machines", *IEEE Transactions on Neural Networks*. vol. 13, nº 2. págs. 415-425, marzo de 2002.

[9] S. Khan y M. Shah, "A multiview approach to tracing people in crowded scenes using a planar homography constraint", en *Proceedings of the 9th European Conference on Computer Vision (ECCV)*, vol. 4, Graz, Austria, mayo de 2006, págs. 133-146.

[10] A. Lanza, L. Di Stefano, J. Berclaz, F. Fleuret, y P. Fua, "Robust multiview change detection," en *British Machine Vision Conference (BMVC)*, Warwick, RU, septiembre de 2007.

[11] J. Munkres, "Algorithms for the assignment and transportation problems". en *SIAM J. Control*, vol. 5, 1957, págs. 32-38.

[12] Q. Ye, Q. Huang, S. Jiang, Y. Liu, y W. Gao, "Jersey number detection in sports video for athlete identification", en *Proceedings of the SPIE, Visual Communications y Image Processing*, vol. 5960, Beijing, China, julio de 2005, págs. 1599-1606.

REIVINDICACIONES

1. Un método basado en ordenador para producción autónoma de un vídeo editado a partir de múltiples corrientes de vídeo capturadas por una pluralidad de cámaras distribuidas alrededor de una escena de interés para seleccionar, como una función del tiempo, puntos de vista óptimos para ajustar la resolución de visualización y otras preferencias de usuario, y para suavizar estas secuencias para una narrativa continua y elegante, comprendiendo el método:

• detectar objetos de interés en las imágenes de las corrientes de vídeo,

• seleccionar para cada localización/posición de cámara prevista, un campo de visión obtenido:

- o bien recortando la imagen capturada por una cámara fija, definiendo mediante ello parámetros de recorte,

- o bien seleccionando los parámetros de panorámica-inclinación-zoom de una cámara motorizada o una cámara virtual, aproximando dicha cámara virtual una imagen en una posición arbitraria con un campo de visión arbitrario aprovechando una red distribuida de dichas cámaras,

estando seleccionado el campo de visión basándose en procesamiento conjunto de las posiciones de múltiples objetos de interés que se han detectado, en el que la selección se hace de una manera que equilibra las métricas de completitud y cercanía como una función de preferencias de usuario individuales, en el que la completitud cuenta el número de objetos de interés que se incluyen y son visibles en el punto de vista visualizado, y la cercanía mide el número de píxeles que están disponibles para describir los objetos de interés,

• montar el vídeo editado seleccionando y concatenando segmentos de vídeo proporcionados mediante una o más cámaras individuales, en el que el montaje se hace de una manera que equilibra las métricas de completitud y cercanía a lo largo del tiempo, mientras se suaviza la secuencia de dichos parámetros de recorte y/o panorámica-inclinación-zoom asociados a segmentos concatenados, en el que el proceso de suavizado se implementa basándose en mecanismo de filtrado temporal de paso bajo lineal o no lineal.

2. El método basado en ordenador de la reivindicación 1, en el que la selección se hace de manera que maximizan las métricas de completitud y cercanía.

3. El método basado en ordenador de la reivindicación 1 o 2, en el que la métrica de completitud se define como el porcentaje de objetos incluidos mientras que la métrica de cercanía se define como área de píxeles media que se usa para representar.

4. El método de la reivindicación 1, 2 o 3 que comprende adicionalmente puntuar el punto de vista seleccionado en cada vista de cámara de acuerdo con la calidad de su equilibrio de completitud/cercanía, y a su grado de oclusiones.

5. El método de la reivindicación 1, 2 o 3, que comprende adicionalmente seleccionar el campo de visión óptimo en cada cámara, en un instante de tiempo dado, en el que un campo de visión v_k en la $k^{ésima}$ vista de cámara se define mediante el tamaño S_k y el centro c_k de la ventana que se recorta en la $k^{ésima}$ vista para visualización real y se selecciona para incluir los objetos de interés y para proporcionar una descripción a alta resolución de los objetos, y se selecciona un campo de visión óptimo v_k^* para maximizar una suma ponderada de intereses de objetos como sigue

$$\mathbf{v}_k^* = \arg\max_{\{S_k, \mathbf{c}_k\}} \sum_{n=1}^{N} I_n \cdot \alpha(S_k, \mathbf{u}) \cdot m\left(\frac{\left\|\mathbf{x}_{n,k} - \mathbf{c}_k\right\|}{S_k}\right)$$

donde, en la ecuación anterior:

• I_n indica el nivel de interés asignado al $n^{ésimo}$ objeto detectado en la escena,

• $x_{n,k}$ indica la posición del $n^{ésimo}$ objeto en la vista de cámara k,

• la función m(.....) modula los pesos del $n^{ésimo}$ objeto de acuerdo con su distancia al centro de la ventana de punto de vista, en comparación con el tamaño de esta ventana,

• el vector u refleja las preferencias de usuario, en particular, su componente u_{res} define la resolución de la corriente de salida, que está generalmente restringida por el ancho de banda de transmisión o la resolución del dispositivo de usuario final,

• la función $\alpha(.)$ refleja la penalización inducida por el hecho de que la señal nativa capturada mediante la k^{ésima} cámara tiene que sub-muestrearse una vez que el tamaño del punto de vista se hace mayor que la resolución máxima u_{res} permitida mediante el usuario.

6. El método de la reivindicación 5, en el que $\alpha(....)$ se reduce con Sk y la función $\alpha(....)$ es igual a uno cuando $S_k<u_{res}$, y se reduce posteriormente, y en el que $\alpha(....)$ se define opcionalmente mediante:

$$\alpha(S,\mathbf{u})=\left[\min\left(\frac{u_{res}}{S},1\right)\right]^{u_{cercano}}.$$

donde el exponente $u_{cercano}$ es mayor que 1, y aumenta a medida que el usuario prefiere representación de resolución completa de área de acercamiento, en comparación con puntos de vista grandes pero sub-muestreados.

7. El método de cualquiera de las reivindicaciones 4 a 6 en el que la puntuación más alta corresponde a una vista que hace visibles a la mayoría de objetos de interés, y está cerca de la acción.

8. El método de cualquiera de las reivindicaciones 4 a 7, en el que, dado el interés I_n de cada jugador, la puntuación $I_k(v_k, u)$ asociada a la k^{ésima} vista de cámara se define como sigue:

$$I_k(\mathbf{v_k},\mathbf{u})=\sum_{n=1}^{N} I_n\cdot o_k(\mathbf{x}_n|\bar{\mathbf{x}})\cdot h_k(\mathbf{x}_n)\cdot \beta_k(S_k,\mathbf{u})m\left(\frac{\left\|\mathbf{x}_{n,k}-\mathbf{c}_k\right\|}{S_k}\right)$$

donde, en la ecuación anterior:

■ I_n indica el nivel de interés asignado al n^{ésimo} objeto detectado en la escena;

■ x_n indica la posición del n^{ésimo} objeto en el espacio en 3D;

■ $o_k(x_n|\ x)$ mide la relación de oclusión del n^{ésimo} objeto en la vista de cámara *k*, conociendo la posición de todos los otros objetos, estando definida la relación de oclusión de un objeto para que sea la fracción de píxeles del objeto que se ocultan por otros objetos cuando se proyectan en el sensor de la cámara;

■ la altura $h_k(x_n)$ se define para que sea la altura en píxeles de la proyección en la vista k de una altura de referencia de un objeto de referencia localizado en x_n; el valor de $h_k(x_n)$ se calcula directamente basándose en la calibración de la cámara, o cuando la calibración no está disponible, puede estimarse basándose en la altura del objeto detectado en la vista k;

■ la función $\beta_k(.)$ refleja el impacto de las preferencias de usuario en términos de vista de cámara y resolución de visualización.

9. El método de la reivindicación 8, en el que $\beta_k(.)$ se define como:

$$\beta_k(S,\mathbf{u})=u_k\cdot\alpha(S,\mathbf{u}),$$

donde u_k indica el peso asignado a la k^{ésima} cámara, y $\alpha(S, u)$ se define como en la reivindicación 6.

10. El método de cualquiera de las reivindicaciones 1 a 9 que comprende adicionalmente suavizar la secuencia de índices de cámara y parámetros de punto de vista correspondientes, en el que el proceso de suavizado se implementa, por ejemplo, basándose en dos Campos Aleatorios de Markov, mecanismo de filtrado de paso bajo lineal o no lineal, o mediante un formalismo de modelo de grafo, resuelto basándose en el algoritmo Viterbi convencional.

11. Sistema basado en ordenador para seleccionar, como una función del tiempo, puntos de vista óptimos para ajustar la resolución de visualización y otras preferencias de usuario, y para suavizar estas secuencias para una narrativa continua y elegante que comprende un motor de procesamiento y memoria para producción autónoma de un vídeo editado desde múltiples corrientes de vídeo capturadas mediante una pluralidad de cámaras distribuidas

alrededor de una escena de interés, comprendiendo el sistema:

detector para detectar objetos de interés en las imágenes de las corrientes de vídeo;

primeros medios para seleccionar uno o más puntos de vista de cámara obtenidos o bien recortando la imagen capturada mediante una cámara fija, definiendo mediante ello parámetros de recorte, o bien seleccionando los parámetros de panorámica-inclinación-zoom de una cámara motorizada o una cámara virtual, aproximando dicha cámara virtual una imagen en una posición arbitraria con un campo de visión arbitrario aprovechando una red distribuida de dichas cámaras, estando seleccionado el punto de vista de la cámara para incluir los objetos de interés y basándose en procesamiento conjunto de las posiciones de los múltiples objetos de interés que se han detectado, en el que la selección se hace de manera que equilibra las métricas de completitud y cercanía como una función de preferencias de usuario individuales, en el que la completitud cuenta el número de objetos de interés que se incluyen y son visibles en el punto de vista visualizado, y la cercanía mide el número de píxeles que están disponibles para describir los objetos de interés;

segundos medios para seleccionar parámetros de representación que maximizan y suavizan las métricas de cercanía y completitud concatenando segmentos en las corrientes de vídeo proporcionadas mediante una o más cámaras individuales, en el que el montaje se hace de manera que equilibra las métricas de completitud y cercanía a lo largo del tiempo, mientras se suaviza la secuencia de dichos parámetros de recorte y/o panorámica-inclinación-zoom asociados a segmentos concatenados, en el que el proceso de suavizado se implementa basándose en el mecanismo de filtrado temporal de paso bajo lineal o no lineal.

12. El sistema basado en ordenador de la reivindicación 11, en el que la selección se hace de manera que maximiza las métricas de completitud y cercanía.

13. El método basado en ordenador de la reivindicación 11 o 12, en el que la métrica de completitud se define como el porcentaje de objetos incluidos mientras la métrica de cercanía se define como el área de píxeles media como se usa para representación.

14. El sistema de la reivindicación 11, 12 o 13, que comprende adicionalmente terceros medios para seleccionar variaciones de parámetros de cámara y de imagen para la vista de la cámara que representa la acción como una función del tiempo para un conjunto de métricas de cercanía y completitud conjuntas, estando adaptados opcionalmente los terceros medios para seleccionar variaciones de parámetros de cámara y de imagen para recortar en la vista de la cámara de una cámara estática o para controlar los parámetros de control de una cámara dinámica.

15. El sistema de la reivindicación 11 o 14 que comprende adicionalmente medios para mapear imágenes desde todas las vistas de todas las cámaras a las mismas coordenadas temporales absolutas basándose en una referencia temporal única común para todas las vistas de cámara.

16. El sistema of de cualquiera de las reivindicaciones 11 a 15 que comprende adicionalmente cuartos medios para seleccionar las variaciones de parámetros que optimizan el equilibrio entre completitud y cercanía en cada instante de tiempo, y para cada vista de cámara, en el que el equilibrio de completitud/cercanía se mide opcionalmente como una función de las preferencias de usuario.

17. El sistema de cualquiera de las reivindicaciones 11 a 16, que comprende adicionalmente medios para puntuar el punto de vista seleccionado en cada vista de cámara de acuerdo con la calidad de su equilibrio de completitud/cercanía, y a su grado de oclusiones.

18. El sistema de la reivindicación 17, que comprende adicionalmente medios para calcular los parámetros de una cámara virtual óptima que realiza panorámica, zoom y cambia a través de las vistas para conservar altas puntuaciones de puntos de vista seleccionados mientras se minimiza la cantidad de movimientos de cámara virtual, para el segmento temporal disponible.

19. El sistema de la reivindicación 17 o 18, que comprende adicionalmente quintos medios para seleccionar el punto de vista óptimo en cada vista de cámara, en un instante de tiempo dado, en el que los quintos medios para seleccionar el punto de vista óptimo están adaptados, para un punto de vista v_k en la $k^{ésima}$ vista de cámara que se define mediante el tamaño S_k y el centro c_k de la ventana que se recorta en la $k^{ésima}$ vista para visualización real y se selecciona para incluir los objetos de interés y para proporcionar una alta resolución, está adaptado para seleccionar una descripción de los objetos y un punto de vista óptimo v_k^* para maximizar una suma ponderada de los intereses de objetos como sigue:

$$\mathbf{v}_k^* = \arg\max_{\{S_k, \mathbf{c}_k\}} \sum_{n=1}^{N} I_n \cdot \alpha(S_k, \mathbf{u}) \cdot m\left(\frac{\left\|\mathbf{x}_{n,k} - \mathbf{c}_k\right\|}{S_k}\right)$$

donde, en la ecuación anterior:

I_n indica el nivel de interés asignado al n^{ésimo} objeto detectado en la escena. $x_{n,k}$ indica la posición del n^{ésimo} objeto en la vista de cámara *k*,

la función m(.....) modula los pesos del n^{ésimo} objeto de acuerdo con su distancia al centro de la ventana de punto de vista, en comparación con el tamaño de esta ventana,

el vector *u* refleja las preferencias de usuario, en particular, su componente u_{res} define la resolución de la corriente de salida, que está generalmente restringida por el ancho de banda de transmisión o la resolución del dispositivo del usuario final,

la función $\alpha(.)$ refleja la penalización inducida por el hecho de que la señal nativa capturada mediante la k^{ésima} cámara tiene que sub-muestrearse una vez que el tamaño del punto de vista se hace mayor que la resolución máxima u_{res} permitida mediante el usuario.

20. El sistema de la reivindicación 19, en el que $\alpha(....)$ se reduce con Sk y la función $\alpha(....)$ es igual a uno cuando $S_k < u_{res}$ y se reduce posteriormente, en el que $\alpha(....)$ se define opcionalmente mediante:

$$\alpha(S, \mathbf{u}) = \left[\min\left(\frac{u_{res}}{S}, 1\right)\right]^{u_{cercano}}$$

donde el exponente $u_{cercano}$ es mayor que 1, y aumenta a medida que el usuario prefiere representación a resolución completa de área de acercamiento, en comparación con puntos de vista grandes pero sub-muestreados.

21. El sistema de cualquiera de las reivindicaciones 18 a 20, que comprende adicionalmente sextos medios para seleccionar la cámara en un instante de tiempo dado que hace visibles a la mayoría de objetos de interés, y está cerca de la acción, en el cual un índice de cámara óptimo *k** se selecciona de acuerdo con una ecuación que es similar o equivalente a:

$$k^* = \arg\max_{\{k\}} \sum_{n=1}^{N} I_n \cdot o_k\left(\mathbf{x}_n \middle| \bar{\mathbf{x}}\right) \cdot h_k\left(\mathbf{x}_n\right) \cdot \beta_k\left(S_k^*, \mathbf{u}\right)$$

donde, en la ecuación anterior:

- I_n indica el nivel de interés asignado al n^{ésimo} objeto detectado en la escena;
- x_n indica la posición del n^{ésimo} objeto en el espacio en 3D;
- $o_k(x_n|\ x)$ mide la relación de oclusión del n^{ésimo} objeto en la vista de cámara *k*, conociendo la posición de todos los otros objetos, definiéndose la relación de oclusión de un objeto para que sea la fracción de píxeles del objeto que se ocultan por otros objetos cuando se proyectan en el sensor de la cámara;
- la altura $h_k(x_n)$ se define para que sea la altura en píxeles de la proyección en la vista k de una altura de referencia de un objeto de referencia localizado en x_n; el valor de $h_k(x_n)$ se calcula directamente basándose en la calibración de la cámara, o cuando la calibración no está disponible, puede estimarse basándose en la altura del objeto detectado en la vista k,
- la función $\beta_k(.)$ refleja el impacto de las preferencias de usuario en términos de vista de cámara y resolución de visualización.

22. El sistema de la reivindicación 21, en el que $\beta_k(.)$ se define como:

$$\beta_k(S,\mathbf{u}) = u_k \cdot \alpha(S,\mathbf{u}),$$

donde u_k indica el peso asignado a la k^{ésima} cámara, y $\alpha(S,u)$ se define como en la reivindicación 20.

23. El sistema de cualquiera de las reivindicaciones 20 a 22 que comprende adicionalmente medios para suavizar la secuencia de índices de cámara y parámetros de punto de vista correspondientes, en el que los medios para suavizar están adaptados para suavizar basándose en dos campos aleatorios de Markov, mediante un mecanismo de filtrado de paso bajo lineal o no lineal, mediante un formalismo de modelo de grafo, resuelto basándose en el algoritmo Viterbi convencional.

24. Un medio de almacenamiento de señal legible por máquina no transitoria que almacena un producto de programa informático que comprende segmentos de código que cuando se ejecutan en un motor de procesamiento ejecuta el método de cualquiera de las reivindicaciones 1 a 10 o implementa el sistema de acuerdo con cualquiera de las reivindicaciones 11 a 23.

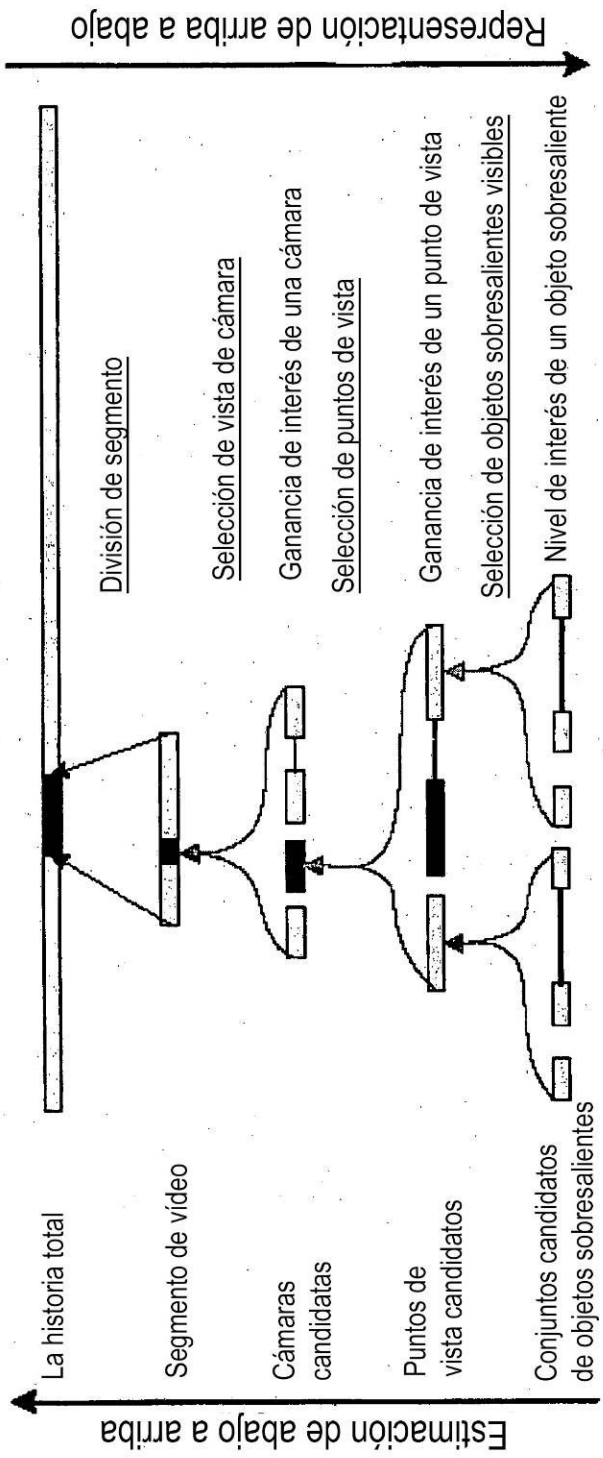


Figura 1: flujo de trabajo jerárquico en nuestra producción personalizada.

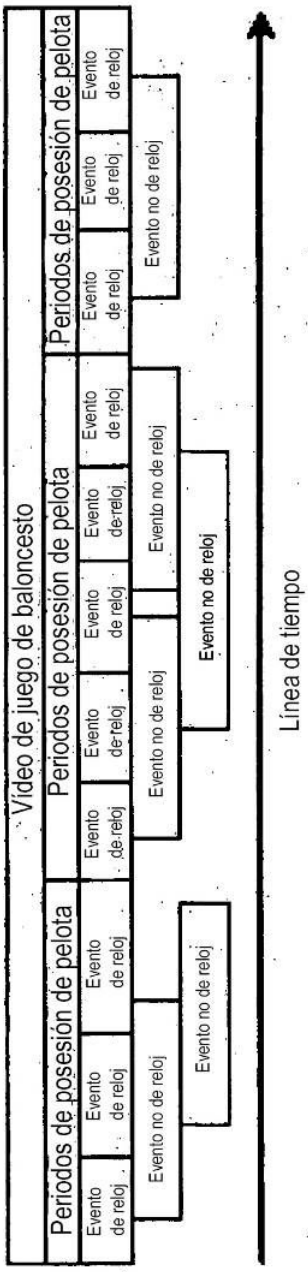
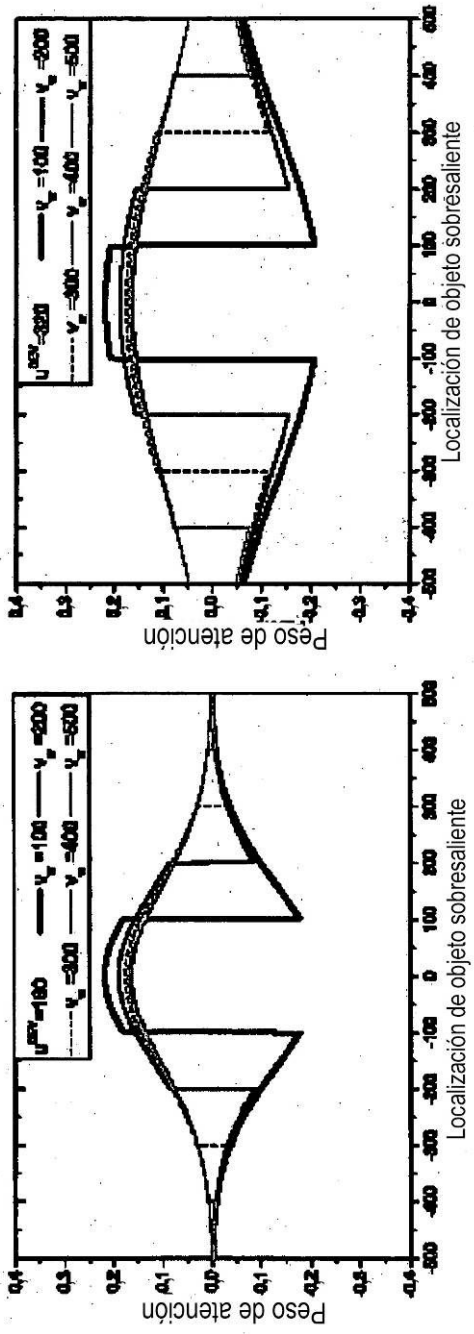


Figura 2: estructura jerárquica de vídeos de baloncesto.



(a) $u^{\mathrm{DEV}} = 160$ (b) $u^{\mathrm{DEV}} = 320$

Figura 3: función de ponderación de atención propuesta en el presente trabajo. Equilibramos entre cercanía y completitud controlando la agudeza de cola y de pico de la función de ponderación.

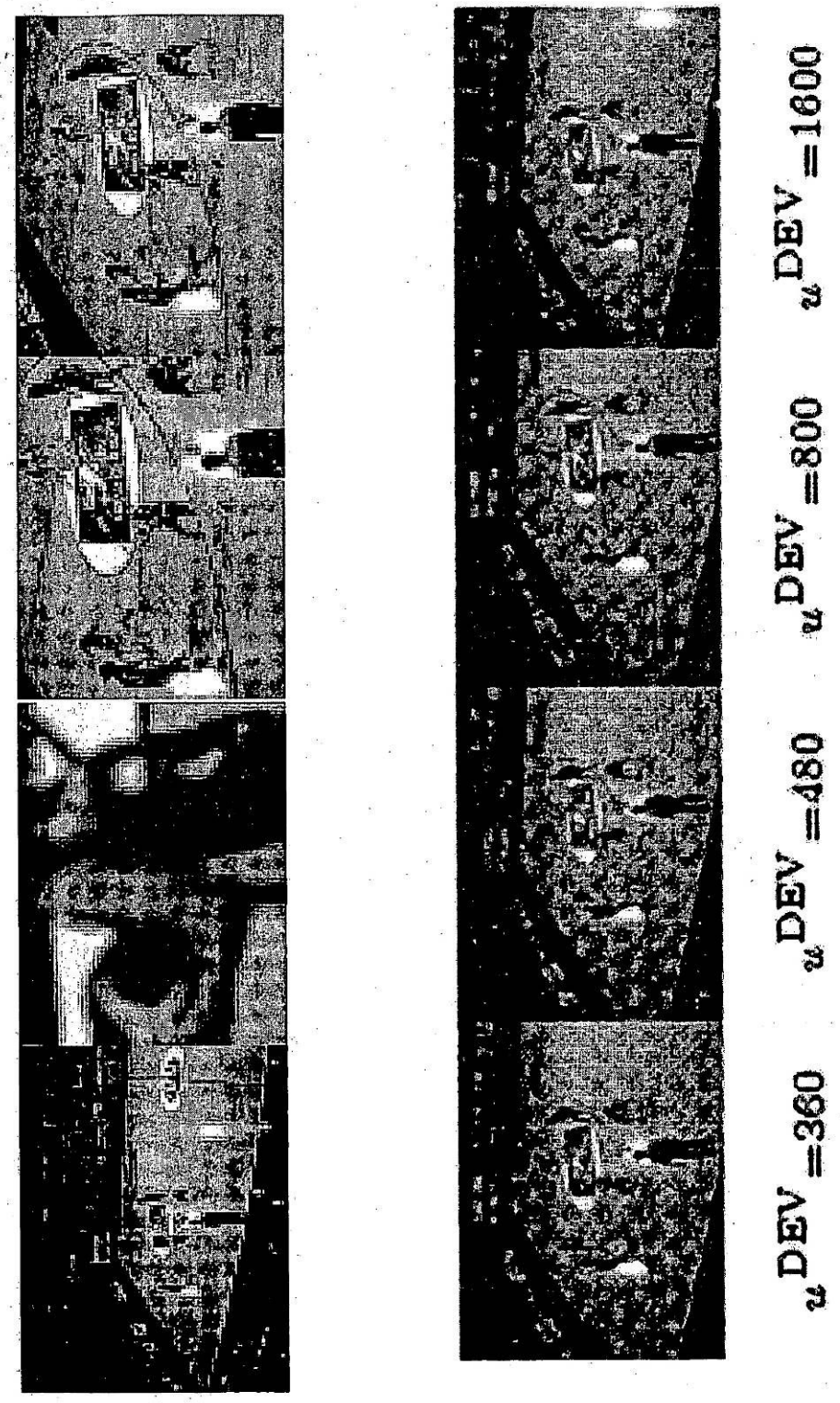


Figura 4: comportamiento de selección de punto de vista bajo diferentes tamaños de visualización

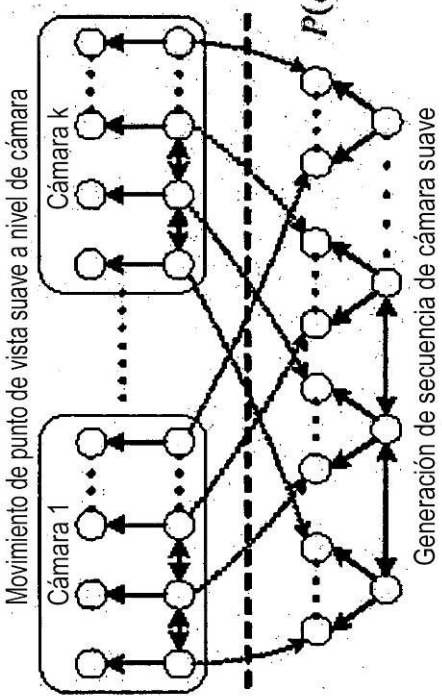


Figura 5: modelo de grafo para estimación de dos etapas de movimiento de punto de vista suave.

Figura 6: planos de cámara al recoger datos de vídeo experimentales.

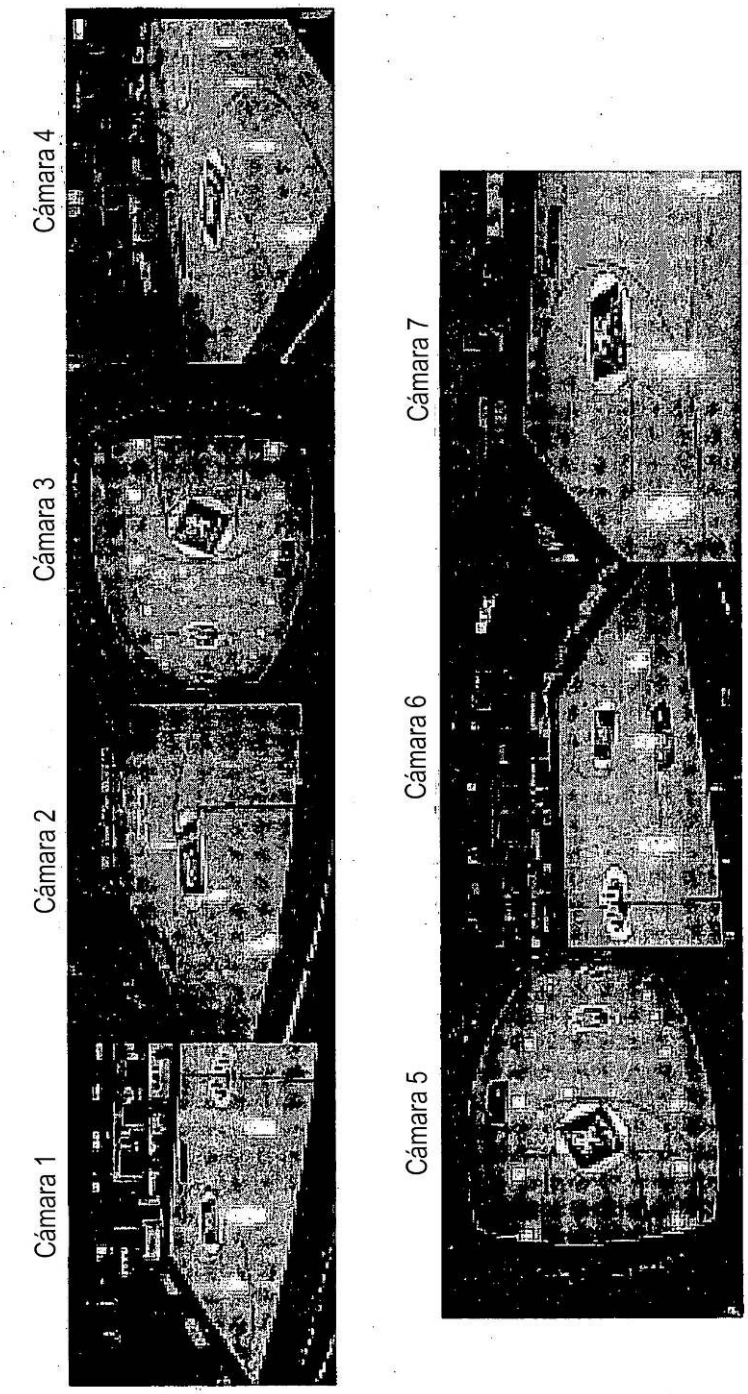


Figura 7: vistas de muestra recogidas mediante diferentes cámaras.

Figura 8: un clip de vídeo corto con 1200 fotogramas se usa para demostrar el sistema. Hay cinco eventos de reloj dentro de este clip. Para cada evento, el momento más destacado se marca en una barra continua roja.

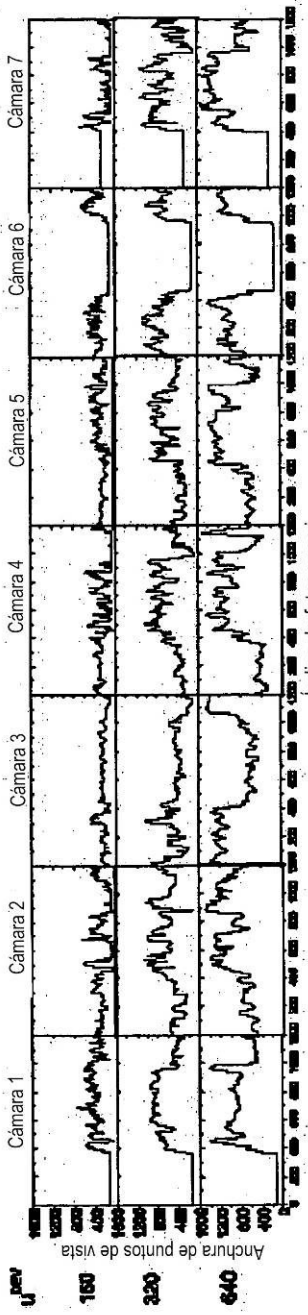


Figura 9: secuencias de punto de vista a nivel de cámara seleccionadas automáticamente bajo tres resoluciones de visualización diferentes.
Se ha aplicado suavizado de punto de vista débil y la intensidad de suavizado se establece a $\sigma_2/\sigma_1 = 4$.

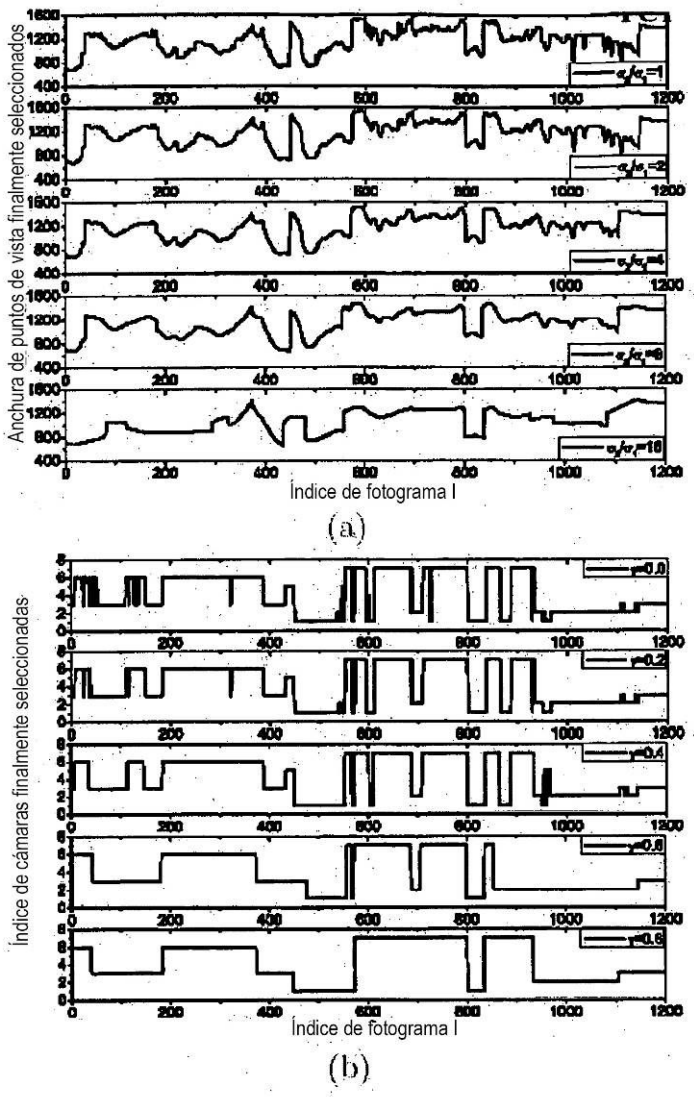


Figura 10: comportamiento de secuencia de cámara/punto de vista optimizados bajo diferentes intensidades de suavizado. Resolución de visualización se establece a $u^{\mathrm{DEV}} = 640$.

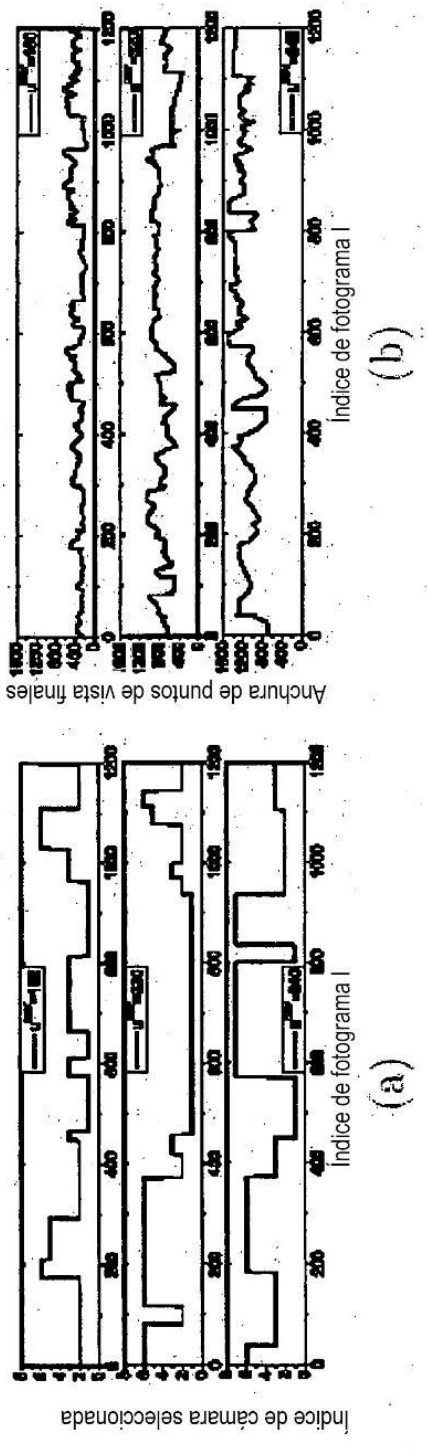


Figura 11: comparación de secuencias de cámara y de punto de vista generadas bajo tres resoluciones de visualización diferentes 160, 320 y 640. Intensidad de suavizado de punto de vista es $\sigma_2/\sigma_1 = 4$ y la intensidad de suavizado de cámara γ se establece a 0,8.

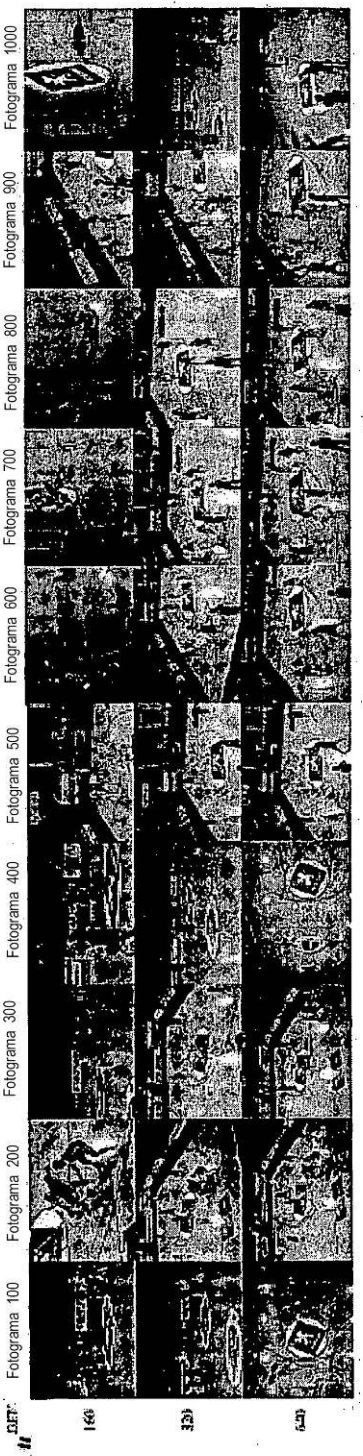


Figura 12: miniaturas de fotogramas en secuencias generadas bajo tres resoluciones de visualización diferentes 160, 320 y 640. Intensidad de suavizado de punto de vista es $\sigma_2/\sigma_1 = 4$ y la intensidad de suavizado de cámara γ se establece a 0,8.

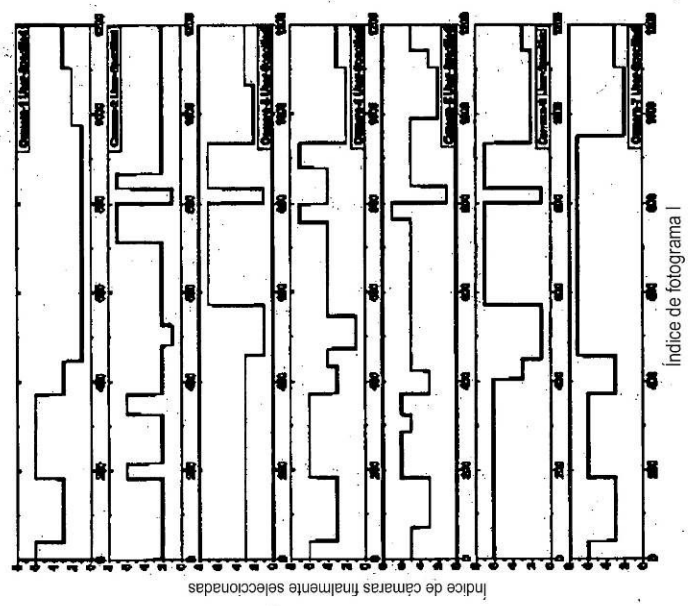


Figura 13: Comparación de secuencias de cámara generadas bajo diferentes preferencias de usuario en vistas de cámara para $u_i^{\mathrm{DEV}} = 6{\cdot}10$. Intensidad de suavizado de punto de vista es $\sigma_2/\sigma_1 = 1$ y la intensidad de suavizado de cámara ? se establece a 0,8.

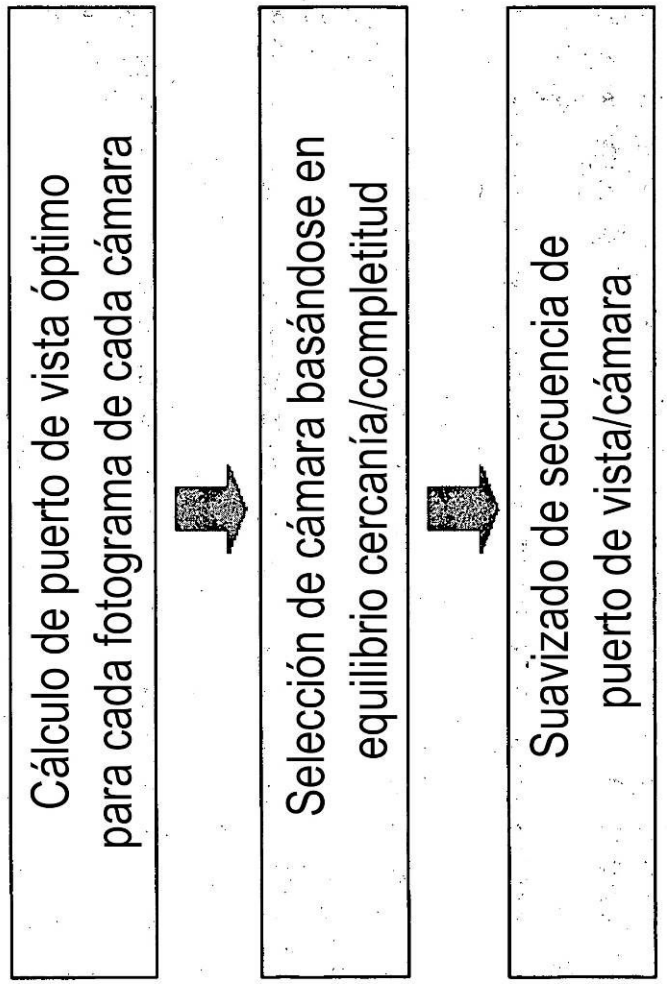


Figura 14: La solución propuesta se implementa como un proceso de 3 etapas. Cabe la pena indicar que el suavizado de puerto de vista podría implementarse también en la vista de cámara correspondiente, antes de la selección de cámara real.

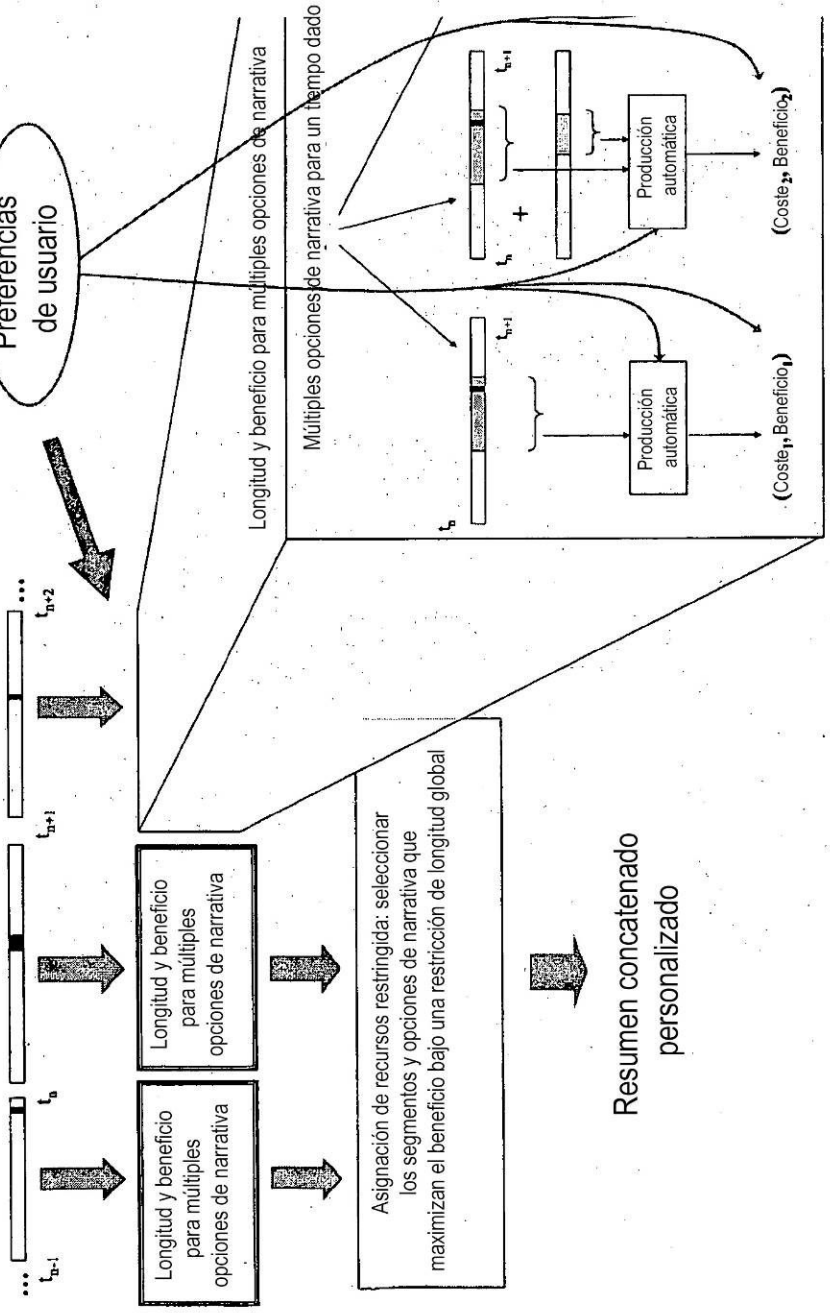

Preferencias de usuario
Longitud y beneficio para múltiples opciones de narrativa
Múltiples opciones de narrativa para un tiempo dado
Producción automática
Producción automática
(Coste1, Beneficio1)
(Coste2, Beneficio2)
Longitud y beneficio para múltiples opciones de narrativa
Longitud y beneficio para múltiples opciones de narrativa
Asignación de recursos restringida: seleccionar los segmentos y opciones de narrativa que maximizan el beneficio bajo una restricción de longitud global
Resumen concatenado personalizado

FIG. 15

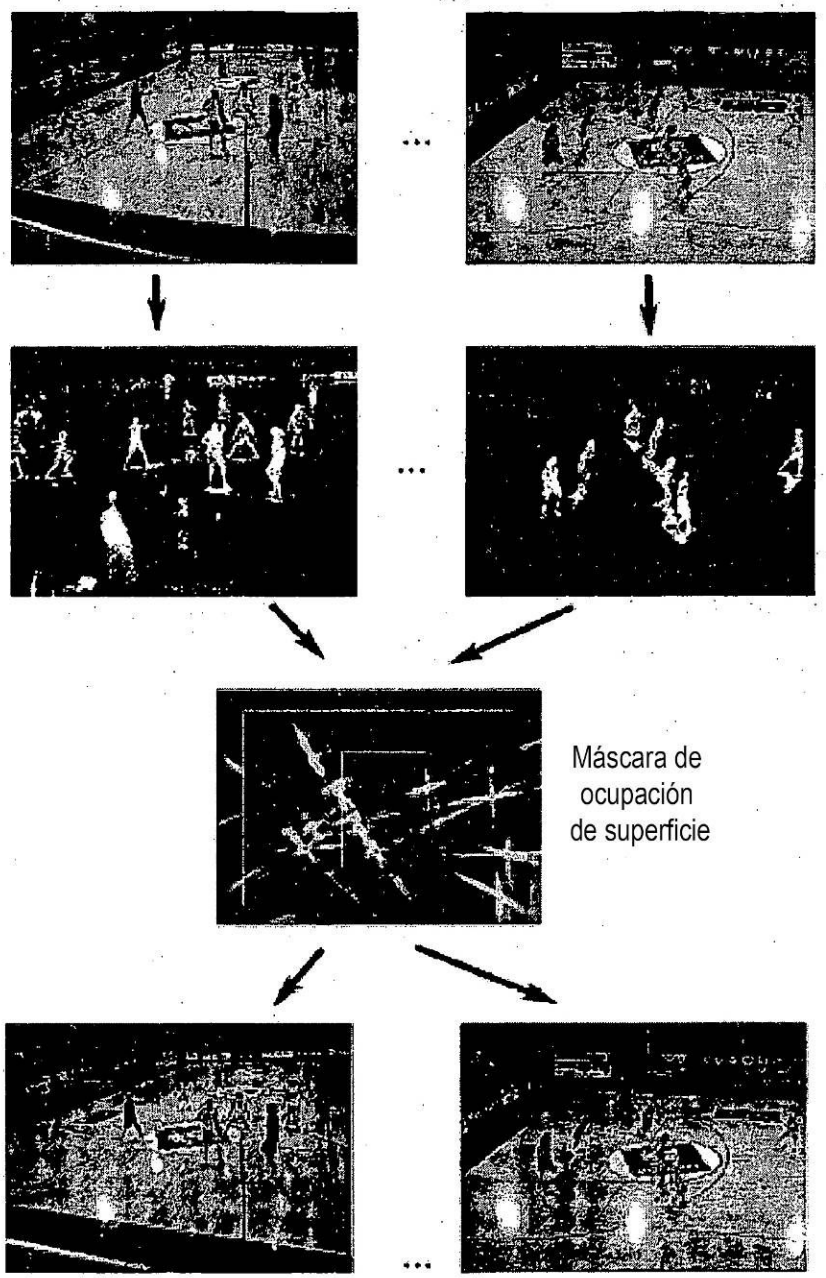


Figura 16: las máscaras de primer plano se unen para crear una máscara de ocupación de superficie