(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 648 368**

(21) Número de solicitud: 201630878

(51) Int. CI.:

H04N 21/466 (2011.01)
H04N 21/25 (2011.01)

(12) SOLICITUD DE PATENTE A1

(22) Fecha de presentación:

29.06.2016

(43) Fecha de publicación de la solicitud:

02.01.2018

(71) Solicitantes:

ACCENTURE GLOBAL SOLUTIONS LIMITED (100.0%)
3 Grand Canal Plaza
Dublin 4 IE

(72) Inventor/es:

BATALLER, Cyrille y
MENDEZ, Luis

(74) Agente/Representante:

CARVAJAL Y URQUIJO, Isabel

(54) Título: **Recomendación de vídeo con base en el contenido**

(57) Resumen:

Los métodos, sistemas, y aparatos incluyen programas de ordenador en un medio de almacenamiento para proporcionar recomendaciones de vídeo. Para cada vídeo, se obtiene un conjunto de imágenes que se incluyen en el vídeo. Para cada imagen respectiva en el conjunto de imágenes de un vídeo, se genera un conjunto de una o más palabras clave que describen el contenido visual. Con base al menos en los conjuntos de palabras clave para al menos algunas de las imágenes, se genera un conjunto de palabras clave que describen el vídeo. Los vídeos se asignan a grupos con base en el conjunto de palabras clave que se generan para cada vídeo. Se recibe una solicitud para una recomendación con base en un primer vídeo. Se proporcionan datos que identifican un segundo vídeo como recomendación con base en el segundo vídeo que se asigna a un mismo grupo como el primer vídeo.

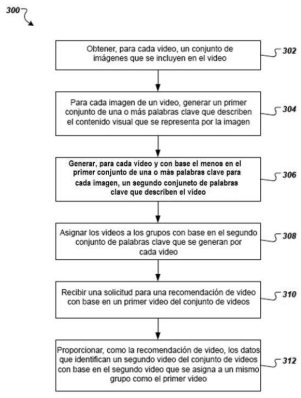

FIG. 3

ES 2 648 368 A1

RECOMENDACIÓN DE VÍDEO CON BASE EN EL CONTENIDO

DESCRIPCIÓN

Campo técnico

Esta memoria se relaciona al aprendizaje automático

Antecedentes

Las redes de comunicación de datos, tales como internet, proporcionan acceso a diversos tipos de información y contenido. Un tipo del contenido disponible sobre internet son los vídeos. Por ejemplo, los sitios web de intercambio de vídeos proporcionan acceso a millones de diferentes vídeos. Adicionalmente, los servicios de transmisión proporcionan acceso a diversas diferentes películas, programas de televisión, y eventos. Debido al gran número de vídeos disponibles en internet, puede ser difícil para los usuarios encontrar vídeos en los cuales los usuarios están interesados.

Descripción breve invención

Esta memoria describe, entre otras cosas, un sistema que genera datos que describen el contenido de los vídeos u otros multimedios y utiliza los datos para identificar los vídeos o los multimedios, por ejemplo, en respuesta a una consulta o para recomendarle a un usuario. Por ejemplo, el sistema puede proporcionar recomendaciones y enlaces de vídeo a los vídeos recomendados a la vez que el usuario observa otro vídeo. Los vídeos recomendados pueden ser vídeos considerados similares al vídeo que está viendo el usuario. La similitud entre los vídeos se puede determinar con base al menos en palabras clave que describan contenido visual que se representa por imágenes en los vídeos. Por ejemplo, las imágenes se pueden obtener a partir de los vídeos y analizarlas usando técnicas de aprendizaje automático para identificar palabras claves que describen las imágenes de los vídeos. Las palabras clave de un vídeo se pueden comparar con las palabras clave de otro vídeo para determinar la similitud entre los vídeos.

En general, un aspecto innovador de la materia descrita en esta memoria, se puede realizar en métodos que incluyen las acciones de, para cada vídeo en un conjunto de vídeos: obtener un conjunto de imágenes que se incluyen en el vídeo; para cada imagen respectiva en el conjunto de imágenes, generar un primer conjunto de una o más palabras clave que describen el contenido visual representado por la imagen respectiva; y generar, con base al menos en los respectivos primeros conjuntos de una o más palabras

clave para al menos algunas de las imágenes, un segundo conjunto de palabras clave que describen el vídeo; asignar los vídeos en el conjunto de vídeos a los grupos con base en el segundo conjunto de palabras clave generadas para cada vídeo; recibir una solicitud para una recomendación de vídeo con base en un primer vídeo del conjunto de vídeos; y proporcionar, como la recomendación de vídeo, los datos que identifican un segundo vídeo del conjunto de vídeos con base en el segundo vídeo que se asigna al mismo grupo como el primer vídeo. Otras realizaciones de este aspecto incluyen los correspondientes sistemas, aparatos, y programas de ordenador, configurados para llevar a cabo las acciones de los métodos, codificados en dispositivos de almacenamiento. Para un sistema de uno o más ordenadores que se configuran para llevar a cabo las operaciones particulares o acciones instaladas en su software, firmware, hardware, o una combinación de ellos. Para uno o más programas por ordenador que se configura para llevar a cabo operaciones particulares de programas que incluyan instrucciones.

Las precedentes y otras realizaciones pueden incluir cada una opcionalmente una o más de las siguientes características, en solitario o en combinación. En algunos aspectos, la solicitud para la recomendación de vídeo se genera en respuesta en (i) una presentación del primer vídeo o (ii) una solicitud para el primer vídeo.

En algunos aspectos, la asignación de vídeos a los grupos con base en el segundo conjunto de palabras clave generadas para cada vídeo, puede incluir usando un proceso de aprendizaje automático, asignar vídeos que tengan al menos un límite similar dentro de un mismo grupo. Varias formas de calcular la similitud se tratan en la descripción detallada. La similitud entre dos vídeos puede ser con base en una similitud entre los respectivos primeros conjuntos de palabras clave generadas para los dos vídeos.

En algunos aspectos, el segundo conjunto de palabras clave para cada vídeo puede definirse en una secuencia con base en una secuencia en la cual las ocurren en el vídeo las imágenes de las cuales las palabras clave fueron generadas. La similitud entre dos vídeos se puede basar en la secuencia de palabras clave para un primer vídeo y la secuencia de palabras clave para un segundo vídeo.

Algunos aspectos pueden incluir identificar, para un primer vídeo, un número de ocurrencias de cada palabra clave en el segundo conjunto de palabras clave para el primer vídeo e identificar, para un segundo vídeo, un número de ocurrencias de cada palabra clave en el segundo conjunto para el segundo vídeo. La similitud entre los dos vídeos se puede basar en una comparación del número de ocurrencias de cada palabra clave.

Algunos aspectos pueden incluir generar, para cada vídeo un tercer conjunto de palabras clave que describen el contenido audible del vídeo. El segundo conjunto de palabras clave que describen el vídeo se pueden generar con base adicionalmente en el tercer conjunto de palabras clave.

En algunos aspectos, la generación del respectivo primer conjunto de una o más palabras clave que describen el contenido visual que se representa por una imagen dada, puede incluir usar un proceso de aprendizaje profundo, para generar al menos una porción del primer conjunto palabras clave. Generar el primer conjunto de una o más palabras clave que describen el contenido visual que se representa por una imagen dada, puede incluir detectar un objeto que se representa por la imagen dada e incluir, en el primer conjunto de palabras clave que describen el contenido visual, una palabra clave que describa el objeto detectado.

En algunos aspectos, generar el primer conjunto de una o más palabras clave que describen el contenido visual que se representa por una imagen dada, puede incluir detectar una persona que se representa por la imagen dada e incluir, en el primer conjunto de palabras clave, una palabra clave que identifica la persona detectada.

En algunos aspectos, la generación, con base en el primer conjunto de palabras clave para algunas de las imágenes, de un segundo conjunto de palabras clave que describen el vídeo, puede incluir la identificación, para cada palabra clave generada, de un número de imágenes del vídeo para las cuales se generó la palabra clave e identificar, para la inclusión en el segundo conjunto de palabras clave, un número preespecificado de palabras clave.

Algunos aspectos pueden incluir la generación de un índice de escenas de vídeo para el conjunto de vídeos. El índice puede incluir, para cada escena de vídeo, un conjunto de palabras clave que describen la escena de vídeo. Se puede recibir una consulta que especifique un término de consulta. Una escena de vídeo en el índice se puede identificar con base en un término de consulta igualando una palabra clave que se incluye en el índice para la escena de vídeo. Los datos que especifican las escenas de vídeo identificadas se pueden proporcionar en respuesta a la solicitud. El segundo conjunto de palabras clave pueden incluir palabras clave seleccionadas a partir de los respectivos primeros conjuntos de palabras clave para al menos algunas de las imágenes.

Se pueden implementar realizaciones particulares para la materia descrita en esta memoria de forma que se realicen una o más de las siguientes ventajas. Los vídeos que pueden ser de interés a un usuario se pueden identificar con más precisión usando el

contenido que se representa por las imágenes del vídeo. Se puede determinar la similitud entre los vídeos con más precisión, comparando las palabras clave que describen el contenido (por ejemplo, objetos, personas, fondos, etc.) que se incluyen en las imágenes obtenidas a partir de los vídeos. Se pueden identificar vídeos similares como recomendaciones a los usuarios para ayudarles a los usuarios a encontrar vídeos que puedan ser de interés a los usuarios. Al proporcionar recomendaciones de vídeos similares de una manera más precisa, el número de consultas de usuario y solicitudes de usuario se pueden reducir, resultando en un menor consumo de banda ancha, menor demanda en los recursos de la red que se usan para transmitir consultas y vídeos, y menores ciclos de procesamiento para un procesador de ordenador.

Al indexar los vídeos y las escenas de vídeos usando palabas clave , los resultados de búsqueda pueden clasificarse de manera más precisa en respuesta a las consultas del usuario. Al proporcionar resultados de búsqueda mejor clasificados, se pueden reducir el número de solicitudes recibidas por un motor de búsqueda, resultando en una menor demanda en los recursos de cálculo y una velocidad mejorada para responder a las consultas. Los usuarios pueden también buscar escenas que tengan contenido particular, por ejemplo, la escena de una película que incluye un tigre, que permite a los usuarios encontrar más rápidamente escenas particulares. Las películas accesibles mediante internet y que tienen un título incorrecto se pueden encontrar en la búsqueda con base en el contenido de la película, permitiendo la identificación de sitios que proporcionan copias fraudulentas de películas que pueden, por ejemplo, incluir títulos de películas falsas, ofuscados o alternativos. En algunas realizaciones, se puede buscar el contenido del vídeo, por ejemplo, para eventos u objetos que se muestran en los vídeos grabados por un sistema de vigilancia por vídeo de circuito cerrado de televisión (CCTV).

Los detalles de realizaciones se describen en los dibujos acompañantes y en la siguiente descripción detallada. Otras características, aspectos, y ventajas de la presente propuesta serán evidentes a la luz de la descripción, los dibujos, y las reivindicaciones.

Breve descripción de los dibujos

La Fig. 1 es un diagrama de bloques de ejemplo en el cual un sistema de vídeo proporciona vídeos y/o datos relacionados con los vídeos.

La Fig. 2 es un diagrama de un generador de palabras clave de ejemplo.

La Fig. 3 es un diagrama de flujo de un proceso de ejemplo para proporcionar una recomendación de vídeo.

La Fig. 4 es un diagrama de flujo de un proceso de ejemplo para proporcionar datos que especifiquen un vídeo o una escena de vídeo.

Al igual que los números y designaciones de referencia en los diversos dibujos indican elementos similares.

Descripción detallada

Esta memoria describe sistemas y técnicas para generar y proporcionar recomendaciones para vídeos, por ejemplo, películas, programas de televisión, vídeos de deportes, eventos, aplicaciones, vídeos musicales, etc., u otro contenido. Por ejemplo, se puede usar una descripción de una o más escenas para identificar otros vídeos que tienen una escena similar o relacionada. Se pueden usar técnicas y sistemas similares para buscar vídeos particulares y/o escenas de vídeo en los vídeos. Por ejemplo, la descripción de una escena u otra porción de vídeo se puede indexar con base en un identificador para el vídeo y un tiempo en el cual ocurre la escena en el vídeo. Cuando se recibe una consulta, la consulta puede compararse con la descripción. Escenas que tengan una descripción que iguale o que sea similar a la consulta puede ser presentada.

En algunas implementaciones, el sistema obtiene imágenes, por ejemplo, capturas de pantalla o cuadros de vídeo, de un vídeo y utiliza las imágenes del vídeo para generar un conjunto de términos o palabras clave que describen el vídeo. El sistema puede obtener una imagen del vídeo a una frecuencia particular. La frecuencia de muestreo puede basarse en el número de cuadros de vídeo por unidad de tiempo o basarse en el tiempo.

El sistema puede generar palabras clave para cada imagen obtenida del vídeo. En algunas implementaciones, el sistema usa procesos automáticos para generar palabras clave que describen el contenido representado por la imagen. Por ejemplo, una imagen que se obtiene a partir de una película puede representar un coche que circula a lo largo de un puente en un día lluvioso. En este ejemplo, el sistema puede generar las palabras clave “coche”, “puente”, y “lluvia” para describir la imagen. En otro ejemplo, una imagen que se obtiene a partir de una película puede representar a Barack Obama de pie en frente de la Casa Blanca. En este ejemplo, el sistema puede generar las palabras clave “Barack Obama” y “Casa Blanca” mediante técnicas de reconocimiento de un objeto que identifican la Casa Blanca en la imagen y técnicas de reconocimiento de una persona que identifican a Barack Obama en dicha imagen. Las técnicas de aprendizaje automáticas usadas para generar las descripciones pueden entrenarse usando imágenes etiquetadas, por ejemplo, por un usuario. Las etiquetas para una imagen pueden describir el contenido de la imagen.

También se pueden usar otros datos para describir cada imagen obtenida. En algunas implementaciones, el audio del vídeo que corresponde a la imagen y/o el audio que ocurre en el vídeo antes o después que la imagen sea analizada para generar palabras clave que describan la imagen. Continuando con el ejemplo anterior del coche y el puente, el audio en la imagen o dentro de un intervalo de uno o dos segundos respecto de la imagen puede incluir el sonido de un trueno. En este ejemplo, la descripción de la imagen puede ser “coche”, “puente”, “lluvia”, “trueno”, y “tormenta”.

El sistema puede generar un conjunto de palabras clave para el vídeo con base en las palabras clave generadas para cada imagen obtenida. El sistema puede incluir en el conjunto de palabras clave, aquellas que ocurren más a menudo. Por ejemplo, las palabras clave para el vídeo pueden incluir las palabras clave superiores-N, clasificadas con base en el número de veces que cada palabra clave ocurre en las palabras clave para las imágenes del vídeo. En otro ejemplo, las palabras clave para el vídeo pueden incluir aquellas palabras clave que ocurren un número límite de veces.

En algunas realizaciones, el conjunto de palabras clave para un vídeo puede estar en forma de un vector de palabras clave. Además, el conjunto de palabras clave puede estar organizado en el vector con base en la secuencia en la cual ocurren las imágenes en el vídeo. Por ejemplo, un primer elemento vector puede incluir las palabras clave para la primera imagen obtenida; el segundo elemento vector del vector puede incluir las palabras clave para la segunda imagen obtenida, y así sucesivamente.

Las palabras clave generadas para un vídeo se pueden usar para identificar otros vídeos similares o relacionados. Por ejemplo, las palabras clave generadas para un vídeo pueden compararse con palabras clave generadas para otros vídeos, para identificar otros vídeos que incluyen las mismas palabras clave o similares. En algunas implementaciones, un proceso de aprendizaje automático puede agrupar los vídeos en grupos según el conjunto de palabras clave generadas. Cuando un usuario observa un vídeo en un grupo en particular, el sistema puede proporcionarle al usuario una lista de otros vídeos del grupo como recomendaciones.

El sistema puede también indexar escenas u otros cortos de un vídeo con las palabras clave generadas para el vídeo. De esta forma, el sistema puede proporcionar escenas de vídeo o cortos en respuesta a las consultas del usuario. Por ejemplo, si un usuario envía una consulta “coche que circula a lo largo de un puente”, el sistema puede comparar los términos de la consulta con el índice. En este ejemplo, el sistema puede identificar el ejemplo de escena anterior, de un coche que circula a lo largo de un puente en un día

lluvioso. En respuesta, el sistema puede proporcionarle al usuario, datos que identifican la escena y/o un enlace a un vídeo que inicie en, o que incluya la escena.

La Fig. 1 es un diagrama de un ambiente 100 de ejemplo en el cual un sistema 130 de vídeo proporciona vídeos y/o datos relacionados a vídeos. El ambiente 100 de ejemplo incluye un dispositivo 110 cliente que le permite a los usuarios descargar, almacenar, y ver vídeos y otro contenido, por ejemplo, otro contenido multimedia. El dispositivo 110 cliente es un dispositivo electrónico que es capaz de solicitar y recibir datos sobre una red 120 de comunicaciones de datos, por ejemplo, una red de área local (LAN), una red de área ancha (WAN), internet, una red móvil, o una combinación de estas. Los dispositivos cliente de ejemplo incluyen ordenadores personales, dispositivos de comunicación móviles, por ejemplo, teléfonos inteligentes y/o dispositivos de tableta de computación, televisores inteligentes, o televisores con internet, por ejemplo, un televisor con una conectividad de red o que está conectado a un conjunto de caja superior que le proporciona al televisor conectividad de red, y otros dispositivos apropiados.

El dispositivo 110 cliente de ejemplo incluye un reproductor 112 de vídeo y un navegador 114 de internet. El navegador 114 de internet facilita el envío y recepción de datos sobre la red 120. El navegador 114 de internet puede permitirle al usuario interactuar con texto, imágenes, vídeos, música, y otra información típicamente ubicada en una página de internet. En algunas implementaciones, el reproductor 112 de vídeo es una aplicación que facilita la descarga, difusión, y vista de vídeos, por ejemplo, a partir de un servicio de difusión de vídeo o servicio de uso compartido de vídeo. Por ejemplo, el reproductor 112 de vídeo puede ser una aplicación nativa desarrollada para una plataforma particular o un tipo particular de dispositivo, por ejemplo, un teléfono inteligente o un teléfono inteligente que incluye un sistema operativo particular. El reproductor 112 de vídeo y/o el navegador 114 de internet pueden proporcionar una interfaz de usuario que permita a los usuarios navegar o buscar vídeos. Por ejemplo, en una implementación de televisión inteligente, el reproductor 112 de vídeo puede proporcionar una guía que se muestra en la pantalla de televisión y que le permite a un usuario navegar o buscar películas, programas, u otros vídeos.

El sistema 130 de vídeo puede proporcionar vídeo y/o datos relacionados con vídeos a los dispositivos 110 cliente sobre la red 120. Por ejemplo, el sistema 130 de vídeo puede ser parte de un servicio de difusión de vídeo o un servicio de vídeo compartido que difunde o descarga vídeos al dispositivo 110 cliente. En otro ejemplo, el sistema 130 de vídeo puede ser un servicio de terceros que proporciona recomendaciones de vídeo, por ejemplo, recomendaciones de películas o programas de televisión, y/o resultados de

búsqueda de vídeo en respuesta a solicitudes para dichos datos. En este ejemplo, un sitio de internet o servicio de difusión de vídeo puede solicitar a partir del sistema 130 de vídeo, resultados de búsqueda o recomendaciones en respuesta a un usuario que observa un vídeo en particular o envía una solicitud para un vídeo.

El sistema 130 de vídeo incluye un generador 140 de palabras clave que describen vídeos almacenados en un sistema 150 de almacenamiento de vídeo, por ejemplo, discos duros y/o discos de estado sólido. El generador 140 de palabras clave puede generar palabras clave para un vídeo con base en el contenido del vídeo. En algunas implementaciones, el generador 140 genera palabras clave para un vídeo con base en el contenido visual que se representa en una o más imágenes obtenidas a partir del vídeo. Por ejemplo, las imágenes pueden ser capturas de pantalla tomadas del vídeo o frames de vídeo del vídeo. Para cada imagen, el generador 140 de palabras clave puede generar un conjunto de una o más palabras clave que describen el contenido visual de la imagen. El contenido visual para el que las palabras clave se generaron, puede incluir contenido de escena general, por ejemplo, exterior, lluvioso, oscuro, etc., objetos representados en la imagen, por ejemplo, coches, edificaciones, etc., y/o personas representadas en la imagen. Como se describe en más detalle a continuación con referencia a la Fig. 2, el generador 140 de palabras clave puede incluir un motor de aprendizaje profundo, un motor de reconocimiento de objetos, y/o un motor de reconocimiento de personas para generar palabras clave que describan el contenido visual representado en las imágenes.

El generador 140 de palabras clave puede también generar palabras clave para un vídeo con base en el contenido de audio del vídeo. Por ejemplo, el generador 140 de palabras clave puede generar palabras clave que describan sonidos, por ejemplo, trueno, choches, pájaros, etc., música, por ejemplo, canciones particulares, y palabras habladas, por ejemplo, usando reconocimiento de voz, que se incluye en el audio del vídeo. El generador 140 de palabras clave puede generar un conjunto de una o más palabras clave para diversos períodos de tiempo del audio. Por ejemplo, el generador 140 de palabras clave puede segmentar el vídeo en una secuencia de porciones de vídeo de un minuto y generar un conjunto de una o más palabras clave que describan el audio para cada porción de vídeo de un minuto. En este ejemplo, un vídeo de diez minutos se puede segmentar en diez segmentos de un minuto y se pueden generar una o más palabras clave para cada segmento con base en el audio que se incluye en el segmento.

El generador 140 de palabras clave puede generar un conjunto de palabras clave que describen un vídeo con base en palabras clave generadas por imágenes obtenidas a partir del vídeo y/o palabras clave generadas con base en el audio del vídeo. Por

ejemplo, el generador 140 de palabras clave puede generar un conjunto agregado de palabras clave para un vídeo con base en las palabras clave generadas por imágenes y las palabras clave generadas para el audio.

En algunas implementaciones, se pueden usar técnicas de selección de palabras clave para seleccionar solo un subconjunto de palabras clave generadas para las imágenes y/o un subconjunto de las palabras clave generadas para el audio. Por ejemplo, el generador 140 de palabras clave puede identificar, para cada palabra clave, el número de ocurrencias de palabras clave en el conjunto de palabras clave generadas para las imágenes y/o el conjunto palabras clave generadas para el audio. En este ejemplo, una palabra clave puede tener múltiples ocurrencias por generarse para describir múltiples diferentes imágenes del vídeo y/o generarse para describir el audio para múltiples diferentes segmentos de vídeo. En otro ejemplo, la palabra clave puede tener múltiples ocurrencias por generarse para describir una imagen del vídeo y para describir el audio de un segmento de vídeo. El generador 140 de palabras clave puede incluir, en el conjunto de palabras clave que describen el vídeo, cada palabra clave que tenga al menos un número límite de ocurrencias o un número particular de palabras clave que tengan el mayor número de ocurrencias, por ejemplo, palabras clave superiores 10, 50, o 100.

El conjunto de palabras clave que describen un vídeo pueden también incluir palabras clave obtenidas de recursos diferentes a las imágenes y al audio. Por ejemplo, el conjunto de palabras clave que describen un vídeo puede también incluir palabras clave que se incluyen en los metadatos para el vídeo, palabras clave que se incluyen en el título del vídeo, palabras clave que se incluyen en la descripción del vídeo, palabras que se incluyen en los datos de subtítulos del vídeo, palabras que se incluyen en los créditos del vídeo, palabras clave que se obtienen de los comentarios o reseñas relacionadas con el vídeo, y/o palabras clave que se obtienen de otras fuentes apropiadas.

El generador 140 de palabras clave puede generar un índice 152 de vídeo que incluye datos de identificación de vídeos y las palabras clave generadas para cada vídeo. Por ejemplo, el índice 152 de vídeo puede incluir, para cada vídeo, un identificador único para el vídeo, por ejemplo, un título único o código numérico, y el conjunto de palabras clave generadas por el generador 140 de palabras clave para describir el vídeo general.

En algunas implementaciones, el índice 152 de vídeo incluye un índice para las escenas de otro tipo de segmentos de vídeo para al menos algunos de los vídeos. Por ejemplo, el índice 152 de vídeo puede incluir, para cada escena de un vídeo, un identificador 161

para el vídeo en el cual ocurre la escena, un identificador 162 único para la escena, un tiempo 163 en el cual ocurre la escena en el vídeo, y palabra(s) 164 clave generadas para la escena. Como se usa aquí, el término “escena” puede referirse a una escena particular de una película o programa de televisión u otro tipo de segmento de vídeo que sea menos que el vídeo completo. Por ejemplo, una escena puede ser una porción de un vídeo para el cual se han generado las palabras clave por el generador 140 de palabras clave. En este ejemplo, el índice 152 de vídeo puede incluir una escena y sus correspondientes palabras clave para cada imagen para la cual el generador 140 de palabras clave genera las palabras clave.

Se considera, por ejemplo, un vídeo de diez minutos. El generador 140 de palabras clave puede obtener una imagen a partir del vídeo cada diez segundos, resultando en un conjunto de sesenta imágenes para el vídeo. El generador 140 de palabras clave puede entonces generar, para cada una de las sesenta imágenes, un conjunto de una o más palabras clave que describen la imagen. En este ejemplo, el índice 152 de vídeo puede incluir, para el vídeo de diez minutos, una entrada para cada imagen resultando en sesenta entradas para el vídeo. La entrada para cada imagen puede incluir un identificador único para la escena que corresponde a la imagen, un tiempo en el cual ocurre la escena en el vídeo, por ejemplo, el tiempo dentro del vídeo en el cual se obtiene la imagen, y la(s) palabra(s) clave generada(s) para la imagen. Adicionalmente, o alternativamente, la entrada para cada imagen puede incluir palabra(s) clave generada(s) con base en el audio del vídeo en el momento que se representa la imagen en el vídeo y/o el audio que ocurre en una cantidad específica de tiempo antes y/o después que la imagen se representa en el vídeo. Por ejemplo, la entrada para una imagen particular puede incluir la(s) palabra(s) clave generada(s) para la imagen y la(s) palabra(s) clave generada(s) para el audio de un segmento de vídeo de diez segundos que empieza cinco segundos antes que la imagen se represente en el vídeo y finaliza cinco segundos después que la imagen se represente en el vídeo.

Las palabras clave generadas para cada vídeo y/o para cada escena de un vídeo pueden usarse para identificar vídeos similares y/o para emerger escenas de vídeo en particular en respuesta a resultados de búsqueda. Por ejemplo, el sistema 130 de vídeo puede incluir un motor 142 de búsqueda, un motor 144 de agrupamiento, y un motor 146 de recomendación. El motor 142 de búsqueda, el motor 144 de agrupamiento, y el motor 146 de recomendación, pueden cada uno implementarse en uno o más servidores, por ejemplo, ubicados en uno o más centros de datos.

En algunas implementaciones, el generador 140 de palabras clave puede también generar palabras clave para colecciones de imágenes y/o secuencias de imágenes y generar un índice para las colecciones y/o secuencias. El índice puede incluir, para cada colección o secuencia, datos que identifiquen la colección o secuencia y/o una o más palabras clave para cada imagen en la colección o secuencia. De esta forma, un usuario puede buscar imágenes particulares en la colección o secuencia. Por ejemplo, una colección de imágenes pueden ser imágenes obtenidas a partir de una cámara de vigilancia. En este ejemplo, un usuario puede buscar imágenes que incluyen un objeto en particular, por ejemplo, un arma, o ropa particular, por ejemplo, una gorra de béisbol.

El motor 142 de búsqueda puede recibir consultas de dispositivos 110 cliente u otras fuentes y proporcionar resultados de búsqueda que identifiquen y/o enlacen los vídeos o escenas en respuesta a sus consultas. El motor 142 de búsqueda puede usar el índice 152 de vídeo para identificar vídeos o escenas que son sensibles a una consulta recibida. Por ejemplo, el motor 142 de búsqueda puede comparar los términos que se incluyen en una consulta a la(s) palabra(s) clave generadas por cada vídeo y/o la(s) palabra(s) clave generadas para cada escena. El motor 142 de búsqueda puede identificar los vídeos y/o las escenas que tengan al menos una palabra clave correspondiente que iguale al menos un término de la consulta. El motor 142 de búsqueda puede entonces proporcionar resultados de búsqueda para al menos una porción de los vídeos o escenas identificadas. Por ejemplo, el motor 142 de búsqueda puede clasificar los vídeos y/o las escenas y proporcionar resultados de búsqueda para un número específico de vídeos y/o escenas mejor clasificadas. Los vídeos y/o escenas pueden clasificarse con base en el número de términos concordantes entre la consulta y los vídeos y/o escenas, la calidad de los vídeos y/o escenas, la popularidad de los vídeos y/o escenas, por ejemplo, en términos del número de veces que se han visto los vídeos o las escenas.

Un resultado de búsqueda puede incluir texto que identifica un vídeo o una escena y/o un texto que describa el vídeo o la escena. El resultado de búsqueda puede también incluir un enlace al vídeo o la escena. Un resultado de búsqueda para una escena o un vídeo en particular puede incluir un enlace a un vídeo que incluye solo la escena en particular. En otro ejemplo, un resultado de búsqueda para una escena en particular puede incluir un enlace al inicio de la escena en particular dentro del vídeo que incluye la escena en particular. Por ejemplo, la interacción del usuario con el enlace o el resultado de búsqueda pueden ocasionar que el dispositivo 110 cliente cargue el vídeo, por ejemplo, en el reproductor 112 de vídeo o en el navegador 114 de internet, e inicie el vídeo en el momento de inicio para la escena en particular.

El motor 142 de búsqueda puede también permitirle a los usuarios buscar un vídeo que puede estar indexado o anunciado en internet usando diferentes títulos. Por ejemplo, al buscar por películas que estén indexadas con base en el contenido de las películas, se pueden encontrar los anuncios fraudulentos de las películas que usan diferentes títulos.

El motor 142 de búsqueda puede también permitirle a los usuarios buscar escenas en particular dentro de un vídeo en particular. Por ejemplo, el motor 142 de búsqueda puede proporcionar una interfaz 114 de usuario dentro del navegador 114 de internet que le permita a los usuarios seleccionar un vídeo e introducir palabras clave dentro de un cajón de búsqueda. El motor 142 de búsqueda puede entonces usar el índice 152 de vídeo para identificar escenas dentro del vídeo seleccionado que son sensibles a la consultaintroducida. Por ejemplo, un usuario puede buscar por la “escena de persecución” en una película en particular. En algunas implementaciones, el usuario puede introducir palabras clave dentro del cajón de búsqueda sin seleccionar primero una película en particular, y puede llevar a cabo una búsqueda con base en las palabras clave a través de diversas películas con base en su contenido. En consecuencia, si un usuario introduce la consulta “escena de persecución”, el motor 142 de búsqueda puede entregar resultados de uno o múltiples vídeos que se determinan que tienen una escena de persecución. Además, los resultados de búsqueda pueden indicar las ubicaciones particulares en los vídeos respectivos en los cuales ocurren las escenas de persecución de manera que, por ejemplo, usuario pueda simplemente hacer clic en una representación de un vídeo dado para conducirlo directamente a la reproducción de la escena relevante que se especificó en el resultado de búsqueda.

El motor 144 de agrupamiento puede identificar vídeos o escenas similares o relacionadas y generar grupos de vídeos o escenas similares o relacionadas. En algunas implementaciones, el motor 144 de agrupamiento, agrupa los vídeos con base en la similitud entre las palabras clave generadas para los vídeos. Por ejemplo, el motor 144 de agrupamiento puede comparar el conjunto de palabras clave generadas por el generador 140 de palabra clave para un primer vídeo del conjunto de palabras clave generadas por el generador 140 de palabras clave para un segundo vídeo, para determinar un nivel de similitud entre el primer vídeo y el segundo vídeo. Si el primer vídeo y el segundo vídeo tienen un nivel de similitud que satisface un nivel límite de similitud, el primer vídeo puede incluirse en un grupo con el segundo vídeo.

El nivel de similitud entre los dos vídeos puede determinarse usando la similitud del coseno. Por ejemplo, cada vídeo puede representarse por un vector que representa las

palabras clave asociadas para el vídeo. La similitud del coseno puede usarse para determinar la similitud entre los vectores para los vídeos.

El nivel de similitud entre los dos vídeos puede basarse en el número de palabras clave comunes para los dos vídeos. Por ejemplo, el nivel de similitud para los dos vídeos puede ser proporcional al número de palabras clave generadas para el primer vídeo que igualen las palabras clave generadas para el segundo vídeo. El nivel de similitud entre los dos vídeos puede basarse en el número de ocurrencias de palabras clave para el primer vídeo y el número de ocurrencias de palabras clave para el segundo vídeo. Por ejemplo, una palabra clave puede ocurrir varias veces para un vídeo si la palabra clave se genera para describir varias imágenes del vídeo. Si los dos vídeos tienen palabras clave comunes que también ocurren varias veces para ambos vídeos o que ocurren un número similar de veces para ambos vídeos, los vídeos pueden tener un nivel elevado de similitud que si las palabras clave que ocurren varias veces para el primer vídeo no ocurren varias veces para el segundo vídeo.

En algunas implementaciones, el nivel de similitud entre los dos vídeos puede basarse en la secuencia en la cual ocurren las palabras clave para los dos vídeos. Por ejemplo, el conjunto de palabras clave para cada vídeo puede estar dispuesto en el orden de las imágenes y/o el audio para el cual se generaron las palabras clave. En particular, la(s) palabra(s) clave generada(s) para la primera imagen que se representa por un vídeo o la primera imagen que se obtiene a partir del vídeo pueden estar dispuestas primero en el conjunto de palabras clave, la(s) palabra(s) clave generada(s) para la segunda imagen que se representa por el vídeo o la segunda imagen que se obtiene a partir del vídeo, pueden estar dispuestas después de las palabras clave generadas para la primera imagen, y así sucesivamente. El motor 144 agrupamiento puede comparar la secuencia de palabras clave para el primer vídeo a la secuencia de palabras clave para el segundo vídeo para determinar el nivel de similitud entre los dos vídeos.

El nivel de similitud entre los dos vídeos puede basarse en el número de palabras clave que ocurren en la misma secuencia o en secuencias similares y/o el número de igualdad de las secuencias similares de al menos un número específico de palabras clave, por ejemplo, al menos tres palabras clave en secuencia. Dos secuencias similares de palabras clave pueden ser secuencias que incluyen las mismas palabras clave, pero que también incluyen no más de un número específico de palabras clave adicionales. Por ejemplo, una secuencia de palabras clave para el primer vídeo puede ser “perro, salto, cerca, calle” con base en las palabras clave generadas para una o más imágenes del primer vídeo. Una secuencia de palabras clave para el segundo vídeo puede ser “perro,

gato, salto, cerca, patio, calle". Estas dos secuencias pueden considerarse secuencias similares debido a la secuencia común de "perro, salto, cerca, calle" en las palabras clave para ambos vídeos, aunque las palabras clave para el segundo vídeo incluyan palabras clave adicionales de "gato" y "patio" que no se incluyen en las palabras clave para el primer vídeo.

Dos secuencias de palabras clave pueden considerarse secuencias similares si menos de un número específico de palabras clave está fuera de secuencia. Por ejemplo, la secuencia de palabras clave para el primer vídeo puede ser "perro, salto, cerca, calle" con base en las palabras clave generadas para una o más imágenes del primer vídeo. Una secuencia de palabras clave para el segundo vídeo puede ser "perro, cerca, salto, calle". Estas dos secuencias pueden considerarse secuencias similares debido a que las secuencias incluyen cuatro palabras clave en común aunque dos de las palabras clave estén transpuestas.

En algunas implementaciones, el motor 144 de agrupamiento utiliza un proceso de aprendizaje automático para asignar vídeo o escenas a grupos con base en la similitud entre los conjuntos de palabras clave y/o la secuencia de palabras clave que se generan para describir los vídeos o las escenas.

El motor 144 de agrupamiento puede generar un índice 154 de grupo que incluye datos relacionados a los grupos de los vídeos o escenas similares. El índice 154 de grupo puede incluir, para cada grupo, un identificador 166 único para el grupo y los datos que especifican los vídeos o las escenas 167 asignadas al grupo.

El motor 146 de recomendación puede proporcionar recomendaciones de vídeo con base en los grupos generados por el motor 144 de agrupamiento. El motor 146 de recomendación puede recomendar vídeo o escenas que sean similares a un vídeo o una escena que es visto por un usuario o los vídeos o escenas que sean similares a un vídeo o escena solicitada por un usuario. Por ejemplo, el reproductor 112 de vídeo o el navegador 114 de internet pueden presentar recomendaciones de vídeo en una interfaz de usuario en la cual se reproduce un vídeo. Cuando el dispositivo 110 cliente solicita un vídeo, por ejemplo, a partir del sistema 130 de vídeo, el motor 146 de recomendación puede acceder al índice de grupo para identificar uno o más grupos en los cuales el vídeo solicitado sea miembro. El motor 146 de recomendación puede entonces seleccionar uno o más vídeos que estén incluidos en el(los) grupo(s) identificado(s) para recomendarlos al usuario del dispositivo 110 cliente del cual se recibió la solicitud. Un proceso de ejemplo

para proporciona una recomendación de vídeo se ilustra en la Fig. 3 y se describe a continuación.

El motor 142 de búsqueda puede también permitirle a un usuario buscar vídeos similares usando el índice 154 de grupo. Por ejemplo, el motor 142 de búsqueda puede proporcionar una interfaz de usuario dentro del navegador 114 de internet que le permita a un usuario seleccionar un vídeo y solicitar vídeos que sean similares al vídeo seleccionado. En otro ejemplo, el reproductor 112 de vídeo puede incluir un ícono que, cuando se selecciona, envía una solicitud al motor 142 de búsqueda por vídeos que sean similares a un vídeo que se presenta en el dispositivo 110 cliente. El motor 142 de búsqueda puede acceder al índice 154 de grupo para identificar otros vídeos que se incluyen en el(los) mismo(s) grupos(s) que el vídeo seleccionado o el vídeo que se presenta en el dispositivo 110 cliente. El motor 142 de búsqueda puede entonces proporcionar datos que especifiquen los vídeos similares para presentación en el dispositivo 110 cliente.

La Fig. 2 es un diagrama del generador 140 de palabras clave de ejemplo de la Fig. 1. El generador 140 de palabras clave incluye un extractor 220 de imagen y un extractor 250 de audio que reciben a la vez un vídeo 210. El extractor 220 de imagen puede obtener un conjunto de imágenes 225 a partir del vídeo 210 y proporcionar el conjunto de imágenes 225 a uno o más motores 230 de análisis de imagen. El extractor 220 de imagen puede obtener el conjunto de imágenes 225 tomando capturas de pantalla del vídeo 210 a una frecuencia de muestreo o extrayendo los cuadros de vídeo del vídeo 210 con base en la frecuencia de muestreo. La frecuencia de muestreo puede basarse en el número de cuadros de vídeo por unidad de tiempo o basarse en el tiempo. Por ejemplo, la frecuencia puede ser cada cuadro, cada dos cuadros, cada cinco cuadros. En otro ejemplo, la frecuencia puede ser cada segundo, cada dos segundos, cada diez segundos.

El generador 140 de palabra clave de ejemplo incluye un motor 232 de aprendizaje profundo, un motor 234 de reconocimiento de objetos, y un motor 236 de reconocimiento de personas. Otras implementaciones pueden incluir solo uno o dos de los motores 232-236 o motores adicionales que no se ilustran en la Fig. 2. El motor 232 de aprendizaje profundo puede usar una o más técnicas de aprendizaje profundas, por ejemplo, de una pila de aprendizaje profunda, para generar o seleccionar una o más palabras clave que describan las características visuales de una imagen. En algunas implementaciones, el motor 232 de aprendizaje profundo puede generar palabras clave que describan en general o en un nivel superior el contenido visual de la imagen. Por ejemplo, el motor 232 de aprendizaje profundo puede analizar las características visuales de una imagen para

identificar las características ambientales, por ejemplo, adentro, afuera, claro, oscuro, lluvia, nieve, etc., y/o las características de ubicación, por ejemplo, ciudad, playa, montañas, granja, etc.

El motor 234 de reconocimiento de objetos puede usar una o más técnicas de reconocimiento de objetos para identificar objetos en una imagen y generar palabras clave que describan el objeto. Por ejemplo, el motor 234 de reconocimiento de objetos puede usar técnicas de detección de bordes, técnicas de trasformada de característica de invariante en escala (SIFT), técnicas de bolsas de palabras, y otras técnicas apropiadas para detectar objetos en las imágenes. Para cada objeto detectado, el motor 232 de reconocimiento de objetos puede generar una o más palabras clave que describan el objeto.

El motor 236 de reconocimiento de personas puede usar una o más técnicas de reconocimiento de personas para identificar personas en una imagen y generar palabras clave que identifiquen y/o describan las personas. Por ejemplo, el motor 236 de reconocimiento de personas puede usar técnicas de reconocimiento facial para detectar personas conocidas en una imagen. El motor 236 de reconocimiento de personas puede también analizar características visuales de una imagen para determinar o predecir el género, edad, u otras características de una persona no reconocida y generar palabras clave que describan estas características.

Cada uno de los motores 232-236 puede entrenarse usando datos de entrenamiento etiquetados. Los datos de entrenamiento etiquetados pueden incluir imágenes que tengan etiquetas que describan las imágenes. Por ejemplo, un usuario puede etiquetar las imágenes con base en lo que el usuario observa en la imagen. Los motores 232-236 pueden entonces entrenarse para generar palabras clave que describan correctamente lo que representan otras imágenes. Los motores 232-236 pueden entrenarse hasta que cada uno genere palabras clave que describan correctamente al menos un porcentaje límite de las imágenes de prueba.

Para un vídeo 210, cada uno de los motores 232-236 puede analizar cada imagen en el conjunto de imágenes 225 y generar un conjunto de una o más palabras clave que describan la imagen con base en su análisis respectivo de la imagen. Cada una de las palabras 240 clave generadas por los motores 232-236 pueden entonces proporcionarse a un agregador 270 de palabras clave que agrega las palabras clave para el vídeo, como se describe a continuación.

El extractor 250 de audio puede extraer audio 255 a partir del vídeo 210. El audio extraído puede ser una corriente continua de audio para la totalidad del vídeo o un conjunto de segmentos de audio. Por ejemplo, el extractor 250 de audio puede segmentar el vídeo 210 dentro de una secuencia de un minuto, dos minutos, o porciones de vídeo sucesivas de tres minutos y extraer el audio a partir de cada porción de vídeo. En otro ejemplo, el extractor 250 de audio puede extraer el audio a partir del vídeo para cada imagen. En este ejemplo, el audio para una imagen puede incluir audio que ocurre en el vídeo 210 antes del punto en el vídeo en el cual la imagen ocurre y el audio que ocurre en el vídeo 210 después del punto en el vídeo en el cual ocurre la imagen. Por ejemplo, si se obtiene una imagen a partir del vídeo en un punto a dos minutos del comienzo del vídeo, el audio de la imagen puede incluir el audio que inicia en un minuto y cincuenta segundos del comienzo del vídeo y finaliza en dos minutos y diez segundos del comienzo del vídeo.

El extractor 250 de audio puede proporcionar el audio extraído a un motor 260 de análisis de audio. El motor 260 de análisis de audio puede analizar el audio 255 para identificar los sonidos que se incluyen en el audio 255. Por ejemplo, el motor 260 de análisis de audio puede comparar el audio con sonidos conocidos para detectar sonidos en el audio extraído. En otro ejemplo, el motor 260 de análisis de audio puede usar el reconocimiento de voz para detectar palabras habladas en el audio.

El extractor 260 de audio puede generar uno o más conjuntos de palabras 265 clave con base en los sonidos detectados en el audio extraído. Por ejemplo, si el audio es una corriente continua simple, el extractor 260 de audio puede generar un conjunto de palabras clave con base en los sonidos detectados mediante el audio extraído. El conjunto de palabras clave puede estar dispuesto en orden con base en el orden en el cual ocurren los sonidos en el audio y en consecuencia, el orden en el cual ocurren los sonidos en el vídeo 210. Si el audio se segmenta, por ejemplo, con base en las porciones de imagen o vídeo, el conjunto de palabras clave puede incluir un subconjunto de una o más palabras clave para cada porción de imagen o vídeo. Los subconjuntos pueden también estar dispuestos en el orden en el cual ocurren las porciones de imágenes o vídeo en el vídeo 210. El motor 260 de análisis de audio puede proporcionar las palabras 265 clave al agregador 270 de palabras clave.

El agregador 270 de palabras clave puede agregar las palabras 240 clave que se reciben a partir de los motores 230 de análisis de imagen y/o las palabras 265 clave recibidas a partir del motor 265 de análisis de audio dentro de un conjunto de palabras clave que describen el vídeo. En algunas implementaciones, el agregador 270 de palabras clave

puede generar una lista combinada de palabras clave que incluye cada palabra clave generada por los motores 230 de análisis de imagen para el vídeo 210 y cada palabra clave generada por el motor 265 de análisis de audio para el vídeo 210.

En algunas implementaciones, el agregador 270 de palabras clave incluye solo un subconjunto de las palabras 240 clave y las palabras 265 clave en el conjunto de palabras clave que describen el vídeo 210. Por ejemplo, el agregador 265 de palabras clave puede incluir las palabras clave más populares para el vídeo 210 en el conjunto de palabras clave que describen el vídeo 210. En este ejemplo, el agregador 270 de palabras clave puede identificar, para cada palabra clave que se genera por un motor 230 de análisis de imagen o el motor 260 de análisis de audio para el vídeo 210, por ejemplo, cada palabra clave que se incluye en las palabras clave 240 y las palabras 265 clave para el vídeo 210, un número de ocurrencias de la palabra clave en las palabras 240 clave y las palabras clave 265. Por ejemplo, una palabra clave puede tener tres ocurrencias si la palabra clave se generó por los motores 230 de análisis de imagen para dos imágenes diferentes y la palabra clave se generó por el motor 260 de análisis de audio para un segmento de vídeo. El agregador 270 de palabras clave puede entonces seleccionar, para inclusión en el conjunto de palabras clave para el vídeo 210, las palabras clave que tengan al menos un número límite de ocurrencias o un número particular de palabras clave que tengan el número más elevado de ocurrencias. Por ejemplo, el conjunto de palabras clave para el vídeo puede incluir las palabras clave que tengan al menos tres ocurrencias en las palabras 240 clave y/o las palabras 265 clave. En otro ejemplo, el agregador 270 de palabras clave puede clasificar las palabras clave con base en el número de ocurrencias de cada palabra clave en las palabras 240 clave y/o las palabras 265 clave y seleccionar las mejores diez, veinte, u otro número apropiado de palabras clave en la clasificación.

El agregador 270 de palabras clave puede también agregar las palabras 240 clave recibidas a partir de los motores 230 de análisis de imagen y las palabras 265 recibidas a partir del motor 265 de análisis de audio dentro de un conjunto de palabras clave que describe cada imagen obtenida a partir del vídeo 210. Por ejemplo, como se describió anteriormente, el extractor 250 de audio puede extraer audio a partir del vídeo para cada imagen y el motor 260 de análisis de audio puede generar una o más palabras clave que describen el audio para cada imagen. El agregador 270 de palabras clave puede generar, para cada imagen, un conjunto agregado de palabras clave para las imágenes que incluyen todas o al menos una porción de palabras clave que se generaron por la imagen por los motores 230 de análisis de imagen y el motor 260 de análisis de audio.

Como se describe anteriormente, el generador 140 de palabras clave puede generar o poblar un índice, por ejemplo, un índice 152 de vídeo de la Fig. 1 con datos relacionados con vídeos. Por ejemplo, el índice puede incluir datos que identifican el vídeo, por ejemplo, un identificador único, y el conjunto de palabras clave que describen el vídeo que se generó por el agregador 270 de palabras clave. El índice puede también incluir, para cada imagen, los datos que identifican una escena que corresponde a la imagen y las palabras clave generadas por la imagen. El índice puede usarse entonces para buscar vídeos y escenas de vídeo en respuesta a las consultas recibidas.

La Fig. 3 es un diagrama de flujo de un proceso 300 de ejemplo para proporcionar una recomendación de vídeo. El proceso 300 puede implementarse por uno o más programas de ordenador instalados en uno o más ordenadores. El proceso 300 será descrito como se lleva a cabo por un sistema programado apropiado de uno o más ordenadores, por ejemplo, el sistema 130 de vídeo de la Fig. 1.

El sistema obtiene, para cada vídeo de un conjunto de vídeos, un conjunto de imágenes que se incluyen en el vídeo (302). Por ejemplo, el sistema puede obtener un conjunto de capturas de pantalla o cuadros de vídeo de cada vídeo con base en una frecuencia de muestreo.

Para cada imagen de cada vídeo, el sistema genera un respectivo primer conjunto de una o más palabras clave que describen el contenido visual que se representa por la imagen (304) respectiva. Por ejemplo, cada imagen puede analizarse usando técnicas profundas de aprendizaje, técnicas de reconocimiento de objetos, técnicas de reconocimiento de personas, y/u otras técnicas de análisis de imagen para generar palabras clave que describan el contenido visual representado por la imagen. El conjunto de una o más palabras clave para una imagen dada puede incluir las palabras clave que se generaron con base en cada uno de los análisis. En algunas implementaciones, el conjunto de una o más palabras clave para una imagen puede también incluir palabras clave que se generaron con base en el audio del vídeo que ocurre al mismo tiempo que la imagen en el vídeo o el audio que ocurre dentro de un tiempo específico antes y después de que ocurra la imagen.

El sistema genera, para cada vídeo en el conjunto de vídeos, un segundo conjunto de palabras clave que describen el vídeo (306). El segundo conjunto de palabras clave para un vídeo dado, puede incluir al menos una porción de palabras clave que se incluyen en el primer conjunto de palabras clave que se generaron para al menos algunas de las imágenes obtenidas del vídeo. Por ejemplo, el sistema puede seleccionar algunas

palabras clave de los primeros conjuntos de palabras clave generadas para al menos algunas de las imágenes obtenidas del vídeo dado. Como se describe anteriormente, el conjunto de palabras clave para un vídeo puede seleccionarse con base en el número de ocurrencias de las palabras clave en los conjuntos de palabras clave generadas para las imágenes del vídeo.

En algunas implementaciones, el segundo conjunto de palabras clave para un vídeo dado puede también incluir palabras clave que se incluyen en los metadatos para el vídeo dado, palabras clave que se incluyen en el título del vídeo dado, palabras clave que se incluyen en la descripción del vídeo dado, palabras que se incluyen en los datos de subtítulos para el vídeo dado, palabras que se incluyen en los créditos para el vídeo dado, palabras que se obtienen de comentarios o reseñas relacionadas con el vídeo dado, y/o palabras clave que se obtienen de otras fuentes apropiadas. El sistema puede asignar vídeos que tengan al menos un nivel límite de similitud con otro de un grupo.

El sistema asigna los vídeos en el conjunto de vídeos a los grupos con base en el segundo conjunto de palabras clave para cada vídeo (308). Por ejemplo, el sistema puede usar una o más técnicas de aprendizaje automático para asignar los vídeos a los grupos con base en la similitud entre los segundos conjuntos de palabras clave para los vídeos.

El nivel de similitud entre dos vídeos puede basarse en una comparación del segundo conjunto de palabras clave para los dos vídeos independientemente de la secuencia. Por ejemplo, el nivel de similitud entre dos vídeos puede basarse en una comparación del número de ocurrencias de cada una de las palabras clave en el segundo conjunto de palabras clave para un primer vídeo de los dos vídeos y el número de ocurrencias de cada una de las palabras clave en el segundo conjunto de palabras clave para un segundo vídeo de los dos vídeos.

En otro ejemplo, el nivel de similitud entre dos vídeos puede basarse en una comparación de la secuencia de palabras clave en el segundo conjunto de palabras clave para los dos vídeos. Por ejemplo, si el segundo conjunto de palabras clave para cada vídeo tiene secuencias de palabras clave similares, los vídeos pueden tener secuencias similares de escenas que indican que los vídeos son similares.

El sistema recibe una solicitud para una recomendación de vídeo con base en un primer vídeo del conjunto de vídeos (310). La solicitud para la recomendación de vídeo puede transmitirse a partir de un dispositivo cliente al sistema en respuesta a la presentación del primer vídeo. Por ejemplo, las recomendaciones de vídeo pueden presentarse

adyacentes a una ventana o un área de visualización en la cual se presenta el primer vídeo. Cuando se presenta el primer vídeo, un reproductor de vídeo o navegador de internet que presenta el vídeo puede ocasionar que el dispositivo cliente envíe una solicitud al sistema para recomendaciones de vídeo.

La solicitud para la recomendación del vídeo puede transmitirse a partir del dispositivo cliente al sistema en respuesta a una solicitud para el primer vídeo. Por ejemplo, si un usuario interactúa con, por ejemplo, selecciona, un enlace para el primer vídeo, el dispositivo cliente puede transmitir la solicitud al sistema para obtener recomendaciones de vídeo para presentarlas con el primer vídeo.

El sistema proporciona datos que identifican un segundo vídeo en el conjunto de vídeos con base en el segundo vídeo que se asigna a un mismo grupo como el primer vídeo (312). Por ejemplo, el sistema puede identificar cada grupo en el cual el primer grupo es un miembro en respuesta a la solicitud para la recomendación de vídeo. El sistema puede entonces seleccionar el segundo vídeo de uno de los grupos identificados, o el grupo si el primer vídeo se asigna a solo un grupo, y proporcionar datos que identifican el segundo vídeo al dispositivo cliente del cual se recibió la solicitud.

El sistema puede también proporcionar un enlace al segundo vídeo. De esta forma, si el usuario está interesado en el segundo vídeo recomendado, el usuario puede fácilmente acceder al segundo vídeo. Por ejemplo, la interacción del usuario con el enlace puede ocasionar que el reproductor de vídeo o el navegador de internet que presentan el primer vídeo, naveguen del primer vídeo al segundo vídeo.

En algunas implementaciones, el sistema puede proporcionar como datos de recomendaciones, identificar múltiples vídeos. Por ejemplo, el primer vídeo puede asignarse a uno o más grupos que incluyen múltiples diferentes vídeos. El sistema puede proporcionar datos que identifican al menos una porción de los múltiples diferentes vídeos como recomendaciones.

La Fig. 4 es un diagrama de flujo de un proceso 400 de ejemplo para proporcionar datos que especifican un vídeo o una escena de vídeo. El proceso 400 puede implementarse por uno o más programas de ordenador instalados en uno o más ordenadores. El proceso 400 se describirá como se realiza por un sistema programado apropiadamente de uno o más ordenadores, por ejemplo, el sistema 130 de vídeo de la Fig. 1.

El sistema genera, para cada vídeo de un conjunto de vídeos, un conjunto de una o más palabras clave para cada escena del vídeo (402). El conjunto de una o más palabras

clave para una escena dada puede generarse con base en una o más imágenes obtenidas del vídeo y que se representan por el vídeo durante la escena. Por ejemplo, cada una o más imágenes para una escena se pueden analizar usando técnicas profundas de aprendizaje, técnicas de reconocimiento de objetos, técnicas de reconocimiento de personas, y/u otras técnicas de análisis de imágenes para generar una o más palabras clave que describen el contenido visual que se representa por la imagen.

El conjunto de una o más palabras clave que se generan para una escena pueden también generarse con base en el contenido audible que ocurre en el vídeo durante la escena. Por ejemplo, el contenido audible puede compararse con sonidos conocidos y/o se puede usar el reconocimiento de voz para generar una o más palabras clave que describen los sonidos que ocurren durante la escena.

El sistema genera un índice de escenas de vídeo para los vídeos que usan la(s) palabra(s) clave generada(s) para cada escena (404). Por ejemplo, el índice puede incluir, para cada escena, un identificador para la escena, un identificador para el vídeo en el cual ocurre la escena, y el conjunto de una o más palabras clave generadas para la escena. El índice puede también incluir, para cada vídeo, un identificador para el vídeo y la(s) palabra(s) clave generada(s) para cada escena en el vídeo.

El sistema recibe una consulta para un vídeo o una escena (406) de vídeo. La consulta se puede recibir de un dispositivo cliente. Por ejemplo, un usuario del dispositivo cliente puede enviar una consulta para un vídeo. La consulta puede incluir uno o más términos de consulta.

El sistema identifica, en el índice, un vídeo o una escena de vídeo que tenga al menos una palabra clave que iguale al menos un término de consulta de la consulta (408). Por ejemplo, el sistema puede comparar el(los) término(s) de consulta para la consulta de las palabras clave que se incluyen en el índice. Si múltiples vídeos o escenas de vídeo tienen una palabra clave que iguale un término de consulta de la consulta, el sistema puede seleccionar un vídeo o una escena de múltiples vídeos o escenas de vídeo. El sistema puede hacer la selección con base en el número de términos concordantes entre la consulta y los vídeos y/o las escenas, la calidad de los vídeos /o las escenas, la popularidad de los vídeos y/o las escenas, por ejemplo, en términos del número de veces que se han visto los vídeos o las escenas.

El sistema proporciona, en respuesta a la consulta, datos que especifican los vídeos o las escenas (410) de vídeo identificados. Por ejemplo, el sistema puede proporcionar, al dispositivo cliente del cual se recibe la consulta, un resultado de búsqueda que especifica

el vídeo identificado o la escena de vídeo. El resultado de búsqueda puede también incluir un enlace al vídeo o a la escena. Si el usuario interactúa con el enlace, el dispositivo cliente puede solicitar el vídeo o la escena y presentarle el vídeo o la escena al usuario.

REIVINDICACIONES

1. Un método implementado por ordenador para realizar recomendaciones de vídeo según su contenido, comprendiendo el método:

obtener, para cada vídeo de un conjunto de vídeos, un conjunto de imágenes que se incluyen en el vídeo;

para cada imagen en el conjunto de imágenes, generar un primer conjunto con una o más palabras clave que describen el contenido visual que representa la imagen usando técnicas de reconocimiento de objetos y personas; y

generar, con base al menos en los primeros conjuntos de una o más palabras clave para al menos algunas de las imágenes, un segundo conjunto de palabras clave que describen el vídeo;

asignar los vídeos en el conjunto de vídeos a grupos con base en el segundo conjunto de palabras clave que se generaron para cada vídeo usando un proceso de aprendizaje automático para asignar vídeos que tengan al menos una similitud límite mayor que un valor umbral dentro de un mismo grupo, donde la similitud entre los dos vídeos se basa en la similitud entre los respectivos primeros conjuntos de una o varias palabras clave que se generaron para los dos vídeos;

recibir una solicitud para una recomendación de vídeo con base en un primer vídeo del conjunto de vídeos; y

proporcionar, como recomendación de vídeo, datos que identifican un segundo vídeo del conjunto de vídeos con base en el segundo vídeo que se asigna al mismo grupo como el primer vídeo.

2. El método de la reivindicación 1, en donde la solicitud para la recomendación de vídeo se genera en respuesta a al menos uno de (i) presentación del primer vídeo o (ii) una solicitud para el primer vídeo.

3. El método de la reivindicación 1, en donde:

el segundo conjunto de palabras clave para cada vídeo se dispone formando una secuencia con base en la secuencia en la cual las imágenes, para las cuales se generaron las palabras clave, ocurren en el vídeo; y

la similitud entre dos vídeos se basa en una similitud entre la secuencia de palabras clave para un primer vídeo y la secuencia de palabras clave para un segundo vídeo.

4. El método de la reivindicación 1, que además comprende:

identificar, para un primer vídeo de dos vídeos, un número de ocurrencias de cada palabra clave en el segundo conjunto de palabras clave para el primer vídeo;

identificar, para un segundo vídeo de dos vídeos, un número de ocurrencias de cada palabra clave en el segundo conjunto de palabras clave para el segundo vídeo,

en donde la similitud entre los dos vídeos se basa en una comparación del número de ocurrencias de cada palabra clave en el segundo conjunto de palabras clave para el primer vídeo y el número de ocurrencias de cada palabra clave en el segundo conjunto de palabras clave para el segundo vídeo.

5. El método de la reivindicación 1, que además comprende generar, para cada vídeo, un tercer conjunto de palabras clave que describen el contenido audible del vídeo, en donde el segundo conjunto de palabras clave que describen el vídeo se genera con base además en el tercer conjunto de palabras clave.

6. El método de la reivindicación 1, en donde la generación del respectivo primer conjunto de una o más palabras clave que describen el contenido visual que se representa por una imagen dada, comprende usar un proceso de aprendizaje profundo para generar al menos una porción del respectivo primer conjunto de una o más palabras clave.

7. El método de la reivindicación 1, en donde la generación del respectivo primer conjunto de una o más palabras clave que describen el contenido visual que se representa por una imagen dada, comprende:

detectar un objeto que se representa por la imagen dada; e

incluir, en el respectivo primer conjunto de una o más palabras clave que describen el contenido visual que se representa por la imagen dada, una palabra clave que describe el objeto detectado.

8. El método de la reivindicación 1, en donde la generación del respectivo primer conjunto de una o más palabras clave que describen el contenido visual que se representa por una imagen dada, comprende:
detectar una persona que se representa por la imagen dada; e

incluir, en el primer conjunto de una o más palabras clave que describen el contenido visual que se representa por la imagen dada, una palabra clave que identifica la persona detectada.

9. El método de la reivindicación 1, en donde la generación, con base al menos en el respectivo primer conjunto de una o más palabras clave para al menos algunas de las imágenes, de un segundo conjunto de palabras clave que describen el vídeo comprende:
identificar, para cada palabra clave generada para al menos una imagen del vídeo, un número de imágenes del vídeo para las cuales se generó la palabra clave: e
identificar, para inclusión en el segundo conjunto de palabras clave, un número previamente especificado de las palabras clave con base en el número de imágenes para las cuales se generó cada palabra clave.

10. El método de la reivindicación 1, que comprende además:

generar un índice de escenas de vídeo para el conjunto de vídeos, en donde el índice incluye, para cada escena de vídeo, un conjunto de palabras clave que describen la escena de vídeo;
recibir una consulta que especifique al menos un término de consulta;

identificar una escena de vídeo en el índice con base en al menos un término de consulta que iguale al menos una palabra clave que se incluye en el índice para la escena de vídeo; y

proporcionar, en respuesta a la consulta, datos que especifiquen la escena de vídeo identificada.

11. El método de la reivindicación 1, en donde el segundo conjunto de palabras clave incluyen palabras clave seleccionadas de los respectivos primeros conjuntos para al menos algunas de las imágenes.

12. El método de la reivindicación 1, que comprende además:

recibir una solicitud para un vídeo que es similar a un vídeo dado:

identificar al menos un grupo que incluye el vídeo dado:

seleccionar, de al menos un grupo, uno o más vídeos; y

proporcionar datos que especifiquen el uno o más vídeos.

13. Un sistema para realizar recomendaciones de vídeo según su contenido, que comprende:

un aparato de procesamiento de datos; y

un aparato de almacenamiento de memoria en comunicación de datos con el aparato de procesamiento de datos, el aparato de almacenamiento de memoria almacena instrucciones que se ejecutan por el aparato de procesamiento de datos y que sobre dicha ejecución ocasionan que el aparato de procesamiento de datos lleve a cabo operaciones que comprenden:

para cada vídeo en el conjunto de vídeos:

obtener un conjunto de imágenes que se incluyen en el vídeo;

para cada imagen en el conjunto de imágenes, generar un respectivo primer conjunto de una o más palabras clave que describen el contenido visual que representa la imagen usando técnicas de reconocimiento de objetos y personas; y

generar, con base al menos en los respectivos primeros conjuntos de una o más palabras clave para al menos algunas de las imágenes, un segundo conjunto de palabras clave que describen el vídeo;

asignar los vídeos en el conjunto de vídeos a grupos con base en el segundo conjunto de palabras clave que se generan para cada vídeo usando un proceso de aprendizaje automático para asignar vídeos que tengan al menos una similitud límite mayor que un valor umbral dentro de un mismo grupo, donde la similitud entre los dos vídeos se basa en la similitud entre los respectivos primeros conjuntos de una o varias palabras clave que se generaron para los dos vídeos;

recibir una solicitud para una recomendación de vídeo con base en un primer vídeo del conjunto de vídeos; y

proporcionar, como recomendación de vídeo, datos que identifiquen un segundo vídeo del conjunto de vídeos con base en el segundo vídeo que se asigna a un mismo grupo como el primer vídeo.

14. El sistema de la reivindicación 13, en donde la solicitud para la recomendación de vídeo se genera en respuesta a al menos uno de (i) presentación del primer vídeo o (ii) una solicitud para el primer vídeo.

15. El sistema de la reivindicación 13, en donde:
el segundo conjunto de palabras clave para cada vídeo se define formando una secuencia con base en la secuencia en la cual las imágenes, de las cuales se generan las palabras clave, ocurren en el vídeo; y

la similitud entre los dos vídeos se basa en una similitud entre la secuencia de palabras clave para un primer vídeo y la secuencia de palabras clave para un segundo vídeo.

16. El sistema de la reivindicación 13, en donde las operaciones comprenden además:

identificar, para un primer vídeo de dos vídeos, un número de ocurrencias de cada palabra clave en el segundo conjunto de palabras clave para el primer vídeo;

identificar, para un segundo vídeo de dos vídeos, un número de ocurrencias de cada palabra clave en el segundo conjunto de las palabras clave para el segundo vídeo,

en donde la similitud entre los dos vídeos se basa en una comparación del número de ocurrencias de cada palabra clave en el segundo conjunto de palabras clave para el

primer vídeo, y el número de ocurrencias de cada palabra clave en el segundo conjunto de palabras clave para el segundo vídeo.

17. El sistema de la reivindicación 13, en donde las operaciones comprenden además la generación, para cada vídeo, de un tercer conjunto de palabras clave que describen el contenido audible del vídeo, en donde el segundo conjunto de palabras clave que describen el vídeo se genera adicionalmente con base en el tercer conjunto de palabras clave.

18. Un producto de programa por ordenador, codificado en uno o más medios de almacenamiento por ordenador no transitorios, que comprenden instrucciones que cuando se ejecutan por uno o más ordenadores, ocasionan que uno o más ordenadores lleven a cabo operaciones que comprenden:
para cada vídeo en un conjunto de vídeos:

obtener un conjunto de imágenes que se incluyen en el vídeo;
para cada imagen en el conjunto de imágenes, generar un primer conjunto con una o más palabras clave que describen el contenido visual que representa la imagen usando técnicas de reconocimiento de objetos y personas; y

generar, con base al menos en los primeros conjuntos de una o más palabras clave para al menos algunas de las imágenes, un segundo conjunto de palabras clave que describen el vídeo;

asignar los vídeos en el conjunto de vídeos a los grupos con base en el segundo conjunto de palabras clave que se generan para cada vídeo usando un proceso de aprendizaje automático para asignar vídeos que tengan al menos una similitud límite mayor que un valor umbral dentro de un mismo grupo, donde la similitud entre los dos vídeos se basa en la similitud entre los respectivos primeros conjuntos de una o varias palabras clave que se generaron para los dos vídeos;

recibir una solicitud para una recomendación de vídeo con base en un primer vídeo del conjunto de vídeos; y

proporcionar, como recomendación de vídeo, datos que identifiquen un segundo vídeo del conjunto de vídeos con base en el segundo vídeo que se asigna a un mismo grupo como el primer vídeo.

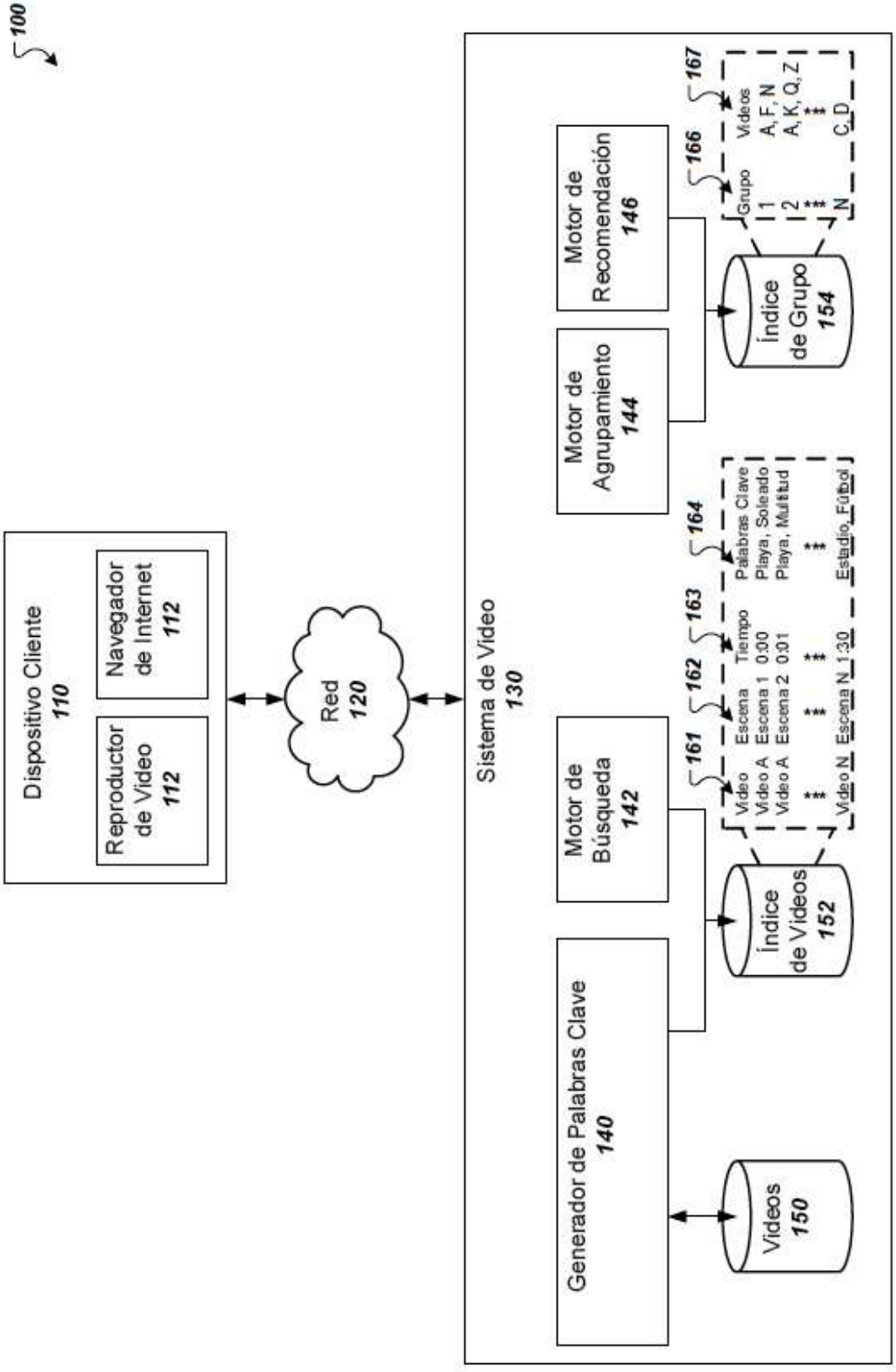
100
Dispositivo Cliente
110
Reproductor de Video
112
Navegador de Internet
112
Red
120
Sistema de Video
130
Generador de Palabras Clave
140
Motor de Búsqueda
142
Motor de Agrupamiento
144
Motor de Recomendación
146
Vídeos
150
Índice de Vídeos
152
Índice de Grupo
154
161
162
163
164
166
167
Vídeo
Escena
Tiempo
Palabras Clave
Video A Escena 1 0:00 Playa, Soleado
Video A Escena 2 0:01 Playa, Multitud
Video N Escena N 1:30 Estadio, Fútbol
Grupo
Videos
1 A, F, N
2 A, K, Q, Z
N C, D

FIG. 1

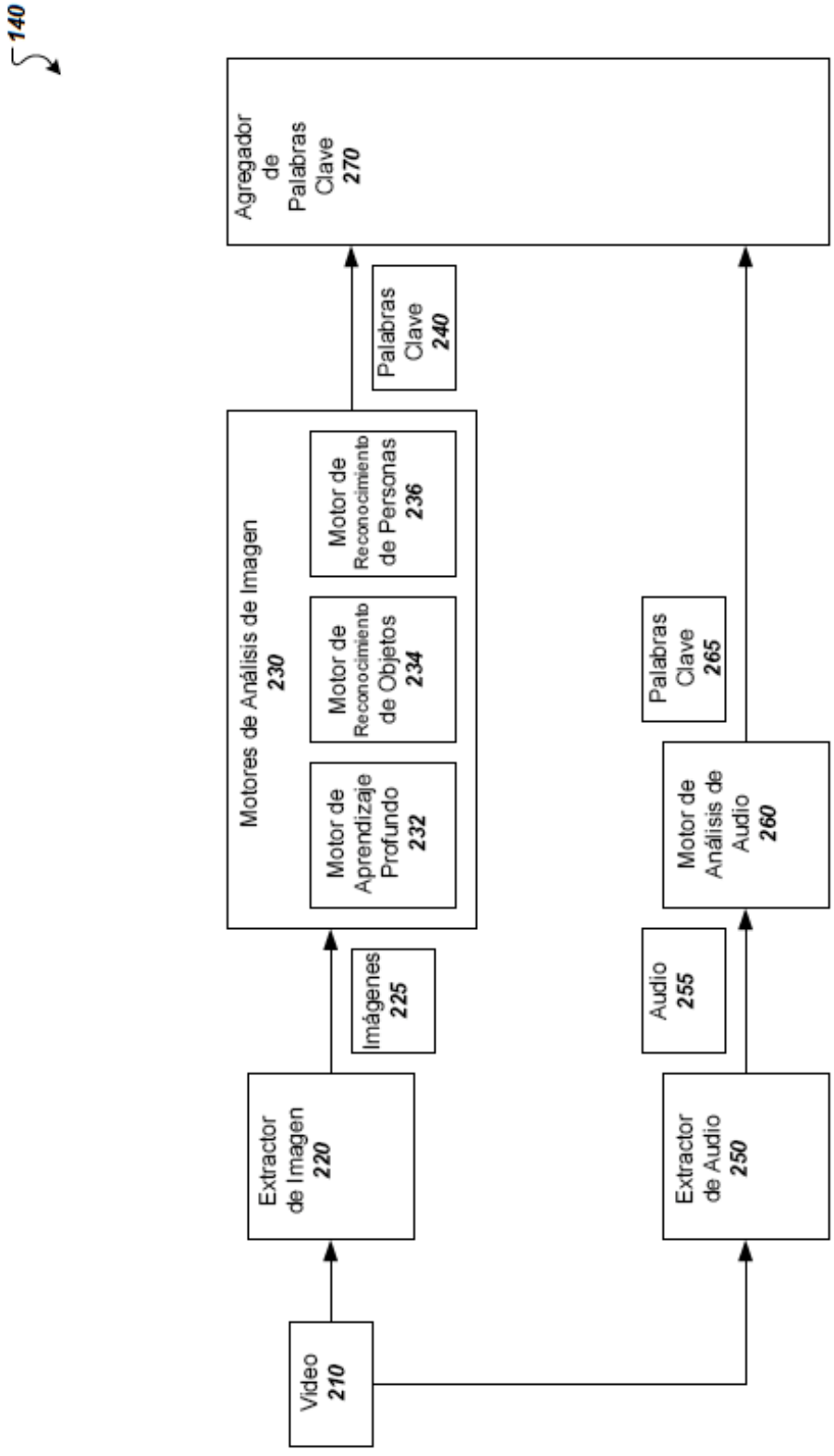
140
Agregador de Palabras Clave 270
Palabras Clave 240
Motores de Análisis de Imagen 230
Motor de Aprendizaje Profundo 232
Motor de Reconocimiento de Objetos 234
Motor de Reconocimiento de Personas 236
Palabras Clave 265
Motor de Análisis de Audio 260
Imágenes 225
Audio 255
Extractor de Imagen 220
Extractor de Audio 250
Video 210

FIG. 2

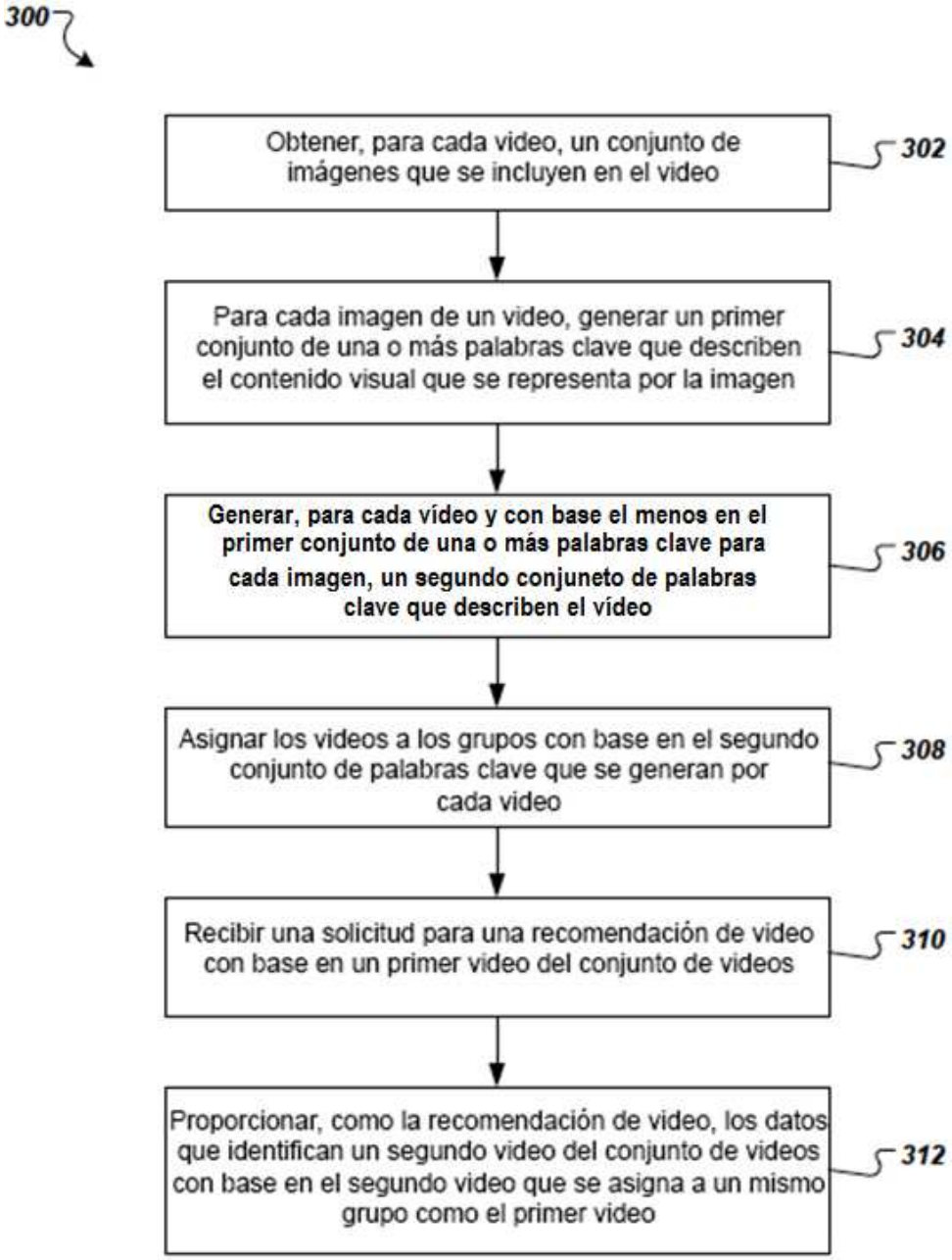
300
Obtener, para cada video, un conjunto de imágenes que se incluyen en el video
302
Para cada imagen de un video, generar un primer conjunto de una o más palabras clave que describen el contenido visual que se representa por la imagen
304
Generar, para cada vídeo y con base el menos en el primer conjunto de una o más palabras clave para cada imagen, un segundo conjuneto de palabras clave que describen el vídeo
306
Asignar los videos a los grupos con base en el segundo conjunto de palabras clave que se generan por cada video
308
Recibir una solicitud para una recomendación de video con base en un primer video del conjunto de videos
310
Proporcionar, como la recomendación de video, los datos que identifican un segundo video del conjunto de videos con base en el segundo video que se asigna a un mismo grupo como el primer video
312

FIG. 3

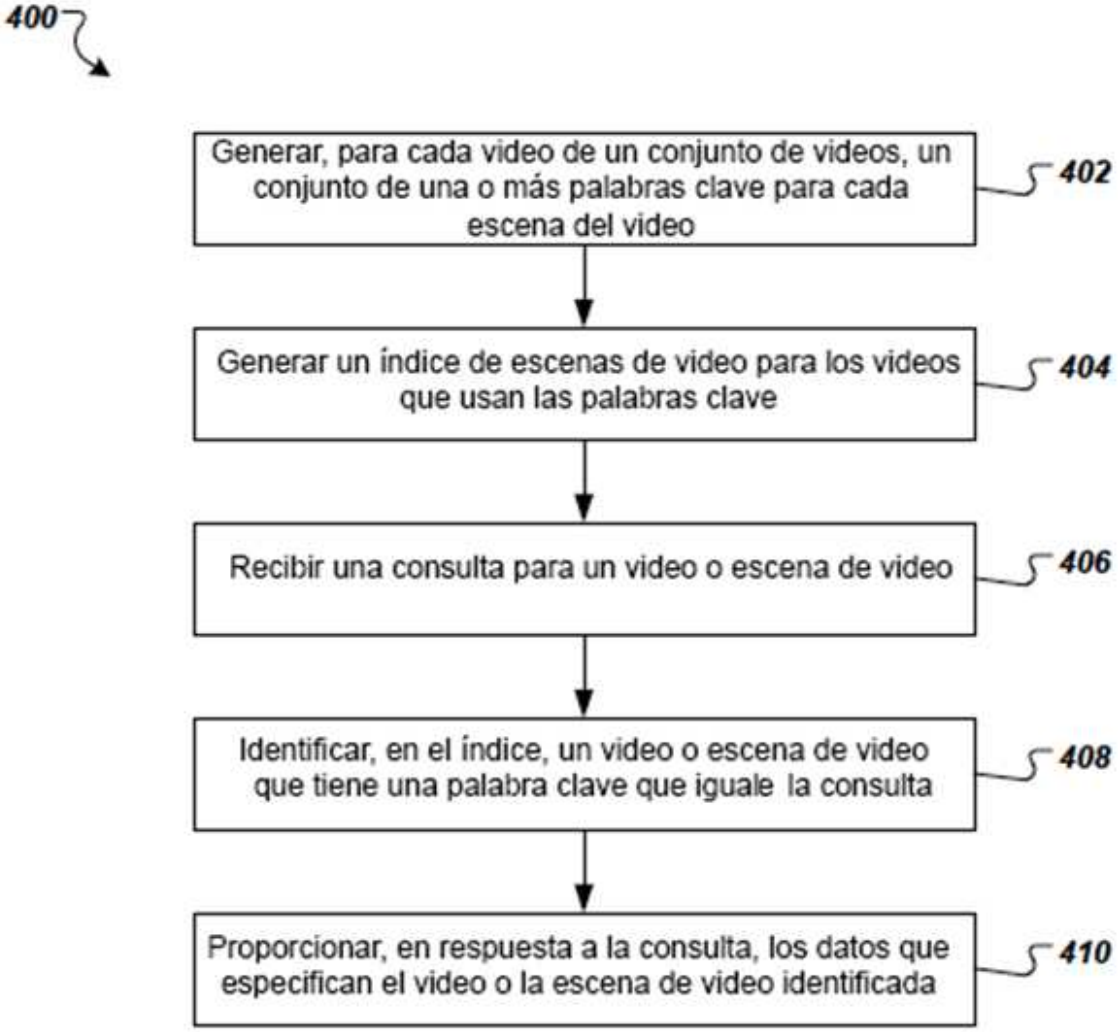
400
Generar, para cada video de un conjunto de videos, un conjunto de una o más palabras clave para cada escena del video
402
Generar un índice de escenas de video para los videos que usan las palabras clave
404
Recibir una consulta para un video o escena de video
406
Identificar, en el índice, un video o escena de video que tiene una palabra clave que iguale la consulta
408
Proporcionar, en respuesta a la consulta, los datos que especifican el video o la escena de video identificada
410

FIG. 4

OFICINA ESPAÑOLA
DE PATENTES Y MARCAS

ESPAÑA

(21) N.º solicitud: 201630878

(22) Fecha de presentación de la solicitud: 29.06.2016

(32) Fecha de prioridad:

INFORME SOBRE EL ESTADO DE LA TECNICA

(51) **Int. Cl. :** ***H04N21/466*** (2011.01)
H04N21/25 (2011.01)

DOCUMENTOS RELEVANTES

Categoría	**(56) Documentos citados**	**Reivindicaciones afectadas**
Y	US 2015082330 A1 (YUN SUNGRACK et al.) 19/03/2015, resumen; figuras 2, 9, 10, 14; párrafos [6, 7, 31 - 51, 59, 63 - 73, 88 - 94, 99];	1 - 18
Y	Jiangfan Feng Y Wenwen Zhou: "An Efficient Method for Automatic Video Annotation and Retrieval in Visual Sensor Networks"; College of Computer Science and Technology, Chongqing University of Posts And Telecommunications, Chongqing 400065, China; Artículo publicado 31/03/2014; Hindawi Publishing Corporation - International Journal of Distributed Sensor Networks Volumen 2014, ID Artículo 832512, 8 páginas URL:// http://journals.sagepub.com/doi/full/10.1155/2014/832512	1 - 18

Categoría de los documentos citados

X: de particular relevancia
Y: de particular relevancia combinado con otro/s de la misma categoría
A: refleja el estado de la técnica
O: referido a divulgación no escrita
P: publicado entre la fecha de prioridad y la de presentación de la solicitud
E: documento anterior, pero publicado después de la fecha de presentación de la solicitud

El presente informe ha sido realizado

☒ para todas las reivindicaciones ☐ para las reivindicaciones nº:

Fecha de realización del informe	**Examinador**	**Página**
23.10.2017	B. Pérez García	1/7

OFICINA ESPAÑOLA
DE PATENTES Y MARCAS

ESPAÑA

(21) N.º solicitud: 201630878

(22) Fecha de presentación de la solicitud: 29.06.2016

(32) Fecha de prioridad:

INFORME SOBRE EL ESTADO DE LA TECNICA

(51) **Int. Cl. :** ***H04N21/466*** (2011.01)
H04N21/25 (2011.01)

DOCUMENTOS RELEVANTES

Categoría	(56) **Documentos citados**	**Reivindicaciones afectadas**
A	US 2015037009 A1 (WANG HAOHONG) 05/02/2015, resumen; figuras 4, 5; párrafos [39 - 54, 64, 65, 84];	1 - 18
A	US 2003101104 A1 (DIMITROVA NEVENKA et al.) 29/05/2003, Resumen; párrafos [8 - 10, 24, 31 - 56, 60]; figuras 4 - 6.	1 - 18

Categoría de los documentos citados
X: de particular relevancia
Y: de particular relevancia combinado con otro/s de la misma categoría
A: refleja el estado de la técnica
O: referido a divulgación no escrita
P: publicado entre la fecha de prioridad y la de presentación de la solicitud
E: documento anterior, pero publicado después de la fecha de presentación de la solicitud

El presente informe ha sido realizado

☒ para todas las reivindicaciones ☐ para las reivindicaciones nº:

Fecha de realización del informe	**Examinador**	**Página**
23.10.2017	B. Pérez García	2/7

Documentación mínima buscada (sistema de clasificación seguido de los símbolos de clasificación)

H04N, G06F

Bases de datos electrónicas consultadas durante la búsqueda (nombre de la base de datos y, si es posible, términos de búsqueda utilizados)

INVENES, EPODOC, WPI, INSPEC

Fecha de Realización de la Opinión Escrita: 23.10.2017

Declaración

Novedad (Art. 6.1 LP 11/1986)	Reivindicaciones	1 - 18	**SI**
	Reivindicaciones		**NO**
Actividad inventiva (Art. 8.1 LP11/1986)	Reivindicaciones		**SI**
	Reivindicaciones	1 - 18	**NO**

Se considera que la solicitud cumple con el requisito de aplicación industrial. Este requisito fue evaluado durante la fase de examen formal y técnico de la solicitud (Artículo 31.2 Ley 11/1986).

Base de la Opinión.-

La presente opinión se ha realizado sobre la base de la solicitud de patente tal y como se publica.

1. Documentos considerados.-

A continuación se relacionan los documentos pertenecientes al estado de la técnica tomados en consideración para la realización de esta opinión.

Documento	Número Publicación o Identificación	Fecha Publicación
D01	US 2015082330 A1 (YUN SUNGRACK et al.)	19.03.2015
D02	An efficient method for automatic video annotation and retrieval in visual sensor networks	31.03.2014
D03	US 2015037009 A1 (WANG HAOHONG)	05.02.2015
D04	US 2003101104 A1 (DIMITROVA NEVENKA et al.)	29.05.2003

2. Declaración motivada según los artículos 29.6 y 29.7 del Reglamento de ejecución de la Ley 11/1986, de 20 de marzo, de Patentes sobre la novedad y la actividad inventiva; citas y explicaciones en apoyo de esta declaración

Se considera D01 el documento del estado de la técnica anterior más cercano al objeto de la invención. Siguiendo la redacción de la primera reivindicación, D01 describe un método implementado por ordenador para realizar recomendaciones de vídeo según su contenido *(ver resumen)*, comprendiendo el método:
- generar palabras clave que describen el contenido visual que representa la imagen usando técnicas de reconocimiento de objetos y personas *(párrafos 36, 37: "análisis de voz, de habla, huella digital de audio, huella digital de vídeo, análisis de escenas, de reconocimiento facial…"; párrafo 38:" la unidad de análisis del canal 212 puede analizar los contenidos de audio y vídeo de los programas del canal y generar etiquetas de audio y vídeo en intervalos de tiempo predeterminados, p.ej. cada k segundos");*
- generar, con base al menos en los primeros conjuntos de una o más palabras clave para al menos algunas de las imágenes, un segundo conjunto de palabras clave *(content tags)* que describen el vídeo;
- asignar los vídeos en el conjunto de vídeos *(214)* a grupos con base en el conjunto de palabras clave que se generaron para cada vídeo cuando tengan al menos una similitud límite mayor que un valor umbral dentro de un mismo grupo, donde la similitud entre los dos vídeos se basa en la similitud entre los respectivos primeros conjuntos de una o varias palabras clave que se generaron para los dos vídeos;
- recibir una solicitud para una recomendación de vídeo *(párrafos 44, 40)* con base en un primer vídeo del conjunto de vídeos; y
- proporcionar, como recomendación de vídeo, datos que identifican un segundo vídeo del conjunto de vídeos con base en el segundo vídeo que se asigna al mismo grupo como el primer vídeo *(párrafo 45)*.

Existen dos diferencias entre D01 y la primera reivindicación; en ésta se señala explícitamente que:
- para cada vídeo de un conjunto de vídeos se obtiene un conjunto de imágenes que se incluyen en el vídeo y que para cada imagen obtenida, se genera un conjunto de palabras clave (primer conjunto de palabras clave) que posteriormente permitirán seleccionar las palabras clave del vídeo (segundo conjunto de palabras clave).
- que se utiliza un proceso de aprendizaje automático para asignar los vídeos;

Respecto a la primera reivindicación, D01 no realiza tal apreciación, simplemente define que se generan un conjunto de palabras clave o etiquetas del vídeo para posteriormente realizar recomendaciones en base al contenido, sin entrar en detalle respecto a la extracción de imágenes del vídeo y creación de primer y segundo conjunto de palabras clave.
No obstante, aunque no se mencione como tal esta división del vídeo en tramas para analizar el contenido del vídeo y realizar el etiquetado, esta etapa resulta evidente para un experto en la materia ya que es la primera fase que se realiza en cualquier sistema de anotación o etiquetado automático de vídeo (ver D02, introducción y apartado a modo de ilustración). Por tanto, esta diferencia no produce un efecto técnico diferente o inesperado para un experto en la materia.

La segunda diferencia sí que produce un efecto técnico: D01 es capaz de realizar recomendaciones en tiempo real en función del contenido que se está visualizando y a partir de una BD de referencia *214*. Sin embargo, esta BD se puede haber creado mediante un proceso de aprendizaje automático, manual o híbrido. No obstante, D02 describe métodos de etiquetado de vídeos que pueden ser mediante aprendizaje automático o basado en búsquedas y multitud de modelos de clasificación de los vídeos.
Un experto en la materia podría utilizar el sistema de recomendación de vídeos basado en el contenido descrito en D01 y añadirle un sistema de aprendizaje automático y clasificación de vídeos con segmentación de imágenes, como los citados en D02 para obtener el objeto de la primera reivindicación. Por tanto, se considera que la primera reivindicación no cumple el requisito de actividad inventiva, según el Art. 8 de la Ley Española de Patentes.

La segunda reivindicación aclara que la recomendación de vídeo se genera en respuesta a una presentación del primer vídeo o una solicitud para el primer vídeo.
D01 realiza una recomendación cuando se está presentando el primer vídeo o bien por preferencias del usuario. Sin actividad inventiva.

La reivindicación 3 indica que el segundo conjunto de palabras clave se dispone formando una secuencia con base en el orden de la secuencia de las imágenes para las cuales se generaron las palabras clave y que la similitud entre dos vídeos se basa en una similitud entre la secuencia de palabras clave para un primer vídeo y la secuencia de palabras clave para un segundo vídeo.
La reivindicación 4 añade que para dos vídeos, se identifica el número de ocurrencias de cada palabra clave en el segundo conjunto de palabras clave de cada uno de los vídeos, tal que la similitud entre los dos vídeos se basa en una comparación del número de ocurrencias de cada palabra clave en el segundo conjunto de palabras clave de cada vídeo.
D01 permite comparar vídeos y sus etiquetas para determinar si se realiza una recomendación, en base a criterios de visualización y etiquetas de contenido… y establecer prioridades para realizar las recomendaciones *(ver figura 14).* Por tanto, se pueden programar o priorizar los criterios para realizar la comparación de las etiquetas o para establecer la similitud de los vídeos. Definir un número de ocurrencias, un número de palabras clave, un orden de presentación… no son parámetros de carácter técnico, si no, criterios del usuario o del programador. No se considera que estas reivindicaciones impliquen un esfuerzo inventivo para un experto en la materia y por tanto, no tienen actividad inventiva.
La quinta reivindicación explica que se genera también un tercer conjunto de palabras clave para cada vídeo que describen el contenido audible del vídeo, en donde el segundo conjunto de palabras clave que describen el vídeo se genera con base además en el tercer conjunto de palabras clave.
Es decir, se añade un etiquetado con base al audio del vídeo. Esta reivindicación queda anticipada por el *párrafo 37* y los módulos *1010, 1020 y 1030* de D01.

La reivindicación número 6 especifica que la generación del respectivo primer conjunto de una o más palabras clave que describen el contenido visual que se representa por una imagen dada, comprende usar un proceso de aprendizaje profundo para generar al menos una porción del respectivo primer conjunto de una o más palabras clave.
Un sistema de aprendizaje profundo para generar las palabras clave (ver D02, *apartado 2.1)* se utiliza como base para realizar cualquier sistema de anotación/etiquetado automático de vídeo. Sin actividad inventiva.

Las reivindicaciones 7 y 8 mencionan algunas alternativas para generar el primer conjunto de palabras al detectar un objeto/persona en una imagen y añadir una palabra clave que describa/identifique ese objeto/persona.
Estas alternativas aparecen en el párrafo 37 de D01. Carecen de actividad inventiva.

La reivindicación 9 menciona que la generación del segundo conjunto de palabras clave que describen el vídeo comprende:
- identificar, para cada palabra clave generada para al menos una imagen del vídeo, un número de imágenes del vídeo para las cuales se generó la palabra clave;
- identificar, para inclusión en el segundo conjunto de palabras clave, un número previamente especificado de las palabras clave con base en el número de imágenes para las cuales se generó cada palabra clave.
D01 es capaz de crear palabras clave o etiquetas a partir del análisis de imágenes/audio/textos… del vídeo/audio. Definir unos valores o umbrales para añadir la etiqueta como palabra clave del vídeo es una cuestión de diseño o programación que no implica superar una dificultad técnica añadida. No se considera que suponga actividad inventiva para un experto en la materia.

La reivindicación número 10 establece que se genera un índice de escenas de vídeo para el conjunto de vídeos, en donde el índice incluye un conjunto de palabras clave que describen cada escena de vídeo y que permite consultar el índice e identificar la escena del vídeo.
D01 no explica la realización de dicho índice como tal, si bien es capaz de realizar la recomendación en tiempo real de escenas de vídeos que tengan unas etiquetas semejantes a la del vídeo que se está visualizando.
El efecto técnico que crea esta diferencia es la de poder acceder directamente y únicamente a la escena de interés por la coincidencia de las palabras clave.
El problema técnico objetivo es por tanto, cómo acceder a la escena de interés directamente.
Esto aparece resuelto en el propio D01 ya que realiza el análisis en tiempo real y por tanto, presenta las recomendaciones que puedan ser de interés al observador. No tiene actividad inventiva.

La reivindicación 11 aclara que el segundo conjunto de palabras clave incluyen palabras clave seleccionadas de los respectivos primeros conjuntos para al menos algunas de las imágenes, lo cual está implícito en un sistema de anotación/etiquetado automático de vídeo. Sin actividad inventiva.

La reivindicación 12 indica que el método comprende además recibir una solicitud para un vídeo que es similar a un vídeo dado, identificar al menos un grupo que incluye el vídeo dado, seleccionar uno o más vídeos y proporcionar datos que especifiquen el uno o más vídeos.
Estas características están anticipadas por D01 *(ver figura 14).* Sin actividad inventiva.

Las reivindicaciones 13 – 17 se refieren al sistema que implementa el método anterior y la reivindicación 18 al programa de ordenador. No aportan características técnicas adicionales a las definidas para el método, y por tanto, corren la misma suerte que éste. No tienen actividad inventiva.

En resumen, la solicitud presentada no cumple el requisito de actividad inventiva para un experto en la materia, según el Art. 8 de la Ley Española de Patentes.