(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 390 295**

(51) Int. CI.:

G10L 17/00	(2006.01)
G10L 11/02	(2006.01)
G06K 9/46	(2006.01)
H04N 7/15	(2006.01)
G06K 9/62	(2006.01)
H04N 7/14	(2006.01)
G10L 21/02	(2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Número de solicitud europea: **07835698 .7**

(96) Fecha de presentación: **13.02.2007**

(97) Número de publicación de la solicitud: **2035799**

(97) Fecha de publicación de la solicitud: **18.03.2009**

(54) Título: **Identificación de personas usando múltiples tipos de entradas**

(30) Prioridad:
22.06.2006 US 425967

(45) Fecha de publicación de la mención BOPI:
08.11.2012

(45) Fecha de la publicación del folleto de la patente:
08.11.2012

(73) Titular/es:
MICROSOFT CORPORATION (100.0%)
ONE MICROSOFT WAY
REDMOND, WA 98052-6399, US

(72) Inventor/es:
ZHANG, CHA;
VIOLA, PAUL A.;
YIN, PEI;
CUTLER, ROSS G.;
SUN, XINDING y
RUI, YONG

(74) Agente/Representante:
CARPINTERO LÓPEZ, Mario

DESCRIPCIÓN

Identificación de personas usando múltiples tipos de entradas

Antecedentes

Existe una gran diversidad de situaciones en la cuales es deseable identificar personas, incluyendo personas que están hablando, usando sistemas que están, al menos, automatizados. Algunos sistemas existentes que identifican hablantes usan audio-por ejemplo, podrían usar localización de fuentes sonoras", que incluye procesar la entrada a partir de múltiples micrófonos en diferentes localizaciones para intentar identificar la dirección o las direcciones a partir de las cuales se origina el habla. Algunos otros sistemas intentan mejorar la precisión de procedimientos similares a la localización de la fuente de sonido llevando a cabo "la fusión del nivel de decisión", donde se combinan los datos procedentes de múltiples entradas en el punto en que se toman las decisiones acerca de la detección de personas o hablantes.

El documento US 204/263636 A1 se refiere a un sistema y un procedimiento para realizar teleconferencias y grabar reuniones. Una matriz de micrófonos puede estar integrada con una cámara de 360 grados. El sistema puede entonces capturar la señal de audio de toda de la sala de reuniones, usar localización de fuente sonora (SSL) para encontrar la dirección del hablante, y el sistema de reunión distribuido puede usar tanto SSL basado en audio como seguimiento de personas basado en visión para detectar hablantes. Asimismo se describe el seguimiento multiseñal para el seguimiento y detección de personas.

Sumario

Un objeto de la presente invención es proporcionar un procedimiento y un sistema para la identificación de personas, incluyendo hablantes.

Este objeto se resuleve por la materia objeto de las reivindicaciones independientes.

Se proporcionan realizaciones en las reivindicaciones dependientes.

Identificación de personas usando múltiples tipos de entradas

En lo sucesivo se presenta un sumario simplificado de la divulgación con el fin de proporcionar una comprensión básica al lector. Este sumario no es una visión general de la divulgación y no identifica elementos clave o críticos de la invención o delimita el alcance de la invención. Su único propósito es presentar algunos conceptos divulgados en el presente documento de una forma simplificada como un preludio de la descripción más detallada que se presentará más tarde.

En el presente documento se describen varias tecnologías y técnicas dirigidas a la identificación de personas, incluyendo los hablantes. Tales tecnologías y técnicas incluyen la identificación de un grupo de "características" de identificación a partir de múltiples tipos de entrada, o modalidades (trayectorias a través de las cuales un sistema informático puede reconocer la entrada), que incluye tanto entrada de audio como de vídeo; y la generación de un "clasificador" que incluye un subconjunto de características del grupo de características donde el subconjunto de características es seleccionado de manera que el clasificador identifica eficientemente regiones donde las personas o hablantes pueden existir.

Descripción de los dibujos

La figura 1 ilustra un diagrama ejemplar generalizado que muestra un sistema en el que se puede llevar a cabo la detección de personas.
La figura 2 ilustra una representación gráfica de una imagen ejemplar así como regiones ejemplares que se pueden identificar por contener personas o hablantes.
La figura 3 ilustra un flujo operativo ejemplar generalizado que incluye varias operaciones que se puede llevar a cabo cuando se identifica una persona.
La figura 4 ilustra un diagrama ejemplar generalizado que muestra algunas características ejemplares que pueden ser identificadas y usadas en algunas implementaciones.
La figura 5 ilustra algunas características ejemplares de vídeo.
La figura 6 ilustra un rectángulo de características ejemplares representativas que incluye algunas características ejemplares de vídeo.
La figura 7 ilustra un diagrama ejemplar generalizado que muestra un sistema en el que se puede llevar a cabo la generación de un clasificador para detección de personas o hablantes.
La figura 8 ilustra un diagrama ejemplar generalizado que muestra un sistema en el que se puede llevar a cabo la detección de personas o hablantes.
La figura 9 ilustra algunas representaciones ejemplares de ventanas de detección que se pueden usar como parte del procedimiento de detección de personas o hablantes.
La figura 10 ilustra un dispositivo informático ejemplar en el que se pueden aplicar las diversas tecnologías descritas en el presente documento

Descripción detallada

La presente invención se extiende a varias tecnologías y técnicas dirigidas a la identificación de personas incluidos hablantes. Mas en particular, en el presente documento se describen, entre otras cosas, procedimientos y sistemas que facilitan la identificación de personas usando múltiples tipos de entradas donde los múltiples tipos de entradas son considerados un principio del procedimiento de detección, en lugar de combinarse al final del procedimiento de detección.

Volviendo ahora a la figura 1, en la misma se ilustra un diagrama ejemplar generalizado que muestra un sistema 100 en el que se puede llevar a cabo la detección de personas. Esta descripción de la figura 1 se hace con referencia a la figura 10. Sin embargo, cabe entender que los elementos descritos con referencia a la figura 1 no están destinados a limitarse a ser usados con los elementos descritos con referencia a esta otra figura. Asimismo, aunque el diagrama ejemplar de la figura 1 indica elementos particulares, en algunas implementaciones no todos estos elementos pueden existir, y en algunas implementaciones pueden existir elementos adicionales.

Incluido en la figura 1 se encuentran uno o más dispositivos de entrada de vídeo 110, uno o más dispositivos de entrada de audio 120, uno o u otros más dispositivos de entrada 130, datos de vídeo 140, datos de audio 150, otros datos 160, un detector de personas 170 aplicado en un dispositivo detector 165, un dispositivo auxiliar 175, y la salida del detector de personas, cualesquiera personas o hablantes 180 detectados.

El detector 170 acepta la entrada, que puede entonces usar para intentar identificar una o más personas 180, incluyendo personas que están hablando, o “hablantes”. El detector puede usar una variedad de mecanismos para intentar identificar personas incluyendo los mencionados más en detalle en el presente documento. En algunas implementaciones, el detector puede utilizar mecanismos de detección determinados en otro lugar, mientras que en otras implementaciones el detector puede tanto determinar cómo ejecutar los mecanismos de detección. El detector puede usar una variedad de entradas, incluyendo datos de vídeo 140, datos de audio 150, y otros datos 160.

El o los dispositivos de entrada de vídeo 110 pueden comprender una variedad de dispositivos de entrada de vídeo, incluyendo una variedad de cámaras y tipos de cámaras con una gama de funcionalidad. En una implementación, los dispositivos de entrada de vídeo 110 pueden incluir múltiples cámaras situadas en una disposición circular para de este modo proporcionar una visión de 360º. En otras implementaciones, la misma visión de 360º se puede proporcionar por una única cámara, quizás con una sola lente. En otras implementaciones adicionales, el o los dispositivos de entrada de vídeo pueden proporcionar una visión que cubre un intervalo inferior a 360º.

Al menos parte de la salida del o los dispositivos de entrada de vídeo 110 son los datos de vídeo 140. Estos datos pueden incluir múltiples tramas individuales de datos de vídeo, donde cada trama comprende una imagen constituida por múltiples píxeles. Por ejemplo, una cámara que es capaz de producir vídeo a una velocidad de 30 tramas de vídeo por segundo puede producir como salida 30 imágenes cada segundo. En algunas implementaciones, cada imagen producida por la cámara puede ser conocida como la “imagen de base” (para diferenciarla de otras imágenes calculadas, como las imágenes diferenciales de corto plazo y las imágenes promedio de largo plazo explicadas en lo sucesivo). Obsérvese que el o los dispositivos de entrada de vídeo 110 pueden proporcionar datos de varias formas, incluyendo formas en las que todos los píxeles de cada trama de vídeo no son transmitidos explícitamente desde el dispositivo de entrada de vídeo. Por ejemplo, la salida del o los dispositivos de entrada de vídeo 110 pueden comprender una sola trama de vídeo inicial donde se proporcionan los valores para todos los píxeles de la trama, y la salida para al menos algunas tramas posteriores adicionales solo pueden incluir cambios de la trama inicial. En este caso, la representación pixel a pixel para cualquier trama posterior se puede determinar aplicando los cambios en la trama original. En cualquier caso, se puede considerar que cada imagen de base producida por la cámara puede incluir una imagen plena pixel a pixel.

Además, los datos de vídeo 140 también pueden comprender datos adicionales calculados. Por ejemplo, en algunas implementaciones puede ser útil calcular una “diferencia de corto plazo” usando múltiples tramas de vídeo. Tal diferencia de corto plazo puede ser útil, por ejemplo, para identificar movimiento. Aunque una diferencia de corto plazo se puede calcular de varias maneras, un posible procedimiento es, para cada pixel de la imagen, sustraer el valor del pixel en la trama inmediatamente anterior, del valor del pixel en la trama actual. Al mismo tiempo, el procedimiento puede también sustraer el valor del pixel en la segunda trama inmediatamente anterior del valor del pixel en la trama actual. Entonces, el mínimo de las dos operaciones de sustracción puede ser tomado como el valor actual del pixel. Para pixeles en los que no existe movimiento –es decir, para áreas donde la imagen permanece fija – este procedimiento tenderá a producir valores próximos a cero. Para los píxeles en los que ha habido movimiento reciente, este procedimiento puede producir valores que son en algunos casos muy superiores a cero. Este procedimiento específico puede ser representado por la siguiente ecuación, donde M_t es la imagen diferencia de corto plazo en el tiempo τ e I_t es la imagen de la cámara en el tiempo t.

$$M_t = \min(|I_t - I_{t-1}|, |I_t - I_{t-2}|)$$

Obsérvese que, dependiendo de la velocidad de trama de la cámara, las tramas “anteriores” usadas para este cálculo

pueden ser algo más que las dos tramas inmediatamente anteriores. Por ejemplo, cuando se usa una cámara con una velocidad de trama de 30 tramas por segundo, se pueden usar las tramas de las 10 tramas anteriores y de las 20 tramas anteriores en lugar de las dos tramas inmediatamente anteriores.

En las mismas o distintas implementaciones puede ser útil calcular un "promedio de largo plazo" de tramas de vídeo, que también pueden ser una parte de los datos de vídeo 140. Un promedio de plazo largo puede identificar porciones de la región capturada por el o los dispositivos de entrada de vídeo 110 donde ha existido previamente movimiento, incluso si el movimiento no se produjo recientemente. Aunque un promedio de plazo largo se puede calcular de varias maneras, un procedimiento posible es calcular un promedio de funcionamiento de las imágenes diferenciales de corto plazo, quizás incluyendo las producidas por el procedimiento de diferencia de corto plazo descrito anteriormente. Usando tal procedimiento, la trama de vídeo de promedio de largo plazo puede actualizarse continuamente para que de este modo cada pixel de la trama comprenda el valor promedio de ese pixel de todas o muchas tramas anteriores de imágenes diferenciales de corto plazo. Para áreas de la región capturada por las cámaras donde ha habido poco o nulo movimiento en el transcurso de la captura del vídeo, este procedimiento puede tender a producir valores que están próximos de cero. Por el contrario, para áreas donde ha habido movimiento en algún punto del pasado, que a menudo incluye áreas de la región que contiene personas, este procedimiento puede tender a producir valores distintos de cero.

Asimismo, en algunas implementaciones, en lugar de considerar imágenes calculadas, como la diferencia de corto plazo y el promedio de largo plazo, de la trama más reciente de datos de vídeo, puede ser útil considerarla como que también incluyen al menos algunos datos "futuros". Por ejemplo, la diferencia de corto plazo puede usar la trama actual, la trama anterior más reciente, y la "siguiente trama" como entrada, esperando hasta que la siguiente trama de vídeo es capturada y a continuación calcular la diferencia de corto plazo usando estas tres tramas identificadas. Cualquier operación puede aumentar la latencia de al menos esta parte del procedimiento de detección de personas en el momento necesario para capturar los datos adicionales "futuros", pero en algunos casos esta latencia aumentada puede ser desviada por los datos finalmente representados por la imagen calculada.

Los datos de vídeo 140 pueden comprender cualquier imagen o todas las imágenes mencionadas anteriormente, así como imágenes o tramas de vídeo adicionales. Estas imágenes pueden ser proporcionadas a partir de, o si fuese necesario, calculadas en, varias localizaciones, incluyendo el o los dispositivos de entrada de vídeo 110, el detector 170, o cualquier otro dispositivo. Además, aunque esta discusión se refiere a "vídeo", es importante entender que se puede usar cualquier cámara capaz de producir imágenes, incluyendo las cámaras no consideradas tradicionalmente como "cámaras de vídeo". Por ejemplo una cámara "fotográfica" capaz de tomar numerosas fotografías en secuencia puede ser usada en algunas implementaciones. Además, si la detección de movimiento no es considerada importante, en algunas implementaciones se puede usar una sola imagen fija. Además, en algunos casos se pueden usar datos adicionales. Por ejemplo, el detector puede usar color de piel como medio adicional para identificar regiones que pueden contener una persona.

El o los dispositivos de entrada de audio 120 pueden comprende varios dispositivos de entrada de audio, incluyendo una variedad de micrófonos y tipos de micrófonos con una gama de funcionalidad. En algunas implementaciones, el o los dispositivos de audio pueden incluir una matriz de micrófonos constituida por múltiples micrófonos situados en diferentes posiciones. Usando una variedad de información de tal conjunto de micrófonos, incluyendo quizás el reconocimiento de las diferentes posiciones de los micrófonos y diferencias en amplitud y tiempos de llegada para los sonidos detectados por los micrófonos, el o los dispositivos de entrada de audio pueden proporcionar datos que incluyen direcciones desde las cuales se han originado sonidos. Tal entrada es a veces incluida como parte de una técnica denominada "localización de fuente sonora" (SSL). En algunos casos, tal información direccional puede ser útil cuando se determinan hablantes.

Los datos de audio 150 pueden en algunas implementaciones, con alguna cantidad de procesamiento, incluir una "función de distribución de probabilidad" que proporciona valores de probabilidad que representan la probabilidad de que este sonido, incluyendo quizás la voz de un hablante, proceda de cualquier dirección particular. Por ejemplo, si se puede usar información del o los dispositivos de entrada de audio 110 para localizar un sonido procedente de cualquier dirección, la función de distribución de probabilidad, también denominada como la función de probabilidad SSL en el presente documento, puede contener un valor de probabilidad par diferentes acimuts, o direcciones. Para las direcciones en las cuales se ha detectado poco o nulo sonido, el valor de probabilidad puede ser bajo, mientras que las direcciones en las cuales se detecta más sonido, el valor de probabilidad puede ser alto.

En algunas implementaciones, quizás dependiendo de las capacidades de los dispositivos de entrada de audio 120, los datos de audio 150 pueden incluir información adicional. Por ejemplo, en algunas implementaciones los datos de audio pueden incluir el alcance o distancia a las fuentes de sonido y/o la elevación de las fuentes de sonido. En algunas implementaciones, estos datos –como el alcance a las fuentes de sonido y/o la elevación de las fuentes de sonido- pueden asociarse con las funciones de distribución de probabilidad.

Los datos de audio 150 pueden comprender cualquiera o todos los datos mencionados anteriormente, así como datos adicionales. Estos datos pueden ser proporcionados a partir de, o si fuese necesario, calculados en, una variedad de localizaciones, incluyendo en el hardware asociado al o los dispositivos de entrada de audio 120, el detector 170 o cualquier otro dispositivo. Por ejemplo, en algunas implementaciones, la localización de fuente sonora, que produce quizás una función de probabilidad de SSL, se puede llevar a cabo usando hardware asociado al o los dispositivos de entrada de audio, se puede llevar a cabo usando hardware asociado al detector, o se puede llevar a cabo usando otro

hardware o en alguna otra localización.

En algunas implementaciones, los datos de vídeo 140 y los datos de audio 150 se pueden unir de alguna manera para que las direcciones asociadas a los datos de vídeo puedan ser correlacionados a direcciones asociadas a los datos de audio. Por ejemplo, en tal implementación, la región de una función de probabilidad de SSL de un acimut a otro se puede correlacionar con una región particular en una o más tramas de vídeo, identificadas quizá por localizaciones horizontales de píxeles. Por ejemplo, en una implementación, la región de por ejemplo 10º a 20º, se puede correlacionar con píxeles situado, por ejemplo, entre las localizaciones horizontales de píxeles 100 a 200. Usando tal correlación, se puede usar información del o los dispositivos de entrada de audio 120 cuando se identifican regiones particulares en las imágenes proporcionadas por el o los dispositivos de vídeo 110, y viceversa. Para los datos de audio 150 que incluyen información adicional como la elevación, la información adicional también se puede correlacionar con regiones particulares en la imagen. Por ejemplo, la información de elevación puede correlacionarse con localizaciones verticales de píxeles. También pueden existir tipos similares de correlación con cualquiera del o los otros dispositivos de entrada 130, dependiendo de la naturaleza y el funcionamiento de tales dispositivos.

En algunas implementaciones, pueden existir tipos adicionales de entrada y usarse como parte del procedimiento de detección. En algunos casos, los tipos adicionales de entrada pueden originarse en el u otros dispositivos de entrada 130 y producir al menos parte de los otros datos 160. Por ejemplo, otro posible dispositivo de entrada debe puede incluir una cámara tridimensional que es capaz de proporcionar alguna medida de la distancia o profundidad a los elementos en una imagen.

El detector 170 se puede aplicar en una variedad de dispositivos informático, incluyendo como se muestra en una dispositivo detector 165. En algunas implementaciones, este dispositivo detector puede contener el hardware necesario para aplicar la detección de personas y puede conectarse por ejemplo, a uno o más dispositivos de entrada de vídeo o uno o más dispositivos de entrada de audio, mediante una variedad de medios de conexión, tales como USB, cualquier variedad de red incluyendo redes inalámbricas, etcétera, como lo apreciará el experto en la técnica. En otras implementaciones, el detector puede aplicarse en un dispositivo detector que incluye uno o más dispositivos de entrada de vídeo y uno o más dispositivos de entrada de audio, tales como por ejemplo uno o más dispositivos de entrada de vídeo 110 y uno o más dispositivos de entrada de audio 120. Cualquier dispositivo detector puede incluir una variedad de elementos de procesamiento, incluyendo unidades centrales de procesamiento universales (COU) y/o unidades procesadoras de señales digitales (DPS). Un entorno informático ejemplar en el que se puede aplicar un detector se describe en lo sucesivo con referencia a la figura 10.

Con independencia de si el dispositivo detector 165 contiene o está conectado a elementos similares al o los dispositivos de entrada de vídeo 110. El o los dispositivos de entrada de audio 120, y otros dispositivos de entrada 130, el dispositivo detector puede igualmente en algunas implementaciones conectarse a uno o más dispositivos auxiliares 175. En este contexto, un dispositivo auxiliar puede ser cualquier dispositivo que proporciona una funcionalidad adicional que se puede asociar a o ser útil para el dispositivo detector 165. Por ejemplo, en algunas implementaciones un dispositivo auxiliar puede comprender un ordenador portátil que contiene un disco duro en el que el dispositivo detector puede almacenar vídeo, audio y posiblemente regiones en las que se han detectado personas o hablantes. En las mismas u otras implementaciones, el dispositivo auxiliar puede proporciona ciclos de procesamiento informático al dispositivo detector para que, por ejemplo, el dispositivo detector pueda descargar algo de o todo su procesamiento de detección en el dispositivo auxiliar. En otras implementaciones más, un dispositivo auxiliar puede comprende solo un medio de almacenamiento –puede, por ejemplo, ser un disco duro en una carcasa USB. En general, un dispositivo auxiliar puede conectarse al dispositivo detector usando cualquier medio de conexión, incluyendo USB, cualquier forma de red, etc.

En algunas implementaciones, puede ser importante que los datos de diferentes dispositivos de entrada se sincronicen. Por ejemplo, la entrada del o los dispositivos de entrada de vídeo 110 se puede sincronizar con la entrada del o los dispositivo de audio 120.

Volviendo ahora a la figura 2, se muestra en la misma una representación gráfica de una imagen ejemplar 200 así como regiones ejemplares que se pueden identificar que contienen personas o hablantes. Incluidas en la representación gráfica se encuentra una primera región 210 asociada a una primera persona identificada, una segunda región 220 asociada a una segunda persona identificada, y una tercera región 230. Esta descripción de la figura 2 se hace con referencia a la figura 1. Sin embargo, cabe entender que los elementos descritos con referencia a la figura 2, no están destinado a limitarse a su uso con los elementos descritos con referencia a esta otra figura. Asimismo, aunque el diagrama ejemplar de la figura 2 indica elementos particulares, en algunas implementaciones no todos estos elementos pueden existir, y en algunas implementaciones pueden existir elementos adicionales.

La imagen ejemplar 200 puede representar una trama de vídeo producida por uno o más dispositivos de entrada de vídeo, incluyendo quizás el o los dispositivos de entrada de vídeo 110 descritos anteriormente con referencia a la figura 1.

En algunas implementaciones, un detector, quizás similar al detector 170 de la figura 1, puede indicar personas o hablantes identificados usando localizaciones horizontes y físicas por píxeles que indican un rectángulo o alguna otra forma. Por ejemplo, un detector puede indicar que la primera región 210 puede tener una alta probabilidad de asociarse a una persona o hablante. Asimismo, y posiblemente al mismo tiempo, puede indicar que la segunda región 220 puede

tener también una alta probabilidad de asociarse a una persona o hablante. Como se puede apreciar examinando la figura 2, en el caso de la primera región 210 y la segunda región 220, tal detector sería correcto, porque cada región contiene una persona. Un detector puede identificar también la tercera región 230 que tiene una alta probabilidad de ser asociada a una persona –quizás, por ejemplo, a causa de la reflexión del sonido a partir de una pared u otra superficie. Como un detector puede indicar solo probabilidades de que una región particular está asociada a una persona, en algunos casos las regiones identificadas por un detector pueden no contener realmente una persona. El umbral o nivel al cual un detector considera que una región contiene una persona puede cambiar y se puede definir dependiendo de la aplicación o uso del detector. Por ejemplo, en algunas implementaciones tal umbral se puede establecer en algún valor relativamente alto, que limitaría presumiblemente el número de regiones que pueden asociarse a una persona a la vez que también limita el número de regiones que finalmente son erróneas.

La imagen ejemplar 200 está destinada a un fin ilustrativo y no cabe interpretarla para limitar el alcance de cualquier invención reivindicada. Asimismo, la representación de persona identificada y regiones erróneas ilustran solo un medio gráfico para mostrar regiones identificadas y erróneas. Se puede usar cualquier medio para representar o ilustrar regiones.

Volviendo ahora la figura 3, en ella se muestra un flujo operativo ejemplar generalizado 300 que incluye varias operaciones que puede ser llevadas a cabo cuando se identifica una persona. La siguiente descripción de la figura 3 se realiza con referencia a figuras adicionales, que incluyen la figura 1, figura 4, la figura 8 y la figura 9. Sin embargo, cabe entender que el flujo operativo descrito con referencia a la figura 3 no está destinado a limitarse a su uso con los elementos descritos con referencia a estas otras figuras. Además, aunque el flujo operativo ejemplar de la figura 3 indica un orden particular de ejecución, en una o más realizaciones alternativas, las operaciones se pueden ordenar de manera diferente. Asimismo, aunque el flujo operativo ejemplar contiene múltiples etapas, cabe reconocer que en algunas implementaciones al menos algunas de estas operaciones se pueden combinar o ejecutar al mismo tiempo.

En una implementación de la operación 310, se identifica una serie de características. La serie de características se pueden usar entonces como entrada cuando se ejecuta la operación de generación del clasificador 315. En este contexto, una característica es una entidad asociada a uno o más tipos de entrada que sirve para cuantificar algún elemento de la entrada o entradas en un momento particular. Puede haber características de audio, características de vídeo y otras características asociadas a otros tipos de entrada. Por ejemplo, en el caso de entrada de audio que incluye una función de probabilidad SSL, se puede definir una característica, al menos en parte, mediante algún tipo de comparación de los valores mínimos y máximos “locales” de la función de probabilidad SSL, en comparación con los valores mínimos y máximos “globales” de la misma función de probabilidad SSL (donde “locales” puede referirse a los valores para un subconjunto de toda la función de probabilidad SSL mientras que "globales" puede referirse a valores para toda la función de probabilidad SSL). Dependiendo de los valores de la función de probabilidad SSL, diferentes características de audio producirán diferentes resultados numéricos. Algunos detalles específicos aplicables a algunas implementaciones relativas a las características que se pueden usar, usando más información acerca de las características específicas de las entradas de audio y vídeo, se mencionan más en detalle en lo sucesivo, por ejemplo, con referencia a la figura 4.

El medio por el cual la serie de características puede ser identificada puede variar dependiendo de la naturaleza de las características y la o las entradas con las cuales están asociados. La identificación de características, así como el modo en que las características pueden ser generadas, es a menudo la tarea de uno o más diseñadores con conocimientos especializados aplicables al área objeto para la cual se ha de generar la serie de características. Por ejemplo, la creación de una característica de audio que es definida, al menos en parte, por una función que produce un número cuando valores dados de una función de probabilidad SSL puede requerir reflexión por parte de un diseñador humano que concibe la característica de audio.

En algunos casos, las características se pueden elegir debido a que se supone que proporcionan alguna información acerca de la existencia de una persona o hablante. Sin embargo, es importante destacar que una característica no tiene necesariamente que proporcionar un resultado siempre preciso o particularmente “bueno”. La operación de generación de clasificador 315, mencionada más adelante, se puede usar para seleccionar, a través de otro procedimiento, las características más apropiadas para la detección de personas y hablantes.

En una implementación ejemplar de tal operación de generación de clasificador 315, un subconjunto de características identificadas en la operación 310 puede ser seleccionado para formar un “clasificador”. Tal como se usa en el presente documento, el término “clasificador” se refiere a una entidad que, cuando se presenta con entradas –incluyendo, en algunas implementaciones, entradas de audio y vídeo similares a las mencionadas en cualquier otro lugar en esta solicitud- puede proporcionar un resultado aproximado que proporciona alguna estimación de si una región particular en una imagen contiene una persona o hablante.

Los clasificadores a menudo se construyen o crean usando un procedimiento automatizado. Por ejemplo, en algunas implementaciones los clasificadores se pueden crear usando algún tipo de “algoritmo de aprendizaje”, que comprende un procedimiento que toma alguna entrada y produce una salida que puede clasificar o responder a cuestiones particulares. El clasificador generado consiste en general en algún subconjunto de las características identificadas en la operación 310, donde las características en el subconjunto han sido seleccionadas por el algoritmo de aprendizaje para responder

a la cuestión asociada al clasificador. Dependiendo de varias necesidades, las características seleccionadas pueden responder a la cuestión más precisamente, más eficientemente, etc. En algunas implementaciones, las características que forman parte del clasificador pueden localizarse en el clasificador para de este modo mejorar la operación del clasificador cuando se usa para la detección. Por ejemplo, se pueden ordenar características preferibles para que sean evaluadas antes en el clasificador si su evaluación requiere relativamente menos recursos informáticos, o si tales recursos se correlacionan en mayor medida con una persona o un hablante que otras características. Tal ordenamiento se puede llevar a cabo ponderando las características preferibles mientras se genera el clasificador, seleccionando las características en el clasificador después de haber generado el clasificador, o a través de otros medios. Algunos detalles específicos aplicables en algunas implementaciones relativas a la generación de clasificadores que usan algoritmos de aprendizaje se describen en más detalle en lo sucesivo, por ejemplo con referencia a la figura 7.

Una vez generado el clasificador en la operación 315, se puede usar, en la operación 320, para identificar personas o hablantes. En general, una aplicación de la operación 320 proporciona entrada, tal como audio y vídeo, al clasificador, que usa la entrada para determinar la probabilidad de que una persona o un hablante esté presente. En algunas implementaciones, una o más tramas de vídeo pueden ser proporcionadas como entrada y se pueden subdividir lógicamente en regiones de varias dimensiones y a continuación el clasificador puede ser evaluado en cada una de las regiones subdivididas. Tal como se usa en el presente documento, cada región subdividida puede ser conocida como una “ventana de detección”. Para cada ventana de detección, un detector puede evaluar algún número de características en el clasificador, determinando finalmente con algún nivel de confianza, si la región particular contiene una persona o un hablante. En algunas implementaciones, después de ser evaluada la ventana de detección para personas o hablantes, las regiones más prometedoras –en algunos casos, más probables- pueden ser identificadas y producidas como regiones que contienen una persona o hablante. Las regiones más probables pueden ser identificadas, en parte, eligiendo regiones que tienen un número relativamente grande de ventanas positivas de detección. Algunos detalles específicos aplicables a algunas implementaciones relativas al uso de un clasificador para identificar personas o hablantes, incluyendo las ventanas de detección, se describen más en detalle en lo sucesivo, por ejemplo con referencia a la figura 8 y la figura 9.

Es importante destacar que las operaciones ilustradas con referencia la figura 3 se puede aplicar o ejecutar en varios dispositivos o plataformas diferentes informáticas, incluyendo el uso de múltiples dispositivo informático en la misma aplicación. Por ejemplo, la operación 310 de identificación de características y la operación 315 de generación de clasificador se pueden ejecutar en asociación con uno o más dispositivos de ordenador personal, mientras que el clasificador evaluado para la operación de detección 320 se puede ejecutar sobre un dispositivo separado del dispositivo o dispositivos asociados con, por ejemplo, la operación de generación de clasificador. Esto incluye, en al menos una implementación ejemplar, un dispositivo similar al dispositivo detector 165 ilustrado en la figura 1. Es importante entender que algunas operaciones se puede llevar a cabo más o menos veces que otras operaciones. Por ejemplo, en algunas implementaciones puede ser habitual que la operación de identificación de características 310 y la operación de generación de clasificador 315 sean ejecutadas algún número de veces, hasta que se encuentra un clasificador apropiado. A continuación el código ejecutable que aplica la detección usando ese clasificador, como se ejemplifica mediante la operación 320, se puede aplicar usando algún otro dispositivo –incluyendo, por ejemplo, un dispositivo de cámara apropiado para su uso en una sala de conferencia- y a continuación se ejecuta repetidamente para detectar realmente personas o hablantes. En otras implementaciones, la operación de generación de clasificador 315 y la operación de evaluación de clasificador para detección 320 se pueden aplicar ambas en el mismo dispositivo. En tales implementaciones o en otras implementaciones, la operación de generación de clasificador se puede ejecutar para cada nuevo espacio o región donde se utiliza el dispositivo, y puede producir distintos clasificadores para cada nuevo espacio o región.

Volviendo ahora a la figura 4 se ilustra en la misma un diagrama ejemplar generalizado que muestra algunas características ejemplares que pueden ser identificadas y usadas en algunas implementaciones. Esta descripción de la figura 4 se realiza con referencia a la figura 5 y la figura 6. Sin embargo, cabe entender que los elementos descritos con referencia a la figura 4 no están destinados a limitarse a su uso con los elementos descritos con referencia a las otras figuras. Asimismo, aunque el diagrama ejemplar de la figura 4 indica elementos particulares, en algunas implementaciones no todos estos elementos existen, y en algunas implementaciones pueden existir elementos adicionales.

El diagrama ejemplar 400 incluye una serie de características 100 que pueden contener características de audio 420 y características de vídeo 430 y otras características 450.

En general, una característica de audio es una característica asociada a algún tipo de entrada de audio. Las características de audio se pueden crear para reflejar cualquier número de una variedad de parámetros de audio, incluyendo la amplitud de una señal de audio, frecuencia de una señal de audio, etc.

En un entorno en el cual los datos de audio incluyen una función de probabilidad SSL, las características de audio pueden usar algún conjunto de información asociado a la función de probabilidad SSL. En algunas implementaciones, un conjunto de características de audio basadas en una función de probabilidad SSL puede usar valores de la función de probabilidad SSL asociada a cada ventana de detección, junto con valores globales para toda la función de probabilidad SSL. Se puede usar estos valores en puntos discretos en el momento, por ejemplo, el momento actual (el tiempo en el que los datos SSL son la información disponible más reciente) o en cualquier momento en, por ejemplo, el último minuto

–así como agregado a algún periodo de tiempo.

Por ejemplo, supongamos que el máximo global de la función de probabilidad SSL, su mínimo global y su promedio global se calculan como sigue: el máximo global (L^g_{max}) es el valor máximo de la función de probabilidad SSL en toda la función de probabilidad SSL; el mínimo global (L^g_{min}) es el valor mínimo de la función de probabilidad SSL en toda la función de probabilidad SSL; y el promedio global (L^g_{avg}) es el valor medio de la función de probabilidad SSL en toda la función de probabilidad SSL.

Supongamos también que, para cada ventana de detección, se calculan algunos valores locales, usando la región de la función de probabilidad SSL que corresponde a la ventana de detección particular (que puede requerir la conversión del espacio de coordenadas usado por la imagen y/o ventana de detección al espacio de coordenadas- posiblemente en grados- usado por la función de probabilidad SSL): el máximo local (L^l_{max}) es el valor máximo de la función de probabilidad SSL en la ventana de detección; el mínimo local (L^l_{min}) es el valor mínimo de la función de probabilidad SSL en la ventana de detección, el promedio local (L^l_{avg}) es el valor medio de la función de probabilidad SSL en la ventana de detección; y la salida media local ($L^l mid$) es el valor de la función de probabilidad SSL en el punto medio de la ventana de detección –por ejemplo, si la ventana de detección comprende los grados de 10º a 20º, la salida media local se puede calcular como el valor de la función de probabilidad SSL en el grado 15. Supongamos también la existencia de un valor máximo de "reposo" (L^{rest}_{max}), que es el valor máximo de la función de probabilidad SSL fuera de la ventana particular de detección.

Dado estos valores, se puede ocupar parte de la serie de características 410 añadiendo características de audio 420 definidas al menos en parte por funciones similares a las de la siguiente lista:

1. $$\frac{L^l_{\max} - L^g_{\min}}{L^g_{\max} - L^g_{\min}}$$

2. $$\frac{L^l_{\min} - L^g_{\min}}{L^g_{\max} - L^g_{\min}}$$

3. $$\frac{L^l_{avg} - L^g_{\min}}{L^g_{\max} - L^g_{\min}}$$

4. $$\frac{L^l_{mid} - L^g_{\min}}{L^g_{\max} - L^g_{\min}}$$

5. $$\frac{L^l_{\max}}{L^l_{\min}}$$

6. $$\frac{L^l_{\max}}{L^l_{avg}}$$

7. $$\frac{L^l_{\min}}{L^l_{avg}}$$

8. $\frac{L^l_{mid}}{L^l_{avg}}$

9. $\frac{L^l_{max} - L^l_{min}}{L^l_{avg}}$

10. $\frac{L^l_{max}}{L^g_{max}}$

11. $\frac{L^l_{min}}{L^g_{max}}$

12. $\frac{L^l_{avg}}{L^g_{max}}$

13. $\frac{L^l_{mid}}{L^g_{max}}$

14. $\frac{L^l_{max} - L^l_{min}}{L^g_{max}}$

15. $L^g_{max} - L^l_{max} < \varepsilon$

(una característica binaria que prueba si la ventana de detección contiene el pico global de la función de probabilidad SSL)

16. $\frac{L^l_{max}}{L^{resl}_{max}}$

Otros medios para crear características de audio pueden usar los datos de la función de probabilidad SSL de una manera similar a la explicada anteriormente, pero pueden usar datos de la función de uno o más periodos de tiempo "anteriores" en lugar de solo el periodo de tiempo "actual". Por ejemplo, además de la creación de un conjunto de características de audio definidas en parte por las funciones indicadas anteriormente donde los datos usados por las funciones son los datos más recientes producidos por la función de probabilidad SSL, características adicionales se pueden crear donde los datos usados por la función son de uno o más periodos de tiempo anteriores. Por ejemplo, el valor máximo global (L^g_{max}) puede seguir siendo el valor máximo de la función de probabilidad SSL en toda la función de probabilidad SSL, pero en un tiempo diferente, quizás, por ejemplo 1/60ª de segundo anterior –usando los segundos valores de función de probabilidad SSL más recientes. Características adicionales similares pueden ser creadas para un número arbitrario de periodos de tiempo anteriores. Por ejemplo, en un entorno que proporciona una nueva función de probabilidad SSL cara 1/60 de segundo, se pueden crear características que usan las 60 funciones de probabilidad SSL

inmediatamente anteriores –si se crea una característica para cada una de las dieciséis (16) funciones indicadas anteriormente, esto puede dar como resultado novecientos sesenta (960) características de audio SSL.

Asimismo para usar el valor de la función de probabilidad SSL en puntos discretos en el tiempo, también se pueden crear algunas características que usan algún valor agregado derivado de múltiples funciones de probabilidad SSL anteriores. Por ejemplo, en algunas características, el valor máximo global (L^g_{max}) puede definirse como el valor máximo absoluto de la función de probabilidad SSL que se produjo en, por ejemplo, el segundo valor anterior junto antes del valor máximo de la función de probabilidad SSL proporcionado por el caso más reciente de la función de probabilidad SSL. Asimismo, por ejemplo el promedio global (L^g_{avg}) se puede definir como el valor medio de la función de probabilidad SSL en toda la función de probabilidad SSL para algún periodo de tiempo anterior.

Además de usar datos para funciones de probabilidad SSL anteriores, puede también ser posible usar datos de funciones de probabilidad SSL "futuras", si la mayor latencia causada esperando estos datos futuros es aceptable.

Cualquiera o todas estas características adicionales podrían entonces añadirse a la misma serie de características 410 y usarse en el procedimiento de generación de clasificador. Se puede también incluir otras características basándose al menos en parte en la función de probabilidad SSL, o evidentemente otras características basándose en otros datos de audio, o datos de audio combinados con otros datos de otras entradas.

Otro conjunto de características puede formar parte de la serie de características 410 son características de vídeo 420. En general, una característica de vídeo puede ser cualquier característica asociada a algún tipo de entrada de vídeo. Una característica de vídeo puede, por ejemplo, realizar algún tipo de operación matemática sobre algunos o todos los píxeles de una imagen, incluyendo la imagen de base así como otras imágenes, quizás similares a las imágenes calculadas de diferencia de corto plazo y de promedio de largo plazo. Algunos detalles específicos aplicables a algunas implementaciones relativas a la definición de imágenes de vídeo se describen más en detalle en lo sucesivo, por ejemplo con referencia a la figura 5 y la figura 6.

Igualmente incluidas en la serie de características 410 pueden estar otras características 450. Tales otras características 450 comprenden cualesquiera características adicionales identificadas como útiles para su estudio cuando se genera un clasificador. En algunas implementaciones, en entornos en los cuales hay dos tipos de entrada, características asociadas a otros tipos de entrada pueden formar parte de las otras características. Por ejemplo, en un entorno que incluye una entrada de una cámara tridimensional, tal como la medida de la distancio o la profundidad respecto de elementos en una imagen, las otras características pueden incluir características que cuantifican estos datos adicionales, bien en forma aisladas de otras entradas, o quizás en combinación con otras entradas juntas –por ejemplo algunas características pueden usar tanto entrada de audio como entrada de video juntas, en la misma o mismas características.

En implementaciones donde las entradas proporcionan una visión de 360º, al menos algunas característica pueden ser aplicadas para que se "envuelvan" –es decir, para que algunas características tengan en cuenta la entrada de, por ejemplo, tanto el "inicio" y el "final" de los datos proporcionado por entradas particulares. Por ejemplo, en un entorno que incluye una entrada de audio que proporciona una visión de 360º, al menos algunas características puede incorporar una entrada de, por ejemplo un acimut de 355º, a por ejemplo un acimut de 5º. Tales características pueden en algunos casos capturar personas o hablantes que se suponen que están situados en el límite entre el "inicio" y el "final" de los datos proporcionados por las entradas.

Volviendo ahora a la figura 5, se muestran en la misma algunas características de vídeo ejemplares. Esta descripción de la figura 5 se realiza con referencia la figura 6 que menciona algunas maneras de usar las características de vídeo. Sin embargo, cabe entender que los elementos descritos con referencia a las figuras 5 no están destinados a limitarse a su uso con los elementos descritos con referencia a esta otra figura. Asimismo, aunque el diagrama ejemplar en la figura 5 indica elementos particulares, en algunas implementaciones no existen todos estos elementos, y en algunas implementaciones pueden existir elementos adicionales.

Aunque las características de vídeo pueden comprender cualquier entidad que es capaz de cuantificar algún elemento de entrada de vídeo en un tiempo particular, un tipo útil de características de vídeo es el constituido en parte por uno o más rectángulos. En general, se suman los valores asociados a los píxeles en uno o más rectángulos o se manipulan matemáticamente para determinar un valor numérico asociados con una característica particular de vídeo rectangular. Por ejemplo, en una imagen en blanco y negro donde cada pixel está activado o desactivado (es decir, un uno (1) o cero (0) binarios), el valor numérico asociado a una característica de vídeo puede ser, por ejemplo, la suma de los píxeles que están activos, o tienen el valor uno (1), en el rectángulo particular. En la figura 5, el rectángulo 550 y el rectángulo 560 ilustran gráficamente dos posibles características de vídeo de rectángulo único. En las imágenes de escala de grises o en color se puede manipular de manera similar el valor numérico asociado a píxeles específicos. Por ejemplo, en una imagen de escala de grises donde un valor numérico asociado a cada pixel varía entre cero (0) y dos cientos cincuentaicinco (255), se puede asociar una característica con la suma de los valores de escala de grises para los píxeles en un rectángulo. Obsérvese que aunque los rectángulos se ilustran y menciona en el presente documento, la

región o regiones asociadas con una característica de vídeo pueden tener cualquier forma, y no se limitan a los rectángulos.

Otro tipo de característica de vídeo puede usar dos o más subrectángulos en el interior de un rectángulo matriz. El rectángulo 510, rectángulo 520, rectángulo 530 y rectángulo 540 son ejemplos gráficos de características de vídeo que usa subrectángulos. En tal característica de vídeo, el valor numérico asociado a la característica se puede calcular por ejemplo, sumando los valores de los píxeles en ambos subrectángulos y a continuación sustrayendo una de las sumas resultantes de la otra suma. En tal implementación, dependiendo de la localización y la orientación de los subrectángulos, el valor numérico resultante puede ser diferente, incluso cuando las características se aplican a la misma sección de la imagen. Por ejemplo, los subrectángulos en el rectángulo matriz 510 están orientados horizontalmente mientras que los subrectángulos del rectángulo matriz 530 están orientados verticalmente, y de este modo el valor numérico resultante asociado a las características de vídeo que usan estos rectángulos puede ser diferente, incluso cuando los rectángulo se aplican a la misma parte de una imagen. En algunos casos este tipo de característica puede favorecer la identificación de regiones de contraste relativamente alto –incluyendo el contraste que puede existir, por ejemplo, entre los ojos en una cara (generalmente oscuros) y la piel circundante (generalmente menos oscura).

Aunque la figura 5 ilustra representaciones gráfica de características de vídeo que incluyen dos subrectángulos. Es posible asimismo definir características de vídeo que incluyen tres rectángulos, cuatro rectángulos, etc. El valor numérico asociado a tales características de vídeo se pueden calcular de varias maneras, incluso tomando la diferencia entre los recuentos de píxeles en diferentes subrectángulos.

Volviendo ahora a la figura 6, en la misma se muestra un rectángulo de características ejemplar representativo 610 que incluye algunas características de vídeo ejemplares. Esta descripción de la figura 6 se realiza con referencia a la figura 1, la figura 4 y la figura 6. Sin embargo, cabe entender que los elementos descritos con referencia a la figura 6 no están destinados a limitarse a su uso con los elementos descritos con referencia a estas otras figuras. Además, aunque el diagrama ejemplar de la figura 6 indica elementos particulares, en algunas implementaciones pueden no existir todos estos elementos, y en algunas implementaciones pueden existir elementos adicionales.

Aunque las ilustraciones de la figura 5 muestran rectángulos (y en algunos casos, subrectángulos) asociados a características de vídeo ejemplares, la ilustración de la figura 5 no muestra explícitamente cómo los rectángulo y las características de vídeo correspondientes se pueden usar para generar o evaluar un clasificador. Un mecanismo para identificar las características de vídeo a incluir en una serie de características es tomar una variedad de características que están asociadas a una variedad de formas, incluyendo unas similares a las descritas anteriormente con referencia a la figura 5, y para variar la localización y la dimensión de tales formas a través de un rectángulo representativos de características 610. El rectángulo representativo de características, y la localización del rectángulo de características de vídeo en su interior, se pueden evaluar entonces en regiones particulares en una imagen en varios momentos o con varios fines, incluyendo el ser parte de un procedimiento de detección.

Dentro del rectángulo representativo de características 610, la localización y dimensión de las formas asociadas a las características de vídeo pueden variar. Por ejemplo, como se muestra, el rectángulo matriz 620 asociado a una característica de vídeo particular ocupa la esquina izquierda superior del rectángulo representativo de características. Además de la localización y dimensión particulares ilustradas por el rectángulo matriz 620, el rectángulo matriz (y sus subrectángulos) puede desplazarse tanto en horizontal como en vertical dentro del rectángulo representativo de características, definiendo cada vez una nueva característica de vídeo. En algunas implementaciones, la localización del rectángulo matriz se puede cambiar una serie de veces para de este modo asegurar el rectángulo de características representativo se ha cubierto. En algunas u otras implementaciones, cuando la localización del rectángulo matriz cambia, la nueva localización puede solapar el rectángulo matriz de la característica de vídeo definida anteriormente o solapar los rectángulos matriz de características de vídeo ya definidas o que están por definir.

Asimismo, la dimensión del rectángulo matriz se puede modificar también para definir nueva características de vídeo. Por ejemplo, el rectángulo matriz 630, el rectángulo matriz 640, y el rectángulo matriz 650 muestran el uso de diferentes dimensiones, cuando se comparan con el rectángulo matriz 620. En algunos casos, es posible que un rectángulo matriz pueda ser incrementado hasta que ocupe todo el rectángulo representativo de características.

En algunas implementaciones, puede ser útil representar la existencia de una característica de vídeo con un rectángulo matriz particular usando simetría bilateral. Es decir, cuando existe una característica de vídeo con un rectángulo matriz en una localización particular, puede ser útil definir otra característica de vídeo con un rectángulo matriz que es una imagen reflejo exacto del rectángulo matriz de la primera característica de vídeo. Un caso ejemplar en el que esto ocurre es el ilustrado con el rectángulo matriz 630 y el rectángulo matriz 640.

En algunas implementaciones se pueden generar múltiples características de vídeo que se aplica a diferentes imágenes, incluyendo los tipos de imágenes descritas anteriormente con referencia a la figura 1. Por ejemplo, algunas características de vídeo pueden ser generadas para aplicarse a la imagen de base mientras que otras características de vídeo se aplican a la imagen de diferencia de corto plazo y otras más a la imagen promedio de de largo plazo.

Después de varias varios factores que incluyen la localización de los rectángulos asociados a las características de vídeo, la dimensión de los rectángulos asociados a las características de vídeo, y las imágenes a las cuales se aplican las características de vídeo, y generar distintas características de vídeo para cualquier combinación o todas las combinaciones de estos factores, no es infrecuente tener literalmente cientos de características de vídeo que pueden formar parte de la serie de características como la serie de características 410 descritas anteriormente con referencia a la figura 4. En algunas implementaciones, alguno conjunto de características de vídeo puede ser seleccionado entre este gran número de características de vídeo durante el procedimiento de generación de una clasificador.

Es importante resaltar de nuevo que la figura 5 y la figura 6 no muestran todas las posibles características de vídeo. En muchas implementaciones las características de vídeo cubrirán, en conjunto, una imagen o imágenes enteras. Las características de vídeo mostradas en la figura 5 y la figura 6 sirven solo para demostrar cómo se pueden definir algunas características de vídeo.

Volviendo ahora a la figura 7, en la misma se muestra un diagrama ejemplar generalizado que muestra un sistema 700 en el cual se puede llevar a cabo la generación de un clasificador para la detección de personas o hablantes. Esta descripción de la figura 7 se realiza con referencia a la figura 3, la figura 4 y la figura 10. Sin embargo, cabe entender que los elementos descritos con referencia a la figura 7 no están destinados a limitarse a su uso con los elementos descritos con referencia a estas otras figuras. Además, aunque el diagrama ejemplar de la figura 7 indica elementos particulares, en algunas implementaciones pueden no existir todos estos elementos, y en algunas implementaciones pueden existir elementos adicionales.

El sistema 700 puede incluir una serie de características 710, datos de entrada de formación 720, etiquetas para datos de entrada 730, un módulo de formación 740 que está asociado a un algoritmo de aprendizaje 745, y un clasificador 755.

Como se ha presentado anteriormente en el flujo operativo descrito con referencia a la figura 3, dada una serie de características, tal como la serie de características 710, puede ser posible generar un clasificador que se pueda usar para aplicar la detección de personas o hablantes. El sistema ilustrado en la figura 7 demuestra algunos mecanismos mediante los cuales se puede generar tal clasificador.

Un módulo ejemplar de formación 740 puede usar entradas particulares para generar un clasificador, tal como el clasificador 755. El módulo de formación se puede aplicar en uno o más dispositivos informáticos, incluyendo el dispositivo ejemplar informático descrito en lo sucesivo con referencia a la figura 10.

En general, el módulo de formación puede estar asociado a alguna forma de algoritmo de aprendizaje. El algoritmo de aprendizaje comprende un procedimiento automatizado que produce un clasificador. Algunos algoritmos de aprendizaje producen un clasificador aceptando una serie de características 710, datos de entrada de formación 720, y etiquetas para datos de entrada 730. La serie de características 710 puede ser un conjunto de entidades que pueden cuantificar algún elemento o elementos de datos de entrada, incluyendo los datos de entrada de formación 720. En algunas implementaciones, la serie de características puede incluir características similares a las mencionadas anteriormente con referencia a la figura 4 y otras figuras relevantes. Los datos de entrada de formación 720 pueden consistir en general en datos de entrada similares a los datos de entrada que se darán al clasificador, una vez generado. En algunas implementaciones, los datos de entrada de formación pueden incluir un conjunto de tramas de vídeo, a partir de las cuales –como la imagen de base, la imagen de diferencia de corto plazo, y la imagen promedio de largo plazo- se puede recuperar o calcular, así como información de audio a partir de la cual se puede generar una función de probabilidad SSL. Las etiquetas para datos de entrada 730 pueden en general consistir en la respuesta “correcta” que un clasificador ideal debería producir cuando se dan los datos de entrada de formación. Por ejemplo, para cada trama de vídeo y cada conjunto de entradas de audio, las etiquetas para datos de entrada pueden identificar regiones particulares dentro de la trama de vídeo donde existen personas o hablantes.

Dada la serie de características 710, los datos de entrada de formación 720, y las etiquetas de entrada 730, el módulo de formación 740 puede usar su algoritmo de aprendizaje asociado 745 para generar un clasificador. La operación del algoritmo de aprendizaje que varía dependiendo del algoritmo de aprendizaje asociado usado, es en general conocida en la técnica y no necesita ser explicada muy en mayor detalle en esta solicitud. Por ejemplo, si el algoritmo de aprendizaje es una forma del algoritmo de Adaboost, la operación del algoritmo de aprendizaje puede incluir la selección de una serie de características tal que la precisión del clasificador resultante mejora a medida que se ejecuta el algoritmo Adaboost. Si el algoritmo de aprendizaje es distinto del algoritmo Adaboost, como por ejemplo una red neuronal, la operación del algoritmo de aprendizaje puede ser diferente.

La salida final del módulo de aprendizaje 740 y el algoritmo de aprendizaje 745 puede incluir un clasificador que, cuando se evalúa en una región particular o ventana de detección, devuelve alguna estimación de la probabilidad de que la región particular incluya una persona o hablante. El propio clasificador puede en general estar constituido por un subconjunto de características que han sido seleccionadas por el módulo de formación. El conjunto de las características seleccionadas se lleva a cabo en general de una manera algo más precisa que las características que no han sido seleccionadas. En algunos casos los elementos del clasificador, incluyendo el subconjunto de características

denominadas “nodos”, donde por ejemplo, cada característica seleccionada está asociada a un único nodo del clasificador.

Diferentes características en el clasificador 755 pueden requerir diferentes cantidades de tiempo de cálculo para evaluar o calcular durante la detección. Por ejemplo, algunas características –como al menos las características de audio en algunas implementaciones- pueden ser capaces de ser evaluadas o calculadas más rápidamente que otras características- como al menos algunas características de vídeo en algunas implementaciones. Debido a las diferencias en la velocidad de evaluación, puede ser útil en algunas implementaciones ordenar características particulares en el clasificador general para que de este modo una característica que requiere menos tiempo de evaluación se ordene antes que una característica que requiere más tiempo de evaluación.

Alguna de las características seleccionadas en el clasificador 755 puede realizar un trabajo relativamente mejor o identificación de una persona o hablante en una ventana de detección que otras características. Por ejemplo, puede ocurrir que una característica de audio o vídeo esté más correlacionada con la detección de una persona o hablante que otra característica de audio o video. En algunas implementaciones puede ser útil ordenas las características del clasificador para que de este modo una característica que está más correlacionada con la detección de personas se lleve a cabo antes que una característica menos precisa.

Independientemente de si se refiere a la velocidad de evaluación, el grado de precisión, o alguna otra propiedad, se pueden ordenar características particulares antes de otras características usando una variedad de mecanismos. En algunas realizaciones, el propio algoritmo de aprendizaje puede tomar en cuenta atributos deseable o preferibles –incluyendo la velocidad de evaluación y el grado de precisión- cuando se genera el clasificador, quizás ponderando de manera más importante las características particulares o preferibles que otras características, lo cual puede dar como resultado las características particulares que tienden a producirse antes en el clasificador generado. En la misma u otras implementaciones, las características en el clasificador generado se pueden reordenar o seleccionar después de que el algoritmo de aprendizaje haya generado un clasificador.

En general cuanto más se usan los datos de entrada de formación 720 para generar el clasificador 755, más preciso será el clasificador resultante. Sin embargo, la producción de datos de entrada de formación requiere tiempo y esfuerzo –por ejemplo, entre otras cosas, las respuestas “correctas”, en forma de etiquetas para datos de entrada 730, pueden necesitar ser generadas para cada trama de vídeo. Un procedimiento para aumentar la cantidad de datos de entrada de formación que puede requerir relativamente menos trabajo que la producción de datos de entrada de formación totalmente nuevos es crear imágenes reflejos exactos de los datos de entrada de formación preexistentes y las etiquetas para los datos de entrada. Por ejemplo, dada una trama de vídeo y una función de probabilidad SSL, se puede crear una nueva trama de vídeo que es la imagen reflejo exacto de la trama de vídeo original y también reflejan la función de probabilidad SSL y la etiquetas para los datos de entrada.

En al menos algunas implementaciones, se pueden seleccionar algunas características, al menos en parte, para que de este modo los “falsos positivos” son, en muchos casos, asociados con otras personas y no con un objeto o entidad que no es una persona. Es decir, en los casos donde la persona o hablante deseado no es detectado. Las características pueden ser seleccionadas para que de este modo, en muchos casos, otra persona es detectada en lugar de algún objeto o entidad que no es una persona. Por ejemplo, las características de vídeo se pueden seleccionar para que de este modo, donde el hablante no es detectado, en muchos casos se detecta una persona que no habla.

Volviendo ahora a la figura 8, se muestra en la misma un diagrama ejemplar generalizada que muestra un sistema 800 en el que se puede llevar a cabo la detección de persona o hablantes. La descripción de la figura 8 se realiza con referencia a la figura 1, figura 3, figura 7, figura 9 y figura 10. Sin embargo, cabe entender que los elementos descritos con referencia a la figura 8 no están destinados a limitarse a su uso con los elementos descritos con referencia a las otras figuras. Además, aunque el diagrama ejemplar de la figura 8 indica elementos particulares, en algunas implementaciones pueden no existir todos estos elementos, y en algunas implementaciones pueden existir elementos adicionales.

El sistema 800 puede incluir datos de entrada 810, un módulo detector 840 asociado a un clasificador 855, y resultados de detección 865.

Como se ha presentado anteriormente en el flujo operativo descrito con referencia a la figura 3, dado un clasificador 855, que incluye uno como el clasificador generado 755 de la figura 7, un detector, quizás aplicado en un módulo detector 840, puede examinar datos de entrada y usar el clasificador para producir resultados de detección 865. El sistema ilustrado en la figura 8 demuestra algunos mecanismos mediante los cuales tal clasificador se puede usar para detectar personas o hablantes. El módulo detector puede ser aplicado en uno o más dispositivos informáticos, incluyendo el dispositivo detector 165 descrito anteriormente con referencia a la figura 1, y el dispositivo ejemplar informático descrito en lo sucesivo con referencia a la figura 10.

Los datos de entrada 810 pueden incluir una gran variedad de datos de entrada. En algunas implementaciones los datos

de entrada pueden incluir datos de entrada similares a los descritos anteriormente, por ejemplo con referencia a la figura 1, que incluyen una serie de tramas de vídeo a partir de las cuales se puede determinar una serie de imágenes de base, imágenes de diferencia de corto plazo, e imágenes promedio de largo plazo. Los datos de entrada también incluyen datos de audio similares a una serie de funciones de probabilidad SSL que están asociado a una o más tramas de vídeo. Los datos de entrada pueden incluir también otros tipos de datos, incluyendo los descritos anteriormente, por ejemplo con referencia a la figura 1.

El módulo detector 840 puede usar entonces el clasificador 855 para determinar regiones de los datos de vídeo de entrada que pueden incluir una persona o un hablante. En algunas implementaciones esto se puede conseguir subdividiendo al menos parte de los datos de entrada en una serie de regiones menores, denominadas ventanas. Las ventanas de detección pueden definirse de varias maneras, incluyendo algunos procedimientos mencionados anteriormente más en detalle con referencia a la figura 9.

Para cada ventana de detección, el módulo detector 840 puede evaluar el clasificador 855 respecto de los datos de entrada para esa ventana de detección. La evaluación del clasificador puede producir en general alguna estimación de la probabilidad de que exista una persona o hablante en la ventana particular de detección. Esta estimación de probabilidad, puede, al menos en algunas implementaciones, formar parte de los resultados de detección 865.

Una vez que algunas, o todas las ventanas de detección han sido evaluadas, en algunas implementaciones de fusión puede llevarse a cabo para determinar regiones particulares de los datos de entrada que de manera especial contienen probablemente personas o hablantes. Esto se puede conseguir en algunas implementaciones eligiendo regiones que tienen un número relativamente grande de ventanas de detección que a su vez tienen una alta probabilidad de contener personas o hablantes. Estas regiones identificadas pueden también al menos en algunas implementaciones, formar parte de los resultados de detección 865.

En algunas implementaciones, todos los elementos, o nodos, de un clasificador pueden ser evaluados antes de que se determine la probabilidad de que la ventana particular de detección contenga una persona o hablante. En algunas implementaciones puede ser posible acortar el tiempo requerido para evaluar el clasificador para algunas ventanas de detección usando una técnica denominada “recorte”.

Cuando se usa el recorte, la evaluación del clasificador se puede parar antes de que todos los nodos del clasificador hayan sido evaluados. La evaluación del clasificador se puede detener, por ejemplo, si se puede determinar que los resultados ya calculados proporcionan algún nivel de certeza de que una ventana particular de detección contiene o no una persona o hablante. Por ejemplo, se puede saber si, en su caso, los cuatro primeros nodos del clasificador son evaluados para resultados particulares, la ventana de detección siempre contiene una persona (al menos para los datos usados para capacitar el clasificador). En este caso, durante el procedimiento de detección, la evaluación del clasificador se puede parar antes de que todos los nodos hayan sido evaluados, y la ventana de detección se puede determinar para contener una persona o hablante.

En algunas implementaciones, se pueden ejecutar subregiones particulares de los datos de entrada a partir de las regiones consideradas para la detección de personas o hablantes. Por ejemplo, una sala puede tener una pantalla de televisión o de proyección que puede en algunos casos visualizar personas o hablantes que no debería ser identificada como personas o hablantes por el detector. En este caso ejemplar, se puede ejecutar una subregión de los datos de entrada asociados con la pantalla de televisión o de proyección a partir de las regiones consideradas para la detección de personas o hablantes. Esto se puede conseguir de varias maneras incluyendo por ejemplo, por la definición de ventanas de detección que comprende las subregiones que se han de excluir.

Volviendo ahora a la figura 9, en la misma se muestran algunas representaciones de ventanas de detección que se pueden usar como parte del procedimiento de detección de personas o hablantes. Esta descripción de la figura 9 se realiza con referencia a la figura 6 y se relaciona con la mención proporcionada en la figura 8. Sin embargo, cabe entender que los elementos descritos con referencia a la figura 9 no están destinados a limitarse a su uso con los elementos descritos con referencia a las otras figuras. Además, aunque el diagrama ejemplar de la figura 9 indica elementos particulares, en algunas implementaciones no todos los elementos pueden existir, y en algunas implementaciones pueden existir elementos adicionales.

En algunas implementaciones datos de entrada, como una trama de vídeo, o una imagen o imágenes derivadas de una trama de vídeo, se pueden subdividir en múltiples ventanas de detección que se usan como parte del procedimiento para detectar personas o hablantes.

Como se muestra en la figura 9, una imagen ejemplar 905 puede contener múltiples ventanas ejemplares de detección, incluyendo la ventana de detección 910, la ventana de detección 970, la ventana de detección 930, la ventana de detección 940 y la ventana de detección 950. Cada ventana de detección ocupa alguna porción de la imagen. Es importante resaltar que no se muestran todas las ventanas de detección que pueden existir en la imagen 905. En muchas implementaciones las ventanas de detección cubrirán, en conjunto, la imagen entera. Las ventanas de detección

mostradas en la figura 9 sirven solo para demostrar como las ventanas de detección se pueden definir. Asimismo, aunque las ventanas de detección se muestran como rectángulos, las ventanas de detección se pueden definir en cualquier forma. Igualmente, aunque las ventanas de detección se describen con referencia a una "imagen", las ventanas de detección se pueden aplicar también a entradas no visuales, incluyendo entrada de audio, como se ha descrito anteriormente. Por ejemplo, una ventana de detección para una función de probabilidad SSL asociada a entrada de audio puede incluir algún subconjunto de la función de probabilidad SSL.

La ventana de detección ejemplar 910 ocupa la esquina superior derecha de la imagen 905. La ventana ejemplar de detección 920 y la ventana ejemplar de detección 930 muestran una manera en la que las ventanas de detección se pueden extender para cubrir más regiones de la imagen. Aunque no se muestran, se pueden definir ventanas de detección que continua en la dirección representada por la flecha 960. Tales ventanas de detección puede cubrir toda la porción superior de la imagen.

Igualmente, la ventana ejemplar de detección 940 muestra cómo se pueden extender las ventanas de detección verticalmente para cubrir regiones adicionales de la imagen. La flecha 870 ilustra una dirección que tales ventanas de detección pueden seguir para cubrir toda la parte izquierda de la imagen.

Extendiendo la ventana ejemplar de detección 940 a la derecha, de manera que haya ventanas de detección por debajo de la ventana ejemplar de detección 920, la ventana ejemplar de detección 930, y en la dirección mostrada por la flecha 960, ilustra una manera de que las ventanas de detección se pueden definir para que cubran toda la imagen 905.

Las ventanas de detección pueden solaparse en cualquier medida. Por ejemplo, como se muestra, la mitad de la ventana de detección 920 solapa la ventana de detección 910. Además del solapamiento mostrado, en las imágenes panorámicas que representan una vista de 360º, las ventanas de detección también pueden solaparse fuera de los extremos de la imagen 905. Por ejemplo, una ventana de detección, no mostrada, puede ocupar el lado muy a la derecha de la imagen y el lado muy a la izquierda de la imagen.

En algunas implementaciones, se pueden usar ventanas de detección de varias dimensiones. Por ejemplo, la ventana de detección 950 es mayor que la ventana de detección 910. En algunas implementaciones se pueden usar ventanas de detección de dimensiones muy diferentes. Por ejemplo, en una implementación, se pueden usar ventanas de detección de 10 dimensiones diferentes. Cada conjunto de ventanas de detección de dimensiones iguales se puede extender para cubrir toda la imagen 905, usando por ejemplo, el mismo procedimiento que el explicado anteriormente con referencia a la ventana de detección 910 y otras ventanas de detección que son de la misma dimensión que la ventana de detección 910.

Algunas características de vídeo pueden usar un rectángulo de características representativo, como se ha mencionado anteriormente con referencia a la figura 6. Durante el procedimiento, el rectángulo de características representativo se puede adaptar para encajar en la ventana de dirección, y cualquier característica de vídeo asociada al rectángulo de características representativo se puede adaptar proporcionalmente al mismo tiempo. Por ejemplo, supongamos un rectángulo de características representativo ejemplar de una dimensión de 50 píxeles de ancho por 50 píxeles de alto que contiene, entre muchas características de vídeo, una característica de vídeo con un rectángulo con una dimensión de 10 píxeles de ancho por 20 píxeles de alto. Si se usa este rectángulo de características representativo con una ventana de detección de la misma dimensión, el rectángulo de características de vídeo puede también ser de la misma dimensión. Si se usa el rectángulo de características representativo con una ventana de detección que es cuatro veces más grande –por ejemplo, con una ventana de detección con una dimensión de 100 píxeles de ancho por 100 píxeles de alto-entonces el rectángulo de características representativo y su rectángulo de características de vídeo asociado también pueden adaptarse para encajar en la ventana de detección mayor. En este ejemplo, el rectángulo de características de video se puede adaptar a una dimensión de 20 píxeles de ancho por 40 píxeles de alto.

En algunas implementaciones, la dimensión y/o la orientación del espacio o el sitio donde se usa un detector puede influir en la dimensión de las ventanas de detección usadas en el procedimiento de detección. Por ejemplo, en un espacio pequeño, las características físicas asociadas a las personas o hablantes –como las caras o los torsos) pueden tienden a aumentar cuando se ven desde la perspectiva de uno o más dispositivos de entrada, o pueden tender a variar en tamaño en menor medida que las características físicas asociadas a las personas o hablantes en un espacio grande. Esto puede ocurrir debido a que, en un espacio pequeño, las personas o hablantes solo pueden estar a una distancia relativamente corta del o los dispositivos de entrada, dando quizás como resultado caras más grandes, por ejemplo – mientras que en un espacio mayor las personas o hablantes pueden estar cerca de o lejos del o los dispositivos de entrada, y de este modo la dimensión de las características físicas asociadas pueden variar en gran medida. En consecuencia, en algunas implementaciones, las ventanas de detección se pueden usar para que, en espacios pequeños por ejemplo, las ventanas de detección se puedan limitar a dimensiones mayores, y quizás puedan variar en dimensión en cantidades relativamente pequeñas. Por el contrario, en espacios mayores, las ventanas de detección pueden variar desde cantidades pequeña a cantidades grandes para intentar capturar una variación más amplia en las dimensiones de las características físicas.

Ejemplo de Entorno informático

Volviendo ahora a la figura 10, esta figura y la discusión relacionada están destinadas a proporcionar una descripción breve y general de un entorno informático ejemplar en el cual las diversas tecnologías descritas en el presente documento se pueden aplicar. Aunque no es necesario, las tecnologías se describen en el presente documento, al menos en parte, en el contexto general de instrucciones ejecutables por ordenador, tal como módulos de programa que son ejecutados por un controlador, procesador, ordenador personal u otro dispositivo informático tal como el dispositivo informático 1000 ilustrado en la figura 10.

En general, los módulos de programa incluyen rutinas, programas, objetos, componentes, interfaces de usuario, estructura de datos, etc., que llevan a cabo tareas particulares información particular de visualización, o aplican tipos de datos abstractos particulares. Las operaciones realizadas por los módulos de programa se han descrito anteriormente con la ayuda de uno o más programas y diagramas de flujo operativo.

El experto en la técnica puede aplicar la descripción, diagramas de bloques y diagramas de flujo en forma de instrucciones ejecutables por ordenador, que se pueden materializar en una o más formas de medios legibles por ordenador. Como se usa en el presente documento, los medios legibles por ordenador pueden ser cualquier medio que pueda almacenar o materializar información que está codificada de manera que se pueda acceder a ella y entender mediante un ordenador. Formas típicas de medios legibles por ordenador incluyen, sin limitación, memoria tanto volátil como no volátil, dispositivos de almacenamiento de datos, incluyendo medios removibles y/o no removibles, y medios de comunicación.

Los medios de comunicación materializan información legible por ordenador en una señal modulada de datos, tal como una onda portadora u otro mecanismo de transporte, e incluye cualquier medio de suministro de información. El término “señal modulada de datos” significa una señal que tiene una característica o más de sus características establecidas o cambiada para codificar información en la señal. A título de ejemplo, y sin limitación, los medios de comunicación incluyen medios alámbricos tales como una red alámbrica o una conexión directa alámbrica, y medios inalámbricos tales como medios acústicos, RF, infrarrojos y otros medios inalámbricos.

El dispositivo informático 1000 ilustrado en la figura 10, en su configuración más básica, incluye al menos una unidad de procesamiento 1002 y la memoria 1004. En algunas implementaciones, la unidad de procesamiento 1002 puede ser una unidad de procesamiento central universal (CPU), como la que existe, por ejemplo en varios ordenadores, incluyendo ordenadores de sobremesa y portátiles. En otras implementaciones, la unidad de procesamiento puede también ser un procesador de señales digitales (DSP) que puede ser especialmente apropiado para tareas de procesamiento de señales digitales, incluyendo las realizadas, por ejemplo por un dispositivo detector como el dispositivo detector 165 descrito anteriormente con referencia a la figura 1. Dependiendo de la configuración exacta y del tipo de dispositivo informático, la memoria 1004 puede ser volátil (tal como una RAM), no volátil (tal como una ROM, memoria ultrarrápida, etc.) o alguna combinación de las dos. Su configuración más básica se ilustra en la figura 10 mediante la línea de punto 1006. Asimismo, el dispositivo informático 1000 puede también tener funcionalidad y características adicionales. Por ejemplo, el dispositivo informático 1000 también puede incluir almacenamiento adicional (removible y/o no removible) incluyendo, pero si n limitarse a, discos o cintas magnéticos u ópticos. Tal almacenamiento adicional se ilustra en la figura 10 mediante el almacenamiento removible 108 y el almacenamiento no removible 1010.

El dispositivo informático 1000 puede también contener una o más conexiones de comunicaciones 1012 que permite que el dispositivo informático 1000 comunique con otros dispositivos y servicios. Por ejemplo, el dispositivo informático puede tener una o más conexiones a otros dispositivos informáticos, incluyendo, por ejemplo, el dispositivo auxiliar 175 descrito anteriormente con referencia a la figura 1. El dispositivo informático 1000 puede también tener uno o más dispositivos de entrada 1014 tal como un dispositivo de entrada de imágenes, etc. Uno o más dispositivos de salida 1016 tal como una pantalla de visualización, unos altavoces, una impresora etc., también pueden incluirse en el dispositivo informático 1000.

El experto en la técnica apreciará que las tecnologías descritas en el presente documento, se pueden poner en práctica con dispositivos informáticos distintos del dispositivo informático 1000 ilustrado en la figura 10. Por ejemplo, y sin limitación, las tecnologías descritas en el presente documento pueden igualmente ponerse en práctica en dispositivos portátiles incluyendo teléfonos móviles, agenda personales digitales, sistemas de multiprocesadores electrónica de consumo programable o basada en microprocesador, ordenadores personales en red, miniordenadores, ordenadores centrales , y similares. Cada uno de estos dispositivos informáticos puede ser descrito, en algún nivel de de detalle, por el sistema de la figura 10, o se puede describir de manera distinta.

Las tecnologías descritas en el presente documento se pueden aplicar en entornos distribuidos informáticos donde las operaciones son realizadas por dispositivos remotos de procesamiento que están unidos a través de una red de comunicación. En un entorno informático distribuido, los módulos de programa se pueden situar tanto en dispositivos locales como en dispositivos remotos.

Mientras que lo descrito en el presente documento se ha aplicado en software, se apreciará también que las tecnología

descritas en el presente documento se puede aplicar de manera alternativa totalmente o en parte en forma de hardware, microprogramas, o varias combinaciones de software, hardware y/o microprogramas.

Aunque se han ilustrado algunas implementaciones particulares de procedimientos y sistemas en los dibujos anexos y descritos en el texto anterior, cabe entender que los procedimientos y sistemas mostrados y descritos no se limitan a las implementaciones particulares descritas, sino que son capaces de numerosas redisposiciones, modificaciones y sustituciones sin salirse del alcance H establecido y definido por las siguientes reivindicaciones.

REIVINDICACIONES

1.- Un procedimiento para detección de hablantes que comprende:

identificar (310) una serie de características (410, 470) que comprende al menos una característica (420, 430, 450) de un primer tipo de entrada y al menos una característica (420, 430, 450) de un segundo tipo de entrada donde el segundo tipo es diferente del primer tipo, en el cual el primer tipo de entrada o el segundo tipo de entrada incluye un entrada de audio (120) y en el cual se calcula una característica para cuantificar algún elemento del tipo correspondiente de entrada en un tiempo particular; y

generar (315) un clasificador (785, 855) para la detección de hablantes usando un algoritmo de aprendizaje (745), en el cual el clasificador está constituido por un subconjunto de características, siendo denominado el subconjunto de características como nodo del clasificador y seleccionado por el algoritmo de aprendizaje, asegurándose también de que los nodos que requieren menos cálculo están situados en el clasificador de tal manera que son evaluados antes que los nodos que requieren más calculo ponderando los nodos mientras se genera el clasificador.

2.- El procedimiento de la reivindicación 1, que comprende, además:

evaluar (320) el clasificador para detectar una persona (320).

3.- El procedimiento de la reivindicación 1 en el cual la serie de características incluye una característica de audio (420) asociad a una localización de fuente sonora, SSL, entrada que proporciona una función de probabilidad SSL, y en el que la característica de audio (420) es calculada con una función seleccionada a partir de las siguientes funciones:

$$\frac{L^l_{\max} - L^g_{\min}}{L^g_{\max} - L^g_{\min}}, \quad \frac{L^l_{\min} - L^g_{\min}}{L^g_{\max} - L^g_{\min}}, \quad \frac{L^l_{avg} - L^g_{\min}}{L^g_{\max} - L^g_{\min}}, \quad \frac{L^l_{mid} - L^g_{\min}}{L^g_{\max} - L^g_{\min}}, \quad \frac{L^l_{\max}}{L^l_{\min}}, \quad \frac{L^l_{\max}}{L^l_{avg}}, \quad \frac{L^l_{\min}}{L^l_{avg}}, \quad \frac{L^l_{mid}}{L^l_{avg}},$$

$$\frac{L^l_{\max} - L^l_{\min}}{L^l_{avg}}, \quad \frac{L^l_{\max}}{L^g_{\max}}, \quad \frac{L^l_{\min}}{L^g_{\max}}, \quad \frac{L^l_{avg}}{L^g_{\max}}, \quad \frac{L^l_{mid}}{L^g_{\max}}, \quad \frac{L^l_{\max} - L^l_{\min}}{L^g_{\max}}, \quad L^g_{\max} - L^l_{\max} < \varepsilon, \quad \text{y} \quad \frac{L^l_{\max}}{L^{resi}_{\max}},$$

en las que L^g_{max} es el valor máximo de la función de probabilidad SSL en toda la función de probabilidad SSL, L^g_{min} es el valor mínimo de la función de probabilidad SSL en toda la función de probabilidad SSL, y para cada ventana de detección, L^l_{max} es el valor máximo de la función de probabilidad SSL en la ventana de detección, L^l_{min} es el valor mínimo de la función de probabilidad SSL en la ventana de detección, L^l_{avg} es el valor medio de la función de probabilidad SSL en la ventana de detección, L^l_{mid} es el valor de la función de probabilidad SSL en el punto medio de la ventana de detección, y L^{rest}_{max} es el valor máximo de la función de probabilidad SSL fuera de la ventana de detección.

4.- El procedimiento de la reivindicación 1 en el cual el primer tipo de entrada o el segundo tipo de entrada incluye una entrada de vídeo (110) y la serie de características incluye una característica de vídeo (430) definida por un rectángulo.

5.- El procedimiento de la reivindicación 1 en el cual el algoritmo de aprendizaje (745) comprende el algoritmo AdaBoost.

6.-Un medio legible por ordenador que almacena instrucciones ejecutables por ordenador que, cuando se aplican por un procesador, hacen que el procesador lleve a cabo el procedimiento de una de las reivindicaciones 1 a 5.

7.- Un sistema para la detección de hablantes que comprende:

un dispositivo de entrada de vídeo (110) que produce datos de vídeo (140);

un dispositivo de entrada de audio (120) que produce datos de audio (150); y

un dispositivo detector (165) que incluye un detector (170) configurado para aceptar los datos de vídeo y los datos de audio y evaluar un clasificador para detectar una persona donde el clasificador ha sido creado:

identificando una serie de características (310) que comprende al menos una característica asociada a una primera entrada que corresponde a los datos de vídeo y al menos una característica asociada a una segunda entrada que corresponde a los datos de audio, en el cual una característica se calcula en un tiempo particular para cuantificar algún elemento del tipo correspondiente de entrada; y

generando el clasificador usando un algoritmo de aprendizaje en el cual el clasificador está constituido por un subconjunto de características de la serie de características, siendo el subconjunto de características denominado como nodos del clasificador y siendo seleccionado por el algoritmo de aprendizaje, incluyendo asegurar también que los nodos que requieren menos cálculo están situados en el clasificador de tal manera que son evaluados antes que los nodos que requieren más calculo ponderando los nodos mientras se genera el clasificador.

8.- El sistema de la reivindicación 7 que comprende, además:

un dispositivo auxiliar (175) que proporciona almacenamiento para al menos una porción de los datos de vídeo o al menos una porción de los datos de audio.

9.- El sistema de la reivindicación 7 en el que los datos de audio incluyen datos de localización de fuente sonora y la serie de características incluye una característica de audio (420) asociada a una función seleccionada a partir de las siguientes funciones:

$$\frac{L^l_{\max}-L^g_{\min}}{L^g_{\max}-L^g_{\min}},\quad \frac{L^l_{\min}-L^g_{\min}}{L^g_{\max}-L^g_{\min}},\quad \frac{L^l_{avg}-L^g_{\min}}{L^g_{\max}-L^g_{\min}},\quad \frac{L^l_{mid}-L^g_{\min}}{L^g_{\max}-L^g_{\min}},\quad \frac{L^l_{\max}}{L^l_{\min}},\quad \frac{L^l_{\max}}{L^l_{avg}},\quad \frac{L^l_{\min}}{L^l_{avg}},\quad \frac{L^l_{mid}}{L^l_{avg}},$$

$$\frac{L^l_{\max}-L^l_{\min}}{L^l_{avg}},\quad \frac{L^l_{\max}}{L^g_{\max}},\quad \frac{L^l_{\min}}{L^g_{\max}},\quad \frac{L^l_{avg}}{L^g_{\max}},\quad \frac{L^l_{mid}}{L^g_{\max}},\quad \frac{L^l_{\max}-L^l_{\min}}{L^g_{\max}},\quad L^g_{\max}-L^l_{\max}<\varepsilon,\quad \text{y}\quad \frac{L^l_{\max}}{L^{resi}_{\max}},$$

en las que L^g_{max} es el valor máximo de la función de probabilidad SSL en toda la función de probabilidad SSL, L^g_{min} es el valor mínimo de la función de probabilidad SSL en toda la función de probabilidad SSL, y para cada ventana de detección, L^l_{max} es el valor máximo de la función de probabilidad SSL en la ventana de detección, L^l_{min} es el valor mínimo de la función de probabilidad SSL en la ventana de detección, L^l_{avg} es el valor medio de la función de probabilidad SSL en la ventana de detección, L^l_{mid} es el valor de la función de probabilidad SSL en el punto medio de la ventana de detección, y L^{rest}_{max} es el valor máximo de la función de probabilidad SSL fuera de la ventana de detección.

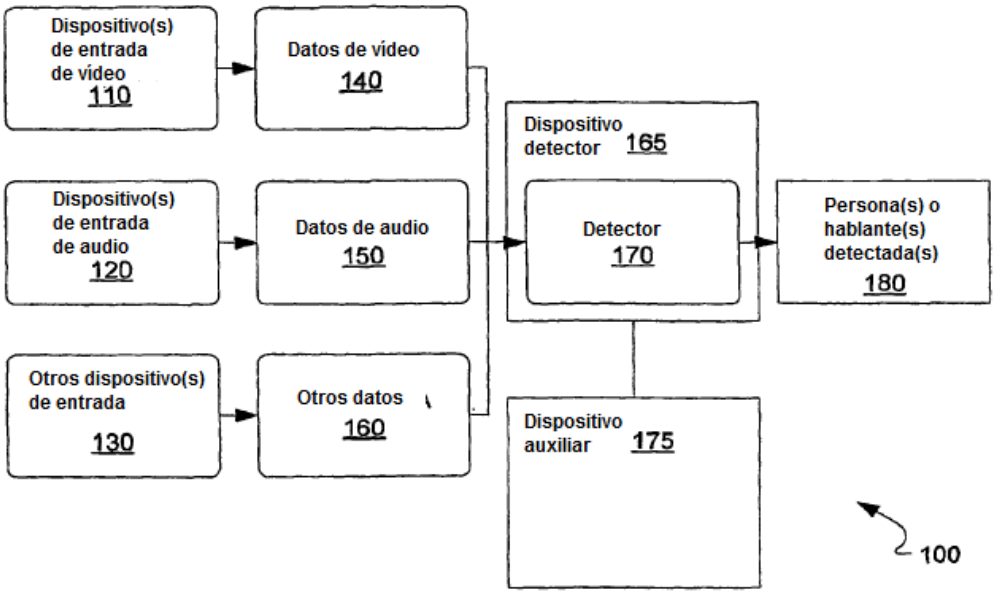
Dispositivo(s) de entrada de vídeo 110
Datos de vídeo 140
Dispositivo(s) de entrada de audio 120
Datos de audio 150
Otros dispositivo(s) de entrada 130
Otros datos 160
Dispositivo detector 165
Detector 170
Dispositivo auxiliar 175
Persona(s) o hablante(s) detectada(s) 180
100

FIG. 1

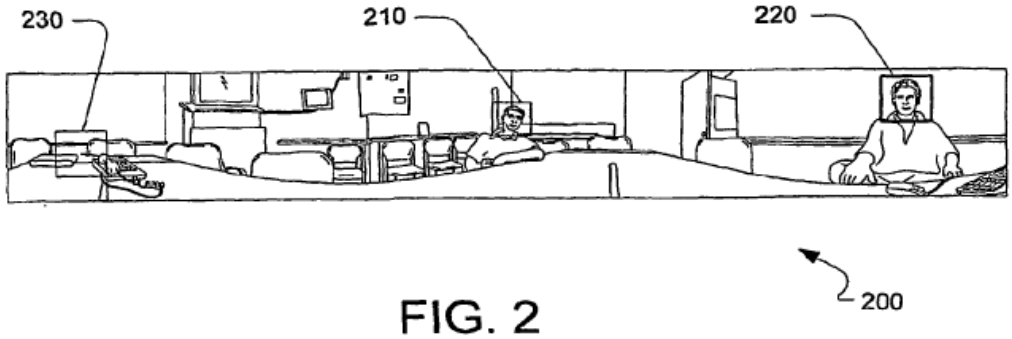
230
210
220
200

FIG. 2

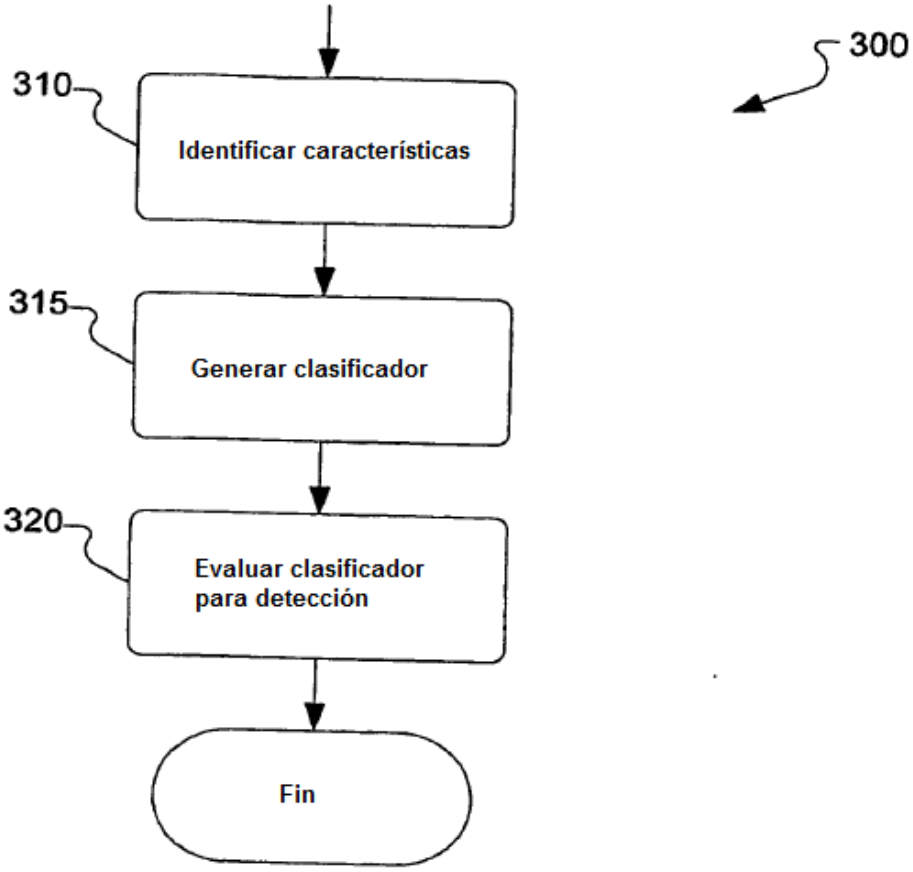
300
310
Identificar características
315
Generar clasificador
320
Evaluar clasificador para detección
Fin

FIG. 3

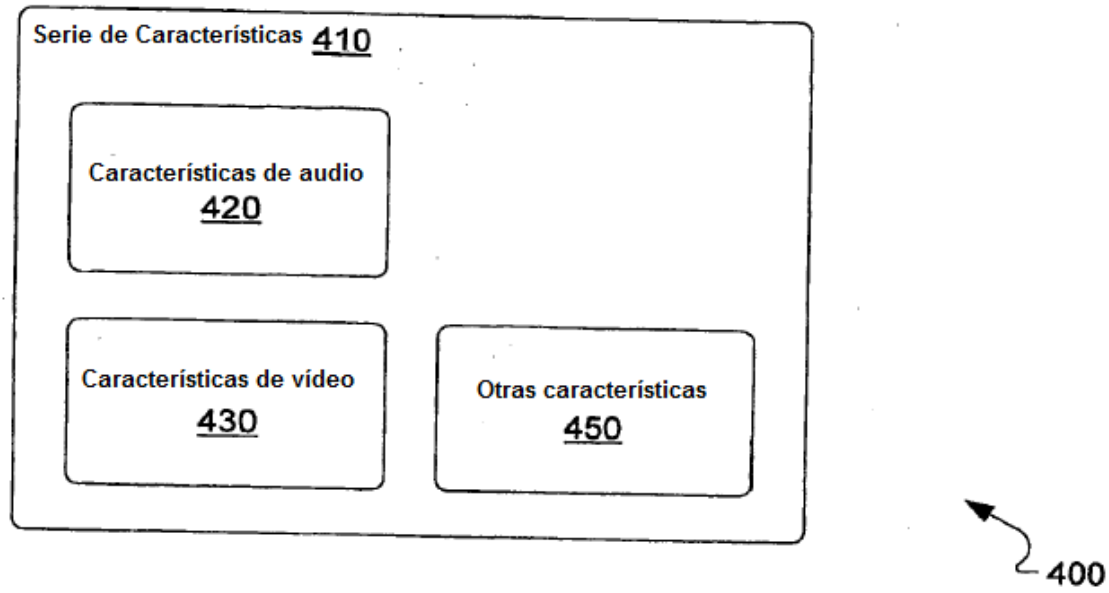
Serie de Características 410
Características de audio 420
Características de vídeo 430
Otras características 450
400

FIG. 4

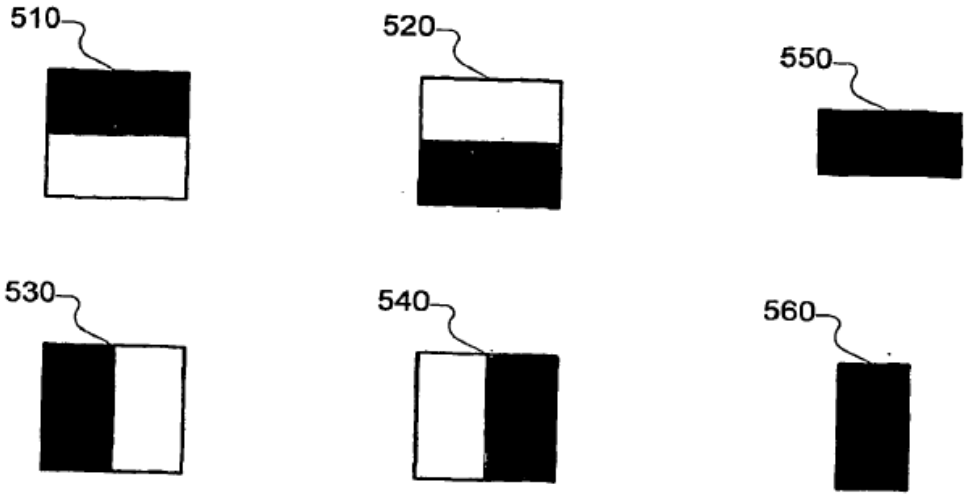

FIG. 5

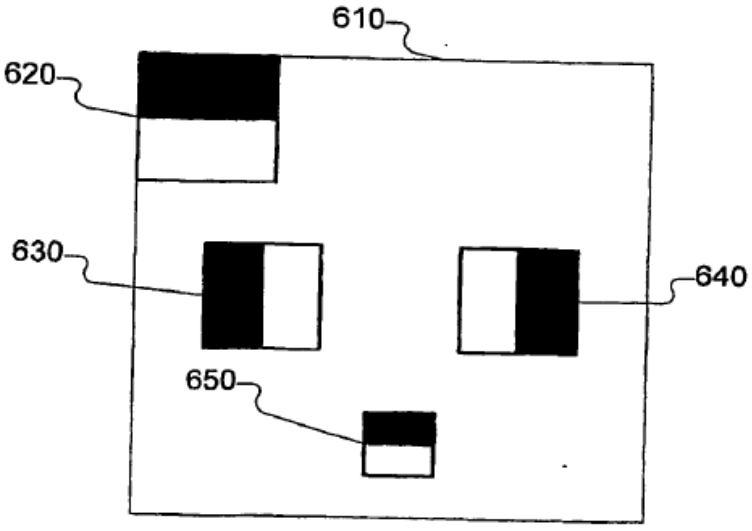

FIG. 6

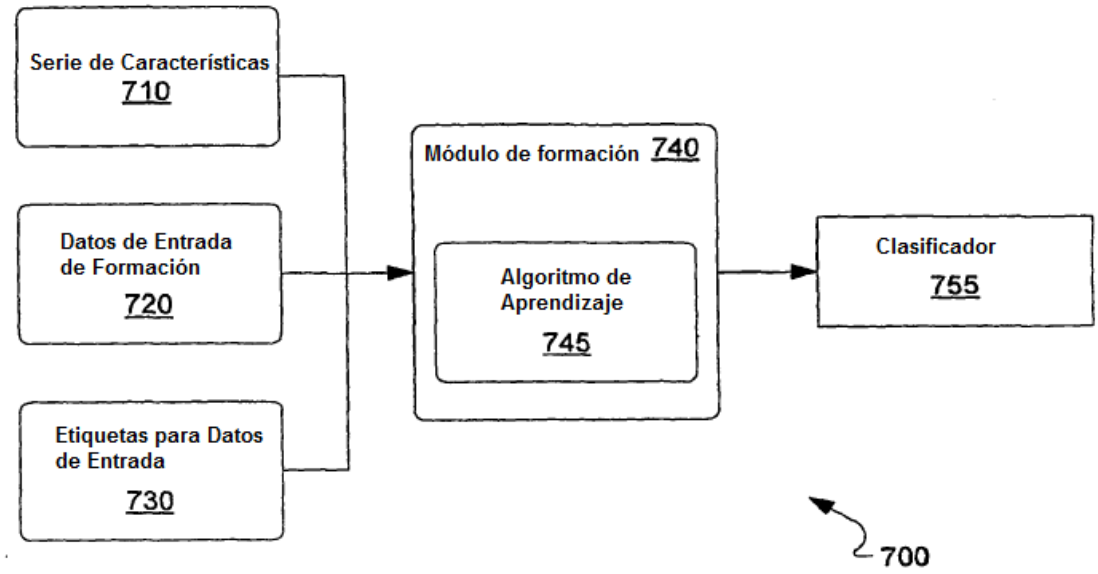
Serie de Características
710
Datos de Entrada de Formación
720
Etiquetas para Datos de Entrada
730
Módulo de formación 740
Algoritmo de Aprendizaje
745
Clasificador
755
700

FIG. 7

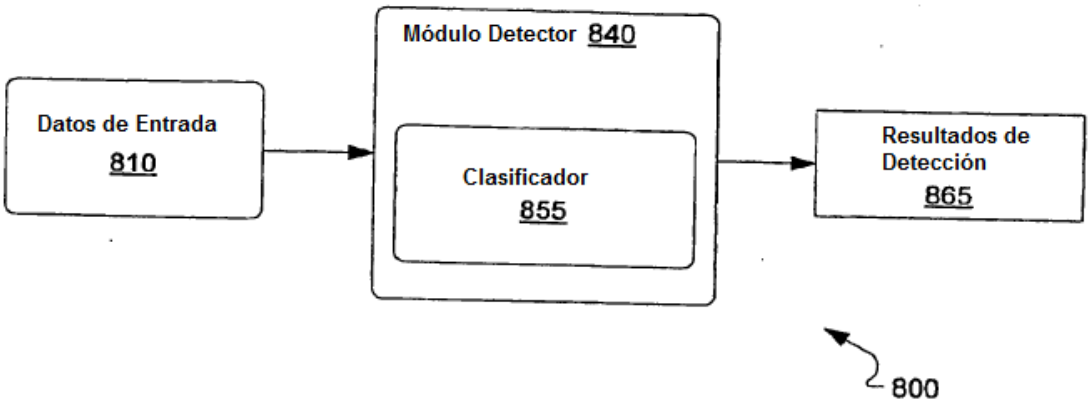
Módulo Detector 840
Datos de Entrada
810
Clasificador
855
Resultados de Detección
865
800

FIG. 8

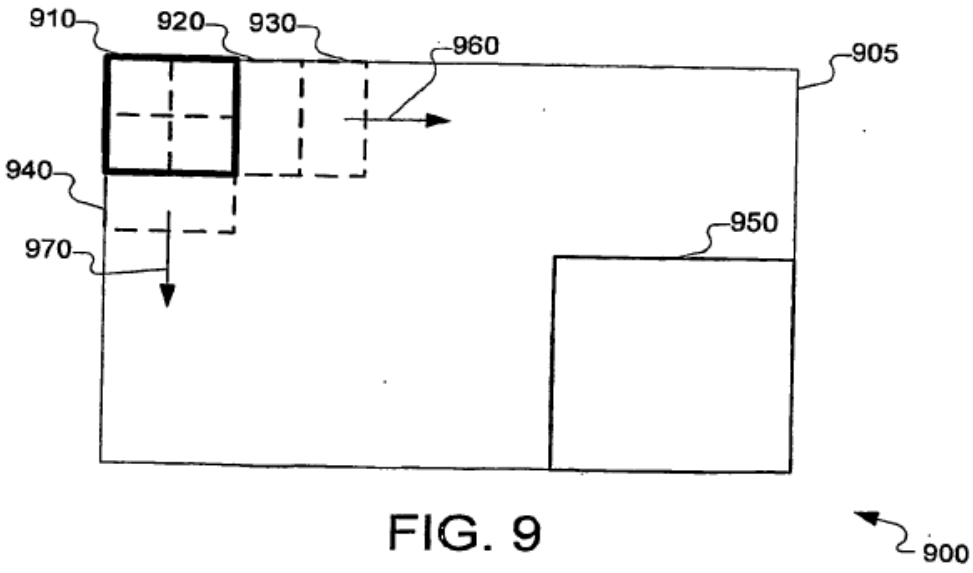
910
920
930
960
905
940
950
970
900

FIG. 9

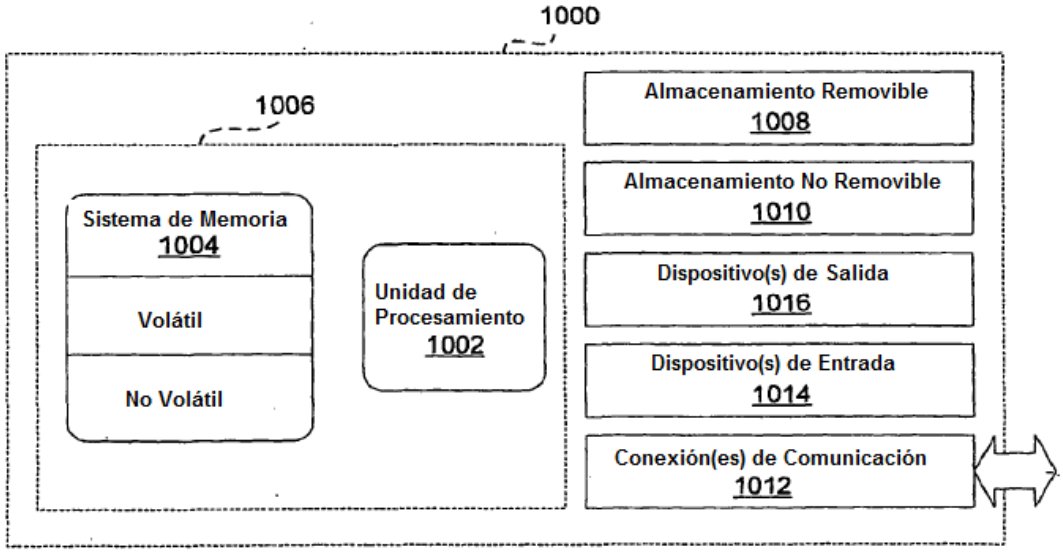
1000
1006
Sistema de Memoria
1004
Volátil
No Volátil
Unidad de Procesamiento
1002
Almacenamiento Removible
1008
Almacenamiento No Removible
1010
Dispositivo(s) de Salida
1016
Dispositivo(s) de Entrada
1014
Conexión(es) de Comunicación
1012

FIG. 10