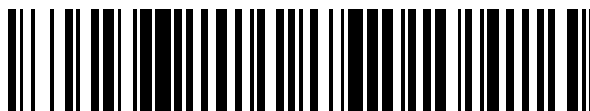


19



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA



11 Número de publicación: **2 932 079**

51 Int. Cl.:

G06F 3/01 (2006.01)

G06K 9/62 (2012.01)

G06T 7/246 (2007.01)

G06T 7/73 (2007.01)

G06V 40/18 (2012.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

96 Fecha de presentación y número de la solicitud europea: **28.03.2019** **E 19165768 (3)**

97 Fecha y número de publicación de la concesión europea: **19.10.2022** **EP 3557377**

54 Título: **Entrenamiento de redes neuronales para la predicción de mirada tridimensional (3D) con parámetros de calibración**

30 Prioridad:

30.03.2018 US 201815942008

45 Fecha de publicación y mención en BOPI de la traducción de la patente:
11.01.2023

73 Titular/es:

TOBII AB (100.0%)
Box 743
182 17 Danderyd, SE

72 Inventor/es:

LINDÉN, ERIK

74 Agente/Representante:

DEL VALLE VALIENTE, Sonia

ES 2 932 079 T3

Aviso: En el plazo de nueve meses a contar desde la fecha de publicación en el Boletín Europeo de Patentes, de la mención de concesión de la patente europea, cualquier persona podrá oponerse ante la Oficina Europea de Patentes a la patente concedida. La oposición deberá formularse por escrito y estar motivada; sólo se considerará como formulada una vez que se haya realizado el pago de la tasa de oposición (art. 99.1 del Convenio sobre Concesión de Patentes Europeas).

DESCRIPCIÓN

Entrenamiento de redes neuronales para la predicción de mirada tridimensional (3D) con parámetros de calibración

5 Campo técnico

La presente solicitud se refiere a sistemas y métodos de detección de mirada. En un ejemplo, tales sistemas y métodos se basan en sistemas de aprendizaje profundo, tales como redes neuronales para detectar la mirada tridimensional (3D).

10 Antecedentes

La interacción con los dispositivos informáticos es una acción fundamental en el mundo actual. Los dispositivos informáticos, tales como ordenadores personales, tabletas, teléfonos inteligentes, se encuentran a lo largo de la vida cotidiana. Además, los dispositivos informáticos que se pueden llevar puestos, tales como los dispositivos de auriculares ponibles (p. ej., auriculares de realidad virtual y auriculares de realidad aumentada), son cada vez más populares. Los sistemas y métodos para interactuar con tales dispositivos definen cómo se usan y para qué se usan.

Los avances en la tecnología de seguimiento ocular han hecho posible interactuar con un dispositivo informático utilizando la información de mirada de una persona. En otras palabras, la localización en una pantalla que el usuario está mirando. Esta información puede utilizarse únicamente para interacción o en combinación con una técnica de interacción basada en contacto (p. ej., utilizando un dispositivo de entrada de usuario, tal como un teclado, ratón, pantalla táctil u otra interfaz de entrada/salida).

En la patente de los Estados Unidos 6.204.828, en la solicitud de patente de los Estados Unidos 20130169560, en la patente de los Estados Unidos 7.113.170, en la publicación de solicitud de patente de los Estados Unidos 20140247232 y en la patente de los Estados Unidos US-9.619.020 pueden encontrarse técnicas de interacción propuestas anteriormente que utilizan información de mirada.

Generalmente, las técnicas de interacción basadas en la mirada se basan en detectar la mirada de un usuario en un punto de mirada. Los sistemas y métodos existentes pueden detectar con precisión la mirada bidimensional (2D). Recientemente, se han implementado redes neuronales para detectar tales miradas 2D.

Se han realizado intentos para ampliar las técnicas existentes que se basan en redes neuronales a la mirada tridimensional (3D). Sin embargo, la precisión de la predicción no es tan buena como la de la mirada 2D. En ausencia de un seguimiento preciso de la mirada 3D, la compatibilidad con pantallas estereoscópicas y aplicaciones 3D es significativamente limitada. Además, incluso en el dominio 2D, una red neuronal generalmente se entrena para una cámara específica y una configuración de pantalla (p. ej., resolución de imagen, distancia focal, distancia a una pantalla de un dispositivo informático, un tamaño de la pantalla y similares). Por lo tanto, cada vez que cambia la configuración (p. ej., diferente resolución de imagen, diferente tamaño de pantalla y similares), la red neuronal ya no puede predecir la mirada 2D con una precisión aceptable. Sería necesario volver a entrenar la red neuronal para la nueva configuración.

ADIBA AMALIA I Y COL.: “Seguimiento de la mirada en espacio 3D con una red neuronal de convolución” “Mira lo que veo”, TALLER DE RECONOCIMIENTO DE PATRONES DE IMÁGENES APLICADAS (AIPR) IEEE 2017, IEEE, 10 de octubre de 2017, páginas 1-7, XP033399287, describe el uso de una convolution neural network (red neuronal de convolución - CNN) para estimar vectores de mirada bajo movimientos de cabeza sin restricciones. La CNN tiene cinco capas y una capa de salida totalmente conectada.

50 Sumario

La invención se define mediante las reivindicaciones independientes adjuntas. Se definen otras realizaciones por las reivindicaciones dependientes.

Las realizaciones adicionales se analizan en la Descripción detallada, y allí se proporciona una descripción adicional.

Breve descripción de los dibujos

La naturaleza y las ventajas de diversas realizaciones pueden comprenderse mejor haciendo referencia a las figuras que siguen. En las figuras adjuntas, componentes o características similares pueden tener el mismo número de referencia. Además, diversos componentes del mismo tipo pueden distinguirse por medio del número de referencia seguido de un guión y un segundo número que lo distingue de componentes similares. Si en la presente memoria descriptiva se utiliza únicamente el primer número de referencia, la descripción es aplicable a cualquiera de los componentes similares que tengan el mismo primer número de referencia independientemente del segundo número de referencia.

La Fig. 1 muestra un sistema de seguimiento ocular, según una realización.

La Fig. 2 muestra un ejemplo de una imagen de un ojo capturada por un sensor de imagen, según una realización.

5 La Fig. 3 es un diagrama de bloques que ilustra un sistema informático especializado, según una realización.

La Fig. 4 muestra un ejemplo de un dispositivo informático portátil que implementa componentes de un sistema de seguimiento ocular, según una realización.

10 La Fig. 5 ilustra un entorno informático de ejemplo para predecir la mirada 3D basándose en un sistema de aprendizaje profundo, según una realización.

La Fig. 6 ilustra componentes de ejemplo de un sistema de aprendizaje profundo para predecir la mirada 3D, según una realización.

15 La Fig. 7 ilustra una arquitectura de red de ejemplo para una red neuronal, según una realización.

La Fig. 8 ilustra una normalización de imagen de ejemplo, según una realización.

20 La Fig. 9 ilustra una predicción de mirada 3D de ejemplo, según una realización.

La Fig. 10 ilustra predicciones para vectores de mirada 2D y corrección de distancia generadas por una red neuronal, según una realización.

25 La Fig. 11 ilustra un flujo de ejemplo para la predicción de mirada 3D basándose en un sistema de aprendizaje profundo, según una realización.

La Fig. 12 ilustra una estrategia para usar imágenes de entrenamiento que están asociadas con una diversidad de ubicaciones de puntos de mirada en relación con planos 2D de cámaras que generaron las imágenes de entrenamiento, según una realización.

30

La Fig. 13 ilustra un flujo de ejemplo para entrenar una red neuronal, según una realización.

La Fig. 14 ilustra un flujo de ejemplo para entrenar una red neuronal utilizando parámetros de calibración incorporados, según una realización.

35

Descripción detallada

Las realizaciones de la presente descripción se refieren a la detección de mirada tridimensional (3D) basándose en un sistema de aprendizaje profundo. En un ejemplo, se utiliza una red neuronal. Esta red neuronal se puede utilizar independientemente de la configuración de la cámara y la pantalla. En otras palabras, independientemente de la cámara, la pantalla y la combinación de configuración de cámara y pantalla específicas, la red neuronal está debidamente entrenada para predecir la información de mirada en apoyo de la detección de mirada 3D.

40

Más particularmente, una cámara captura una imagen bidimensional (2D) de un usuario que mira fijamente un punto en el espacio 3D. A partir de esta imagen 2D se estima una distancia aproximada entre la cámara y los ojos del usuario. La imagen 2D se normaliza para generar imágenes distorsionadas centradas alrededor del ojo del usuario a alta resolución y una imagen distorsionada alrededor del rostro del usuario se genera a baja resolución. Estas imágenes distorsionadas se entran a la red neuronal que, a su vez, predice un vector de mirada 2D por ojo y una corrección de distancia para la distancia aproximada. La posición del ojo en el espacio 3D se estima basándose en la distancia aproximada y la corrección de distancia, y la posición de la cámara en el espacio 3D. Basándose en una función de mapeo entre el espacio 2D y el espacio 3D, cada vector de mirada 2D se asigna a una dirección de mirada 3D. La posición 3D de un ojo y su dirección de mirada 3D indican la mirada 3D asociada con ese ojo.

45

Debido a que se usa una imagen 2D normalizada y debido a que la red neuronal predice una corrección de distancia y vectores de mirada 2D, las predicciones de la red neuronal no dependen de la configuración de la cámara y la pantalla. Por ende, esa misma red neuronal entrenada se puede usar en diferentes sistemas de seguimiento ocular, incluidos los integrados con diferentes tipos de teléfonos inteligentes, tabletas, ordenadores portátiles, dispositivos de auriculares ponibles (p. ej., auriculares de realidad virtual y realidad aumentada) y sistemas de seguimiento ocular independientes. Además, debido a que se detecta la mirada 3D, se pueden admitir visualizaciones estereoscópicas y aplicaciones 3D.

55

El entrenamiento de la red neuronal se basa generalmente en imágenes de entrenamiento que son diversas con respecto a las ubicaciones de los puntos de mirada en relación con los planos 2D de las cámaras utilizadas para capturar las imágenes de entrenamiento (p. ej., para cada cámara, un plano X, Y perpendicular a la línea de visión de la cámara). En particular, algunas de las imágenes de entrenamiento muestran ojos de usuarios que miraban

65

puntos de mirada ubicados en los planos 2D de las cámaras, mientras que otras imágenes de entrenamiento muestran ojos de usuarios que miraban puntos de mirada fuera de los planos 2D. Durante el entrenamiento, la red neuronal busca ángulos de mirada entre los ojos del usuario-cámaras y usos de usuarios-puntos de mirada y distancias ojo-cámara. Debido a que se utiliza la diversidad, la red neuronal puede encontrar correctamente los ángulos y las distancias. Las líneas de mirada (también conocidas como rayos de mirada) se predicen a partir de los ángulos de mirada y se calculan las distancias entre los puntos de mirada y las líneas de mirada (distancias de punto de mirada a línea de mirada). La función de pérdida de la red neuronal implica distancias de punto de mirada a línea de mirada. Durante el entrenamiento, los parámetros de la red neuronal (p. ej., pesos de conexión entre nodos en las diferentes capas) se actualizan para minimizar la función de pérdida al minimizar la diferencia entre las distancias de línea de punto a punto de mirada reales y predichas. En aras de la brevedad, de aquí en adelante, una imagen se refiere a una imagen 2D, salvo se indique lo contrario.

Estas y otras características del entrenamiento y el uso de la red neuronal para la predicción de la mirada 3D independientemente de la configuración de la cámara y la pantalla se describen con más detalle a continuación. A continuación se describen varios componentes y configuraciones de los sistemas de seguimiento ocular para proporcionar una mejor comprensión de las técnicas de predicción de mirada 3D.

Además, durante el entrenamiento, la red neuronal aprende “n” parámetros de calibración para las personas que se muestran en las imágenes de entrenamiento. Un diseñador de la red neuronal solo necesita especificar el número “n” en lugar de especificar los parámetros de calibración. Estos parámetros se entran en las capas de la red neuronal y se convierten en parte del entrenamiento. Al finalizar el entrenamiento, un usuario final que opera un sistema de seguimiento sigue típicamente un proceso de calibración. Como parte de este proceso, las imágenes de calibración (p. ej., donde se le pide al usuario final que observe puntos específicos) se entran a la red neuronal que después genera los “n” parámetros de calibración para el usuario. Estos parámetros se utilizan en la predicción de mirada para el usuario final.

La Fig. 1 muestra un sistema 100 de seguimiento ocular (que también puede denominarse sistema de seguimiento de la mirada) según una realización. El sistema 100 comprende unos iluminadores 111 y 112 para iluminar los ojos de un usuario y un sensor 113 de imagen para capturar imágenes de los ojos del usuario. Los iluminadores 111 y 112 pueden ser, por ejemplo, diodos emisores de luz que emiten luz en la banda de frecuencias del infrarrojo o en la banda de frecuencias del infrarrojo cercano. El sensor 113 de imagen puede ser, por ejemplo, una cámara, tal como una cámara de complementary metal oxide semiconductor (semiconductor de óxido metálico complementario - CMOS) o una cámara de charged coupled device (dispositivo de carga acoplada - CCD). La cámara no se limita a ser una cámara de infrarrojo ni una cámara de profundidad ni una cámara de campo claro. El mecanismo obturador del sensor de imagen puede ser un obturador rodante o un obturador global.

Un primer iluminador 111 está dispuesto de forma coaxial con el (o cerca del) sensor 113 de imagen para que el sensor 113 de imagen pueda capturar imágenes de pupila luminosas de los ojos del usuario. Debido a la disposición coaxial del primer iluminador 111 y del sensor 113 de imagen, la luz reflejada procedente de la retina de un ojo vuelve a través de la pupila al sensor 113 de imagen, de modo que la pupila aparece como más brillante que el iris que la rodea en imágenes donde el primer iluminador 111 ilumina el ojo. Un segundo iluminador 112 está dispuesto de forma no coaxial con el (o más alejado del) sensor 113 de imagen para capturar imágenes de pupila oscuras. Debido a la disposición no coaxial del segundo iluminador 112 y del sensor 113 de imagen, la luz reflejada procedente de la retina de un ojo no llega al sensor 113 de imagen, y la pupila parece más oscura que el iris que la rodea en imágenes donde el segundo iluminador 112 ilumina el ojo. Los iluminadores 111 y 112 pueden por ejemplo turnarse para iluminar el ojo, de manera que cada primera imagen sea una imagen de la pupila brillante y cada segunda imagen sea una imagen de la pupila oscura.

El sistema 100 de seguimiento ocular comprende también sistema 120 de circuitos (que incluye p. ej. uno o más procesadores) para procesar las imágenes capturadas por el sensor 113 de imagen. Por ejemplo, el sistema 120 de circuitos puede estar conectado al sensor 113 de imagen y a los iluminadores 111 y 112 por una conexión por cable o inalámbrica. En otro ejemplo, el sistema 120 de circuitos en forma de uno o más procesadores puede proporcionarse en una o más capas apiladas debajo de la superficie fotosensible del sensor 113 de imagen.

La Fig. 2 muestra un ejemplo de una imagen de un ojo 200 capturada por el sensor 113 de imagen. El sistema 120 de circuitos puede emplear, por ejemplo, procesamiento de imágenes (tal como procesamiento de imágenes digitales) para extraer las características de la imagen. El sistema 120 de circuitos puede emplear, por ejemplo, un seguimiento ocular por reflexión entre la córnea y el *pupil center cornea reflection* (centro de la pupila - PCCR) para determinar dónde está mirando el ojo 200. En el seguimiento ocular PCCR, el procesador 120 estima la posición del centro de la pupila 210 y la posición del centro de un reflejo 220 en el ojo 200. El reflejo 220 es causado por el reflejo de la luz procedente de uno de los iluminadores 111 y 112. El procesador 120 calcula dónde está el usuario en el espacio utilizando el reflejo 220 y dónde está apuntando el ojo 200 del usuario utilizando la pupila 210. Puesto que de forma típica existe un desplazamiento entre el centro óptico del ojo 200 y la fovea, el procesador 120 realiza la calibración del desvío de la fovea para poder determinar dónde está mirando el usuario. Las direcciones de mirada obtenidas del ojo izquierdo y del ojo derecho pueden combinarse entonces para formar una dirección de mirada estimada combinada (o dirección de visión). Como se describirá más adelante, muchos factores distintos pueden

afectar a cómo deben ponderarse las direcciones de mirada para los ojos izquierdo y derecho entre sí cuando se forma esta combinación.

En la realización descrita con referencia a la Fig. 1, los iluminadores 111 y 112 se disponen en un módulo 110 de seguimiento ocular colocado debajo de un sistema de representación observado por el usuario. Esta disposición solo sirve como ejemplo. Se apreciará que puede emplearse, más o menos, cualquier número de iluminadores y sensores de imagen para el seguimiento ocular, y que tales iluminadores y sensores de imagen pueden distribuirse de muchas formas distintas respecto a pantallas observadas por el usuario. Se apreciará que el esquema de seguimiento ocular descrito en la presente descripción puede emplearse, por ejemplo, para el seguimiento ocular remoto (por ejemplo en un ordenador personal, un teléfono inteligente o integrado en un vehículo) o para el seguimiento ocular ponible (tal como en gafas de realidad virtual o gafas de realidad aumentada).

La Fig. 3 es un diagrama de bloques que ilustra un sistema 300 informático especializado en el que pueden aplicarse realizaciones de la presente descripción. Este ejemplo ilustra un sistema 300 informático especializado tal como el que puede utilizarse, en su totalidad, en parte o con diversas modificaciones, para proporcionar las funciones de los componentes descritos en la presente memoria.

Se muestra un sistema 300 informático especializado que comprende elementos de hardware que pueden estar acoplados eléctricamente a través de un bus 390. Los elementos de hardware pueden incluir una o más unidades 310 de procesamiento central, uno o más dispositivos 320 de entrada (p. ej., un ratón, un teclado, un dispositivo de seguimiento ocular, etc.) y uno o más dispositivos 330 de salida (p. ej., un dispositivo de visualización, una impresora, etc.). El sistema 300 informático especializado puede incluir también uno o más dispositivos 340 de almacenamiento. A modo de ejemplo, el dispositivo(s) 340 de almacenamiento pueden ser unidades de disco, dispositivos de almacenamiento óptico, dispositivo de almacenamiento en estado sólido tales como una memoria de acceso aleatorio ("RAM") y/o una memoria de solo lectura ("ROM"), que pueden ser programables, actualizables por memoria flash y/o similares.

El sistema 300 informático especializado puede incluir de forma adicional un lector 350 de medios de almacenamiento legibles por ordenador, un sistema 360 de comunicaciones (p. ej., un módem, una tarjeta de red (inalámbrica o por cable), un dispositivo de comunicación infrarroja, un dispositivo Bluetooth™, un dispositivo de comunicación celular, etc.) y una memoria funcional 380, que puede incluir dispositivos RAM y ROM, como se ha descrito anteriormente. En algunas realizaciones, el sistema informático especializado 300 puede incluir también una unidad 370 de aceleración del procesamiento, que puede incluir un procesador de señales digitales, un procesador especial y/o procesadores similares.

La Fig. 4 muestra un ejemplo de un dispositivo 400 informático ponible que implementa algunos o todos los componentes anteriores de un sistema de seguimiento ocular como se describe en relación con las Figs. 1-2. El dispositivo 400 informático ponible puede ser un auricular VR o un auricular AR que puede ser usado por un usuario. Como se ilustra, el dispositivo 400 informático ponible incluye un conjunto de lentes 410, tales como lentes de Fresnel, un conjunto de cámaras 420, un conjunto de espejos calientes 430 (p. ej., como se ilustra en las Figs. 12-14, el conjunto incluye dos espejos calientes para cada ojo en diversas realizaciones), y un conjunto de pantallas 440. La cámara 420 puede incluir los sensores 113 de imagen de la Fig. 1. Aunque no se muestra en la Fig. 4, el dispositivo 400 informático ponible puede incluir también un conjunto de iluminadores y sistema de circuitos de procesamiento. Estos y otros componentes se pueden integrar dentro de un alojamiento 450 del dispositivo 400 informático ponible. De esta manera, después de que el usuario ponga el dispositivo 400 informático ponible en su cabeza, el conjunto de lentes 410 estaría relativamente cerca de los ojos del usuario y el conjunto de pantallas estaría relativamente lejos del ojo del usuario, y los componentes restantes pueden estar ubicados en el medio. La disposición de estos componentes permite la detección del punto de mirada del usuario en un espacio virtual o real tridimensional.

A continuación en la presente memoria, se describe el uso de un sistema de aprendizaje profundo para la predicción de mirada 3D. En aras de la claridad de la explicación, este sistema se describe en relación con una cámara, una pantalla y dos ojos del usuario (p. ej., la cámara captura imágenes, algunas o todas de las que muestran los dos ojos del usuario). El sistema de aprendizaje profundo se puede usar con una cámara arbitraria y una configuración de pantalla para el seguimiento ocular que usa luz visible, infrarrojo pasivo, infrarrojo activo de pupila brillante (BP) y similares. Sin embargo, las realizaciones de la presente descripción no están limitadas como tales.

Por ejemplo, las realizaciones se aplican similarmente a un sistema de seguimiento ocular que usa una cámara por ojo de usuario, como en el contexto de un auricular de realidad virtual o un auricular de realidad aumentada. Los cambios en cómo se implementa el sistema de aprendizaje profundo para un seguimiento de una cámara por ojo deberían ser evidentes para un experto en la técnica a la luz de la presente descripción. Por ejemplo, en lugar de introducir dos imágenes distorsionadas, cada una de las que se enfoca en uno de los ojos del usuario, solo se usa una única imagen distorsionada del ojo del usuario que está asociada con la cámara. Asimismo, no se puede ser introducir ninguna imagen distorsionada alrededor del rostro del usuario. Durante el entrenamiento, la red neuronal aprende a predecir una corrección de distancia a partir de la imagen distorsionada en lugar de predecir esta corrección basándose en imágenes distorsionadas alrededor de los ojos y el rostro del usuario. Después, este

sistema generaría un vector de mirada 2D para el ojo del usuario asociado con la cámara y una corrección de distancia para la distancia aproximada entre la cámara y el ojo. Además, en el caso de un dispositivo VR o AR, se puede predefinir una distancia aproximada y no es necesario estimarla basándose en una imagen generada por la cámara.

La Fig. 5 ilustra un entorno informático de ejemplo para predecir la mirada 3D basándose en un sistema de aprendizaje profundo, según una realización. En general, la información de mirada 2D se refiere a una posición de mirada X, Y en un plano 2D. En comparación, 3D se refiere no solo a la posición de mirada X, Y, sino también a la posición de mirada Z. En un ejemplo, la mirada 3D se puede caracterizar por una posición del ojo en el espacio 3D como origen y una dirección de mirada 3D desde el origen.

Como se ilustra en la Fig. 5, un usuario 510 opera un dispositivo informático 520 que rastrea la mirada 3D 512 del usuario 510. Para hacerlo, el dispositivo informático 520 está, en un ejemplo, en comunicación con un ordenador servidor 530 que aloja un sistema 532 de aprendizaje profundo. El dispositivo informático 520 envía, al ordenador servidor 530 a través de una red de datos (no mostrada), una imagen 2D 550 que muestra los ojos del usuario mientras el usuario 510 está mirando. El ordenador servidor 530 entra esta imagen 2D 550 al sistema 532 de aprendizaje profundo que, en respuesta, predice la mirada 3D 512. El ordenador servidor 530 envía información 560 sobre la mirada 3D 512, como la posición del ojo 3D y la dirección de mirada 3D de vuelta al dispositivo informático 520 a través de la red de datos. El dispositivo informático 520 utiliza esta información 560 para proporcionar un servicio informático 3D basándose en la mirada al usuario 510.

Aunque la Fig. 5 muestra el servidor informático 530 que aloja el sistema 532 de aprendizaje profundo, las realizaciones de la presente descripción no se limitan como tales. Por ejemplo, el dispositivo informático 530 puede descargar código y alojar una instancia del sistema 532 de aprendizaje profundo. De esta forma, el dispositivo informático 520 se basa en esta instancia para predecir localmente la mirada 3D 512 y no necesita enviar la imagen 2D 550 al ordenador servidor 530. En este ejemplo, el ordenador servidor 530 (o algún otro sistema informático conectado a ella a través de una red de datos) puede entrenar el sistema 532 de aprendizaje profundo y proporcionar una interfaz (p. ej., una interfaz web) para descargar el código de este sistema 530 de aprendizaje profundo para dispositivos informáticos, alojando así instancias del sistema 530 de aprendizaje profundo en estos dispositivos informáticos.

En un ejemplo, el dispositivo informático 520 incluye una cámara 522, una pantalla 524 y una aplicación 526 de mirada 3D. La cámara 522 genera la imagen 2D 550 que es una representación 2D 540 del rostro del usuario. Esta imagen 2D 550 muestra los ojos del usuario mientras mira en el espacio 3D. Se puede definir un sistema 528 de coordenadas 3D en asociación con la cámara 522. Por ejemplo, la cámara 522 está en el origen de este sistema 528 de coordenadas 3D. Los planos X e Y pueden ser un plano perpendicular a la línea de visión de la cámara 522. En comparación, la imagen 2D 550 tiene un plano 2D que se puede definir alrededor de un sistema 542 de coordenadas 2D local a la representación 2D 540 del rostro del usuario. La cámara 522 está asociada con un mapeo entre el espacio 2D y el espacio 3D (p. ej., entre los dos sistemas 542 y 528 de coordenadas). En un ejemplo, este mapeo incluye la matriz de retroproyección de la cámara 522 y se almacena localmente en el dispositivo informático 522 (p. ej., en la ubicación de almacenamiento asociada con la aplicación 526 de mirada 3D).

La pantalla 524 puede estar, pero no necesariamente, en el plano X, Y de la cámara 522 (si no, las posiciones relativas entre las dos se determinan basándose en la configuración del dispositivo informático 520). La aplicación 526 de mirada 3D puede procesar la imagen 2D 550 para introducirla en el sistema 530 de aprendizaje profundo (ya sea remoto o local al dispositivo informático 520) y puede procesar la información 560 sobre la mirada 3D para admitir visualizaciones estereoscópicas (si también es compatible con la pantalla 524) y aplicaciones 3D (p. ej., controles 3D y manipulaciones de objetos visualizados en la pantalla 524 basándose en la información 560).

La Fig. 6 ilustra componentes de ejemplo de un sistema 600 de aprendizaje profundo para predecir la mirada 3D, según una realización. Como se ilustra, el sistema de aprendizaje profundo incluye un detector 620 ocular, un estimador 630 de distancia, un generador 640 de imágenes, una red 650 neuronal y un generador 660 de mirada 3D. Algunos o todos estos componentes pueden implementarse como hardware especializado y/o como módulos de software (p. ej., instrucciones específicas legibles por computadora) alojados en hardware de procesamiento general o especializado.

Como se ilustra, se introduce una imagen 2D 610 en el detector 620 ocular. Por ejemplo, esta imagen 2D 610 se genera con una cámara. En respuesta, el detector 620 ocular detecta los ojos del usuario en la imagen 2D 610 y emite información sobre las posiciones 622 de los ojos en la imagen (p. ej., ubicaciones de estos centros de las pupilas en el plano 2D de la imagen 2D 610). En un ejemplo, el detector 620 ocular se implementa como un algoritmo de aprendizaje automático entrenado para la detección de ojos. Muchos algoritmos de aprendizaje automático son posibles y son conocidos por los expertos en la materia.

Las posiciones 622 de los ojos y la imagen 2D 610 se introducen en el estimador 630 de distancia. En respuesta, el estimador 630 de distancia genera una distancia 632 estimada, como una distancia aproximada. Para ello, el estimador 630 de distancia proyecta los ojos detectados en la imagen 2D 610 en un sistema de coordenadas 3D

centrado alrededor de la cámara. Esta proyección utiliza el mapeo espacial 2D-3D de la cámara. El estimador 630 de distancia busca las proyecciones del ojo en el espacio 3D donde la distancia interocular (ID) es aproximadamente la ID humana promedio (p. ej., sesenta y tres milímetros). La distancia entre la cámara y cada una de estas proyecciones oculares puede denominarse distancia proyectada. La distancia aproximada se establece en función de la identificación humana promedio y las distancias proyectadas.

Para ilustrar, supongamos que $d_{aproximada}$ se refiera a la distancia aproximada y K a la matriz intrínseca de la cámara, y $e_{izquierdo}$ y $e_{derecho}$ los ojos detectados en la imagen 2D 610. K es un componente del mapeo espacial 2D-3D de la cámara. Los ojos izquierdo y derecho proyectados $e'_{izquierdo, proy.}$ y $e'_{derecho, proy.}$ se calculan como $e'_{izquierdo, proy.} = K^{-1}e_{izquierdo}$ y $e'_{derecho, proy.} = K^{-1}e_{derecho}$ y representan vectores de ojos proyectados desde ojos detectados en la imagen 2D 610 en el sistema de coordenadas 3D centrado alrededor de la cámara, como se muestra en la Fig. 8. Estos ojos proyectados se normalizan como $e_{izquierdo, proy.} = e'_{izquierdo, proy.} / ||e'_{izquierdo, proy.}||$ y $e_{derecho, proy.} = e'_{derecho, proy.} / ||e'_{derecho, proy.}||$. La distancia aproximada es el ordenador como $d_{aproximada} = ID / ||e_{izquierdo, proy.} - e_{derecho, proy.}||$.

El generador 640 de imágenes recibe la imagen 2D 610 y, en respuesta, genera imágenes 642 distorsionadas alrededor de los ojos del usuario y una imagen 644 distorsionada alrededor del rostro del usuario. En un ejemplo, el generador 640 de imágenes accede desde la memoria local a una primera distancia s predefinida en píxeles a las imágenes 642 distorsionadas de los ojos del usuario y a una segunda distancia s predefinida en píxeles a la imagen 644 distorsionada del rostro del usuario. Estas distancias s predefinidas son diferentes distancias interoculares en píxeles s , de tal forma que las imágenes 642 distorsionadas de los ojos del usuario y la imagen 644 distorsionada del rostro del usuario están en diferentes proyecciones en el espacio 3D y tienen una resolución diferente (la resolución de las imágenes 642 distorsionadas de los ojos del usuario es mayor que la de la imagen 644 distorsionada del rostro del usuario usando una primera distancia s predefinida más pequeña relativa a la segunda distancia s predefinida). El generador 640 de imágenes utiliza estas distancias s predefinidas para generar las imágenes distorsionadas de tal forma que se normalicen con respecto a la cámara. En particular, cada imagen distorsionada representa una proyección de la imagen 2D 610 en el sistema de coordenadas 3D centrada alrededor de la cámara, una rotación de la imagen 2D 610 alrededor del eje X (de tal forma que cualquier inclinación de la cabeza se rota para estar en la posición horizontal), una escala de la imagen 2D 610 (basándose en las distancias s predefinidas), y una distorsión (de tal forma que los ojos y el rostro del usuario estén en el centro de cada imagen distorsionada y no estén sesgados geométricamente). En otras palabras, cada imagen distorsionada representa una imagen normalizada que se puede introducir a la red 650 neuronal, donde la normalización desacopla la dependencia de esta entrada a la configuración de la cámara (p. ej., resolución de imagen, distancia focal, distancia a una pantalla, tipo de cámara tal como estenopeicas y no estenopeicas, y similares).

Para ilustrar, el generador 640 de imágenes genera una matriz R de rotación que rota los puntos del espacio 3D real (p. ej., el sistema de coordenadas 3D centrado alrededor de la cámara) a un espacio 3D normalizado (p. ej., un sistema de coordenadas 3D centrado también alrededor de la cámara pero rotado en relación con el espacio real 3D de tal forma que el vector entre los ojos del usuario en la imagen 2D 610 son horizontales), como se muestra además en la Fig. 8. El generador 640 de imágenes genera también una matriz M de escala basándose en una distancia s predefinida (una matriz M se genera para los ojos del usuario y otra matriz M se genera para el rostro del usuario). Por ejemplo, M se genera como una matriz diagonal, donde $M = \text{diag}([1, 1, f])$, donde f es una distancia focal seleccionada para hacer que la distancia interocular entre los ojos del usuario sea igual a la distancia s predefinida. Una transformación T se define como una función de la matriz K intrínseca, la matriz R de rotación, y la matriz M de escala y se usa para normalizar la imagen 2D 610 en una imagen normalizada (para el ojo del usuario y para el rostro del usuario dependiendo de la matriz M). Por ejemplo, la transformación T se expresa como $T = MRK^{-1}$ y se aplica a la imagen 2D 610 para generar una primera imagen normalizada a la primera distancia s predefinida para los ojos del usuario y una segunda imagen normalizada a la segunda distancia s predefinida para el rostro del usuario. Cada una de estas imágenes normalizadas es una rotación y una proyección de la imagen 2D 610 desde el espacio 3D real al espacio 3D normalizado, donde la imagen rotada y proyectada está a la distancia s predefinida en el espacio 3D normalizado. El generador 640 de imágenes genera una imagen 642 distorsionada de los ojos del usuario distorsionando la imagen normalizada del ojo del usuario usando interpolación bilineal y recortando una región $W \times H$ centrada alrededor de uno de los ojos del usuario. Similarmente, el generador 640 de imágenes genera la imagen 644 distorsionada del rostro del usuario distorsionando la imagen normalizada del rostro del usuario usando interpolación bilineal y recortando una región $W \times H$ centrada alrededor del punto medio entre los ojos del usuario.

Además, una de las imágenes 642 distorsionadas de los ojos se refleja de tal forma que la imagen reflejada resultante se alinea con la otra imagen distorsionada en el ojo (no reflejada). Por ejemplo, se refleja la imagen distorsionada alrededor del ojo izquierdo. Como resultado, el canto interno del ojo izquierdo se alinea con el canto interno del ojo derecho, y el canto externo del ojo izquierdo se alinea con el canto externo del ojo derecho, como se muestra en la imagen reflejada del ojo izquierdo y la imagen distorsionada del ojo derecho en la Fig. 10. Al tener esta alineación, la entrada de la imagen distorsionada de los ojos a la red neuronal es la misma en términos de orientación, lo que simplifica la arquitectura y el entrenamiento de la red neuronal.

Las imágenes 642 distorsionadas de los ojos y la imagen 644 distorsionada del rostro se introducen en la red 650 neuronal. En un ejemplo, la imagen 2D 610 no se introduce a esta red 650. En respuesta, la red 650 neuronal emite

una corrección 652 de distancia y vectores 654 de mirada 2D (uno por ojo). La corrección 652 de distancia es un factor de corrección multiplicativo que puede multiplicado por la distancia 632 estimada para corregir esta estimación y generar una distancia corregida. Cada uno de los vectores 654 de mirada 2D tiene un origen de mirada (p. ej., el centro de la pupila o un reflejo) y una dirección de mirada en la imagen recortada del ojo correspondiente (el origen y la dirección están en el plano 2D de esta imagen y se pueden seguir de vuelta al espacio 3D real).

En un ejemplo, la red 650 neuronal es una red neuronal convolucional que incluye múltiples subredes (p. ej., a lo largo de ramas paralelas de la red 650 neuronal). Estas subredes (y, de forma equivalente, toda la red neuronal convolucional) se pueden entrenar en conjunto. El ejemplo de una arquitectura de red se ilustra en la Fig. 7. Los ejemplos del entrenamiento se describen además en relación con las Figs. 12 y 13. Cada una de las imágenes 642 de ojos distorsionadas se introduce en una subred. Estas imágenes 642 se pueden introducir en conjunto (p. ej., las dos imágenes como entradas paralelas) o por separado (p. ej., una entrada de imagen a la vez, donde la primera subred predeciría la dirección de mirada de esa imagen). En respuesta, la subred genera vectores 654 de mirada 2D por ojo (p. ej., correspondiente al ojo del usuario que se muestra en la imagen de entrada). Cada vector de mirada 2D se puede expresar como un origen o_{2D} de mirada 2D (p. ej., la ubicación 2D del ojo del usuario en el plano de la imagen) y una dirección d_{2D} de mirada 2D. La imagen 644 distorsionada del rostro se introduce también en una subred que, en respuesta, la corrección c 652 de distancia. Por ende, la salida de la red neuronal incluye cinco componentes: un origen o_{2D} de mirada 2D y una dirección d_{2D} de mirada 2D por ojo y una corrección c de distancia.

La corrección 652 de distancia, la distancia 632 estimada y los vectores 654 de mirada 2D se introducen en el generador 660 de mirada 3D. En respuesta, el generador 660 de mirada 3D genera y emite una mirada 3D 662. En un ejemplo, la mirada 3D 662 incluye una dirección de mirada 3D por ojo de usuario (que puede expresarse en el espacio real 3D) y una posición 3D del ojo del usuario (que también puede expresarse en el espacio real 3D).

Para ilustrar, una distancia d_{corr} corregida se genera como $d_{corr} = d_{aproximada} \times c$. Con referencia a uno de los ojos y su vector de mirada 2D, el generador 660 de mirada 3D calcula su mirada 3D basándose en la distancia corregida y el mapeo espacial 2D a 3D (p. ej., las matrices relevantes). Por ejemplo, el origen $o_{3D,N}$ de mirada 3D en el espacio 3D normalizado se calcula como $o'_{3D,N} = M^1 o_{2D}$ y $o_{3D,N} = o'_{3D,N} / \|o'_{3D,N}\| d_{corr}$. Para generar la dirección de mirada 3D en el espacio 3D normalizado, primero se genera un vector (X,Y,Z) base normalizado, donde $Z' = M^1 o_{2D}$, $Y' = [0,1,0]^T \times Z'$, y $X' = Y' \times Z'$ y $X = X' / \|X'\|$, $Y = Y' / \|Y'\|$, y $Z = Z' / \|Z'\|$. Una dirección $d_{3D,N}$ de mirada 3D normalizada se genera como $d'_{3D,N} = [X, Y] d_{2D} - Z$ y $d_{3D,N} = d'_{3D,N} / \|d'_{3D,N}\|$. El original de mirada 3D y la dirección de mirada 3D se asignan desde el espacio 3D normalizado al espacio 3D real basándose en la matriz R de rotación. Por ejemplo, el origen o_{3D} de mirada 3D en el espacio 3D real se calcula como $o_{3D} = R^1 o_{3D,N}$. Del mismo modo, la dirección d_{3D} de mirada 3D en el espacio 3D real se calcula como $d_{3D} = R^1 d_{3D,N}$.

La Fig. 7 ilustra una arquitectura de red de ejemplo para una red neuronal, según una realización. Como se ilustra en la Fig. 7, se introducen en la red tres imágenes 710-714: una imagen 710 distorsionada del rostro a baja resolución, una imagen 712 distorsionada del ojo izquierdo a alta resolución y una imagen 714 distorsionada del ojo derecho a alta resolución (donde, “baja” y “alta” son términos relativos, de tal forma que “baja resolución” se refiere a ser más baja que la “alta resolución”). Las imágenes 712 y 714 distorsionadas de los ojos se centran en las detecciones del ojo, $x = e_{izquierdo}$ o $e_{derecho}$ escaladas a una distancia interocular de $s = 320$ píxeles y recortada a 224×112 píxeles. La imagen 714 del ojo derecho es especular a 716 modificando la matriz R de rotación. Esto proporciona a la red una apariencia consistente. La imagen 710 distorsionada del rostro se centra en el punto medio entre las detecciones de los ojos, se escala a $a = 84$ píxeles y se recorta a 224×56 píxeles.

La red incluye redes 720-726 neuronales convolucionales (CNN) separadas para los ojos (que se muestran como CNN 722 para el ojo izquierdo y CNN 724 para el ojo derecho, con ponderaciones vinculadas) y el rostro (que se muestra como CNN 720). Ambas son la parte convolucional de ResNet-18, similarmente a lo que se describe en K. He, X. Zhang, S. Ren y J. Sun. “Aprendizaje residual profundo para el reconocimiento de imágenes”, CoRR, abs/1512.03385, 2015. La salida de todas las CNN, 720-726 de ambos ojos y del rostro, se concatenan como una capa 730 de concatenación y se alimentan a un módulo 740 completamente conectado, que predice una corrección c de distancia.

La salida de cada ojo CNN 722 y 724 se concatena con un conjunto de “n” parámetros de calibración personal y la corrección c de distancia. La concatenación para el ojo izquierdo y el ojo derecho se realiza a través de una capa 732 de concatenación y una capa 734 de concatenación, respectivamente. El vector de características combinado que resulta de cada concatenación se alimenta a un módulo completamente conectado (mostrado como un módulo 742 totalmente conectado para el ojo izquierdo y un módulo 744 totalmente conectado para el ojo derecho).

Los módulos 720-724 totalmente conectados se pueden describir como: FC(3072)-BN-ReLU-DO(0,5)-FC(3072)-BN-ReLU-DO(0,5)-FC({4, 1}). El resultado final es el origen de mirada 2D y la dirección de mirada 2D, o la corrección c de distancia.

No es necesario proporcionar los “n” parámetros de calibración al componente de estimación de distancia (p. ej., la capa 739 de concatenación), ya que normalmente es difícil detectar errores de distancia a partir de los datos de calibración recopilados a una distancia, que es lo que suele estar disponible.

La red está entrenada para minimizar la distancia perdida mínima media entre las líneas de mirada predichas y los puntos de estímulo verdaderos fundamentales, como se ilustra en las siguientes figuras. También se penalizan los orígenes de la mirada que están fuera de las imágenes oculares y la corrección de distancia superior al cuarenta por ciento en cualquier dirección.

La Fig. 8 ilustra una normalización de imagen de ejemplo, según una realización. Como se ilustra, un espacio 3D real está centrado alrededor de una cámara que genera una imagen 830 real. El espacio 3D 810 se puede definir como un sistema de coordenadas 3D. Un espacio 3D 820 normalizado se centra también alrededor de la cámara y se genera a partir del espacio 3D 810 real basándose en una matriz R de rotación. Por ejemplo, el espacio 3D 820 normalizado corresponde a una rotación del espacio 3D 810 real a lo largo de uno del eje (p. ej., el eje Y) del espacio 3D 810 real. Los centros entre los ojos del usuario forman una línea. Esa línea es paralela a un eje en el espacio 3D 820 normalizado (p. ej., el eje X). Una imagen 840 normalizada (que se muestra como correspondiente a una imagen distorsionada del ojo izquierdo) se proyecta desde la imagen 830 real en el espacio 3D normalizado (que se muestra a una distancia y paralela a la imagen 830 real). En el caso en que esta imagen 840 normalizada corresponda a una ubicación donde distancia interocular entre los dos ojos sea la ID humana promedio de 63 mm, la distancia entre la cámara y el centro del ojo en la imagen 840 normalizada es la distancia aproximada.

La Fig. 9 ilustra una predicción de mirada 3D de ejemplo, según una realización. Como se ilustra, se utiliza el espacio 3D 820 normalizado de la Fig. 8, aunque la predicción se puede mapear en el espacio 3D 810 real usando la inversa de la matriz R de rotación como se describe en relación con la Fig. 6. Un origen 910 de mirada 3D se deriva de un origen 920 de mirada 2D basándose en la inversa de una matriz M de escala y una distancia 930 corregida. El origen 920 de mirada 2D y una dirección 940 de mirada 2D representan un vector de mirada 2D en el plano de la imagen 840 normalizada. Una dirección 950 de mirada 3D se deriva del origen 920 de mirada 2D y de la dirección 940 de mirada 2D basándose en la inversa de la matriz M de escala como se describe en relación con la Fig. 6.

Durante el entrenamiento, se proyecta una línea de mirada 3D (o rayo de mirada 3D) desde el origen 910 de mirada 3D a lo largo de la dirección 950 de mirada 3D. Se conoce un punto 960 de estímulo para la mirada (p. ej., un punto de mirada conocido). La distancia 970 (p. ej., la distancia más corta) entre el punto 96 de estímulo y la línea de mirada 3D es un parámetro de distancia de la función de pérdida. La función de pérdida se minimiza minimizando esta distancia 970 (p. ej., si la red neuronal predice correctamente la mirada 3D, la distancia 970 sería cero y el punto 960 de estímulo caería en la línea de mirada 3D).

La Fig. 10 ilustra predicciones para vectores de mirada 2D y corrección de distancia generadas por una red neuronal, según una realización. La Fig. 10 muestra tres imágenes 1010, 1020, 1030 que se introdujeron en la red neuronal. Las imágenes 1010 y 1020 son imágenes distorsionadas de los ojos, mientras que la imagen 1030 es una imagen distorsionada del rostro. Además, la imagen 1020 (p. ej., del ojo izquierdo) se refleja de tal forma que las dos imágenes 1010 y 1020 están alineadas (p. ej., el canto interno del ojo izquierdo se alinea con el canto interno del ojo derecho y el canto externo del ojo izquierdo se alinea con el canto externo del ojo derecho).

Los vectores de mirada 2D predichos se superponen sobre las imágenes 1010 y 1020. Como se muestra en la superposición de la imagen 1010 (y, similarmente, en la superposición de la imagen 1020), el vector de mirada 2D tiene un origen 1012 de mirada (p. ej., un reflejo en el ojo derecho) y una dirección 1014 de mirada. Las distancias 1032 estimadas y corregidas se superponen en la imagen 1030. Como se muestra, la distancia estimada es de seiscientos cincuenta y tres milímetros y la distancia corregida es de cuatrocientos setenta y cinco milímetros, lo que representa una corrección de distancia predicha de alrededor del setenta y dos por ciento.

La Fig. 11 ilustra un flujo de ejemplo para la predicción de mirada 3D basándose en un sistema de aprendizaje profundo como el sistema 600 de aprendizaje profundo, según una realización. Se describe un sistema de seguimiento ocular que realiza las operaciones del flujo de ejemplo. En un ejemplo, el sistema de seguimiento ocular aloja el sistema de aprendizaje profundo. En otro ejemplo, un sistema informático remoto aloja el sistema de aprendizaje profundo y el sistema de seguimiento ocular interactúa con este sistema informático remoto a través de una red de datos para proporcionar una imagen 2D y recibir una predicción de mirada 3D. En otro ejemplo más, el sistema de aprendizaje profundo se distribuye entre el sistema de seguimiento ocular y el sistema informático remoto (p. ej., el sistema informático remoto puede alojar la red neuronal mientras que el sistema de seguimiento ocular puede alojar el componente restante del sistema de aprendizaje profundo).

Las instrucciones para realizar las operaciones del flujo ilustrativo pueden almacenarse como instrucciones legibles por ordenador en un medio no transitorio legible por ordenador del sistema de seguimiento ocular. Tal como están almacenadas, las instrucciones representan módulos programables que incluyen código o datos ejecutables por uno o más procesadores del sistema de seguimiento ocular. La ejecución de dichas instrucciones configura el sistema de seguimiento ocular para realizar las operaciones específicas que se muestran en la figura y se describen en la presente memoria. Cada módulo programable en combinación con el procesador representa un medio para realizar una o varias operaciones respectivas. Si bien las operaciones se ilustran en un orden particular, debe entenderse que no es necesario un orden particular y que una o más operaciones pueden omitirse, saltarse y/o reordenarse.

El flujo de ejemplo comienza en la operación 1102, donde el sistema de seguimiento ocular recibe una imagen (p. ej., una imagen 2D). En un ejemplo, la imagen es generada por una cámara asociada con el sistema de seguimiento ocular (p. ej., una cámara integrada o en interfaz con dicho sistema). La imagen muestra un ojo de usuario (p. ej., cuando la cámara está asociada con dos ojos de usuario, la imagen muestra ambos ojos; cuando la cámara está asociada con un ojo de usuario como en el caso de un visor de realidad virtual, la imagen muestra solo ese ojo de usuario).

En la operación 1104, el sistema de seguimiento ocular genera una distorsión alrededor del ojo del usuario. Por ejemplo, el sistema de seguimiento ocular detecta el ojo del usuario en la imagen y estima una distancia aproximada entre la cámara y el ojo del usuario según se detecta en la imagen. El sistema de seguimiento ocular también proyecta la imagen en un espacio 3D basándose en una matriz K intrínseca de la cámara, rota la imagen proyectada basándose en una matriz R de rotación y escala la imagen rotada basándose en una matriz M de escala. La matriz M de escala se genera basándose en una distancia s predefinida en píxeles y la distancia interocular humana promedio. Después, el sistema de seguimiento ocular deforma la imagen escalada mediante interpretación bilineal le recorta la imagen distorsionada alrededor del ojo del usuario en función de una región predefinida (p. ej., una que tiene una anchura y una altura predefinidas). Esta imagen tiene una primera resolución de imagen que depende de la distancia s predefinida.

En la operación 1106, el sistema de seguimiento ocular genera una segunda imagen distorsionada alrededor de un segundo ojo de usuario a partir de la imagen. Esta operación se realiza si la imagen muestra el segundo ojo del usuario y, de lo contrario, se omite. La operación 1106 es similar a la operación 1104. La segunda imagen distorsionada muestra el segundo ojo del usuario y tiene la primera resolución de imagen.

En la operación 1108, el sistema de seguimiento ocular genera una tercera imagen distorsionada alrededor del usuario a partir de la imagen. Esta operación también se realiza si la imagen muestra el segundo ojo del usuario y, de lo contrario, se omite. La es similar a la operación 1104, donde la escala se basa en una segunda distancia s predefinida (y, por lo tanto, una segunda matriz M) de escala para lograda una segunda resolución de imagen. La tercera imagen distorsionada muestra el rostro del usuario o al menos ambos ojos del usuario y tiene la segunda resolución de imagen más baja que la primera resolución de imagen.

En la operación 1110, el sistema de seguimiento ocular introduce en una red neuronal la imagen distorsionada del ojo del usuario. Si se generan las otras dos imágenes distorsionadas (p. ej., del segundo ojo del usuario y del rostro del usuario), se introducen también a la red neuronal. La red neuronal ya está entrenada y predice (i) una corrección c de distancia y (ii) un origen de mirada 2D y una dirección de mirada 2D por ojo en la imagen de ojo distorsionada asociada. En general, la corrección c de distancia se utiliza para corregir la distancia aproximada estimada y se predice basándose en la imagen distorsionada y, si están disponibles, la segunda y la tercera imágenes distorsionadas. El origen de mirada 2D y la dirección de mirada 2D para el ojo del usuario (y, similarmente, para el segundo ojo del usuario) se genera a partir de la imagen distorsionada (y, similarmente, de la segunda imagen distorsionada) por separado de la segunda y la tercera imagen distorsionada (y, similarmente, por separado de las imágenes distorsionadas y de las terceras imágenes distorsionadas).

En un ejemplo, la red neuronal se entrena también basándose en “ n ” parámetros de calibración que se integran en la red neuronal. Durante una calibración del sistema de seguimiento ocular, este sistema genera una pluralidad de imágenes de calibración instruyendo al usuario para que mire a puntos de mirada conocidos. Las imágenes de calibración también se normalizan y distorsionan y se introducen a la red neuronal. Los “ n ” parámetros de calibración se ajustan de manera que se minimice la función de pérdida de la red neuronal. Esta minimización utiliza los puntos de mirada conocidos como verdad fundamental. Una vez que el sistema de seguimiento ocular recibe la imagen en la operación 1102 y la procesa a través de las operaciones 1106-110, la predicción de la red neuronal en la operación 1110 usa los “ n ” parámetros de calibración.

En la operación 1112, el sistema de seguimiento ocular recibe de la red neuronal la corrección c de distancia. El sistema de seguimiento ocular recibe también de la red neuronal el origen de mirada 2D y la dirección de mirada 2D para el ojo del usuario y, según corresponda, el origen de mirada 2D y la dirección de mirada 2D para el segundo ojo del usuario.

En la operación 1114, el sistema de seguimiento ocular genera una distancia corregida entre el ojo del usuario y la cámara actualizando al menos la distancia aproximada estimada basándose en la corrección c de distancia. Por ejemplo, la corrección c de distancia es un factor de corrección y el sistema de seguimiento ocular multiplica la distancia aproximada estimada por este factor para generar la distancia corregida.

En la operación 1116, el sistema de seguimiento ocular genera información de mirada 3D para el ojo del usuario a partir de la distancia corregida y su origen de mirada 2D y dirección de mirada 2D. La información de la mirada 3D incluye una posición en el espacio 3D del ojo del usuario (p. ej., el origen de mirada 3D) y una dirección de mirada 3D que se origina en el ojo del usuario. En un ejemplo, el origen de mirada 3D y la dirección de mirada 3D se derivan de la distancia corregida, el origen de mirada 2D y la dirección de mirada 2D basándose en la matriz M de

escala y la matriz R de rotación (o su inversa). Esta operación también se puede repetir para generar un origen de mirada 3D y una dirección de mirada 3D para el segundo ojo del usuario basándose en la distancia corregida y su origen de mirada 2D y dirección de mirada 2D. En este caso, la información de la mirada 3D incluye los orígenes de la mirada 3D y las direcciones de la mirada 3D de los dos ojos del usuario..

En la operación 1120, el sistema de seguimiento ocular proporciona la información de mirada 3D generada para el ojo del usuario y, según corresponda, la información de mirada 3D generada para el segundo ojo del usuario a una aplicación de mirada 3D. La aplicación de mirada 3D utiliza la información de mirada 3D para admitir visualizaciones estereoscópicas y aplicaciones 3D (p. ej., controles 3D y manipulaciones de objetos visualizados en una pantalla).

Se entrena una red neuronal de un sistema de aprendizaje profundo en función de imágenes de entrenamiento para predecir una corrección de distancia y un vector de mirada 2D (p. ej., origen de mirada 2D y dirección de mirada 2D) por ojo de usuario. Generalmente, el entrenamiento es iterativo a través de las imágenes de entrenamiento para minimizar una función de pérdida y, en consecuencia, actualizar los parámetros de la red neuronal a través de la propagación hacia atrás (p. ej., una que usa descenso de gradiente). Debido a que la red neuronal debe predecir dos resultados (la corrección de distancia y el vector de mirada 2D), depender de imágenes de entrenamiento que muestren los ojos del usuario mientras observa puntos de mirada dentro de planos 2D de las cámaras asociadas que capturaron estas imágenes es insuficiente. En cambio, se necesita la diversidad de las ubicaciones de los puntos de mirada en relación con los planos 2D para un entrenamiento adecuado. Este problema en el entrenamiento y la solución se describen en más detalle en relación con la Fig. 12.

La Fig. 12 ilustra una estrategia para usar imágenes de entrenamiento que están asociadas con una diversidad de ubicaciones de puntos de mirada en relación con planos 2D de cámaras que generaron las imágenes de entrenamiento, según una realización. En aras de la claridad de la explicación, se hace referencia a un único ojo de usuario en relación con la Fig. 12. Sin embargo, las imágenes de entrenamiento que muestran dos ojos de usuario son igualmente aplicables. Los ojos del usuario capturados en las imágenes de entrenamiento también se denominan aquí “ojos de usuario mostrados”. Las cámaras utilizadas para generar las imágenes de entrenamiento también se denominan en el presente documento “cámaras de entrenamiento”. Como se ilustra, una cámara 1210 genera imágenes de entrenamiento que muestran el ojo de un usuario. Cada imagen de entrenamiento es una imagen 2D capturada mientras el ojo del usuario observaba un punto de mirada.

Por ejemplo, una primera imagen de entrenamiento muestra al usuario mientras observa un punto 1220 de mirada. Si esa primera entrada de entrenamiento fue una entrada a la red neuronal durante su entrenamiento, la red neuronal puede encontrar incorrectamente la dirección de mirada del ojo del usuario al punto 1220 de mirada y la distancia del ojo del usuario a la cámara 1210. Como se ilustra en la Fig. 12, puede haber múltiples soluciones para la dirección y la distancia de mirada (la Fig. 12 muestra dos de estas soluciones). En una primera solución, la red neuronal puede predecir que el ojo del usuario (que se muestra como el elemento 1230 en esta solución) está mirando en un primer ángulo “a1” 1232 con respecto al punto 1220 de mirada y está a una primera distancia “d1” 1234 con respecto a la cámara 1210. En una segunda solución, la red neuronal puede predecir que el ojo del usuario (que se muestra como elemento 1240 en esta solución) está mirando en un segundo ángulo “a2” 1242 con respecto al punto 1220 de mirada y está a una segunda distancia “d2” 1244 con respecto a la cámara 1210. Ambas soluciones son posibles, pero solo una es correcta. Por ende, la red neuronal puede predecir la solución incorrecta, por lo que el entrenamiento no sería adecuado y daría como resultado una red neuronal entrenada que generaría predicciones inexactas.

Para evitar este potencial, se utiliza una estrategia para la diversidad de ubicaciones de puntos de observación. Bajo esta estrategia, se utiliza adicionalmente una segunda imagen de entrenamiento del ojo del usuario. En esta segunda imagen de entrenamiento, se cambia el ángulo de mirada o la distancia entre el ojo del usuario y la cámara. En consecuencia, cuando se utilizan ambas imágenes de entrenamiento, el conjunto de posibles soluciones se reduce a una sola solución (p. ej., la primera solución o la segunda solución en el ejemplo anterior).

Son posibles distintas formas para lograr esta diversidad. Generalmente, las imágenes de entrenamiento pueden incluir dos conjuntos de imágenes de entrenamiento. El primer conjunto es para entrenar imágenes de los ojos del usuario mirando puntos de mirada dentro de los planos 2D de las cámaras. El segundo conjunto es para entrenar imágenes de los ojos del usuario mirando puntos de mirada fuera de los planos 2D de las cámaras. De esta forma, cuando las diversas imágenes de entrenamiento se introducen en la red neuronal, la red neuronal aprende a encontrar las soluciones correctas.

La Fig. 12 ilustra una forma específica. Como se ilustra, la segunda imagen de entrenamiento se genera para el ojo del usuario mientras mira un segundo punto 1250 de mirada a lo largo del mismo ángulo de mirada que en la primera imagen de entrenamiento. De esta forma, la única solución posible es la primera solución (p. ej., el primer ángulo “a1” 1232 y la primera distancia “d1” 1234). Otra forma es usar el mismo punto 1220 de mirada para la segunda imagen de entrenamiento, pero cambiar la distancia entre el ojo del usuario y la cámara 1210 (p. ej., acercando o alejando la cámara del ojo del usuario mientras el ojo del usuario mantiene su mirada en el punto 1220 de mirada). Otra forma más es mantener la misma distancia ojo-cámara del usuario, introducir un segundo punto de mirada (no necesariamente en el mismo ángulo de mirada que en la forma específica ilustrada en la Fig. 12) y

generar la segunda imagen de entrenamiento mientras el ojo del usuario está mirando a ese segundo punto de mirada. Por supuesto, se puede usar una combinación de estas diferentes formas para generar múltiples imágenes de entrenamiento para diferentes ojos de usuario, diferentes ángulos de mirada y diferentes distancias entre el ojo del usuario y la cámara. La misma cámara (o configuración de cámara para varias cámaras) puede usarse, pero no es necesario, para generar las imágenes de entrenamiento porque el entrenamiento no depende de una cámara o configuración de pantalla específicas.

Una vez que se generan las imágenes de entrenamiento, se introducen a la red neuronal para el entrenamiento. En particular, la red neuronal predice los ángulos de mirada y las distancias del ojo del usuario a la cámara a partir de estas imágenes de entrenamiento. La función de pérdida se puede definir basándose en los ángulos de mirada y las distancias entre el ojo del usuario y la cámara. En aras de la claridad, considere la primera imagen de entrenamiento descrita anteriormente en la presente memoria (p. ej., generada cuando el ojo del usuario miraba el punto 1220 de mirada). Basándose en esa imagen de entrenamiento (y la segunda imagen de entrenamiento), la red neuronal predice la primera solución (p. ej., el primer ángulo "a1" 1232 y la primera distancia "d1" 1234). Supongamos que el primer ángulo "a1" 1232 y la primera distancia "d1" 1234 son el ángulo y la distancia reales de la mirada, que se conocen en el entrenamiento (p. ej., son medidas de verdad fundamental que se pueden almacenar en una etiqueta de entrenamiento asociada con la primera imagen de entrenamiento). Supongamos también que la predicción no es completamente precisa (p. ej., el ángulo predicho se desvía del primer ángulo "a1" 1232 y la distancia predicha se desvía de la primera distancia "d1" 1234). La función de pérdida incluye un término de distancia y un término de ángulo. El término de distancia es la diferencia entre la distancia predicha y la distancia verdadera fundamental (p. ej., la primera distancia "d1" 1234 real). El término de ángulo es la diferencia entre el ángulo de mirada predicho y el ángulo de mirada verdadero fundamental (p. ej., el primer ángulo "a1" 1232 real). El objetivo del entrenamiento es actualizar los parámetros de la red neuronal de modo que se minimice su función de pérdida, donde minimizar la función de pérdida incluye minimizar el término de ángulo y el término de distancia de modo que el ángulo predicho esté lo más cerca posible del ángulo verdadero fundamental y la distancia pronosticada esté lo más cerca posible de la distancia verdadera fundamental.

En un ejemplo, este término de ángulo y el término de distancia se pueden reemplazar con un solo término de distancia. Por ejemplo, una línea de mirada predicha se genera a la distancia predicha de la cámara y tiene el ángulo de mirada predicho. Se mide la distancia (p. ej., la distancia más corta) entre el punto 1220 de mirada y la línea de mirada predicha (esto se muestra en la Fig. 9 como la distancia 970). Si la predicción fuera completamente precisa, esta distancia entre el punto de mirada y la línea de mirada predicha sería cero. Sin embargo, en otros casos, la función de pérdida se minimiza (y los parámetros de la red neuronal se actualizan), minimizando la distancia entre el punto de mirada y la línea de mirada predicha, de tal forma que la línea de mirada predicha esté lo más cerca posible del punto 1220 de mirada.

En este ejemplo también, la función de pérdida puede incluir un término de penalización. Más específicamente, cuando el origen de una línea de mirada predicha cae fuera de la imagen de entrenamiento correspondiente (p. ej., volviendo al ejemplo anterior, si "d1" 1234 es lo suficientemente grande como para que el ojo 1230 esté fuera de la primera imagen de entrenamiento), se añade una penalización a la función de pérdida (p. ej., la penalización tiene un valor predefinido). De la misma manera, si una corrección de distancia pronosticada supera un cierto umbral en cualquier dirección (p. ej., cuarenta por ciento en cualquier dirección), se añade la misma penalización u otra a la función de pérdida.

La Fig. 13 ilustra un flujo de ejemplo para entrenar una red neuronal, según una realización. La Fig. 14 ilustra también un flujo de ejemplo adicional para el entrenamiento, donde este entrenamiento utiliza parámetros de calibración incorporados según unas realizaciones. Se describe un sistema informático que realiza las operaciones de los flujos de ejemplo. En un ejemplo, el sistema informático realiza el entrenamiento y almacena el código de la red neuronal. Una vez finalizado el entrenamiento, el sistema informático puede recibir imágenes del sistema de seguimiento ocular y utilizar la red neuronal para responder con predicciones de la mirada 3D. De forma adicional o alternativamente, los sistemas de seguimiento ocular pueden descargar el código de la red neuronal del sistema informático.

Las instrucciones para realizar las operaciones de los flujos ilustrativos pueden almacenarse como instrucciones legibles por ordenador en un medio no transitorio legible por ordenador del sistema informático. Tal como están almacenadas, las instrucciones representan módulos programables que incluyen código o datos ejecutables por uno o más procesadores del sistema informático. La ejecución de dichas instrucciones configura el sistema informático del ojo para realizar las operaciones específicas que se muestran en las figuras y se describen en la presente memoria. Cada módulo programable en combinación con el procesador representa un medio para realizar una o varias operaciones respectivas. Si bien las operaciones se ilustran en un orden particular, debe entenderse que no es necesario un orden particular y que una o más operaciones pueden omitirse, saltarse y/o reordenarse.

El ejemplo de flujo de la Fig. 13 comienza en la operación 1302, donde el sistema informático accede a imágenes de entrenamiento que incluyen un primer conjunto de imágenes de entrenamiento y un segundo conjunto de imágenes de entrenamiento. Algunas o todas las imágenes de entrenamiento en el primer conjunto muestran ojos de usuario asociados con puntos de mirada en un plano de una cámara. Algunas o todas las imágenes de entrenamiento del

segundo conjunto muestran los ojos del usuario asociados con puntos de mirada fuera del plano de la cámara. De esta manera, se logra diversidad de imágenes de entrenamiento que muestran los ojos del usuario mirando a los puntos de mirada dentro y fuera del avión.

En la operación 1304, el sistema informático entrena la red neuronal basándose en imágenes de entrenamiento. Generalmente, el entrenamiento incluye la actualización de parámetros de la red neuronal (p. ej., ponderaciones de conexiones entre nodos a través de capas de la red neuronal) para minimizar una función de pérdida de la red neuronal. La función de pérdida puede usar un término de ángulo y/o un término de distancia, como se ha explicado anteriormente en la presente memoria en relación con la Fig. 12. En un ejemplo, la operación 1304 incluye dos operaciones 1306 y 1308 adicionales.

En la operación 1306, el sistema informático introduce una primera imagen de entrenamiento y una segunda imagen de entrenamiento desde las imágenes de entrenamiento a la red neuronal. La primera imagen de entrenamiento pertenece al primer conjunto de imágenes de entrenamiento. La segunda imagen de entrenamiento pertenece al segundo conjunto de imágenes de entrenamiento. En un ejemplo específico, la primera imagen de entrenamiento muestra el ojo del usuario mientras mira a un punto de mirada según un ángulo de mirada y la segunda imagen de entrenamiento muestra el ojo del usuario mientras mira a otro punto de mirada según el ángulo de mirada. En otro ejemplo específico, la primera imagen de entrenamiento y la segunda imagen de entrenamiento muestran el ojo del usuario mientras mira el punto de mirada en un ángulo de mirada. En este ejemplo, la primera imagen de entrenamiento corresponde a una primera distancia entre la cámara y el ojo del usuario, y la segunda imagen de entrenamiento corresponde a una segunda distancia entre la cámara y el ojo del usuario. En otro ejemplo específico más, la primera imagen de entrenamiento muestra el ojo del usuario mientras mira el punto de mirada según un ángulo de mirada y la segunda imagen de entrenamiento muestra el ojo del usuario mientras mira otro punto de mirada según un ángulo de mirada diferente. En estos ejemplos, la primera y la segunda imagen de entrenamiento se pueden introducir como un par.

Como se usa en la presente memoria en relación con el flujo de la Fig. 13, una imagen de entrenamiento se refiere a un conjunto de imágenes de entrenamiento que incluye una imagen distorsionada que muestra un ojo de una persona y, opcionalmente, una imagen distorsionada del otro ojo de la persona y una imagen distorsionada del rostro de la persona. En otras palabras, las imágenes utilizadas en el entrenamiento son de tipos similares a las imágenes que se utilizarían al finalizar el entrenamiento (p. ej., como se ilustra en relación con la Fig. 9) y se pueden generar siguiendo operaciones similares de proyección, rotación, escalado y recorte.

En la operación 1308, el sistema informático minimiza la función de pérdida de la red neuronal basándose en la distancia entre un punto de mirada y una línea de mirada. El punto de mirada está asociado con una de la primera imagen de entrenamiento o la segunda imagen de entrenamiento. La línea de mirada se predice por la red neuronal para un ojo de usuario a partir de la primera imagen de entrenamiento o la segunda imagen de entrenamiento. Las operaciones 1306-1308 se repiten en muchas, si no en todas, las imágenes de entrenamiento para completar el entrenamiento de la red neuronal.

Volviendo a la Fig. 14, el entrenamiento de la red neuronal puede implicar también parámetros de calibración, de tal forma que cuando la red neuronal predice información de mirada (como se describe anteriormente en la presente memoria), la predicción se basa en tales parámetros. Un diseñador de la red neuronal y/o un sistema de seguimiento ocular no necesita especificar los parámetros de calibración. En cambio, los parámetros de calibración están integrados y es suficiente que el diseñador especifique el número de estos parámetros. Durante el entrenamiento, los parámetros de calibración se inicializan y después se actualizan junto con los parámetros de la red neuronal (p. ej., las ponderaciones entre las conexiones en la red) basándose en la minimización de la función de pérdida de la red neuronal. De esta manera, los parámetros de la red neuronal se actualizan durante el entrenamiento basándose también en los parámetros de calibración. Durante el entrenamiento de la red neuronal, algunas de las operaciones del flujo de la Fig. 14 se puede implementar y utilizar junto con las operaciones de la Fig. 13. Al finalizar el entrenamiento, las operaciones restantes del flujo de la Fig. 14 se puede implementar y utilizar junto con las operaciones de la Fig. 11. En este caso, los parámetros de calibración del usuario se establecen en función de minimizar la función de pérdida dado un conjunto de imágenes de calibración. Estos parámetros de calibración luego se introducen a la red neuronal en apoyo de la predicción de mirada 3D.

Si bien la Fig. 14 describe el aprendizaje de los parámetros de calibración dentro del contexto de una red neuronal que predice la información de mirada, las realizaciones no se limitan como tales. Generalmente, las operaciones de la Fig. 14 se aplican similarmente en relación con una red neuronal que está asociada con un sistema y que está entrenada para una tarea (p. ej., predecir un resultado particular), donde la operación adecuada del sistema implica la calibración. Asimismo, en aras de la claridad de la explicación, el ejemplo de flujo de la Fig. 14 se describe en relación con una imagen de entrenamiento, una imagen de calibración y una imagen. Sin embargo, el flujo de ejemplo se aplica similarmente a múltiples imágenes de entrenamiento, múltiples imágenes de calibración y múltiples imágenes.

Como se ilustra, el ejemplo de flujo de la Fig. 14 comienza en la operación 1402, donde el sistema informático accede a una imagen de entrenamiento asociada con una persona. La imagen de entrenamiento muestra un rostro

y/o un ojo(s) de la persona. Generalmente, la persona tiene un índice “i” y la imagen de entrenamiento también se puede indexar con índice “i” (p. ej., la etiqueta o los datos de metadatos de la imagen de entrenamiento incluyen el índice “i”). La imagen de entrenamiento puede estar disponible desde un almacén de datos y acceder desde allí como parte del proceso de entrenamiento. El número “n” es un número entero mayor que cero que especifica un diseñador de la red neuronal. En un ejemplo, el número “n” está entre dos y diez. En un ejemplo específico, el número “n” es tres.

En la operación 1404, el sistema informático inicializa “n” parámetros de calibración para la persona. Por ejemplo, los valores de estos parámetros de calibración se establecen en cero o en algún otro primer valor.

En la operación 1406, el sistema informático introduce la imagen de entrenamiento y los “n” parámetros de calibración en la red neuronal. Por ejemplo, la imagen de entrenamiento se introduce a la CNN Res18 relevante de la red neuronal, mientras que los “n” parámetros de calibración se introducen a las capas de concatenación asociadas con módulos completamente conectados de la red neuronal donde estos módulos son responsables de predecir los orígenes de mirada 2D y las direcciones de mirada 2D.

En la operación 1408, el sistema informático actualiza los “n” parámetros de calibración y los parámetros de la red neuronal (denominados en la presente memoria como parámetros de red e incluyen, por ejemplo, ponderaciones de conexiones entre nodos de la red neuronal) basándose en la minimización de una función de pérdida de la red neuronal. La función de pérdida se minimiza en función de la imagen de entrenamiento y los “n” parámetros de calibración a través de la retropropagación. En consecuencia, los primeros valores (p. ej., ceros) de los “n” parámetros de calibración se actualizan a segundos valores.

Las operaciones 1402-1408 pueden implementarse como parte del entrenamiento de la red neuronal por parte del sistema informático. Una vez completado el entrenamiento, la red neuronal puede estar disponible para que la use un usuario de un sistema de seguimiento ocular. Cuando se usa el sistema de seguimiento ocular, este sistema puede calibrarse calibrando la red neuronal específicamente para el usuario.

En la operación 1410, el sistema informático inicializa los “n” parámetros de calibración para el usuario. Esta operación es similar a la operación 1404. En un ejemplo, la inicialización puede, pero no necesita ser, a los mismos primeros valores (p. ej., ceros). En otro ejemplo, la inicialización es a los valores actualizados (p. ej., los segundos valores) de los “n” parámetros de calibración como se determina en la operación 1408.

En la operación 1412, el sistema informático genera una imagen de calibración para el usuario. Por ejemplo, el sistema informático indica al usuario que mire a un punto de mirada conocido y genera la imagen de calibración en consecuencia. La imagen de calibración muestra un ojo (o dos ojos, o un rostro como aplicación) del usuario basándose en los datos de imagen generados por una cámara asociada con el sistema de seguimiento ocular.

En la operación 1414, el sistema informático introduce la imagen de calibración y los parámetros de calibración en la red neuronal. Esta operación es similar a la operación 1406.

En la operación 1416, el sistema informático actualiza los “n” parámetros de calibración para el usuario sin actualizar los parámetros de red que se establecieron durante el entrenamiento. Los “n” parámetros de calibración se actualizan al menos minimizando la función de pérdida en función de la imagen de calibración y los “n” parámetros de calibración, donde la minimización no cambia los parámetros de la red. Aquí, la imagen de calibración tiene un punto de calibración conocido. El punto de calibración conocido se utiliza una base para la mirada del usuario en la minimización de la función de pérdida. El sistema informático actualiza iterativamente los “n” parámetros de calibración mientras mantiene los parámetros de la red a través de la retropropagación hasta que se minimiza la función de pérdida. En consecuencia, los valores inicializados (p. ej., los primeros valores) de los “n” parámetros de calibración realizados en la operación 1410 se actualizan a terceros valores. Típicamente, los terceros valores son diferentes de los segundos valores de los “n” parámetros de calibración según se determina en la operación 1408.

Las operaciones 1410-1416 pueden implementarse como parte de la calibración de la red neuronal por parte del sistema informático. Una vez finalizada la calibración, la red neuronal puede estar disponible para generar información de mirada 3D para el usuario.

En la operación 1418, el sistema informático introduce una imagen y los “n” parámetros de calibración actualizados para el usuario en la red neuronal. La imagen muestra el ojo (los dos ojos o el rostro, según corresponda) del usuario basándose en los datos de imagen adicionales generados por la cámara. Esta operación es similar a la operación 1406.

En la operación 1420, el sistema informático recibe una predicción de la red neuronal basándose en la imagen y los “n” parámetros de calibración. La predicción incluye una corrección de distancia, un origen de mirada 2D del ojo del usuario en la imagen y una dirección de mirada 2D del ojo del usuario en la imagen.

La invención se ha descrito ahora detalladamente a efectos de claridad y de comprensión. Sin embargo, se apreciará que pueden hacerse determinados cambios y modificaciones dentro del ámbito de las reivindicaciones adjuntas.

5 La descripción anterior solo proporciona realizaciones ilustrativas, y no se pretende que limite el ámbito, la aplicabilidad o la configuración de la invención. Más bien, la descripción anterior de las realizaciones ilustrativas proporcionará a los expertos en la técnica una descripción que les permitirá aplicar una o más realizaciones ilustrativas. Se entiende que pueden hacerse varios cambios en la función y disposición de los elementos sin alejarse del alcance de la descripción tal como se expone en la presente memoria.

10 Por ejemplo, cualquier detalle analizado con respecto a una realización puede estar presente o no en todas las versiones contempladas de esa realización. De la misma manera, cualquier detalle analizado con respecto a una realización puede estar presente o no en todas las versiones contempladas de otras realizaciones analizadas en la presente memoria. Finalmente, la ausencia del análisis de cualquier detalle con respecto a la realización de la presente memoria será un reconocimiento implícito de que tal detalle puede estar presente o no en cualquier versión de cualquier realización analizada en la presente memoria.

15 En la siguiente descripción, se dan detalles específicos para proporcionar una comprensión completa de las realizaciones. Sin embargo, un experto en la técnica entenderá que las realizaciones pueden ponerse en práctica sin estos detalles específicos. Por ejemplo, los circuitos, sistemas, redes, procesos y otros elementos de la invención pueden mostrarse como componentes en forma de diagrama de bloques para no complicar las realizaciones con detalles innecesarios. En otros casos, los circuitos, procesos, algoritmos, estructuras y técnicas muy conocidos pueden mostrarse sin detalles innecesarios para evitar complicar las realizaciones.

25 Además, cabe señalar que pueden describirse realizaciones individuales como un procedimiento que se ilustra como un flujograma, un diagrama de flujo, un diagrama de flujo de datos, un diagrama estructural o un diagrama de bloques. Aunque un diagrama de flujo puede describir las operaciones como un proceso secuencial, muchas de las operaciones pueden realizarse en paralelo o al mismo tiempo. Un proceso puede terminarse cuando se completan sus operaciones, pero podría tener etapas adicionales no analizadas o incluidas en una figura. Asimismo, no todas las operaciones de cualquier proceso particularmente descrito pueden ocurrir en todas las realizaciones. Un proceso puede corresponder a un método, una función, un procedimiento, una subrutina, un subprograma, etc. Cuando un proceso corresponde a una función, su terminación corresponde a un retorno de la función a la función de llamada o a la función principal.

35 La expresión “medio legible por máquina” incluye, pero no se limita a, dispositivos de almacenamiento transitorios y no transitorios, portátiles o fijos, dispositivos ópticos de almacenamiento, canales inalámbricos y otros diversos medios capaces de almacenar, contener o transmitir una o más instrucciones y/o datos. Un segmento de código o instrucciones ejecutables por máquina pueden representar un procedimiento, una función, un subprograma, un programa, una rutina, una subrutina, un módulo, un paquete de software, una clase o cualquier combinación de instrucciones, estructuras de datos o sentencias de programa. Un segmento de código puede acoplarse a otro segmento de código o a un circuito de hardware al pasar y/o recibir información, datos, argumentos, parámetros o contenido de memoria. La información, los argumentos, los parámetros, los datos, etc. pueden pasarse, enviarse o transmitirse a través de cualquier medio adecuado que incluye compartir la memoria, pasar el mensaje, pasar el identificador, transmisión de red, etc.

45 Además, pueden aplicarse realizaciones de la invención, al menos en parte, de modo manual o automática. Pueden ejecutarse implementaciones manuales o automáticas, o al menos asistidas, mediante el uso de máquinas, hardware, software, firmware, software personalizado, microcódigo, lenguajes descriptivos de hardware, o cualquier combinación de los mismos. Cuando se implementan en software, firmware, software personalizado o microcódigo, el código de programa o los segmentos de código para realizar las tareas necesarias pueden almacenarse en un medio legible por máquina. Uno o más procesadores pueden realizar las tareas necesarias.

REIVINDICACIONES

1. Un método implementado por ordenador que comprende:

5 entrenar, mediante un sistema informático, una red neuronal mediante al menos:

introducir (1402) una imagen de entrenamiento y un primer parámetro de calibración en la red neuronal, mostrando la imagen de entrenamiento un ojo de una persona, y
 10 actualizar (1408) el primer parámetro de calibración y un parámetro de red de la red neuronal basándose en la minimización de una función de pérdida de la red neuronal, en donde la función de pérdida se minimiza basándose en la imagen de entrenamiento y el primer parámetro de calibración;

15 al finalizar el entrenamiento, calibrar a través del sistema informático la red neuronal para un usuario mediante al menos:

introducir (1414) una imagen de calibración y un segundo parámetro de calibración en la red neuronal, mostrando la imagen de calibración un ojo del usuario basándose en datos de imagen generados por una cámara asociada con un sistema de seguimiento ocular del usuario, y
 20 actualizar (1416) el segundo parámetro de calibración sin actualizar el parámetro de red de la red neuronal, en donde el segundo parámetro de calibración se actualiza al menos minimizando la función de pérdida basándose en la imagen de calibración y el segundo parámetro de calibración; y

25 al finalizar la calibración, generar a través del sistema informático información de mirada tridimensional, 3D, para el usuario mediante al menos:

introducir (1418) una imagen y el segundo parámetro de calibración en la red neuronal, mostrando la imagen el ojo del usuario basándose en datos de imagen adicionales generados por la cámara, y
 30 recibir (1420) una predicción de la red neuronal basándose en la imagen y el segundo parámetro de calibración, en donde la predicción comprende una corrección de distancia, un origen de mirada bidimensional, 2D, del ojo del usuario en la imagen, y una dirección de mirada 2D del ojo del usuario en la imagen;
 35 comprendiendo además:
 generar, mediante el sistema informático, una distancia corregida entre el ojo del usuario y la cámara actualizando al menos una distancia estimada entre el ojo del usuario y la cámara basándose en la corrección de distancia, y en donde la información de mirada 3D para el ojo del usuario se genera basándose en el origen de mirada 2D, la dirección de
 40 mirada 2D y la distancia corregida.

2. El método implementado por ordenador de la reivindicación 1, en donde la imagen es una imagen 2D, y que comprende además:

45 determinar, a través del sistema informático, la distancia estimada entre la cámara y el ojo del usuario basándose en la imagen 2D; y
 estimar, a través del sistema informático, una posición del ojo del usuario en un espacio 3D basándose en la distancia corregida y en una posición de la cámara en el espacio 3D, en donde la información de mirada 3D comprende la posición del ojo en el espacio 3D.

3. El método implementado por ordenador de la reivindicación 2, que comprende además:
 55 estimar, mediante el sistema informático, una dirección de mirada 3D desde la posición del ojo en el espacio 3D basándose en el origen de mirada 2D y la dirección de mirada 2D, en donde la información de mirada 3D comprende la dirección de mirada 3D.

4. El método implementado por ordenador de la reivindicación 1, en donde introducir el primer parámetro de calibración comprende inicializar el primer parámetro de calibración a un primer valor, en donde actualizar el primer parámetro de calibración comprende actualizar el primer valor a un segundo valor.

- 60 5. El método implementado por ordenador de la reivindicación 4, en donde introducir el segundo parámetro de calibración comprende inicializar el segundo parámetro de calibración al primer valor, en donde actualizar el primer parámetro de calibración comprende actualizar el primer valor a un tercer valor.

6. El método implementado por ordenador de la reivindicación 4, en donde tanto el entrenamiento como la calibración utilizan “n” parámetros de calibración que incluyen el primer parámetro de calibración y el segundo parámetro de calibración, y en donde “n” es un número entero en el intervalo de dos a diez.
- 5 7. El método implementado por ordenador de la reivindicación 6, en donde “n” es igual a tres.
8. Un sistema informático que comprende:
un procesador; y
una memoria que almacena instrucciones legibles por ordenador que, al ser ejecutadas por el procesador,
10 configuran el sistema informático para realizar operaciones que comprenden:

entrenar una red neuronal mediante al menos:

introducir (1402) una imagen de entrenamiento y un primer parámetro de calibración en la
15 red neuronal, mostrando la imagen de entrenamiento un ojo de una persona, y
actualizar (1408) el primer parámetro de calibración y un parámetro de red de la red
neuronal basándose en la minimización de una función de pérdida de la red neuronal, en
donde la función de pérdida se minimiza basándose en la imagen de entrenamiento y el
primer parámetro de calibración;
20 y
al finalizar el entrenamiento, calibrar a través del sistema informático la red neuronal para un
usuario mediante al menos:

introducir (1414) una imagen de calibración y un segundo parámetro de calibración en la
25 red neuronal, mostrando la imagen de calibración un ojo del usuario basándose en datos
de imagen generados por una cámara asociada con un sistema de seguimiento ocular del
usuario, y
actualizar (1416) el segundo parámetro de calibración sin actualizar el parámetro de red
de la red neuronal, en donde el segundo parámetro de calibración se actualiza al menos
30 minimizando la función de pérdida basándose en la imagen de calibración y el segundo
parámetro de calibración; y

al finalizar la calibración, generar a través del sistema informático información de mirada
tridimensional, 3D, para el usuario mediante al menos:
35 introducir (1418) una imagen y el segundo parámetro de calibración en la red neuronal,
mostrando la imagen el ojo del usuario basándose en datos de imagen adicionales
generados por la cámara, y
recibir (1420) una predicción de la red neuronal basándose en la imagen y el segundo
40 parámetro de calibración, en donde la predicción comprende una corrección de distancia,
un origen de mirada bidimensional, 2D, del ojo del usuario en la imagen, y una dirección
de mirada 2D del ojo del usuario en la imagen;
comprendiendo además:
generar una distancia corregida entre el ojo del usuario y la cámara actualizando al menos
45 una distancia estimada entre el ojo del usuario y la cámara basándose en la corrección de
distancia, y en donde la información de mirada 3D para el ojo del usuario se genera
basándose en el origen de mirada 2D, la dirección de mirada 2D y la distancia corregida.
9. El sistema informático de la reivindicación 8, en donde el entrenamiento de la red neuronal comprende
50 además:
introducir una segunda imagen de entrenamiento en la red neuronal, en donde la imagen de entrenamiento
muestra el ojo de la persona mientras mira fijamente a un punto de mirada según un ángulo de mirada, en
donde la segunda imagen de entrenamiento muestra el ojo de la persona mientras mira fijamente a otro
punto de mirada según el ángulo de mirada.
- 55 10. El sistema informático de la reivindicación 8, en donde el entrenamiento de la red neuronal comprende
además:
introducir una segunda imagen de entrenamiento en la red neuronal, en donde la imagen de entrenamiento
y la segunda imagen de entrenamiento muestran el ojo de la persona mientras mira fijamente un punto de
60 mirada en un ángulo de mirada, en donde la imagen de entrenamiento corresponde a una primera distancia
entre una primera cámara y el ojo de la persona, y en donde la segunda imagen de entrenamiento
corresponde a una segunda distancia entre la primera cámara y el ojo de la persona.
- 65 11. El sistema informático de la reivindicación 8, en donde el entrenamiento de la red neuronal comprende
además:

introducir una segunda imagen de entrenamiento en la red neuronal, en donde la imagen de entrenamiento muestra el ojo de la persona mientras mira fijamente a un punto de mirada según un ángulo de mirada, en donde la segunda imagen de entrenamiento muestra el ojo de la persona mientras mira fijamente a otro punto de mirada según un ángulo de mirada diferente.

12. El sistema informático de la reivindicación 8, en donde el entrenamiento de la red neuronal comprende además:
introducir una segunda imagen de entrenamiento a la red neuronal, en donde la función de pérdida de la red neuronal se minimiza aún más basándose en una distancia entre un punto de mirada y una línea de mirada, en donde el punto de mirada está asociado con una de las imágenes de entrenamiento o la segunda imagen de entrenamiento, y en donde la red neuronal predice la línea de mirada para el ojo de la persona que se muestra en la imagen de entrenamiento.

13. Un medio no transitorio legible por ordenador que almacena instrucciones que, al ejecutarse en un sistema informático, configuran el sistema informático para realizar operaciones que comprenden:

entrenar una red neuronal mediante al menos:

introducir (1402) una imagen de entrenamiento y un primer parámetro de calibración en la red neuronal, mostrando la imagen de entrenamiento un ojo de una persona, y actualizar (1408) el primer parámetro de calibración y un parámetro de red de la red neuronal basándose en la minimización de una función de pérdida de la red neuronal, en donde la función de pérdida se minimiza basándose en la imagen de entrenamiento y el primer parámetro de calibración; y

al finalizar el entrenamiento, calibrar a través del sistema informático la red neuronal para un usuario mediante al menos:

introducir (1414) una imagen de calibración y un segundo parámetro de calibración en la red neuronal, mostrando la imagen de calibración un ojo del usuario basándose en datos de imagen generados por una cámara asociada con un sistema de seguimiento ocular del usuario, y actualizar (1416) el segundo parámetro de calibración sin actualizar el parámetro de red de la red neuronal, en donde el segundo parámetro de calibración se actualiza al menos minimizando la función de pérdida basándose en la imagen de calibración y el segundo parámetro de calibración; y

al finalizar la calibración, generar a través del sistema informático información de mirada tridimensional, 3D, para el usuario mediante al menos:

introducir (1418) una imagen y el segundo parámetro de calibración en la red neuronal, mostrando la imagen el ojo del usuario basándose en datos de imagen adicionales generados por la cámara, y recibir (1420) una predicción de la red neuronal basándose en la imagen y el segundo parámetro de calibración, en donde la predicción comprende una corrección de distancia, un origen de mirada bidimensional, 2D, del ojo del usuario en la imagen, y una dirección de mirada 2D del ojo del usuario en la imagen, comprendiendo además:
generar, mediante el sistema informático, una distancia corregida entre el ojo del usuario y la cámara actualizando al menos una distancia estimada entre el ojo del usuario y la cámara basándose en la corrección de distancia, y en donde la información de mirada 3D para el ojo del usuario se genera basándose en el origen de mirada 2D, la dirección de mirada 2D y la distancia corregida.

14. Un programa informático que comprende instrucciones ejecutables por ordenador para hacer que un sistema informático, cuando las instrucciones ejecutables por ordenador son ejecutadas en una unidad de procesamiento comprendida en el sistema informático, realice operaciones que comprenden:

entrenar una red neuronal mediante al menos:

introducir (1402) una imagen de entrenamiento y un primer parámetro de calibración en la red neuronal, mostrando la imagen de entrenamiento un ojo de una persona, y actualizar (1408) el primer parámetro de calibración y un parámetro de red de la red neuronal basándose en la minimización de una función de pérdida de la red neuronal, en donde la función de pérdida se minimiza basándose en la imagen de entrenamiento y el primer parámetro de calibración; y

al finalizar el entrenamiento, calibrar a través del sistema informático la red neuronal para un usuario mediante al menos:

5 introducir (1414) una imagen de calibración y un segundo parámetro de calibración en la red neuronal, mostrando la imagen de calibración un ojo del usuario basándose en datos de imagen generados por una cámara asociada con un sistema de seguimiento ocular del usuario, y
10 actualizar (1416) el segundo parámetro de calibración sin actualizar el parámetro de red de la red neuronal, en donde el segundo parámetro de calibración se actualiza al menos minimizando la función de pérdida basándose en la imagen de calibración y el segundo parámetro de calibración; y

al finalizar la calibración, generar a través del sistema informático información de mirada tridimensional, 3D, para el usuario mediante al menos:

15 introducir (1418) una imagen y el segundo parámetro de calibración en la red neuronal, mostrando la imagen el ojo del usuario basándose en datos de imagen adicionales generados por la cámara, y
20 recibir (1420) una predicción de la red neuronal basándose en la imagen y el segundo parámetro de calibración, en donde la predicción comprende una corrección de distancia, un origen de mirada bidimensional, 2D, del ojo del usuario en la imagen, y una dirección de mirada 2D del ojo del usuario en la imagen
comprendiendo además:
25 generar, mediante el sistema informático, una distancia corregida entre el ojo del usuario y la cámara actualizando al menos una distancia estimada entre el ojo del usuario y la cámara basándose en la corrección de distancia, y en donde la información de mirada 3D para el ojo del usuario se genera basándose en el origen de mirada 2D, la dirección de mirada 2D y la distancia corregida.

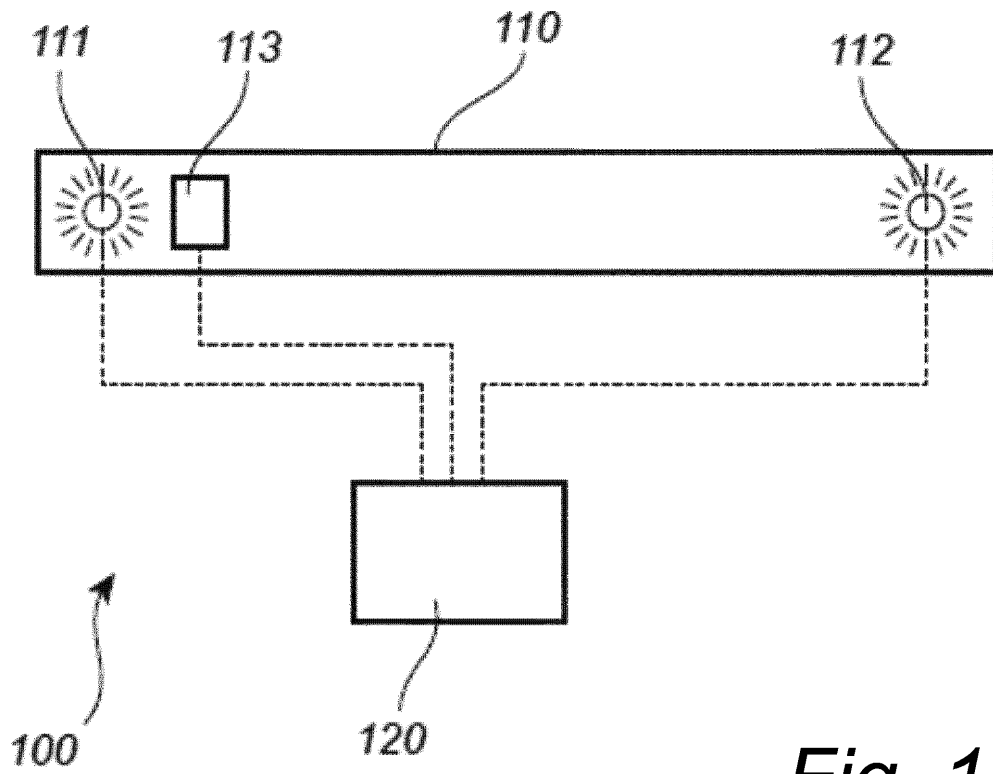


Fig. 1

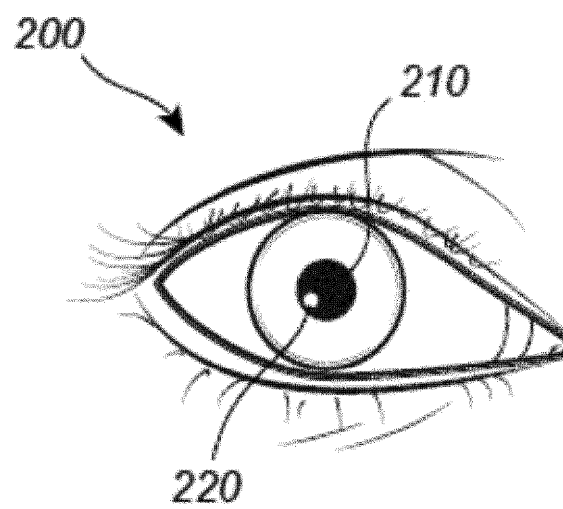


Fig. 2

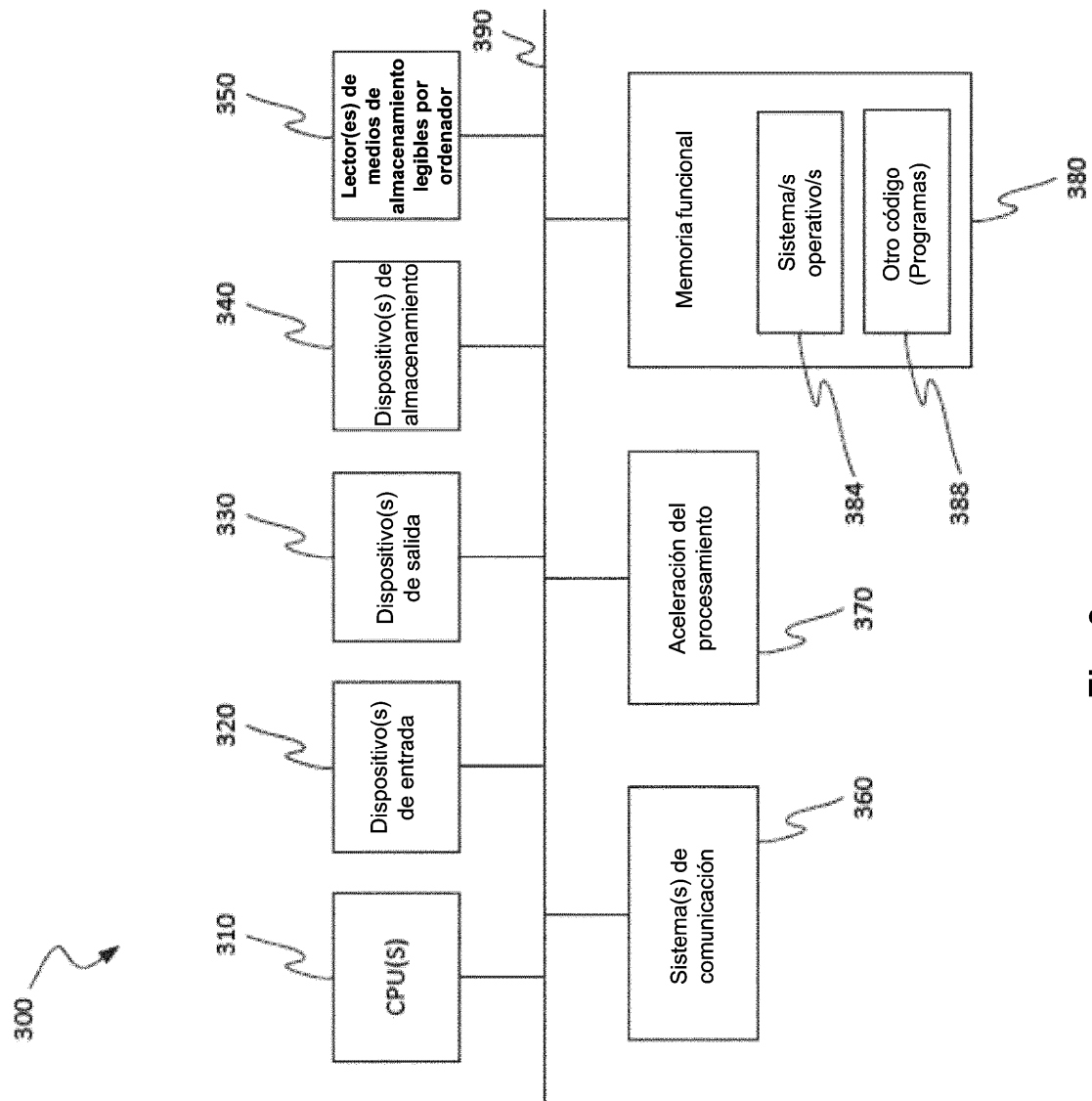


Fig. 3

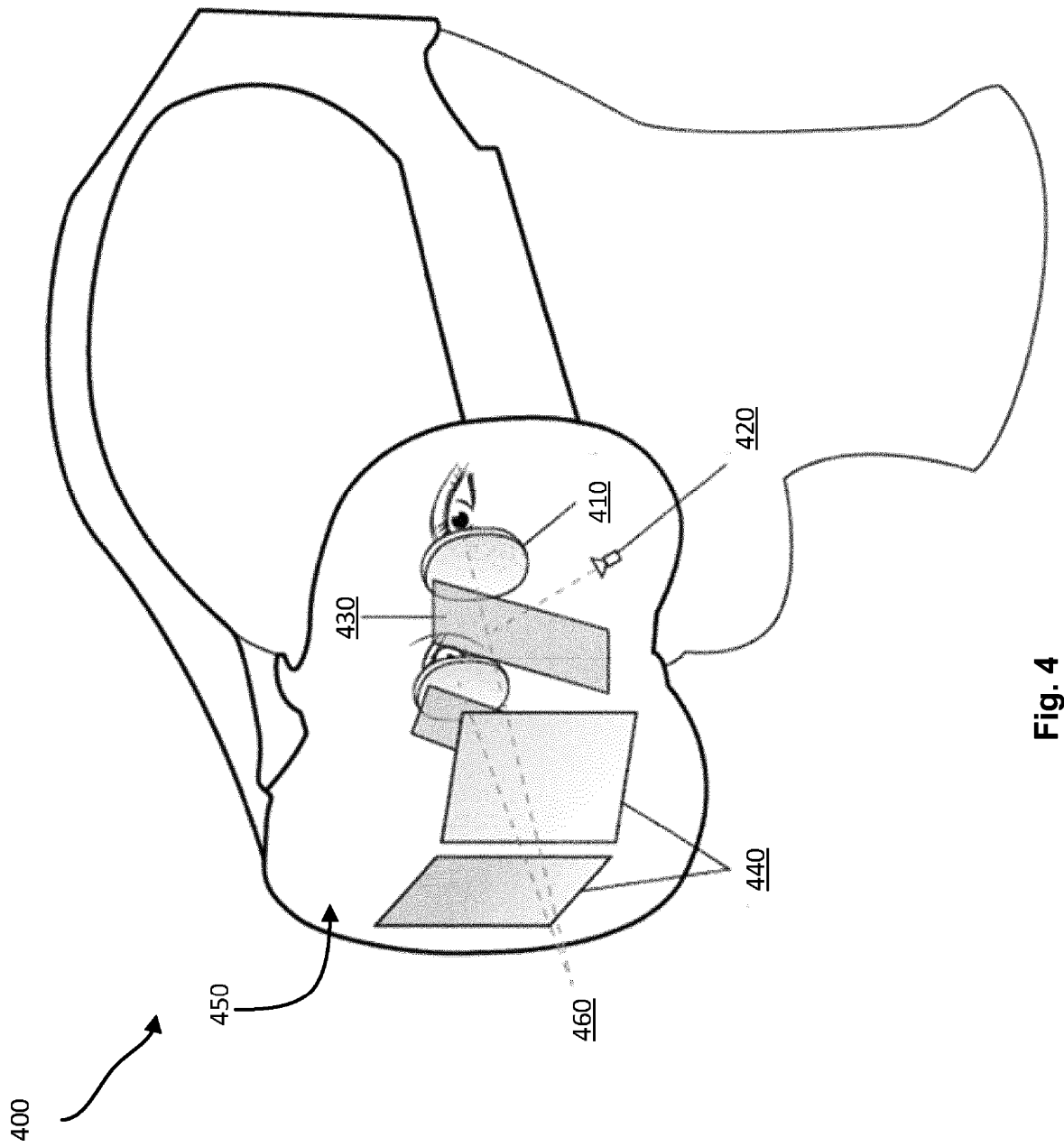


Fig. 4

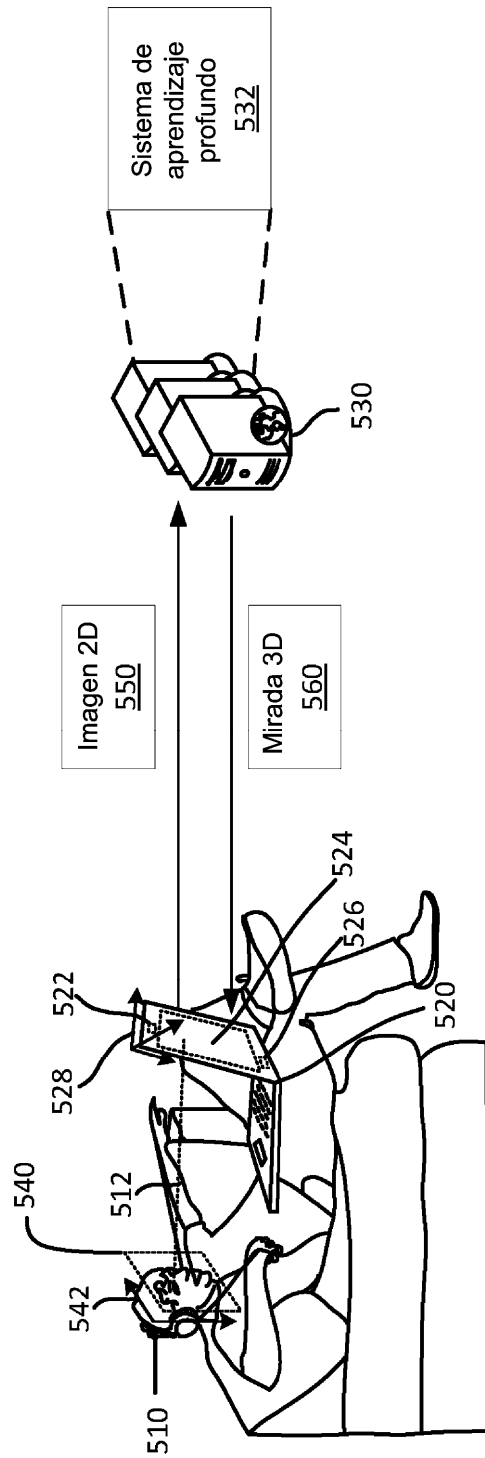


Fig. 5

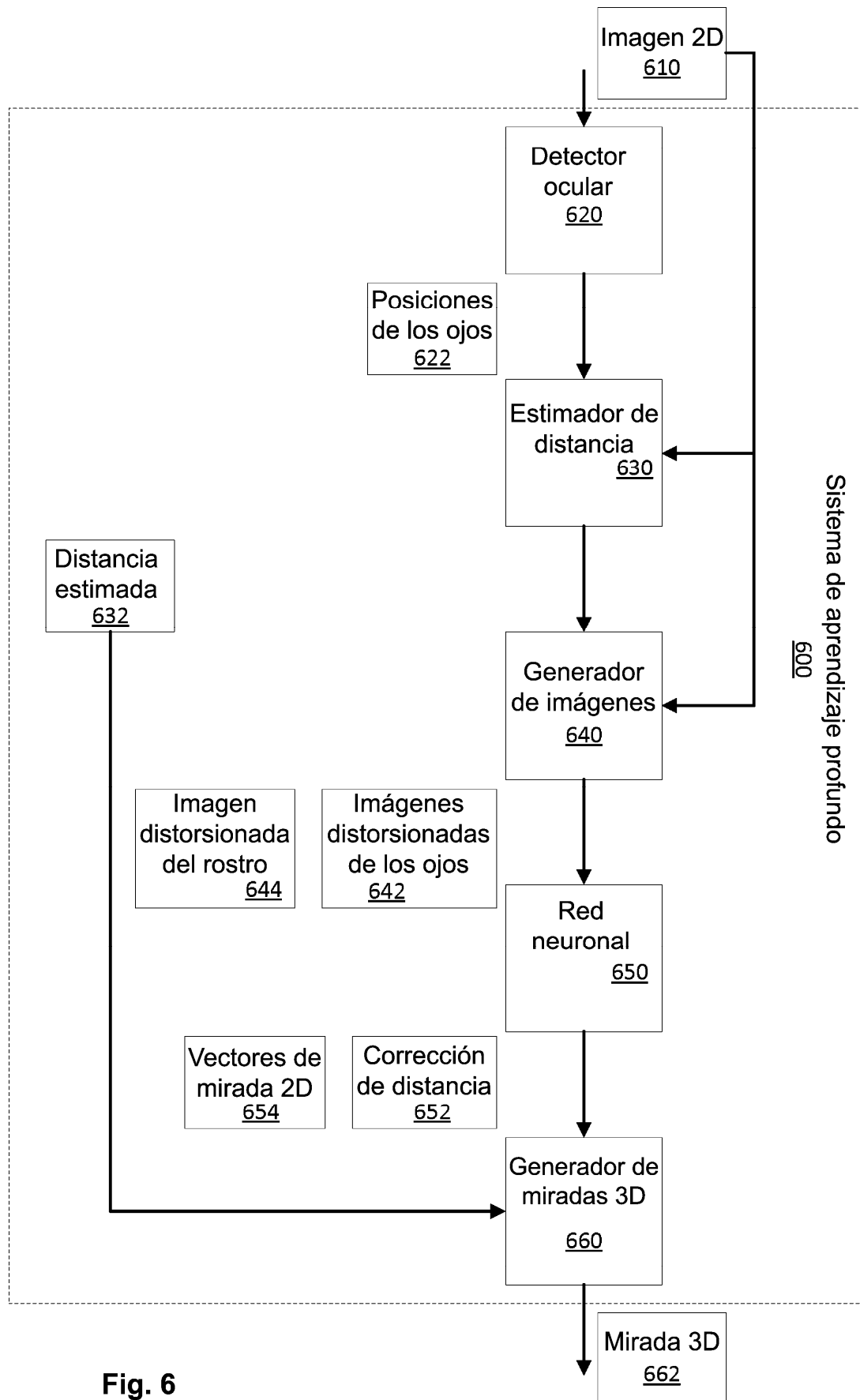


Fig. 6

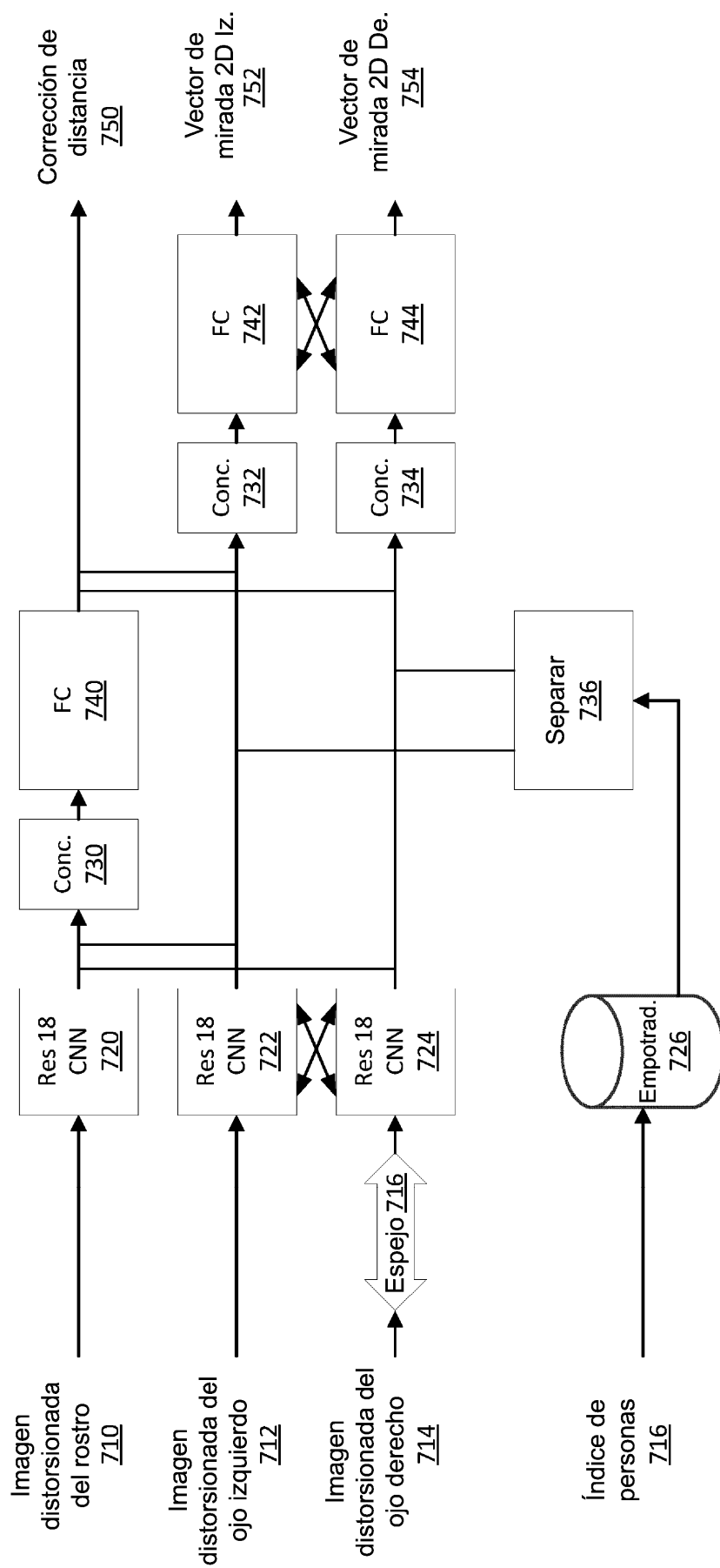


Fig. 7

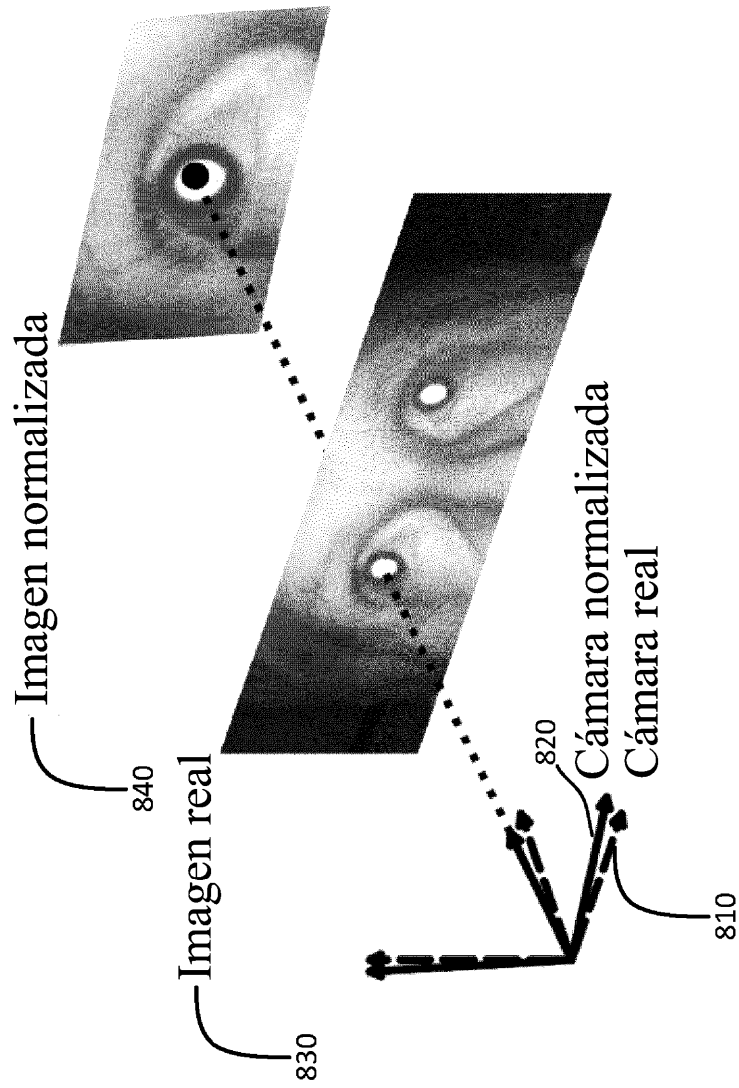


Fig. 8

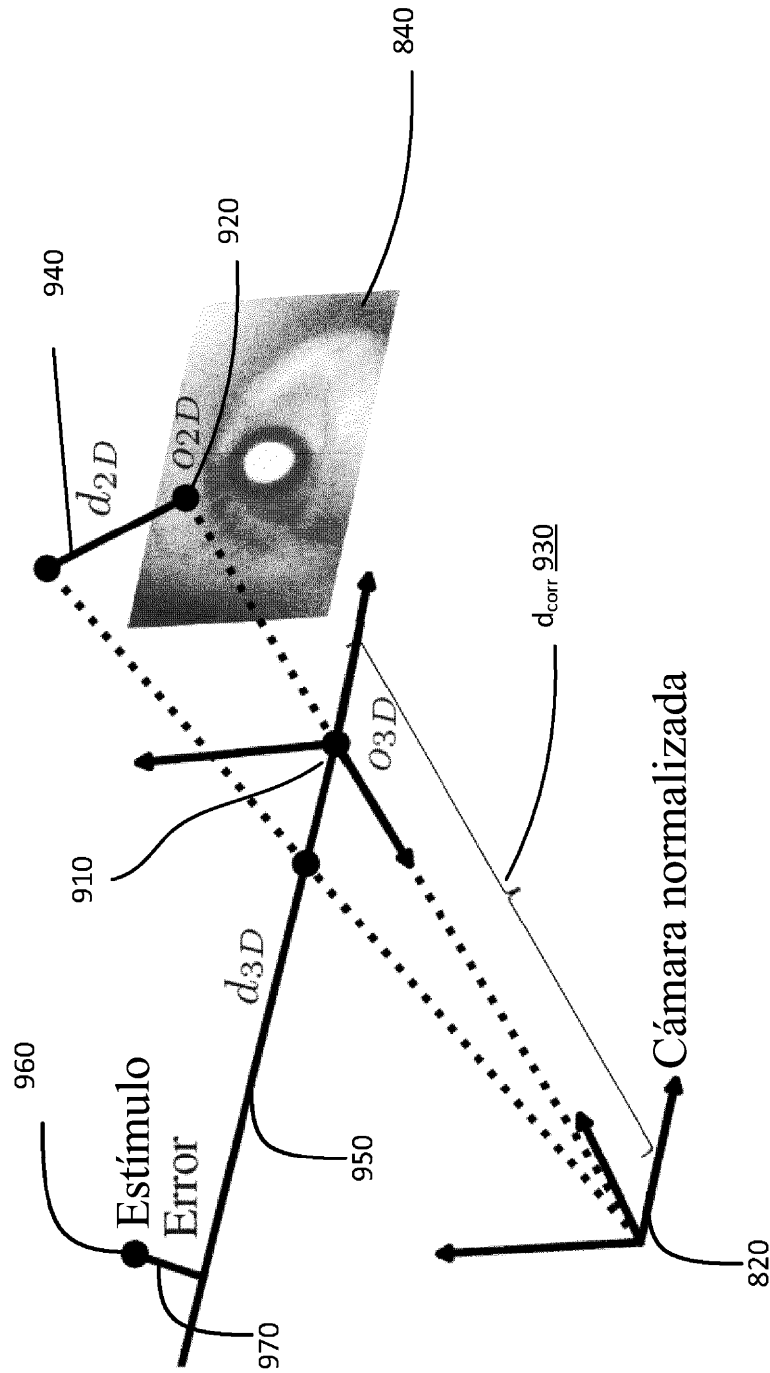


Fig. 9

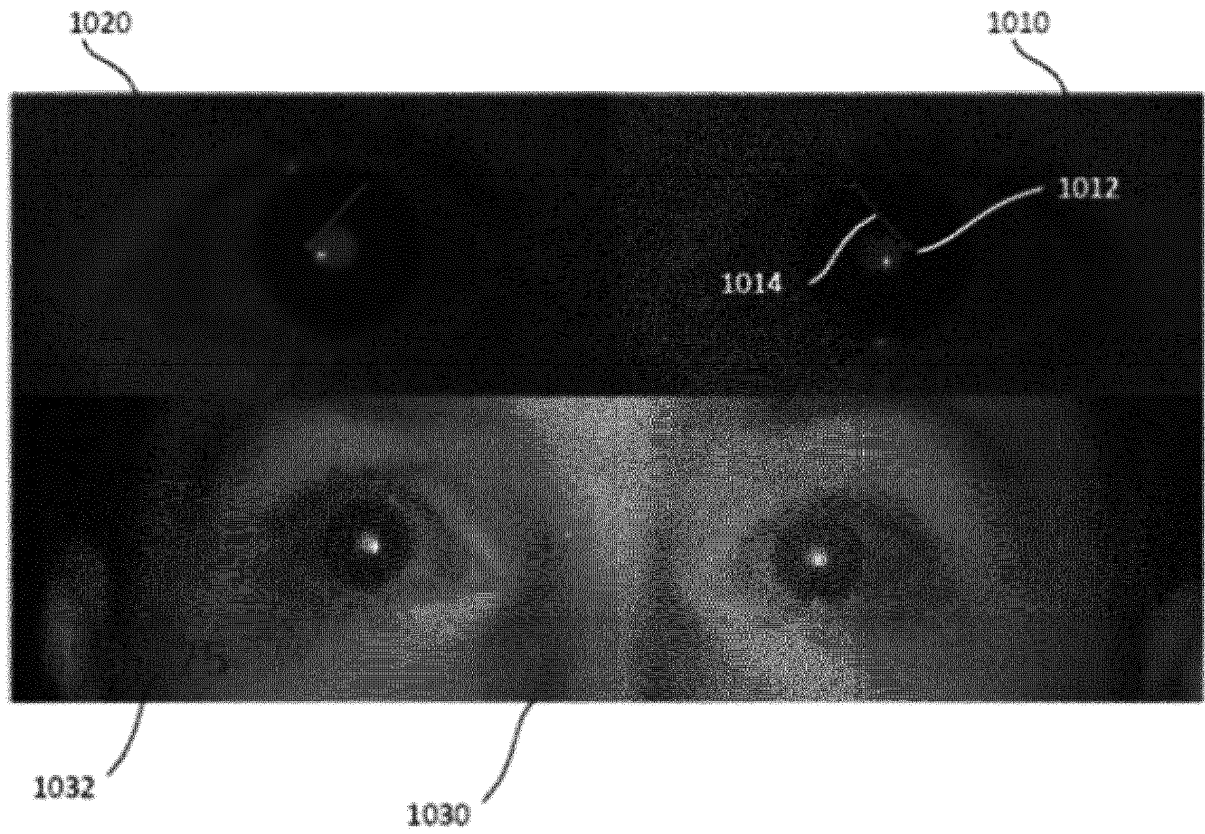


Fig. 10

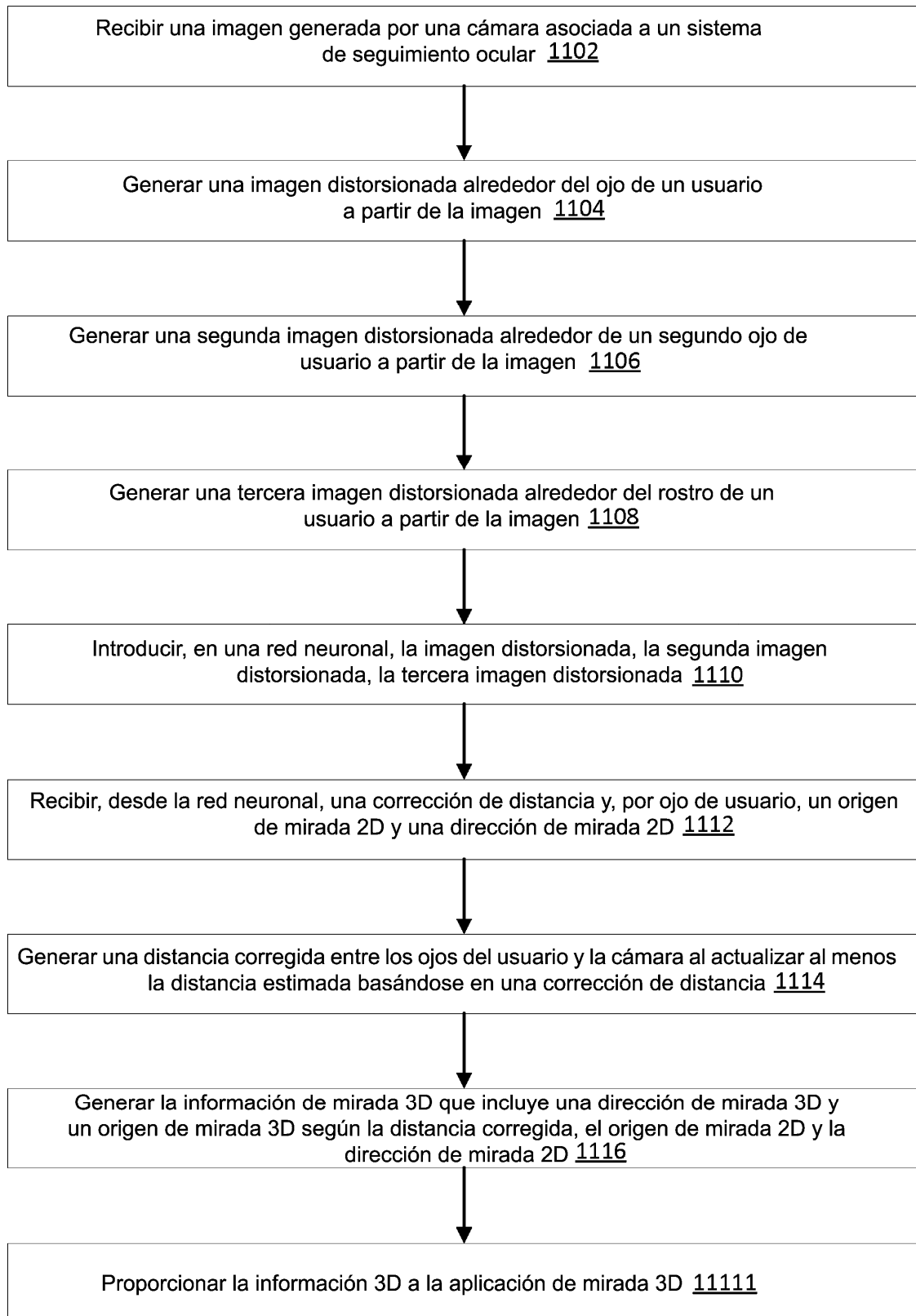


Fig. 11

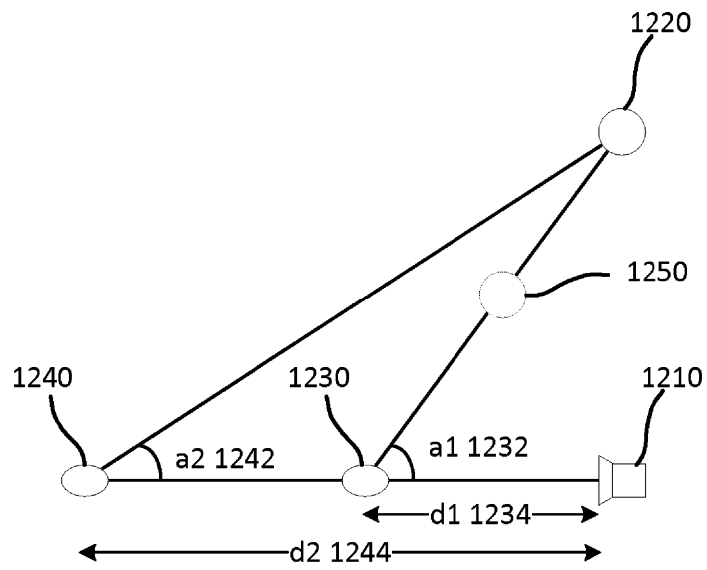


Fig. 12

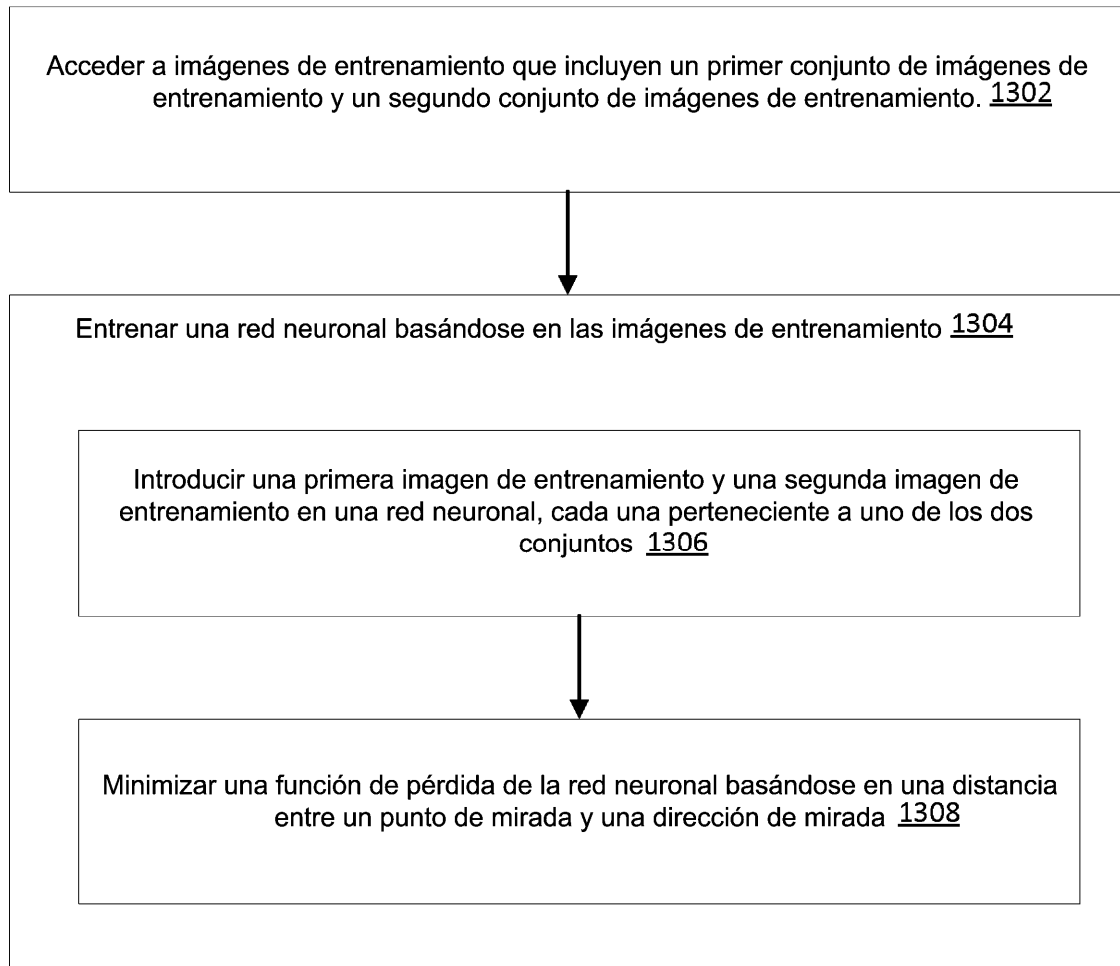


Fig. 13

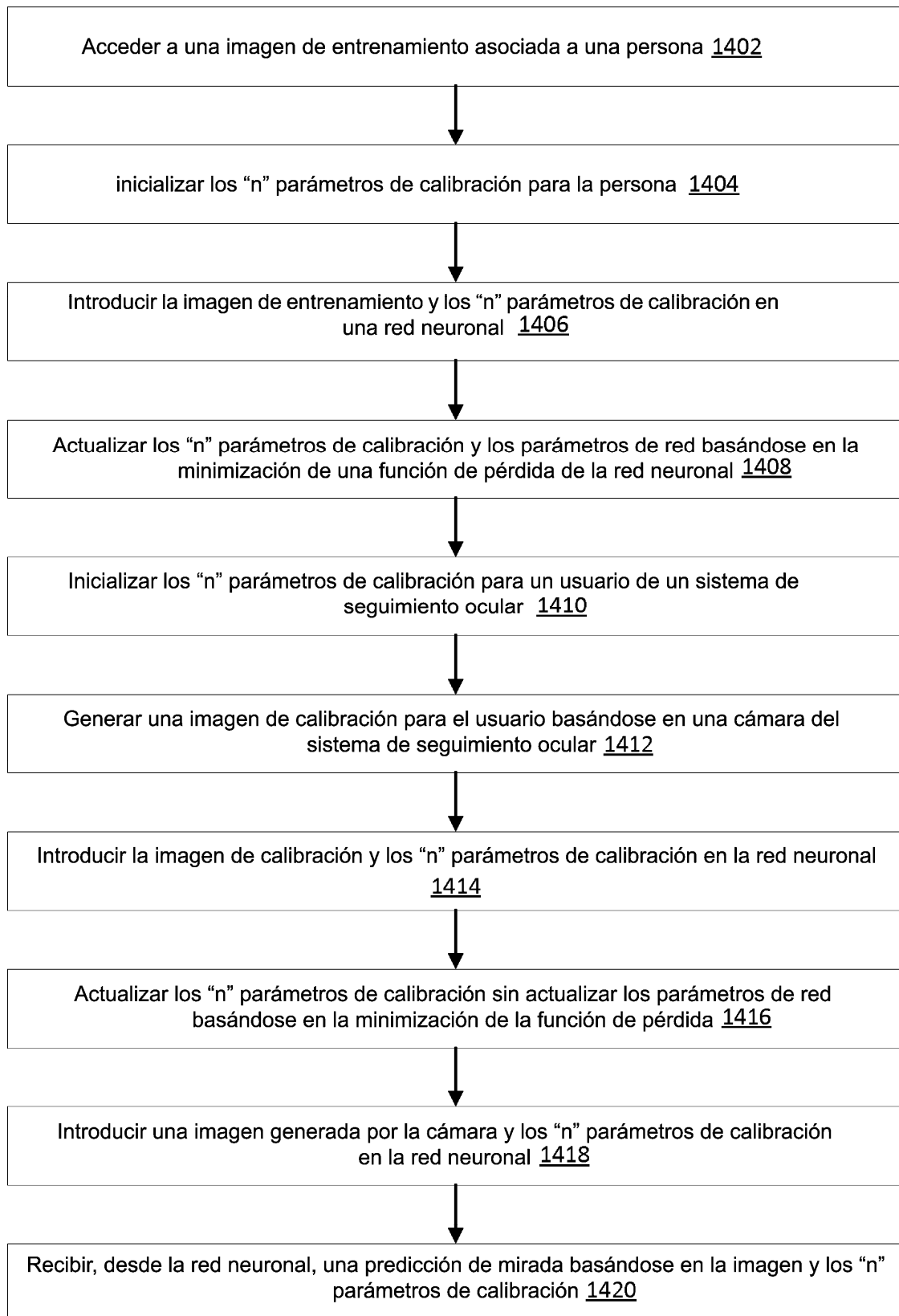


Fig. 14