

DETECCIÓN CUANTIZADA DE CAMBIO DE TRANSICIÓN PARA EL RECONOCIMIENTO DE ACTIVIDAD

CAMPO TÉCNICO

[0001] La presente divulgación se refiere en general a la inteligencia artificial y, más específicamente, al reconocimiento de la actividad humana a partir de una secuencia de video y al procesamiento simbólico.

ANTECEDENTES

[0002] Con el avance de la tecnología, el reconocimiento de las actividades físicas humanas está adquiriendo una enorme importancia. El reconocimiento de las actividades físicas humanas contribuye a diversas aplicaciones, como la vigilancia de un proceso de salida de una tienda minorista que implica un sistema de auto pago (SCO). Dicho sistema permite a los compradores completar un proceso de compra por sí mismos. Otro ejemplo de aplicación del reconocimiento de las actividades físicas humanas es la prestación de asistencia en la videovigilancia mediante la detección de actividades ilegales realizadas por los ladrones de tiendas, como el robo, y de ese modo alertar al personal empleado en el taller para evitar el robo. Además, el reconocimiento de las actividades físicas humanas se emplea en sistemas inteligentes de asistencia al conductor, sistemas de vida asistida para humanos necesitados, videojuegos, fisioterapia, y así sucesivamente. Además, el reconocimiento de las actividades físicas humanas se utiliza activamente en el campo de los deportes, el ejército, la medicina, la robótica, y así sucesivamente.

[0003] Las actividades físicas humanas representan los componentes básicos de la mayoría de los modelos de procesos. Sin embargo, como el comportamiento humano es impredecible, el reconocimiento de tales actividades físicas humanas en un entorno diverso es una tarea difícil. La actividad física humana se puede descomponer típicamente en un conjunto de acciones básicas que involucran varias partes del cuerpo humano, como manos, pies, cara, etc. Además, el conjunto de acciones básicas asociadas con la actividad física humana se extiende a lo largo de una pluralidad de intervalos

de tiempo. Las tareas de reconocimiento de tales actividades enfrentan el problema de resumir la secuencia general de acciones en un intervalo de tiempo variable.

[0004] Las técnicas convencionales de reconocimiento de la actividad física humana son ineficaces para reconocer las actividades físicas humanas, debido a una estructura corporal diferente, una forma corporal diferente, un color de piel diferente, etc. de cada cuerpo humano. Además, el marco de tiempo para una actividad humana presenta una variación importante en el tiempo dependiendo del sujeto y tal vez de otras condiciones ambientales. Además, no todos los movimientos básicos de las partes del cuerpo están relacionados con el propósito de la actividad considerada. Por lo tanto, el proceso de reconocimiento de la actividad enfrenta dos problemas importantes relacionados con la variación del tiempo de las acciones y la variación de la trayectoria física de las partes del cuerpo humano involucradas en la actividad.

[0005] Por lo tanto, a la luz de la discusión anterior, existe la necesidad de superar los inconvenientes antes mencionados asociados con el reconocimiento de las actividades físicas humanas y proporcionar un sistema y método que tenga como objetivo reducir la influencia de la variación en el tiempo y la variedad de los movimientos de las partes del cuerpo en el reconocimiento de actividad utilizando una red neuronal recurrente.

RESUMEN

[0006] La presente divulgación busca proporcionar un sistema para reconocer la actividad humana a partir de una secuencia de video y un método del mismo.

[0007] De acuerdo con un aspecto de la presente divulgación, se proporciona un sistema para reconocer la actividad humana a partir de una secuencia de video capturada por un dispositivo de formación de imágenes. El sistema incluye una memoria para almacenar una o más instrucciones y un procesador acoplado comunicativamente a la memoria. El sistema incluye un

clasificador acoplado comunicativamente al dispositivo de imágenes y configurado para clasificar un cuadro de imagen de la secuencia de video en una o más clases de un conjunto de clases predefinidas, en el que el cuadro de imagen se clasifica en función de la acción del usuario en una región de interés del cuadro de imagen, y generar un vector de probabilidad de clase para el cuadro de imagen basado en la clasificación, en el que el vector de probabilidad de clase incluye un conjunto de probabilidades de clasificación del cuadro de imagen en cada clase predefinida. El sistema incluye además un módulo de filtrado y binarización de datos configurado para filtrar y binarizar cada valor de probabilidad del vector de probabilidad de clase basándose en un valor umbral de probabilidad predefinido. El sistema incluye además un módulo de composición de palabras comprimidas configurado para determinar una o más transiciones de una o más clases en uno o más cuadros de imagen consecutivos de la secuencia de video, basándose en los correspondientes vectores de probabilidad binarizados, y generar una secuencia de palabras comprimidas en base a una o más transiciones determinadas en uno o más cuadros de imagen consecutivos. El sistema incluye además un clasificador dependiente de la secuencia configurado para extraer una o más acciones del usuario mediante el análisis de la secuencia de palabras comprimidas y reconocer la actividad humana a partir de las mismas.

[0008] De acuerdo con otro aspecto de la presente divulgación, se proporciona un método para reconocer la actividad humana a partir de una secuencia de video. El método incluye clasificar mediante un clasificador, un cuadro de imagen de la secuencia de video en una o más clases de un conjunto de clases predefinidas, en el que el cuadro de imagen se clasifica basándose en la acción del usuario en una región de interés del cuadro de imagen. El método incluye además generar un vector de probabilidad de clase para el cuadro de imagen basado en la clasificación, en el que el vector de probabilidad de clase incluye un conjunto de probabilidades de clasificación del cuadro de imagen en cada clase predefinida. El método incluye además binarizar cada valor de probabilidad del vector de probabilidad de clase basándose en un valor umbral de probabilidad predefinido. El método incluye además determinar una o más transiciones de una o más clases en uno o más cuadros de imagen consecutivos

de la secuencia de video, basándose en los correspondientes vectores de probabilidad binarizados. El método incluye además generar una secuencia de palabras comprimidas en base a una o más transiciones determinadas en uno o más cuadros de imagen consecutivos. El método incluye además extraer una o más acciones del usuario analizando la secuencia de palabras comprimidas y reconocer la actividad humana a partir de las mismas.

[0009] De acuerdo con otro aspecto más de la presente divulgación, se proporciona un producto programable por computador para reconocer la actividad humana a partir de una secuencia de video, comprendiendo el producto programable por computador un conjunto de instrucciones. El conjunto de instrucciones cuando es ejecutado por un procesador hace que el procesador clasifique un cuadro de imagen de la secuencia de video en una o más clases de un conjunto de clases predefinidas, en el que el cuadro de imagen se clasifica en función de la acción del usuario en una región de interés del cuadro de imagen, generar un vector de probabilidad de clase para el cuadro de imagen en función de la clasificación, en donde el vector de probabilidad de clase incluye un conjunto de probabilidades de clasificación del cuadro de imagen en cada clase predefinida, binarizar cada valor de probabilidad del vector de probabilidad de clase basado en un valor umbral de probabilidad predefinido, determinar una o más transiciones de una o más clases en uno o más cuadros de imagen consecutivos de la secuencia de video, en base a los correspondientes vectores de probabilidad binarizados, generar una secuencia de palabras comprimidas en base a una o más transiciones determinadas en uno o más cuadros de imagen consecutivos y extraer una o más acciones del usuario analizando la secuencia de palabras comprimidas para extraer una o más acciones del usuario, y reconocer la actividad humana de las mismas.

[0010] La presente divulgación busca proporcionar un sistema para reconocer la actividad humana a partir de una secuencia de video. Tal sistema permite el reconocimiento eficiente y confiable de las actividades humanas a partir de la secuencia de video.

[0011] Se apreciará que las características de la presente divulgación son susceptibles de combinarse en varias combinaciones sin apartarse del alcance de la presente divulgación tal como se define en las reivindicaciones adjuntas.

DIVULGACIÓN DE LOS DIBUJOS

[0012] El resumen anterior, así como la siguiente divulgación detallada de modalidades ilustrativas, se entienden mejor cuando se leen junto con los dibujos adjuntos. Con el fin de ilustrar la presente divulgación, en los dibujos se muestran construcciones ejemplares de la divulgación. Sin embargo, la presente divulgación no se limita a los métodos e instrumentos específicos descritos en el presente documento. Además, los expertos en la técnica comprenderán que los dibujos no están a escala. Siempre que ha sido posible, los elementos similares se han indicado con números idénticos.

[0013] Las modalidades de la presente divulgación se describirán ahora, solo a modo de ejemplo, con referencia a los siguientes diagramas en los que:

[0014] **La FIG. 1** ilustra un entorno, en el que se pueden practicar diversas modalidades de la presente divulgación;

[0015] **La FIG.2** ilustra el sistema de reconocimiento de actividad para reconocer una o más acciones y actividades humanas en la secuencia de video capturada por el dispositivo de formación de imágenes de la figura 1, de acuerdo con una modalidad de la presente divulgación; y

[0016] **La FIG. 3** es un diagrama de flujo que ilustra un método para reconocer la actividad humana a partir de una secuencia de video, de acuerdo con una modalidad de la presente divulgación.

[0017] En los dibujos adjuntos, se emplea un número subrayado para representar un artículo sobre el que está colocado el número subrayado o un artículo al que está adyacente el número subrayado. Un número no subrayado se refiere a un artículo identificado por una línea que vincula el número no subrayado con el artículo. Cuando un número no está subrayado y está

acompañado por una flecha asociada, el número no subrayado se usa para identificar un elemento general al que apunta la flecha.

DIVULGACIÓN DE MODALIDADES

[0018] La siguiente divulgación detallada ilustra modalidades de la presente divulgación y formas en las que se pueden implementar. Aunque se han revelado algunos modos de llevar a cabo la presente divulgación, los expertos en la técnica reconocerán que también son posibles otras modalidades para llevar a cabo o poner en práctica la presente divulgación.

[0019] La FIG. 1 ilustra un entorno 100, en el que se pueden poner en práctica diversas modalidades de la presente divulgación. El entorno 100 incluye un dispositivo 101 de formación de imágenes, un sistema 102 de reconocimiento de actividad y un dispositivo informático 103, acoplados comunicativamente entre sí a través de una red de comunicación 104. La red de comunicación 104 puede ser cualquier red cableada, red inalámbrica, una combinación de estas o cualquier otra red convencional, sin limitar el alcance de la presente divulgación. Pocos ejemplos pueden incluir una red de área local (LAN), una conexión LAN inalámbrica, una conexión a Internet, una conexión punto a punto u otra conexión de red y combinaciones de las mismas.

[0020] El dispositivo de formación de imágenes 101 está configurado para capturar una secuencia de video. En una modalidad de la presente divulgación, el dispositivo de imágenes 101 está configurado para capturar uno o más videos de un proceso de pago minorista que incluye un sistema de auto pago (SCO). Opcionalmente, el dispositivo de formación de imágenes 101 incluye, entre otros, una cámara de protocolo de Internet (IP), una cámara Pan-Tilt-Zoom (PTZ), una cámara termográfica o una cámara infrarroja.

[0021] El sistema 102 de reconocimiento de actividad está configurado para reconocer acciones humanas y actividades humanas en la secuencia de video capturada por el dispositivo 101 de formación de imágenes.

[0022] El sistema 102 de reconocimiento de actividad incluye una unidad 106 de procesamiento central (CPU), un panel 108 de operación y una memoria

110. La CPU 106 es un procesador, computadora, microcontrolador u otro circuito que controla las operaciones de varios componentes tales como el panel de operación 108 y la memoria 110. La CPU 106 puede ejecutar software, firmware y / u otras instrucciones, por ejemplo, que se almacenan en una memoria volátil o no volátil, como la memoria 110, o se proporcionan de otra manera a la CPU 106. La CPU 106 puede estar conectada al panel de operación 108, y la memoria 110, a través de conexiones alámbricas o inalámbricas, tales como uno o más buses del sistema, cables u otras interfaces. En una modalidad de la presente divulgación, la CPU 106 puede incluir un software de servidor de unidad de procesamiento gráfico (GPU) personalizado para proporcionar detección y predicción de objetos en tiempo real, para todas las cámaras en una red local.

[0023] El panel de operación 108 puede ser una interfaz de usuario para el aparato de formación de imágenes 100 y puede tomar la forma de un teclado físico o pantalla táctil. El panel de operación 108 puede recibir entradas de uno o más usuarios relacionadas con funciones seleccionadas, preferencias y / o autenticación, y puede proporcionar y / o recibir entradas de forma visual y / o audible.

[0024] La memoria 110, además de almacenar instrucciones y / o datos para su uso por la CPU 106 en la gestión de la operación del aparato de formación de imágenes 100, también puede incluir información de usuario asociada con uno o más usuarios del aparato formador de imagen 100. Por ejemplo, la información del usuario puede incluir información de autenticación (por ejemplo, pares de nombre de usuario / contraseña), preferencias del usuario y otra información específica del usuario. La CPU 106 puede acceder a estos datos para ayudar a proporcionar funciones de control (por ejemplo, transmitir y / o recibir una o más señales de control) relacionadas con el funcionamiento del panel 108 de operación y la memoria 110.

[0025] El dispositivo de procesamiento de imágenes 101 y el sistema de reconocimiento de actividad 102 pueden ser controlados / operados por el dispositivo informático 103. Los ejemplos del dispositivo informático 103 incluyen un teléfono inteligente, una computadora personal, una computadora portátil y similares. El dispositivo informático 103 permite al usuario / operador ver y guardar los videos capturados por el dispositivo de imágenes 101, y acceder a los videos / imágenes procesados por el sistema 102 de reconocimiento de actividad. El dispositivo informático 103 puede ejecutar una aplicación móvil del sistema 102 de reconocimiento de actividad para permitir que un usuario acceda y procese la secuencia de video capturada por el dispositivo 101 de formación de imágenes.

[0026] En una modalidad, la cámara 101, el sistema de reconocimiento de actividad 102 y el dispositivo informático 103 pueden integrarse en un solo dispositivo, donde el único dispositivo es un teléfono inteligente portátil que tiene una cámara incorporada y una pantalla.

[0027] La figura 2 ilustra el sistema 102 de reconocimiento de actividad para reconocer una o más acciones y actividades humanas en la secuencia de video capturada por el dispositivo 101 de formación de imágenes, de acuerdo con una modalidad de la divulgación presente.

[0028] El sistema de reconocimiento de actividad 102 incluye la CPU 106 que incluye un clasificador 202 que es operable para analizar cada cuadro de la secuencia de video para determinar al menos una región de acción de interés, donde al menos una región de interés comprende al menos un objeto. La región de acción de interés se refiere a un área rectangular en cada cuadro de la secuencia de video, donde al menos se ve un objeto y tienen lugar una o más acciones. En un ejemplo, el al menos un objeto puede ser una persona, objetos como prendas de vestir, comestibles, billetera, etc., y una o más acciones pueden incluir una persona que saca la billetera de su bolsillo, la persona que camina en una fila, la persona que pasa una tarjeta de crédito y cosas por el estilo. Cada acción se puede utilizar como un bloque de construcción para la extracción del modelo de proceso, donde un proceso se puede expresar como una cadena de acciones.

[0029] En una modalidad de la presente divulgación, el clasificador 202 puede ser un clasificador basado en algoritmos, como una red neuronal convolucional (CNN) entrenada para clasificar un cuadro de imagen del video del área de escaneo de SCO (área de escaneo de acción de escaneo)) en clases como mano, objeto en mano, objeto, parte del cuerpo, escáner vacío. Los criterios para la clasificación de un cuadro de imagen en cada clase se mencionan a continuación:

[0030] Mano – el cuadro de imagen muestra la (s) mano (s) humana (s).

[0031] Objeto en mano – El cuadro de imagen muestra un objeto en la mano del usuario.

[0032] Objeto – El cuadro de imagen muestra solo el objeto

[0033] Parte del cuerpo – El cuadro de imagen muestra una parte del cuerpo humano.

[0034] Escáner vacío – El cuadro de imagen muestra solo el escáner vacío

[0035] La CNN como se hace referencia en este documento se define como redes neuronales artificiales profundas entrenadas que se usan principalmente para clasificar el al menos un objeto en la al menos una región de interés. En particular, son algoritmos que pueden identificar rostros, personas, letreros de calles y similares. El término "red neuronal", como se usa en este documento, puede incluir una red altamente interconectada de elementos de procesamiento, cada uno asociado opcionalmente con una memoria local. En un ejemplo, la red neuronal puede ser un mapa de Kohonen, un perceptrón de múltiples capas, etc. Además, los elementos de procesamiento de las redes neuronales pueden ser "unidades neuronales artificiales", "neuronas artificiales", "unidades neuronales", "neuronas", "nodos" y similares. Además, la neurona puede recibir datos de una entrada o de una o más neuronas, procesar los datos y enviar los datos procesados a una salida o incluso a una o más neuronas. La red neuronal o una o más neuronas de la misma se pueden generar en hardware, software o una combinación de hardware y software, y la red neuronal se puede entrenar posteriormente. Se apreciará que la red neuronal convolucional (CNN) consta

de una capa de entrada, una pluralidad de capas ocultas y una capa de salida. Además, la pluralidad de capas ocultas de la red neuronal convolucional consiste típicamente en capas convolucionales, capas agrupadas, capas completamente conectadas y capas de normalización. Opcionalmente, se utiliza un modelo de Visual Geometry Group 19 (VGG 19) como arquitectura de red neuronal convolucional. El modelo VGG 19 está configurado para clasificar el al menos un objeto en el cuadro de la secuencia de video en clases. Se apreciará que las capas ocultas comprenden una pluralidad de conjuntos de capas de convolución.

[0036] En funcionamiento, el clasificador 202 recibe y clasifica un cuadro de imagen de la secuencia de video del área de escaneo SCO (región de interés de la acción de escaneo) en clases tales como mano, objeto en mano, objeto, parte del cuerpo, escáner vacío basado en contenido del cuadro de imagen. En una modalidad de la presente divulgación, el clasificador 202 analiza cada cuadro de imagen estáticamente y para cada cuadro de imagen, genera un vector de probabilidad de clase. P_v tener un componente para cada clase considerada, de modo que, $P_v = \{P_{\text{Hand}}, P_{\text{HandObject}}, P_{\text{Object}}, P_{\text{BodyPart}}, P_{\text{EmptyScanner}}\}$

Dónde P_{Hand} = Probabilidad de que el cuadro de imagen se clasifique en la clase "mano"

$P_{\text{HandObject}}$ = Probabilidad de que el cuadro de imagen se clasifique en la clase "objeto en mano"

P_{Object} = Probabilidad de que el cuadro de imagen se clasifique en la clase "objeto"

P_{BodyPart} = Probabilidad de que el cuadro de imagen se clasifique en la clase "parte del cuerpo"

$P_{\text{EmptyScanner}}$ = Probabilidad de que el cuadro de imagen se clasifique en la clase "escáner vacío"

[0037] En un ejemplo, el clasificador 202 genera seis vectores de probabilidad P_{v1} a P_{v6} para seis cuadros de imagen consecutivos en cinco clases, en un formato que se indica a continuación.

$P_{v1} = \{0.0, 0.0, 0.0, 0.0, 1.0\}$

$P_{v2} = \{0.0, 0.0, 0.28, 0.0, 0.72\}$

$$P_{v3} = \{0.0, 0.0, 0.26, 0.0, 0.74\}$$

$$P_{v4} = \{0.0, 0.0, 0.19, 0.0, 0.81\}$$

$$P_{v5} = \{0.0, 0.0, 0.29, 0.0, 0.71\} \quad P_{v6} = \{0.0, 0.45, 0.14, 0.0, 0.41\}$$

[0038] La CPU 106 incluye además un módulo 204 de generación de firma cuantificada para generar una firma cuantificada para cada acción de exploración determinada por el clasificador 202. Una acción de escaneo es una acción realizada por el usuario para escanear un artículo en una zona de escaneo de un terminal de auto pago (SCO).

[0039] El módulo 204 de generación de firmas cuantificadas incluye un módulo 205 de filtrado y binarización de datos, un módulo 206 de detección de intervalo silencioso y un módulo 207 de composición de palabras comprimidas.

[0040] El módulo 205 de filtrado y binarización de datos está configurado para aplicar un filtro en los vectores de probabilidad de clase generados por el clasificador 202 para minimizar los errores del clasificador 202. Aparece un error de clasificador si el clasificador 202 clasifica un movimiento continuo en el escáner usando una sola clase para toda la secuencia excepto un cuadro aislado. En tal caso, el cuadro aislado puede clasificarse incorrectamente.

[0041] A continuación se muestra una salida de ejemplo de vectores de probabilidad del clasificador 202 para seis cuadros de imagen consecutivos de la secuencia de video, en el que los seis cuadros de imagen consecutivos cubren un movimiento continuo sobre el escáner. Para un cuadro de imagen i_n , cada vector de probabilidad P_{vn} incluye las probabilidades de clasificación del cuadro de imagen en cada una de las cinco clases "mano", "objeto en mano", "objeto", "parte del cuerpo" y "escáner vacío".

$$P_{v1} = \{0.0, 0.0, 0.28, 0.0, 0.72\}$$

$$P_{v2} = \{0.0, 0.0, 0.28, 0.0, 0.72\}$$

$$P_{v3} = \{0.0, 0.0, 0.01, 0.27, 0.72\}$$

$$P_{v4} = \{0.0, 0.0, 0.28, 0.0, 0.72\}$$

$$P_{v5} = \{0.0, 0.0, 0.28, 0.0, 0.72\}$$

$$P_{v6} = \{0.0, 0.0, 0.28, 0.0, 0.72\}$$

[0042] Puede verse claramente que el vector de probabilidad Pv3 del tercer cuadro de imagen de la secuencia de video es diferente, lo que significa que hay un error en la clasificación del tercer cuadro de imagen por el clasificador 202. El módulo 205 de filtrado y binarización de datos rectifica el error en la clasificación del tercer cuadro de imagen basándose en la información de que los seis cuadros cubren información sustancialmente similar. En una modalidad de la divulgación anterior, el módulo 205 de filtrado y binarización de datos rectifica el error eliminando la trama errónea.

[0043] El módulo 205 de filtrado y binarización de datos se configura entonces para binarizar los valores filtrados de los vectores de probabilidad utilizando un valor de umbral heurístico, de modo que a cada componente de un vector de probabilidad se le asigna un valor "1" si es igual a o mayor que el valor umbral heurístico, en caso contrario "0".

[0044] En un ejemplo, cuando el valor de umbral heurístico es 0,2, los vectores de probabilidad filtrados ejemplares Pvf para cinco cuadros de imagen consecutivos pueden representarse como se muestra a continuación:

$$P_{vf1} = \{0.0, 0.0, 0.0, 0.0, 1.0\}$$

$$P_{vf2} = \{0.0, 0.0, 0.28, 0.0, 0.72\}$$

$$P_{vf3} = \{0.0, 0.0, 0.26, 0.0, 0.74\}$$

$$P_{vf4} = \{0.0, 0.0, 0.39, 0.0, 0.71\}$$

$$P_{vf5} = \{0.0, 0.45, 0.14, 0.0, 0.41\}$$

y los vectores de probabilidad binarizados correspondientes Pvb se pueden representar como se muestra a continuación:

$$P_{vb1} = \{0 \ 0 \ 0 \ 0 \ 1\}$$

$$P_{vb2} = \{0 \ 0 \ 1 \ 0 \ 1\}$$

$$P_{vb3} = \{0 \ 0 \ 1 \ 0 \ 1\}$$

$$P_{vb4} = \{0 \ 0 \ 1 \ 0 \ 1\}$$

$$P_{vb5} = \{0 \ 1 \ 0 \ 0 \ 1\}$$

[0045] Cada vector de probabilidad binarizado Pvb es, por tanto, una cadena binarizada de una serie de números binarios, que puede usarse para determinar transiciones de clases en cuadros consecutivos. Por ejemplo, en el primer cuadro de imagen, el valor binario correspondiente a la clase 'objeto' es '0', y en el segundo cuadro de imagen, el valor binario correspondiente a la clase 'objeto' es '1', lo que significa que hay es claramente una transición de clase del primer al segundo cuadro de imagen. De manera similar, en el cuarto cuadro de imagen, el valor binario correspondiente a la clase "objeto en mano" es "0" y el valor binario correspondiente a la clase "objeto" es "1". En el quinto cuadro, el valor binario de "objeto en mano" cambia a "1" y el valor binario de "objeto" cambia a "0". Esto indica claramente que el usuario ha mantenido el objeto en su mano durante la transición del cuarto al quinto cuadro de imagen. Por lo tanto, los vectores de probabilidad binarizados / cuantificados proporcionan información sobre la transición de clases en cuadros de imágenes consecutivos.

[0046] El módulo de detección de intervalo silencioso 206 está configurado para detectar uno o más intervalos silenciosos en la secuencia de video. En una modalidad de la presente divulgación, durante el intervalo de silencio, no se detecta actividad en la zona de exploración durante una duración de tiempo umbral. En un ejemplo, la duración del umbral de tiempo se puede establecer en '0.5 s', y un intervalo de tiempo de más de 0.5 s se marca como 'intervalo silencioso' cuando permanece el valor binario de la clase "escáner vacío" de los cuadros de imagen correspondientes. '1' durante todo el intervalo de tiempo.

[0047] El módulo 207 de composición de palabras comprimidas está configurado para generar una secuencia de palabras comprimidas en base a las cadenas binarizadas generadas por el módulo 205 de filtrado y binarización de datos. Las palabras comprimidas se generan basándose en la transición de clases desde '1' a '0' y '0' a '1' en cuadros de imagen consecutivos.

[0048] En una modalidad de la presente divulgación, cada palabra se compone de letras de un alfabeto que contiene $2 * N$ letras correlacionadas con la semántica de las acciones del proceso, donde N representa el número de clases. En un ejemplo, si el número de clases es 5, entonces cada palabra se

compone de un total de 10 letras. Para cada clase, una transición "0 -> 1" genera una letra "inicial" específica (por ejemplo, 'O' para la clase Objeto), mientras que una transición "1 -> 0" genera una letra "final" (por ejemplo, 'o' para la clase Objeto).

[0049] [0049] Por lo tanto, el alfabeto para cinco clases: "mano", "objeto en mano", "objeto", "parte del cuerpo" y "escáner vacío", contiene las siguientes letras:

```
classHand arriba:H abajo:h
classHandObject arriba:Q abajo:q
classObject arriba:O abajo:o
classBodyPart arriba: B abajo: b
classEmptyScanner arriba: E abajo: e
```

[0050] En una modalidad de la presente divulgación, dos palabras adyacentes están separadas por al menos un cuadro clasificado como "escáner vacío". Esto podría representar o no un intervalo de silencio dependiendo de la longitud de los valores consecutivos de "1" "escáner vacío".

[0051] Un ejemplo de salida cuantificada generada por el módulo de composición de palabras comprimidas 207 se representa a continuación.:

```
Silencio
OoE
Silencio
OQoOqBobE
Silencio
```

[0052] El clasificador 208 dependiente de la secuencia está configurado para recibir la salida cuantificada del módulo 207 de composición de palabras comprimidas y extraer una o más acciones de exploración de la secuencia continua de transiciones representadas como letras del alfabeto. El clasificador dependiente de secuencia 208 incluye un motor basado en aprendizaje automático, como se usa en este documento se relaciona con un motor que es capaz de estudiar algoritmos y modelos estadísticos y usarlos para realizar de manera efectiva una tarea específica sin usar instrucciones explícitas, confiando

en patrones e inferencias. Los ejemplos del clasificador 208 dependiente de la secuencia incluyen una red neuronal recurrente (RNN), un algoritmo de vecino más cercano K (KNN) y un algoritmo de máquina de vectores de soporte (SVM), y así sucesivamente.

[0053] El clasificador 208 dependiente de la secuencia analiza la secuencia de palabras comprimidas para reconocer la actividad humana de la secuencia de video. La secuencia de palabras comprimidas se analiza para determinar las diversas transiciones de las clases en la región de interés. Tal determinación de las transiciones de las clases conduce al reconocimiento de la actividad humana a partir de la secuencia de video. El clasificador 208 dependiente de la secuencia reconoce las transiciones de la señal de entrada binarizada que sugieren acciones básicas.

[0054] Por tanto, el módulo 204 de generación de firmas cuantificadas proporciona un proceso de cuantificación para las señales de entrada procedentes del clasificador 202 que observan una región de interés donde tiene lugar una actividad. El método para la cuantificación de las transiciones tiene como objetivo reducir la influencia de la variación en el tiempo y la variedad de movimientos de las partes del cuerpo en el reconocimiento de la actividad utilizando el clasificador 208 dependiente de la secuencia.

[0055] La FIG. 3 es un diagrama de flujo que ilustra un método 300 para reconocer la actividad humana a partir de una secuencia de video, de acuerdo con una modalidad de la presente divulgación. Se pueden discutir algunos pasos con respecto al sistema como se muestra en la FIG. 2.

[0056] En el paso 302, un cuadro de imagen de la secuencia de video en una o más clases de un conjunto de clases predefinidas se clasifica mediante un clasificador, en el que el cuadro de imagen se clasifica basándose en la acción del usuario en una región de interés del cuadro de imagen. En una modalidad de la presente divulgación, el clasificador es una red neuronal convolucional. En otra modalidad de la presente divulgación, el conjunto de clases predefinidas para una zona de exploración de Self-check out (auto pago) (SCO) incluye

clases tales como mano, objeto en mano, objeto, parte del cuerpo y escáner vacío.

[0057] En el paso 304, se genera un vector de probabilidad de clase para el cuadro de imagen en base a la clasificación, en el que el vector de probabilidad de clase incluye un conjunto de probabilidades de clasificación del cuadro de imagen en cada clase predefinida. En un ejemplo, un vector de probabilidad de clase P_v es representado por:

$$P_v = \{P_{\text{Hand}}, P_{\text{HandObject}}, P_{\text{Object}}, P_{\text{BodyPart}}, P_{\text{EmptyScanner}}\}$$

Dónde P_{Hand} = Probabilidad de que el cuadro de imagen se clasifique en la clase "mano"

$P_{\text{HandObject}}$ = Probabilidad de que el cuadro de imagen se clasifique en la clase "objeto en mano"

P_{Object} = Probabilidad de que el cuadro de imagen se clasifique en la clase "objeto"

P_{BodyPart} = Probabilidad de que el cuadro de imagen se clasifique en la clase "parte del cuerpo"

$P_{\text{EmptyScanner}}$ = Probabilidad de que el cuadro de imagen se clasifique en la clase "escáner vacío"

[0058] En el paso 306, cada valor de probabilidad del vector de probabilidad de clase se binariza en base a un valor umbral de probabilidad predefinido. En un ejemplo, a cada componente de un vector de probabilidad se le asigna un valor "1" si es igual o mayor que el valor umbral heurístico, en caso contrario "0".

[0059] En el paso 308, se determinan una o más transiciones de una o más clases en uno o más cuadros de imagen consecutivos de la secuencia de video, basándose en los correspondientes vectores de probabilidad binarizados. Por ejemplo, si en el primer cuadro de imagen, el valor binario correspondiente a la clase 'objeto' es '0', y en el segundo cuadro de imagen, el valor binario

correspondiente a la clase 'objeto' es '1', lo que significa que hay claramente una transición de clase del primer al segundo cuadro de imagen.

[0060] En el paso 310, se genera una secuencia de palabras comprimidas en base a una o más transiciones determinadas en uno o más cuadros de imagen consecutivos. Las palabras comprimidas se generan en función de la transición de clases de "1" a "0" y de "0" a "1" en cuadros de imagen consecutivos. En una modalidad de la presente divulgación, una palabra comprimida se forma a partir de letras de un alfabeto que contiene un número de letras equivalente al doble del número de clases predefinidas. Además, cada una de las palabras comprimidas de la secuencia de palabras comprimidas comprende al menos un cuadro de no actividad entre ellos. En un ejemplo, si el número de clases es 5, entonces cada palabra se compone de un total de 10 letras. Para cada clase, una transición "0 -> 1" genera una letra "inicial" específica (por ejemplo, 'O' para la clase Objeto), mientras que una transición "1 -> 0" genera una letra "final" (por ejemplo, 'o' para la clase Objeto).

[0061] En el paso 312, se extraen una o más acciones de usuario basándose en el análisis de la secuencia de palabras comprimidas por un clasificador dependiente de la secuencia. La una o más acciones del usuario pueden usarse para reconocer la actividad humana en el área de exploración de SCO (región de interés de la acción de exploración) y transmite los resultados del reconocimiento a un dispositivo informático del usuario. En algunas modalidades, el dispositivo informático del usuario puede configurarse para almacenar o mostrar los resultados del reconocimiento. En una modalidad de la presente divulgación, el clasificador dependiente de la secuencia es una red neuronal recurrente.

[0062] La presente divulgación también se refiere a productos de software grabados en medios de almacenamiento de datos no transitorios legibles por máquina, en los que los productos de software se pueden ejecutar en hardware informático para implementar métodos de reconocimiento de la actividad humana a partir de una secuencia de video.

[0063] Las modificaciones de las modalidades de la invención descritas anteriormente son posibles sin apartarse del alcance de la invención tal como se define en las reivindicaciones adjuntas. Expresiones como "que incluye", "que comprende", "que incorpora", "que consiste en", "tiene", "es", que se utilizan para describir y reivindicar la presente invención, deben interpretarse de manera no exclusiva, es decir, permitiendo que artículos, componentes o elementos no descritos explícitamente también deban estar presentes. También se debe interpretar que la referencia al singular se refiere al plural. Los números incluidos entre paréntesis en las afirmaciones adjuntas tienen la intención de ayudar a comprender las afirmaciones y no deben interpretarse de ninguna manera para limitar la materia objeto de estas reivindicaciones.