(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **1 323 172**

(21) Número de solicitud: 202330532

(51) Int. CI.:

G06Q 50/10 (2012.01)
G06Q 50/12 (2012.01)
G06V 10/82 (2012.01)
A47F 10/02 (2006.01)

(12) SOLICITUD DE MODELO DE UTILIDAD U

(22) Fecha de presentación:

17.09.2020

(43) Fecha de publicación de la solicitud:

09.10.2025

(71) Solicitantes:

UNIVERSIDAD DE MÁLAGA (100.00%)
Avenida Cervantes, 2
29071 Málaga (Málaga) ES

(72) Inventor/es:

MOLINA CABELLO, Miguel Ángel;
DOMÍNGUEZ MERINO, Enrique;
LÓPEZ RUBIO, Ezequiel;
LUQUE BAENA, Rafael Marcos y
PALOMO FERRER, Esteban José

(54) Título: **Sistema para la identificación y cobro automáticos de consumiciones**

ES 1 323 172 U

DESCRIPCIÓN

Sistema y método de identificación y cobro automáticos de consumiciones

Campo de la invención

La presente invención pertenece al campo de la identificación automática de productos y la gestión de pedidos. La invención es aplicable en diversos sectores, tales como en el sector del equipamiento de restauración (hostelería), en particular en la identificación de productos que aparecen en una bandeja de autoservicio y cálculo de su precio.

Antecedentes de la invención

En restaurantes, cantinas y demás establecimientos de restauración, es deseable ofrecer un servicio de calidad, eficiente y lo más rápido posible. Esto es especialmente crítico durante picos de trabajo que se producen a ciertas horas (horas punta). En este sentido, existen algunas propuestas que han tratado de contribuir a la automatización de los servicios de identificación y/o cobro de productos.

Una de estas propuestas se describe en JPH10255169, que se refiere a un sistema compuesto por una cámara y un dispositivo de irradiación estereoscópica, mediante el cual se identifican tipos de recipientes (platos, etc.) sobre una bandeja. De esta forma, estableciendo previamente una correspondencia entre tipos de recipientes y comidas contenidas en cada tipo de recipiente, es posible identificar indirectamente las comidas servidas en la bandeja, para su posterior cobro. Este sistema es ineficiente, ya que no se identifica el tipo de comida en sí, sino el plato (recipiente) al que esa comida va asociada. Por tanto, basta con servir equivocadamente un tipo de comida en un recipiente al que no se ha asociado esa comida, para realizar una identificación incorrecta de la comida de la bandeja.

CN107341508 divulga un método y sistema para clasificar imágenes en dos categorías: las que contienen comida y las que no contienen comida. El método se basa en una red neuronal *Inception-BN* que ha sido previamente entrenada mediante U202330532 28-03-2023 un conjunto de imágenes (incluyendo comida o no), de forma que clasifica una imagen como una clase ("con comida" o "sin comida").

US2016/0063692A1 divulga un sistema y método de reconocimiento de la comida que aparece en una fotografía, mediante técnicas de aprendizaje automático y otras técnicas, tales como algoritmos de detección de características y algoritmos de similaridad. Este sistema precisa de una serie de módulos de aprendizaje (de aprendizaje jerárquico, de aprendizaje contextual, de aprendizaje específico de usuario, o de aprendizaje de co-ocurrencia) que a su vez requieren de feedback por parte del usuario.

A su vez, US2003/0076983A1 divulga un analizador de comida para identificar el tipo de comida de un plato que aparece en una fotografía. El analizador utiliza técnicas de aprendizaje automático. Para una correcta identificación del producto presente en el plato, este analizador necesita un dispositivo de referencia que proporcione unas características de color, tamaño, forma y altura de referencia. Por tanto, sin el dispositivo de referencia, el analizador no es capaz de realizar una identificación correcta del producto.

Por lo tanto, existe la necesidad de un nuevo método y sistema totalmente automáticos de identificación y cobro de productos.

Descripción de la invención

La presente invención proporciona un método y sistema para la detección automática de los productos que un cliente lleva en una bandeja. El método y sistema utilizan una imagen, por ejemplo, tomada por una cámara, tal como una cámara cenital, y técnicas de visión artificial. La detección automática de los productos, por ejemplo, consumiciones, permite además al usuario abonar los productos sin necesidad de un cobrador humano (autopago), teniendo en cuenta que el precio puede variar en función de la combinación de productos existentes en la bandeja. Se consigue así agilizar el proceso de cobro, evitando que se produzcan esperas innecesarias.

El sistema utiliza una cámara de visión artificial que enfoca, normalmente desde arriba, al lugar donde el cliente debe depositar la bandeja de autoservicio para su valoración.

La cámara captura una imagen de la bandeja y transmite la imagen adquirida a un dispositivo dotado con medios de procesado. Los medios de procesado implementan un software de visión artificial, basado en una red neuronal convolucional profunda, que se encarga de detectar y localizar los productos (por ej. consumiciones) presentes en la imagen de la bandeja. Esta red ha sido entrenada previamente con un conjunto de imágenes de platos, con objeto de que pueda generalizar correctamente cuando tenga como entrada una nueva imagen. Así, a partir de la imagen captada, se realiza una clasificación n-aria entre distintos productos establecidos para posteriormente calcular el precio total. Además, se detecta cada plato, su posición y su contenido.

En el contexto de la presente invención, se define un producto como un conjunto de alimentos que se vende de forma unificada por un precio en el establecimiento. Por tanto, un producto se corresponde con una determinada línea de la carta del establecimiento. Normalmente un producto se proporciona de forma unificada en un mismo recipiente. Ejemplos de productos: filete de ternera con guarnición, flan, refresco o pan. No son productos objetos como los cubiertos o vasos sin bebida.

En el contexto de la presente invención, se define una "clase" como un objeto que el sistema de visión puede reconocer directamente. Por ejemplo, filete de ternera con ensalada es una clase, mientras que filete de ternera con patatas fritas es otra clase distinta. Sin embargo, aunque sean diferentes clases, representan al mismo producto (filete de ternera con guarnición) ya que pertenecen a una misma línea de la carta. Del mismo modo, filete de cerdo con ensalada es una clase, mientras que filete de ternera con ensalada es otra clase. En este caso, ambas clases se corresponden a productos distintos dado que se asocian a diferentes líneas de la carta (filete de cerdo con guarnición y filete de ternera con guarnición). Nótese que el precio de un producto podrá variar para distintas clases de un mismo producto, por ejemplo, si varía el precio de su guarnición. Otro ejemplo: la Coca Cola, la Fanta, el Sprite, la Pepsi, son clases distintas, pero son el mismo producto (refresco). No son clases los cubiertos ni los vasos sin bebida, ya que el sistema de visión está diseñado para que no los reconozca. El sistema de la invención detecta cada uno de los productos que aparecen en la bandeja, ya que los productos están asociados a las clases reconocibles por el sistema.

Por último, en el contexto de la presente invención, un "menú" es un conjunto de productos que se vende por un precio menor que la suma de los precios de los productos que lo componen. Ejemplo de menú: producto1 (ensalada - primer plato), producto2 (boquerones con ensalada - segundo plato), producto3 (cerveza), producto4 (pan), y producto5 (natillas).

Finalmente, a partir de la información generada por la red, correspondiente a los productos detectados en la imagen, se calcula el precio total de los mismos (por ejemplo, de las consumiciones de la bandeja, es decir, de la imagen). Esto se realiza preferentemente en los

medios de procesado. El precio total se transmite a otro dispositivo, por ejemplo, a un terminal de punto de venta, para que el cliente pueda realizar el pago.

En un primer aspecto de la presente invención, se proporciona un sistema para la identificación y cobro automáticos de consumiciones, que comprende: una cámara para la captura de una imagen de una bandeja que contiene al menos un producto; unos medios para procesar información configurados para identificar, a partir de la imagen de la bandeja, el al menos un producto contenido en la bandeja y para calcular el precio del al menos un producto; y un dispositivo de cobro conectado a los medios para procesar información, para cobrar automáticamente al usuario el precio calculado por los medios para procesar información. Los medios para procesar información comprenden unos medios de detección de objetos basados en una red neuronal convolucional profunda configurada para, a partir de la imagen de la bandeja, determinar al menos una región de la imagen en la que hay un producto y clasificar el producto identificado en la al menos una región de acuerdo con una clase previamente establecida.

En realizaciones del primer aspecto de la invención, la red neuronal convolucional profunda es una red de tipo detección.

En realizaciones del primer aspecto de la invención, la red neuronal convolucional profunda proporciona un vector por cada producto detectado, donde cada vector define una región que engloba un producto y un valor de probabilidad de pertenencia a una clase.

En realizaciones del primer aspecto de la invención, cada vector define dicha región mediante las coordenadas (x, y) de una esquina de la región y la anchura y altura de la región.

En realizaciones del primer aspecto de la invención, la red neuronal convolucional profunda divide la imagen en una rejilla de S x S celdas, donde en cada celda la red neuronal convolucional profunda está configurada para predecir una clase de un conjunto de C clases diferentes.

En realizaciones del primer aspecto de la invención, la red neuronal convolucional profunda implementa una arquitectura de capas convolucionales y capas de *max-pooling*, e incluye además dos capas totalmente conectadas en la parte final.

En realizaciones del primer aspecto de la invención, la red de tipo detección se basa en una regresión a partir de los pixeles de la imagen de la bandeja hasta obtener dicha al menos una región de la imagen en la que hay un producto.

En realizaciones del primer aspecto de la invención, la red de tipo detección se basa en una red neuronal principal de detección de objetos y una red neuronal auxiliar configurada para proponer regiones candidatas a albergar un producto.

En realizaciones del primer aspecto de la invención, la red neuronal convolucional profunda es una red de tipo clasificación combinada con técnicas de división de la imagen.

En realizaciones del primer aspecto de la invención, los medios para procesar información comprenden medios de cálculo de precios conectados a una base de datos que comprende una relación de clases de productos y los precios asociados a cada clase de productos.

En realizaciones del primer aspecto de la invención, los medios para procesar información comprenden medios de pre-procesamiento configurados para eliminar ruido de la imagen de la bandeja captada por la cámara y para compensar diferencias de iluminación en el establecimiento.

En realizaciones del primer aspecto de la invención, los medios para procesar información comprenden medios de post-procesamiento configurados para eliminar posibles detecciones de consumiciones erróneas.

En realizaciones del primer aspecto de la invención, la red neuronal convolucional profunda ha sido entrenada previamente con un conjunto de imágenes de bandejas de consumiciones, donde las imágenes de entrenamiento incluyen una pluralidad de ejemplos de cada producto a consumir; y donde las imágenes de entrenamiento han sido etiquetadas indicando las regiones de la imagen en las que hay un producto y asociando una clase a dicho producto.

En un segundo aspecto de la presente invención, se proporciona un método para la identificación y cobro automáticos de consumiciones, que comprende: capturar una imagen de una bandeja que contiene al menos un producto; identificar, a partir de la imagen de la bandeja, el al menos un producto contenido en la bandeja, y calcular el precio del al menos un producto; y cobrar automáticamente al usuario el precio calculado. La identificación del al menos un producto contenido en la bandeja se realiza mediante una red neuronal convolucional profunda configurada para, a partir de la imagen de la bandeja, determinar al menos una región de la imagen en la que hay un producto y clasificar el producto identificado en la al menos una región de acuerdo con una clase previamente establecida.

En realizaciones del segundo aspecto de la invención, la red neuronal convolucional profunda es una red de tipo detección.

En realizaciones del segundo aspecto de la invención, la red neuronal convolucional profunda proporciona un vector por cada producto detectado, donde cada vector define una región que engloba un producto y un valor de probabilidad de pertenencia a una clase.

En realizaciones del segundo aspecto de la invención, cada vector define dicha región mediante las coordenadas (x, y) de una esquina de la región y la anchura y altura de la región.

En realizaciones del segundo aspecto de la invención, la red neuronal convolucional profunda divide la imagen en una rejilla de S x S celdas, donde en cada celda la red neuronal convolucional profunda está configurada para predecir una *clase* de un conjunto de C clases diferentes.

En realizaciones del segundo aspecto de la invención, la red neuronal convolucional profunda implementa una arquitectura de capas convolucionales y capas de *max-pooling*, e incluye además dos capas totalmente conectadas en la parte final.

En realizaciones del segundo aspecto de la invención, la red de tipo detección se basa en una regresión a partir de los pixeles de la imagen de la bandeja hasta obtener dicha al menos una región de la imagen en la que hay un producto.

En realizaciones del segundo aspecto de la invención, la red de tipo detección se basa en una red neuronal principal de detección de objetos y una red neuronal auxiliar configurada para proponer regiones candidatas a albergar un producto.

En realizaciones del segundo aspecto de la invención, la red neuronal convolucional profunda es una red de tipo clasificación combinada con técnicas de división de la imagen.

En realizaciones del segundo aspecto de la invención, para calcular el precio del al menos un producto, se accede a una base de datos que comprende una relación de clases productos y los precios asociados a cada clase.

En realizaciones del segundo aspecto de la invención, antes de proporcionar la imagen a la red neuronal convolucional profunda, se elimina ruido de la imagen de la bandeja captada por la cámara y se compensan diferencias de iluminación en el establecimiento.

En realizaciones del segundo aspecto de la invención, tras aplicar la red neuronal convolucional profunda, se eliminan posibles detecciones de consumiciones erróneas.

En realizaciones del segundo aspecto de la invención, la red neuronal convolucional profunda ha sido entrenada previamente con un conjunto de imágenes de bandejas de consumiciones, donde las imágenes de entrenamiento incluyen una pluralidad de ejemplos de cada producto a consumir; y donde las imágenes de entrenamiento han sido etiquetadas indicando las regiones de la imagen en las que hay un producto y asociando una clase a dicho producto.

En un tercer aspecto de la presente invención, se proporciona un producto de programa informático que comprende instrucciones / código de programa informático para llevar a cabo las etapas del método descrito.

En un cuarto aspecto de la presente invención, se proporciona una memoria / soporte legible por ordenador que almacena instrucciones / código de programa para llevar a cabo las etapas del método descrito.

Frente a propuestas anteriores, que utilizan técnicas clásicas de aprendizaje automático (machine learning), visión por computador y análisis estadístico para la detección de elementos en la imagen, el método y sistema de la presente invención aplican técnicas de aprendizaje profundo, en concreto redes convolucionales. Gracias a las técnicas usadas, se evitan limitaciones, como la necesidad de colocar los productos de una determinada manera, o utilizar bandejas o recipientes de determinadas dimensiones y características. Además, el rendimiento en la identificación de las consumiciones es considerablemente mayor que el rendimiento de propuestas basadas en aprendizaje automático.

Además, la posibilidad de cobro sin necesidad de cobrador humano, derivada de la detección automática de los tipos de consumiciones de la bandeja, permite agilizar el proceso de cobro, evitando que se produzcan esperas innecesarias. Lo que, es más, el método y sistema permiten indirectamente recabar estadísticas sobre los hábitos de consumo en el establecimiento.

Ventajas y características adicionales de la invención serán evidentes a partir de la descripción en detalle que sigue y se señalarán en particular en las reivindicaciones adjuntas.

Breve descripción de las figuras

Para complementar la descripción y con objeto de ayudar a una mejor comprensión de las características de la invención, de acuerdo con un ejemplo de realización práctica de la misma, se acompaña como parte integrante de la descripción, un juego de figuras en el que, con carácter ilustrativo y no limitativo, se ha representado lo siguiente:

La figura 1 muestra una vista esquemática con una posible configuración de un sistema de acuerdo con la presente invención.

La figura 2A muestra esquemáticamente un diagrama de bloques de los medios de procesado para la detección de productos y cálculo de su precio, del sistema de la presente invención.

La figura 2B muestra más en detalle los medios de procesado para la detección de productos y cálculo de su precio, del sistema de la presente invención.

La figura 3 muestra una vista aérea detallada de una posible configuración de productos en la bandeja.

La figura 4 muestra una vista aérea detallada de otra posible configuración de productos en la bandeja.

La figura 5 muestra un diagrama de flujo de una etapa de entrenamiento de una red neuronal convolucional profunda, de acuerdo con una posible implementación de la invención.

La figura 6 muestra un ejemplo de resultados obtenidos por el método y sistema de la invención a partir de una imagen de una bandeja.

Descripción de una forma de llevar a cabo la invención

A la vista de las mencionadas figuras, y de acuerdo con la numeración adoptada, se puede observar en ellas un ejemplo de realización preferente de la invención, la cual comprende las partes y elementos que se indican y describen en detalle a continuación.

La figura 1 muestra un esquema de una posible configuración del sistema propuesto para la identificación de productos y cobro de los mismos, de forma automática, es decir, sin intervención humana. El sistema incluye una cámara 4, unos medios para procesar información 5 y un dispositivo de cobro 6, tal como un terminal de punto de venta. En una posible implementación, los medios para procesar información 5, la cámara 4 y el dispositivo de cobro 6 se conectan a través de una red local.

La cámara 4 debe disponerse de forma que enfoque a una ubicación destinada a recibir a una bandeja 1 sobre la que se encuentra un menú formado de forma general por varios productos. Esto se consigue, por ejemplo, con una cámara cenital, es decir, una cámara cuyo punto de vista se encuentra perpendicular respecto del suelo (concretamente, de la bandeja cuyo contenido se va a detectar) y la imagen obtenida ofrece un campo de visión orientado de arriba abajo. Tal y como se ha definido, un producto es un conjunto de alimentos que se vende de forma unificada, y por lo tanto se corresponde con una determinada línea de la carta del establecimiento. Por ejemplo, filete de ternera con guarnición es un producto, flan es otro producto, refresco es otro producto y pan es otro producto. Un menú es un conjunto de productos formado típicamente por primer plato (un producto, por ejemplo, ensalada), segundo plato (otro producto, por ejemplo, boquerones con ensalada), pan (otro producto), bebida (otro producto, por ejemplo, cerveza) y postre (otro producto, por ejemplo, natillas). De esta forma, en función de las preferencias del establecimiento, se habrán definido una serie de productos identificados como "primer plato" (por ejemplo, sopa, macarrones, lentejas, etc.), otra serie de productos identificados como "segundo plato" (por ejemplo, filete de ternera con ensalada, filete de ternera con patatas fritas, merluza con verduras, etc.), otra serie de productos identificados como "bebida" (por ejemplo, agua, vino, refresco, etc.), etc. En el ejemplo de la figura 1, la bandeja 1 muestra una posible configuración de menú, compuesto por un primer plato (en este ejemplo el producto es sopa 12 servida en un cuenco 13), un segundo plato (en este ejemplo el producto es un filete 11 acompañado de ensalada 10, servidos en un plato llano 9), postre (en este ejemplo el producto es un trozo de tarta 7 servido en un plato de postre 8), bollo de pan 2 (otro producto) y bebida (en este ejemplo el producto es una lata de refresco 3). En una posible realización, a cada producto se le asocia una etiqueta para, posteriormente, calcular el precio de los productos de la bandeja 1. Nótese que la figura 6 muestra una vista diferente del mismo menú de la figura 1, a pesar de que los productos aparecen dispuestos en diferente posición y que en la figura 1 no aparecen los cubiertos.

La cámara 4 es una cámara convencional, adecuada para la captura de imágenes de alta resolución. El sensor de visión de la cámara 4 tiene preferentemente alta definición (HD, del inglés High Definition). El sensor de la cámara 4 tiene una resolución mínima de 640x480 pixeles, pudiendo tener mayor resolución (por ejemplo, pero de forma no limitativa, 1024x780 pixeles, 1280x720 píxeles o 1920×1080 pixeles). La cámara 4 tiene además una óptica de alta calidad para la captación de imágenes de acuerdo con, al menos, la citada resolución mínima. La cámara 4 tiene también una interfaz de red, por ejemplo, una interfaz de red local, para la transmisión de las imágenes captadas a los medios para procesar información 5. La interfaz de red puede ser cableada o inalámbrica.

Los medios para procesar información 5 pueden ser cualquier dispositivo dotado con al menos un procesador y medios de almacenamiento de datos. Los medios 5 pueden ser un ordenador personal, ya sea de sobremesa o portátil, o cualquier otro dispositivo portátil, tal como una Tablet o un Smartphone. Los medios 5 deben tener capacidad de memoria y procesado suficiente para implementar un software de visión artificial basado en una red neuronal convolucional profunda, para la detección y localización de los productos presentes en la imagen de la bandeja. A modo de ejemplo, sin carácter limitativo, los medios 5 pueden ser un ordenador personal que tiene un procesador Intel Core i7, 16 G B de memoria RAM y 1TB de disco duro para almacenar las imágenes capturadas y el software necesario, con sistema operativo Linux y con una unidad de procesamiento gráfico (GPU, del inglés Graphics Processing Unit) que permite la ejecución de redes neuronales convolucionales. Esto permite que la detección de las consumiciones esté implementada mediante aceleración de gráficos por hardware. Así, el sistema de visión de la invención reconoce distintas clases. Por ejemplo, filete de ternera con ensalada es una clase, mientras que filete de ternera con patatas fritas es otra clase distinta. Sin embargo, aunque sean diferentes clases, representan al mismo producto (filete de ternera con guarnición).

Los medios para procesar información 5 para la detección de consumiciones a partir de una imagen y cálculo de su precio, se esquematizan en las figuras 2A y 2B. El sistema tiene, embebidos en los medios 5, un primer módulo 51 (módulo de visión artificial), para la detección automática de clases de productos que contiene cada bandeja de autoservicio 1, y un segundo módulo 52 que recibe la información de las clases detectadas por el módulo de visión artificial 52 y que calcula el precio total de los productos detectados. Estos módulos 51, 52 se implementan como un conjunto de algoritmos expresados mediante código informático (software). Desde el punto de vista hardware, el primer y segundo módulos 51, 52 (por ejemplo, el código informático que los implementa) pueden estar almacenados en un mismo sistema de almacenamiento (en general, medio de memoria), debidamente particionada, o en sistemas de almacenamiento separados. Típicamente se ejecutan en uno o varios procesadores, en función de la configuración de los medios 5. El primer módulo 51 (de visión artificial) se encarga de la detección de objetos (clases) y de localizar las consumiciones presentes en la bandeja, para lo que comprende un sub-módulo de detección de objetos 512, configurado para detectar múltiples consumiciones en la imagen adquirida, independientemente de la colocación de las consumiciones dentro de la bandeja y del tipo de bandeja utilizada. Este sub-módulo 512 implementa un software de visión artificial, basado en una red neuronal convolucional profunda, que se encarga de detectar y localizar las consumiciones presentes en la imagen de la bandeja. La red neuronal aprende intrínsecamente las características de cada producto para poder detectar los productos en imágenes. Además, la red es entrenada previamente mediante un algoritmo de aprendizaje con un conjunto elevado de imágenes de platos (incluidos comidas y bebidas), con objeto de que pueda generalizar correctamente cuando tenga como entrada una nueva imagen. Así, a partir de la imagen captada, se realiza una clasificación n-aria entre distintos tipos de comida establecidos para posteriormente calcular el precio total. Además, se detecta cada plato, su posición y su contenido. Puesto que el objetivo es el cálculo del precio final de los productos consumidos por un usuario, la red neuronal no solo detecta el contenido del producto (la comida o bebida en sí misma), sino que determina la clase (por ejemplo, filete con patatas) a la que pertenece cada

producto detectado, para después cotejar con la base de datos del establecimiento para determinar el precio de dicho producto (filete con guarnición) en función de si es primer plato, segundo plato, etc. Preferentemente, el primer módulo 51 tiene además un sub-módulo de pre-procesamiento 511 y un sub-módulo de post-procesamiento 513.

El sub-módulo de pre-procesamiento 511 está configurado para realizar tareas de adecuación de la imagen a analizar. Preferentemente se encarga de la compensación de iluminación externa (iluminación presente en el establecimiento), tal como ajuste automático de brillo y contraste (etapa 511a en la figura 2B); de la eliminación de ruido (etapa 511b en la figura 2B); y del re-escalado de la imagen (etapa 511c en la figura 2B), necesario porque habitualmente la cámara obtiene la imagen con una resolución, mientras que la red necesita como entrada una imagen con una resolución específica, que varía en función de la red utilizada. Por ejemplo, en el caso de la red YOLO, esta resolución es de 416x416 píxeles.

El sub-módulo de post-procesamiento 513 está configurado para eliminar errores, tales como posibles detecciones de productos erróneos. Para saber si un producto es erróneo, pueden hacerse por ejemplo estimaciones de la localización de productos, del tamaño y de la confianza de las detecciones. En particular, el post-procesamiento 513 puede incluir una primera etapa de comprobación de la localización y confianza de los objetos detectados (etapa 513a en la figura 2B), por ejemplo, comprobando que el tamaño del plato es de unas dimensiones razonables, y una segunda etapa de extracción de los productos detectados (etapa 513b en la figura 2B).

El segundo módulo 52 recibe la información de qué consumiciones ha detectado en la bandeja 1 el primer módulo 51 de visión artificial (es decir, recibe la salida de la etapa de post-procesamiento 513, si la hubiera) y calcula el precio total de los productos detectados a partir de un conjunto de datos relativos a descripciones de productos que están a la venta en el establecimiento y precios de los productos correspondientes (etapa 52a en la figura 2B). En función de las preferencias establecidas por el restaurante, los precios pueden establecerse por tipos de producto. Por ejemplo, todas las consumiciones catalogadas como "primer plato" pueden tener un mismo precio, todas las consumiciones catalogadas como "segundo plato" pueden tener un mismo precio, etc. Alternativamente, se puede establecer precios específicos para productos concretos. Los pares de datos de consumiciones (o tipos de consumiciones) y precios se almacenan preferentemente en una base de datos 525 de consumiciones y precios, a la que puede conectarse el segundo módulo 52. Esta base de datos 525 puede almacenarse de forma local al sistema, por ejemplo, en medios de almacenamiento de datos comprendidos en los medios para procesar información 5, o de forma remota al sistema. El segundo módulo 52 se conecta a la base de datos 525 a través de una interfaz de comunicación correspondiente, ya sea una interfaz para conexión local o remota. La interfaz puede ser cableada o inalámbrica.

Por último, el dispositivo de cobro 6, por ejemplo, terminal de punto de venta, permite el cobro de las consumiciones sin intervención de ningún cobrador humano. El dispositivo 6 está conectado a través de una interfaz de comunicación con los medios para procesar información 5, concretamente con el segundo módulo 52, para acceder al precio total de la consumición calculado por este en la etapa 52a. La interfaz de comunicación puede ser, por ejemplo, una red local, o cualquier tipo de red de comunicación dedicada, ya sea cableada o inalámbrica. El dispositivo 6, una vez recibido el precio calculado por el módulo 52, está configurado para generar el ticket correspondiente a las consumiciones contendidas en la bandeja 1. El cliente puede realizar el pago mediante cualquier medio de pago convencional, preferentemente medio de pago electrónico, tal como tarjeta de débito o crédito, teléfono móvil o tarjeta de fidelización. El dispositivo 6 puede imprimir el ticket y dispensarlo al cliente para que lo conserve como resguardo de la transacción económica efectuada.

La figura 3 muestra, mediante una vista aérea, una posible configuración de una bandeja 1 con los diferentes elementos que componen un menú típico de autoservicio en un comercio de hostelería. Pueden observarse los instrumentos de mesa: tenedor 15, cuchillo 16, cuchara grande 17 y cuchara pequeña 14. El primer plato del menú es una sopa 12 (un producto) que viene servida en un cuenco 13. El segundo plato del menú es un filete 11 acompañado de ensalada 10 (otro producto), servidos en un plato llano 9. El postre del menú es una tarta 7 (otro producto) que viene servida en un plato de postre 8. El menú incluye un bollo de pan 2 (otro producto) y una lata con refresco 3 (otro producto). La figura 4 muestra mediante una vista aérea otra posible configuración de una bandeja 1, diferente de la configuración mostrada en la figura 3. Pueden observarse los instrumentos de mesa que se utilizan para comer los alimentos: tenedor 15, cuchillo 16, cuchara grande 17 y cuchara pequeña 14, dispuestos sobre la bandeja 1 de forma diferente a como están dispuestos en la bandeja 1 de la figura 2. El primer plato del menú es una ensalada 10 (un producto) que viene servida en un plato llano 9. El segundo plato del menú es un filete 11 acompañado de patatas fritas 19 (otro producto), servidos en un plato llano 9. El postre del menú es una tarta 7 (otro producto) que viene servida en un plato de postre 8. El menú incluye un bollo de pan 2 (otro producto) y una botella de agua 3 (otro producto). Se proporciona un vaso 18 para poder servirse el agua de una forma más cómoda. En las figuras 3 y 4 aparecen identificados todos los elementos tal y como los entendería una persona; sin embargo, el sistema y método de la invención identifican los productos tal y como se indica por ejemplo en la figura 7. La figura 7 muestra la detección del menú que realiza el sistema propuesto, estando compuesto este ejemplo de menú por: producto1 (sopa - primer plato) 74, producto2 (filete de ternera con verduras - segundo plato) 75, producto3 (refresco) 71, producto4 (pan) 73 y producto 5 (tarta) 72. Concretamente, el módulo 51 de visión artificial reconoce en la imagen de la figura 7 cinco clases que representan los cinco productos 71-75.

El sub-módulo de detección de objetos 512 del módulo 51 de visión artificial implementa una red neuronal convolucional profunda que, dada una imagen de entrada de una bandeja y una serie de productos dispuestos sobre la bandeja, identifica las clases de objetos que hay en la bandeja (por ejemplo, filete con patatas es una clase, filete con ensalada es otra clase; ambas asociadas al producto filete con guarnición) y la localización de los objetos que ha detectado en dicha imagen. Para ello, la red procesa la imagen, detecta los objetos que aparecen en la misma y los clasifica en la clase correspondiente. La red detecta los objetos de la imagen independientemente de la posición y orientación en la que aparezcan. La red clasifica los productos que aparecen en la imagen de la bandeja en una de las posibles clases de productos que se hayan definido previamente, durante el proceso de entrenamiento de la red neuronal, asignando la probabilidad de pertenencia a la clase en la que el producto se ha clasificado. La red neuronal no determina el volumen de los productos, ya que esta información no es necesaria para determinar el tipo de producto.

Concretamente, la red neuronal procesa la imagen de entrada, determinando cada región (RoI, del inglés Region o f Interest) de dicha imagen en la que hay un producto. Cada región en la que hay un producto se determina preferentemente mediante un recuadro (área rectangular, BB, del inglés bouding box). Cada producto detectado, es decir, cada área rectangular mínima detectada, es clasificado en una de las posibles clases de productos previamente definidas (por ejemplo, filete de ternera con ensalada es una clase, mientras que filete de ternera con patatas fritas es otra clase distinta, pero ambas representan al mismo producto: filete de ternera con guarnición), indicándose además la probabilidad estimada de pertenencia del producto a la clase en la que se ha clasificado. Una vez la red finaliza la detección y clasificación de los productos, devuelve como salida una indicación de las regiones detectadas (por ejemplo, posición y tamaño de cada recuadro) y las probabilidades de clase que corresponde con cada región. La entrada a la red es la imagen captada por la cámara 4, preferentemente pre-procesada en el sub-módulo 511. Preferentemente, la salida de la red es un vector por cada objeto (clase) detectado. Cada vector define el recuadro que engloba el objeto correspondiente (bounding box) y un valor de

probabilidad o confianza (entre 0 y 1) de pertenencia a una clase (en este caso, a un tipo de producto). Así, cada vector tiene V elementos, de los que V-1 definen el recuadro y un elemento define la confianza. La red neuronal convolucional profunda implementa una arquitectura de capas convolucionales y capas de max-pooling, añadiendo dos capas totalmente conectadas en la parte final.

En una posible implementación, la red neuronal convolucional profunda es una red de tipo detección. Ejemplos no limitativos de redes neuronales convolucionales profundas de tipo detección que pueden usarse, son:

(a) Redes basadas en considerar el problema de la detección de múltiples objetos como una regresión, directamente desde los pixeles de la imagen de entrada a las coordenadas de los recuadros (bounding boxes) y las probabilidades de clase de salida. Un ejemplo de estas redes es la red YOLO (del inglés, You Look Only Once). En la red YOLO, la imagen de entrada se divide en S x S regiones. Por cada objeto presente en la imagen, una región es responsable de predecir dicho objeto. Cada región puede predecir B recuadros (en inglés, bounding boxes, que son recuadros que pueden englobar un objeto) y C probabilidades de clase.

(b) Redes basadas en utilizar una red neuronal auxiliar para proponer regiones prometedoras (RPN, Region Proposal Network), que comparte rasgos con la red principal de detección de objetos, facilitando así una propuesta de regiones con un coste computacional muy reducido. Este sistema está formado por una red principal y una red auxiliar. La red principal es la que detecta los distintos objetos que hay dentro de la imagen de entrada, mientras que la red auxiliar proporciona ventanas candidatas que parece que podrían delimitar un objeto. Un ejemplo de estas redes es la red *Faster R-CNN* (Faster Region Convolutional Neural Network).

(c) Redes basadas en una deconvolución progresiva para aumentar la resolución de los mapas de rasgos, y una combinación de rasgos entre una vía de convolución y otra vía de deconvolución. Un ejemplo de estas redes es la red *DSSD* (Deconvolutional Single Shot Detector).

Las tres redes anteriores ofrecen como salida los recuadros o regiones (bounding box) en los que hay un producto y la clasificación de los mismos con su probabilidad de acierto.

(d) Redes basadas en extender la Faster R-CNN, añadiendo una rama que predice una máscara de objetos en paralelo con la detección de un recuadro (en inglés, bounding box). En este tipo de redes neuronales, además de proporcionar la información de las clases identificadas, también se proporciona una imagen segmentada, identificando los pixeles que se corresponden con cada uno de los productos detectados. Estas redes ofrecen como salida los pixeles de cada objeto detectado, por lo que ofrecen una solución con más detalle. Un ejemplo de estas redes es la red Mask R-CNN.

En una implementación alternativa, la red neuronal convolucional profunda es una red de tipo clasificación combinada con una división de la imagen, por ejemplo, en una rejilla. En este caso, un algoritmo genera una pluralidad (típicamente, cientos) de regiones o recuadros en posiciones aleatorias dentro de la imagen y de tamaños diversos, para después enviar esta región a la red convolucional con objeto de clasificar dicha región en una de las clases conocidas. Este algoritmo de generación de regiones tiene el mismo objetivo que la red auxiliar de la opción (b). Si en esa región la red no detecta ninguna de las clases entrenadas para ello, la probabilidad será 0 o muy baja para esa clase.

Para que la red neuronal convolucional profunda detecte y clasifique los objetos de forma correcta, es necesario realizar un proceso de entrenamiento previo de la red con un conjunto de

datos lo suficientemente grande y que contiene una muestra representativa de todos los objetos a identificar. Para conseguir un conjunto de imágenes suficientemente representativo de todos los objetos que pueden consumirse en un establecimiento determinado a partir de imágenes originales proporcionadas por dicho establecimiento, se pueden utilizar técnicas de aumentación de datos mediante procesamiento de la imagen y redes neuronales generativas (GANs). En una posible realización, el procedimiento de entrenamiento comprende: Realización de un cierto número de imágenes cenitales de bandejas con consumiciones (preferentemente, un mínimo de 1000 imágenes), con un número mínimo de ejemplos de cada tipo de producto (preferentemente, un mínimo de 50 ejemplos de cada tipo de producto); Etiquetado de cada imagen, indicando la ubicación y tamaño del recuadro mínimo (área mínima rectangular, *bounding box*) que encierra a cada producto presente en la imagen y el tipo de producto que representa dicha área (clase asociada); Creación de un número elevado de imágenes artificiales mediante una red neuronal generativa (GAN) preentrenada, tal como al menos 8 veces, o al menos 10 veces, el número de imágenes realizadas; artificiales por ejemplo, se crean 10.000 imágenes artificiales. La creación de estas, por ejemplo 10.000, imágenes artificiales se corresponde con la aumentación de imágenes. Este proceso incrementa la diversidad de los datos disponibles para el entrenamiento de la red neuronal. Cada imagen artificial proviene de una modificación de una imagen original, por ejemplo, tras recortar, rotar o voltear la imagen original. División del conjunto de datos (las imágenes realizadas de las bandejas con productos, junto con su tamaño de bounding box y clase asociada) en un conjunto de entrenamiento, que sirve para entrenar la red, y uno de test, que sirve para medir el rendimiento de la red, siendo ambos conjuntos balanceados (cada conjunto tiene aproximadamente la misma distribución de clases que la muestra original). El conjunto de entrenamiento tiene preferentemente al menos el 70% de los datos, tal como el 75% o el 80% de los datos. El conjunto de test tiene preferentemente como máximo el 30% de los datos, tal como el 25% o el 20% de los datos.

Los pasos del proceso de entrenamiento se resumen en el diagrama de flujo 60 de la figura 5. En primer lugar (etapa 61), se inicializan o pre-entrenan un conjunto de M capas convolucionales de la red elegida, usando un conjunto de datos (imágenes). Por ejemplo, se usan los datos ImageNet-1000, que es un conjunto de datos formado por imágenes etiquetadas con el nombre del objeto que aparece en ellas (por ejemplo, gato, coche, ...), que provienen de la versión de la base de datos ImageNet que considera 1000 *clases*. En una posible implementación, M=20. Nótese que la eficacia de la detección depende del valor M, de forma que, en general, a mayor M, más eficacia, pero también más tiempo de entrenamiento. El tamaño del conjunto de datos se escala convenientemente, si es necesario, para que se adecúe a la entrada de la red. Esta inicialización d a como resultado "Época 1" (etapa 62). Se alcanza una época cuando todo el conjunto de imágenes ha sido proporcionado a la red neuronal para su entrenamiento de forma iterativa. Normalmente, una red neuronal necesita varias épocas para que la red entrenada proporcione buenos resultados. es decir, a la red neuronal se le proporciona todas las imágenes etiquetadas de forma repetitiva. Este proceso (entrenamiento) converge en un estado de equilibrio, en el cual la red proporciona unos resultados aceptables. A continuación, se vuelve a entrenar la red, pero en este caso con el conjunto de imágenes de bandejas de *menú* con los *productos* que pueden consumirse en el establecimiento. Estas imágenes deben estar etiquetadas. El conjunto de imágenes de entrenamiento debe contener imágenes de todas las combinaciones posibles de productos, es decir, de cada clase (por ejemplo, pollo con ensalada), para que éstas puedan ser reconocidas por la red. En estas imágenes, cada *clase* puede aparecer sola en la bandeja o junto a otras clases en la misma bandeja, con un cierto mínimo de ejemplos por cada *clase* (por ejemplo, mínimo 50 ejemplos) y etiquetando cada *clase*, indicando la ubicación y tamaño del recuadro mínimo (área mínima rectangular, bounding box) que delimita a cada *clase* presente en la imagen. Para cada imagen de entrada, su etiqueta puede comprender un número N de vectores igual al número N de posibles productos a consumir en el establecimiento. Y por cada uno de los N vectores, hay V elementos de vector: un primer elemento de vector relativo a la confianza y V-1 elementos de vector que definen el recuadro que

engloba el objeto correspondiente. Es decir, cada imagen de entrenamiento proporcionada como entrada tiene definido un bounding box por cada producto que haya en dicha imagen y la clase de cada uno de esos bounding boxes. En una posible realización, V=5. En este caso hay V-1=4 elementos del recuadro y pueden representar el valor (x, y) de la esquina superior izquierda y el ancho y el alto del recuadro. En este caso, la etiqueta de cada imagen de entrada tiene el siguiente formato, en el que en cada celda o región los objetos se codifican de la siguiente forma:

(confianza1, x1, y1, w1, h1, confianza2, x2, y2, w2, h2, c1, c2, c3, ..., cN)

donde "confianza" indica la probabilidad de la clase, "x" e "y" es el centroide con respecto al tamaño de la celda o región i (valor entre 0 y 1), y "w" y "h" el ancho y el alto respectivamente del recuadro (BB) i con respecto al tamaño de la imagen (valor entre 0 y 1). Los valores c1, c2, ..., cN se corresponden con la clase detectada, donde sólo uno de estos valores estará a 1 y el resto a 0. Un valor de confianza igual a 0 indica que no se ha detectado ningún objeto en la región y se ignoran el resto de valores. Un objeto pertenece a una celda si su centroide se encuentra en ella, independientemente de si el objeto puede ocupar varias celdas dado su tamaño.

Cada establecimiento cuenta con un conjunto de datos de sus productos (imágenes) para un entrenamiento individualizado. Tras la inicialización (etapas 61, 62), la red se entrena con lotes de imágenes etiquetadas (etapa 63), que han podido seguir técnicas de aumentación de datos (etapa 68) si es necesario, para aumentar su tamaño. Los lotes de imágenes etiquetadas (etapa 63) se van tomando de un conjunto total de imágenes disponibles (referencia 69) de bandejas de menús del establecimiento. A modo de ejemplo, los lotes pueden ser lotes de 64 imágenes. Así, tomando las imágenes modelo en lotes (batches), la red completa se entrena (etapas 64, 65) un cierto número de iteraciones (por ejemplo, más de 100). En cada iteración, se realiza un procesamiento neuronal y un cálculo de errores (función de error) y precisión (etapa 64) y se actualizan los parámetros de la red neuronal (etapa 65). Por ejemplo, y de forma no limitativa, se pueden actualizar los parámetros de la red denominados *momentum*, que sirve para controlar cómo varía la velocidad a la que aprende; y *decay*, que sirve para producir una ralentización progresiva del aprendizaje a medida que avanza el entrenamiento. La tasa de aprendizaje es gradual, normalmente incrementándose lentamente durante las primeras iteraciones (épocas). Posteriormente y hasta el final del entrenamiento (respuesta Sí a la pregunta "¿Fin de entrenamiento"? de la etapa 67) la tasa de aprendizaje va normalmente disminuyendo. Si la respuesta a esta pregunta es NO, se continúa entrenando la red ["Época +1" (etapa 66)]. Tras la última iteración, la red neuronal se d a por entrenada (etapa 70).

La red neuronal convolucional profunda del sub-módulo de detección de objetos 512 no discrimina la configuración de las bandejas 1 por su tamaño ni por la disposición de los elementos que pueden aparecer en ella, tal y como se muestra en las figuras 3 y 4. Así, el tamaño y la forma de la bandeja 1 no tiene que ser especificado previamente y puede ser diferente de una bandeja a otra. Como puede observarse, la red discrimina los instrumentos de mesa (tales como tenedor 15, cuchillo 16, cuchara grande 17 y cuchara pequeña 14) aunque en las bandejas de las figuras 3 y 4 no aparecen dispuestos en el mismo orden. El tamaño de estos instrumentos puede además ser diferente. La comida que aparece en los platos también puede ser diferente. Por ejemplo, la comida del primer plato es diferente en ambas figuras (sopa 12 en la figura 3 y ensalada 10 en la figura 4), pero la red es capaz de determinar a qué clase pertenecen y, por tanto, que ambas cosas son "primer plato". El tamaño del recipiente donde viene servida la comida es indiferente. Por ejemplo, el recipiente del primer plato es diferente en ambas figuras (un cuenco 13 en la figura 3 y un plato llano 9 en la figura 4). De igual forma, un recipiente del mismo tamaño no indica que la comida que contenga sea la misma. Por ejemplo, en la figura 4 los platos utilizados son del mismo tamaño (plato llano 9), aunque uno se identifique con una determinada clase y, por tanto, como primer plato (el que contiene ensalada 10) y el otro con otra determinada clase y, por tanto, como segundo plato (el que contiene el filete 11 con patatas fritas 19). Igualmente,

una misma comida puede ser considerada como primer plato o acompañamiento del segundo. Por ejemplo, la ensalada 10 que aparece en la figura 4 es el primer plato de esa bandeja, mientras que la ensalada 10 que aparece en la figura 3 es el segundo plato de esa bandeja. Esto se consigue con el entrenamiento de la red con imágenes etiquetadas por clases (productos). Aunque una clase puede estar formada por varios objetos (comidas, por ejemplo, pescado con ensalada), la red no aprende esta información (comidas individuales), sino que aprende y detecta clases (productos). Así, la red reconoce clases de las que se le han proporcionado ejemplos para aprender. Por ejemplo, se le han dado ejemplos de ensalada sola (primer plato), que es una clase, y por otro lado se le han dado ejemplos de filete de pollo con ensalada (segundo plato), que es otra clase. El tamaño de la comida no afecta. Por ejemplo, los bollos de pan 8 que aparecen en las figuras 3 y 4 son de diferente tamaño y forma. Además de los instrumentos de mesa, pueden aparecer otros elementos superfluos que no afectan al precio del menú. Por ejemplo, en la figura 4 aparece un vaso 18. Nótese que la red neuronal profunda sólo puede aprender aquellas clases que hayan sido provistas previamente en el entrenamiento, dado que aprende por imitación. es decir, que, si en el entrenamiento no se le pasan imágenes de una clase, esa clase no será reconocida cuando se observe ese producto en una bandeja. Este es el ejemplo de los cubiertos, vasos vacíos u otros elementos superfluos: al no estar entre las clases de entrenamiento, no se reconocerán estos objetos. En la figura 6 se observa cómo cada bounding box delimita un producto que pertenece a una clase reconocida.

A continuación, se describe en detalle una implementación de la red neuronal convolucional profunda del sub-modulo 512, basada en una red YOLO. Esta implementación no debe considerarse limitativa, ya que el sub-módulo 512 puede implementar otras redes neuronales convolucionales profundas como las enumeradas anteriormente.

Como se ha explicado, el objetivo es reconocer los objetos de interés (clases) y, para ello, la red neuronal tiene que detectar y localizar en la imagen esos objetos de interés. La entrada a la red es la imagen captada por la cámara 4, preferentemente pre-procesada en el sub-módulo 511. La salida de la red es un vector por cada objeto (clase) detectado. Cada vector define el recuadro que engloba el objeto correspondiente (bounding box) y un valor de probabilidad o confianza (entre 0 y 1) de pertenencia a una clase (en este caso, a un tipo de producto). Así, cada vector tiene V elementos, de los que V-1 definen el recuadro y un elemento define la confianza. Por ejemplo, cada vector puede contener 5 números reales, 4 de los cuales definen el recuadro (por ejemplo, valor (x, y) de la esquina superior izquierda y ancho y alto del recuadro), y el quinto número indica dicho valor de probabilidad o confianza de pertenencia a una clase.

Debido a que las imágenes capturadas por la cámara 4 están siempre a la misma distancia de la cámara, debido a la disposición fija de cámara y bandeja al hacer la foto, no es necesario incorporar ningún módulo para generar mapas de características piramidales (que definan diferentes tamaños de entrada), con objeto de considerar que un plato o alimento se puede encontrar en la imagen en diferentes tamaños.

Inicialmente, se realiza un redimensionado de la imagen de entrada, por ejemplo, a 416 x 416 pixeles (etapa 511c de la figura 2B). A continuación, la imagen reescalada accede a la red neuronal (bloque 512), donde se divide la imagen en una rejilla de S x S regiones o celdas, donde cada celda puede predecir hasta B objetos de C clases diferentes. A modo de ejemplo, se ha definido S = 7 y B = 2. Así, la red puede predecir, como máximo 7 x 7 x 2 = 98 objetos (clases) en cada imagen, número más que suficiente para identificar las consumiciones de un menú, de la invención. El valor de C es el número de productos diferentes que puede detectar el sistema. Este valor se define previamente, en el entrenamiento de la red, según el conjunto de imágenes de entrenamiento disponible. Cada predicción es un vector de V valores (por ejemplo, 5 valores: x, y, w, h, confianza) al que añadimos un vector de C valores correspondientes a las clases a detectar, donde sólo está activo (valor 1) la clase del objeto detectado. El sistema sólo detecta

aquellos objetos o clases que se hayan indicado en el entrenamiento, descartando otro tipo de objetos que aparezcan en la imagen. Por tanto, la capa de salida tendrá un tamaño de S x S x (B*V + C). En el ejemplo citado, este tamaño es de 7 x 7 x (2*5 + C).

La red neuronal convolucional profunda implementada en el sub-módulo de visión artificial de los medios de procesado del sistema, y en concreto la red YOLO de esta implementación concreta, implementa una arquitectura de capas convolucionales y capas de max-pooling, añadiendo dos capas totalmente conectadas en la parte final. La siguiente tabla muestra en detalle un ejemplo no limitativo de implementación de cada una de estas capas, aplicable a la red YOLO descrita. En la tabla, con respecto a los nombres (primera columna), "Conv # " indica que la capa # es una capa convolucional; "Max Pool #" indica que la capa # es una capa de max-pooling y "FC #" indica que la capa # es una capa totalmente conectada. Además, con respecto a los filtros (segunda columna), los primeros dos valores se corresponden con su tamaño, el tercero indica el número de kernels o núcleos y el siguiente (stride) es el desplazamiento a usar en la aplicación del filtro. La última capa varía en función del número de clases a detectar. Todas las capas utilizan como función de activación una Leaky RELU, excepto la última capa, que tiene una función de activación lineal.

Nombre	Filtro	Dimensión de salida
Conv 1	7 × 7 x 64, stride=2	224 x 224 x 64
Max Pool 1	2 x 2, stride=2	112 x 112 x 64
Conv 2	3×3×192	112x 112x192
Max Pool 2	2 x 2, stride=2	56 X 56 x 192
Conv 3	1× 1 × 128	56 x 56 x 128
Conv 4	3x3 x 256	56 x 56 x 256
Conv 5	1×1×256	56 x 56 x 256
Conv 6	1x1 x 512	56 x 56 x 512
Max Pool 3	2x2, stride=2	28x 28 x 512
Conv 7	1 x 1x 256	28x28 x 256
Conv 8	3×3 × 512	28 x 28 x 512
Conv 9	1×1×256	28 x 28 x 256
Conv 10	3×3× 512	28 x 28 x 512
Conv 11	1x1x 256	28 x 28 x 256
Conv 12	3 x 3 x 512	28 x 28 x 512
Conv 13	1×1×256	28x28 x 256
Conv 14	3x3 × 512	28 x 28 x 512
Conv 15	1 x 1 x 512	28 x 28 x 512
Conv 16	3 x 3 x 1024	28 x 28 x 1024
Max Pool 4	2x2, stride=2	14 x 14x x1024
Conv 17	1 x 1 x 512	14 x 14 x 512
Conv 18	3 x 3 x 1024	14 x 14x x1024
Conv 19	1 x 1 x 512	14 x 14 x 512
Conv 20	3 x 3 x 1024	14 x 14x x1024
Conv 21	3 x 3 x 1024	14 x 14x x1024
Conv 22	3 x 3 x 1024, stride=2	7×7 x1024
Conv 23	3 x 3 x 1024	7×7 x1024
Conv 24	3 x 3 x 1024	7×7 x1024
FC 1	-	4096
FC 2	-	7×7 x (10+C)

A modo de ejemplo no limitativo, la red YOLO detallada con anterioridad ha sido entrenada de la siguiente forma: En primer lugar (etapa 61), se inicializan o pre-entrenan las 20 primeras capas convolucionales de la red, usando el conjunto de datos ImageNet-1000 que tienen un tamaño de imagen de 224x224 pixeles. Este tamaño se escala convenientemente para que se adecúe a la entrada de la red. A continuación, se vuelve a entrenar la red, pero en este caso con el conjunto de imágenes etiquetadas de bandejas de menú con los productos que pueden consumirse en el establecimiento.

Para cada imagen de entrada, su etiqueta tiene la siguiente forma:

(confianza1, x1, y1, w1, h1, confianza2, x2, y2, w2, h2, c1, c2, c3, ..., cN)

donde "confianza" indica la probabilidad de la clase, "x" e "y" es el centroide con respecto al tamaño de la celda o región i (valor entre 0 y 1), y "w" y "h" el ancho y el alto respectivamente del recuadro (BB) i con respecto al tamaño de la imagen (valor entre 0 y 1). Los valores c1, c2, ..., cN se corresponden con la clase detectada, donde sólo uno de estos valores estará a 1 y el resto a 0. Un valor de confianza igual a 0 indica que no se ha detectado ningún objeto en la región y se ignoran el resto de valores. Un objeto pertenece a una celda si su centroide se encuentra en ella, independientemente de si el objeto puede ocupar varias celdas dado su tamaño.

Tras la inicialización, la red YOLO completa se entrena con lotes de 64 imágenes etiquetadas (etapa 63), que han podido seguir técnicas de aumentación de datos (etapa 68) si es necesario, para aumentar su tamaño. Los lotes de 64 imágenes etiquetadas se van tomando de un conjunto total de imágenes disponibles 69 de bandejas de menús del establecimiento, ya que cada establecimiento cuenta con un conjunto de datos de sus productos (imágenes) para un entrenamiento individualizado. Así, tomando las imágenes modelo en lotes (batches de 64 imágenes), la red completa se entrena (etapas 64, 65) unas 135 épocas (iteraciones). En cada iteración, se realiza un procesamiento neuronal y un cálculo de errores (función de error) y precisión (etapa 64) y se actualizan los parámetros de la red neuronal (etapa 65). La tasa de aprendizaje es gradual, incrementándose lentamente durante las primeras 75 iteraciones (épocas) de 0,001 a 0,01. Posteriormente y hasta el final del entrenamiento (respuesta SÍ a la etapa 67) la tasa de aprendizaje va disminuyendo hasta 0.001. Tras la última iteración, la red YOLO se da por entrenada (etapa 70).

De esta forma, una vez entrenada la red, considerando por ejemplo como entrada la imagen de la bandeja 1 mostrada en la figura 6 , la red neuronal implementada en el sub-modulo 512 devuelve como salida la posición y tamaño de los recuadros 71-75, realiza la siguiente asociación: recuadro 71 es un refresco (una clase determinada); recuadro 72 es una tarta (otra clase determinada); recuadro 73 es pan (otra clase); recuadro 74 es sopa (otra clase); y recuadro 75 es filete con ensalada (otra clase); descarta los cubiertos como elementos relevantes, e indica la probabilidad de clase asignada que corresponde con cada recuadro 71-75.

A continuación, tras una etapa opcional de post-procesamiento (sub-módulo 513), el módulo 52 calcula los precios de los productos de los recuadros 71-75 a partir de la asociación de consumiciones y precios previamente establecida.

Frente a técnicas convencionales de detección de imágenes basadas en machine learning, el método y sistema de la invención, basado en técnicas de aprendizaje profundo (redes neuronales convolucionales profundas), obtienen un error en la identificación de productos considerablemente más bajo y, por tanto, un mayor rendimiento.

En suma, el método y sistema propuestos permiten detectar automáticamente, mediante una cámara preferentemente cenital y técnicas de visión artificial, los tipos de consumiciones que el

cliente lleva en su bandeja de autoservicio, una vez el cliente deposita la bandeja en un espacio designado al efecto. Y a partir de la identificación de esos tipos de consumiciones, calcular el precio de las consumiciones de la bandeja para que el cliente pueda abonarlas en un terminal de punto de venta. Se facilita así, a los clientes, el abono de sus consumiciones, sin necesidad de un cobrador humano.

En este texto, el término "comprende" y sus derivaciones (tal como "comprendiendo", etc.) no deben entenderse en un sentido excluyente, es decir, estos términos no deben ser interpretados como que excluyen la posibilidad de que lo que se describe y se define pueda incluir elementos, etapas adicionales, etc.

En el contexto de la presente invención, el término "aproximadamente" y términos de su familia (como "aproximado", etc.) deben interpretarse como indicando valores muy cercanos a aquellos que acompañan a dicho término. Es decir, una desviación dentro de límites razonables con respecto a un valor exacto, deberían aceptarse, porque un experto en la materia entenderá que tal desviación con respecto a los valores indicados puede ser inevitable debido a imprecisiones de medida, etc. Lo mismo aplica a los términos "unos", "alrededor de" y "sustancialmente".

La invención no se limita obviamente a la(s) realización(es) específica(s) descrita(s), sino que abarca también cualquier variación que pueda ser considerada por cualquier experto en la materia (por ejemplo, con relación a la elección de materiales, dimensiones, componentes, configuración, etc.), dentro del alcance general de la invención como se define en las reivindicaciones.

REIVINDICACIONES

1. Un sistema para la identificación y cobro automáticos de consumiciones, que comprende una cámara (4) para la captura de una imagen de una bandeja (1) que contiene al menos un producto, unos medios para procesar información (5) que a su vez comprenden unos medios de detección de objetos (512) basados en una red neuronal convolucional profunda, y un dispositivo de cobro (6) conectado a los medios para procesar información (5).

2. Un sistema según la reivindicación anterior donde los medios para procesar información (5) comprenden medios (52a) de cálculo de precios conectados a una base de datos (525) que comprende una relación de clases de productos y los precios asociados a cada clase de productos.

3. Un sistema según cualquiera de las reivindicaciones anteriores donde los medios para procesar información (5) comprenden medios de pre-procesamiento (511).

4. Un sistema según cualquiera de las reivindicaciones anteriores donde los medios para procesar información (5) comprenden medios de post-procesamiento (513).

5. Un producto de programa informático que comprende instrucciones / código de programa informático para ser ejecutado en un sistema para la identificación y cobro automáticos de consumiciones según cualquiera de las reivindicaciones anteriores 1 a 4.

6. Una memoria / soporte legible por ordenador que almacena instrucciones / código de programa de un producto de programa informático según la reivindicación anterior 5.

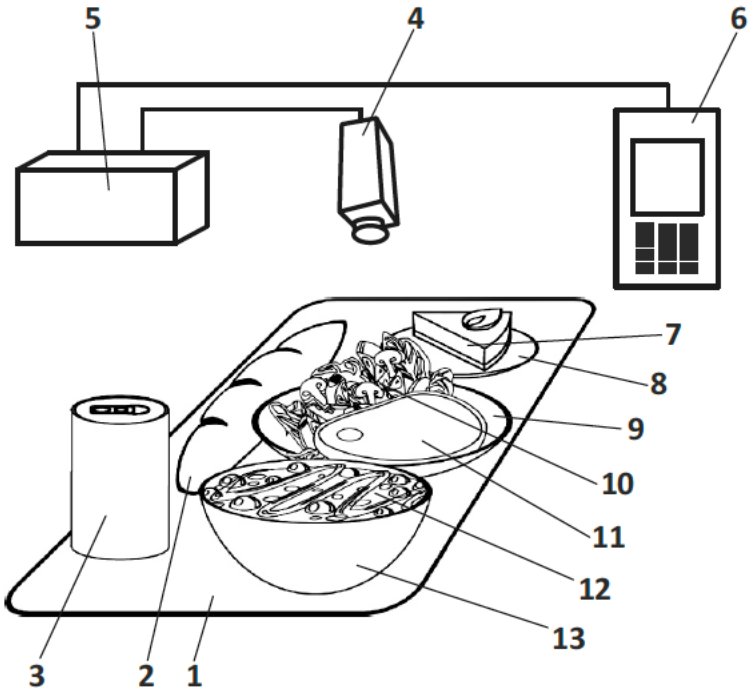

FIG. 1

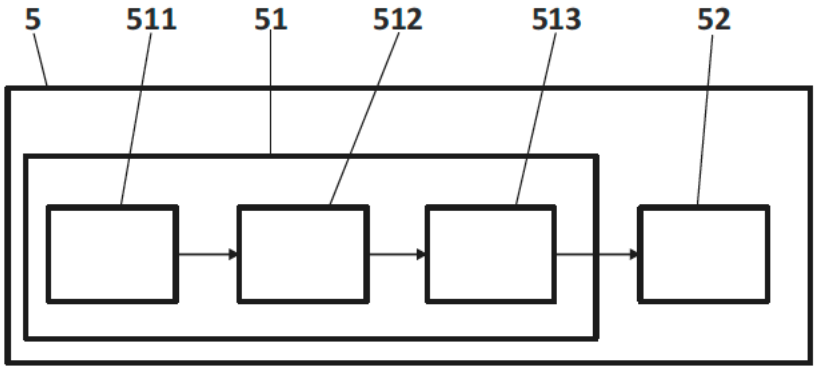

FIG. 2A

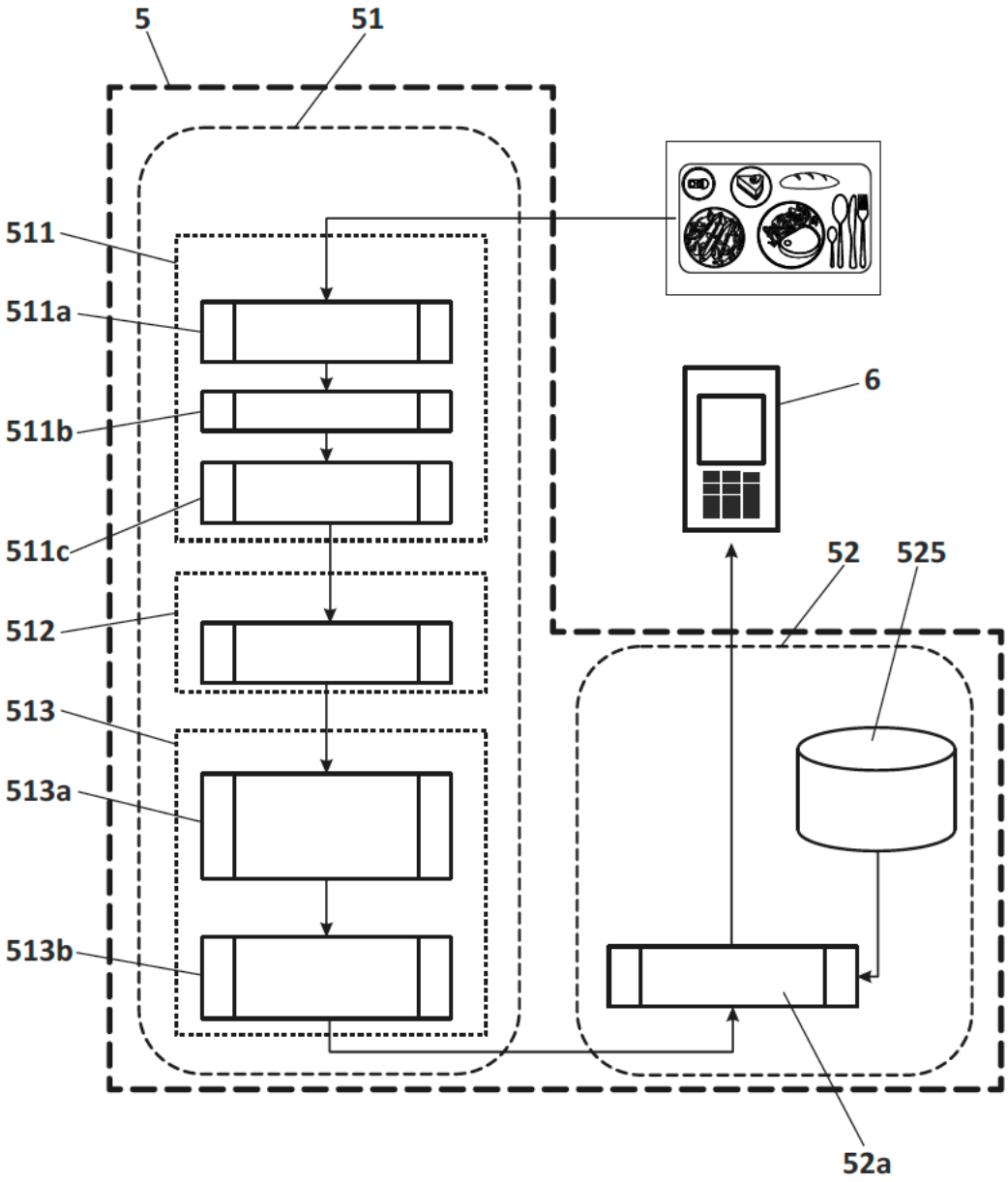

FIG. 2B

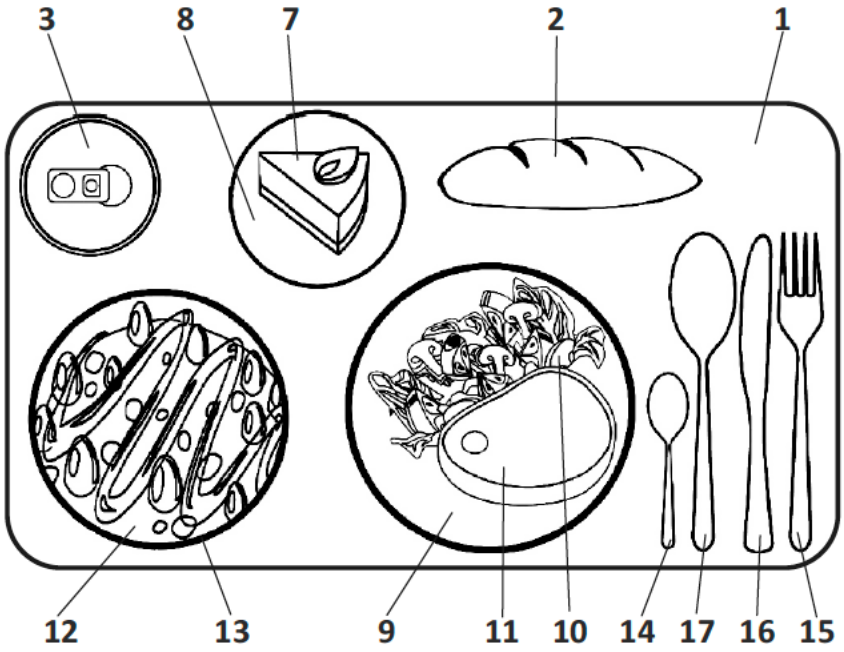

FIG. 3

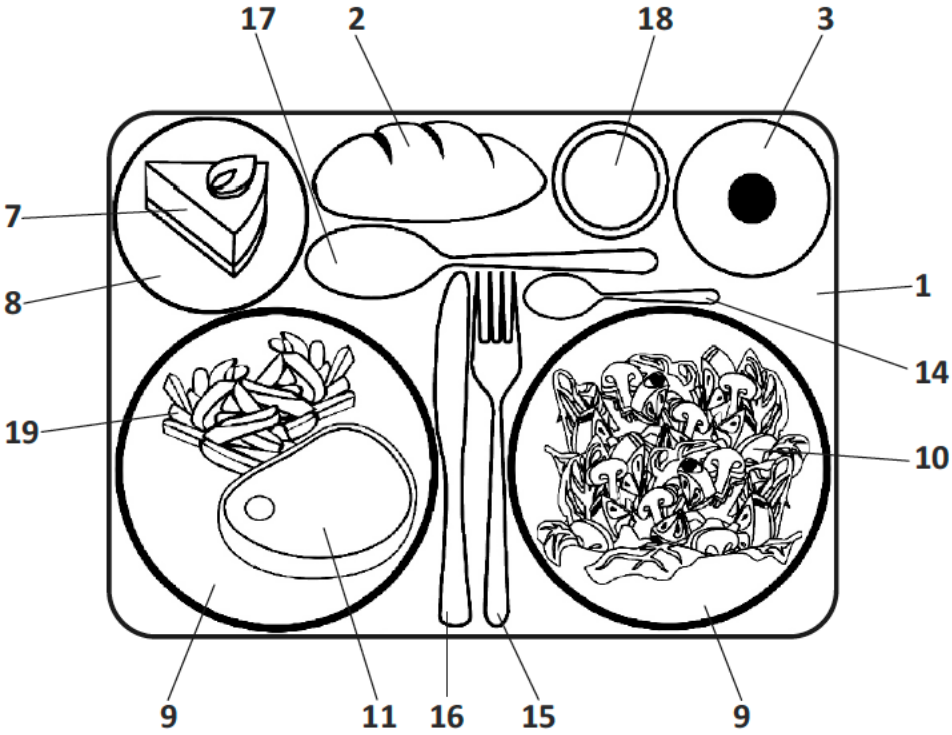

FIG. 4

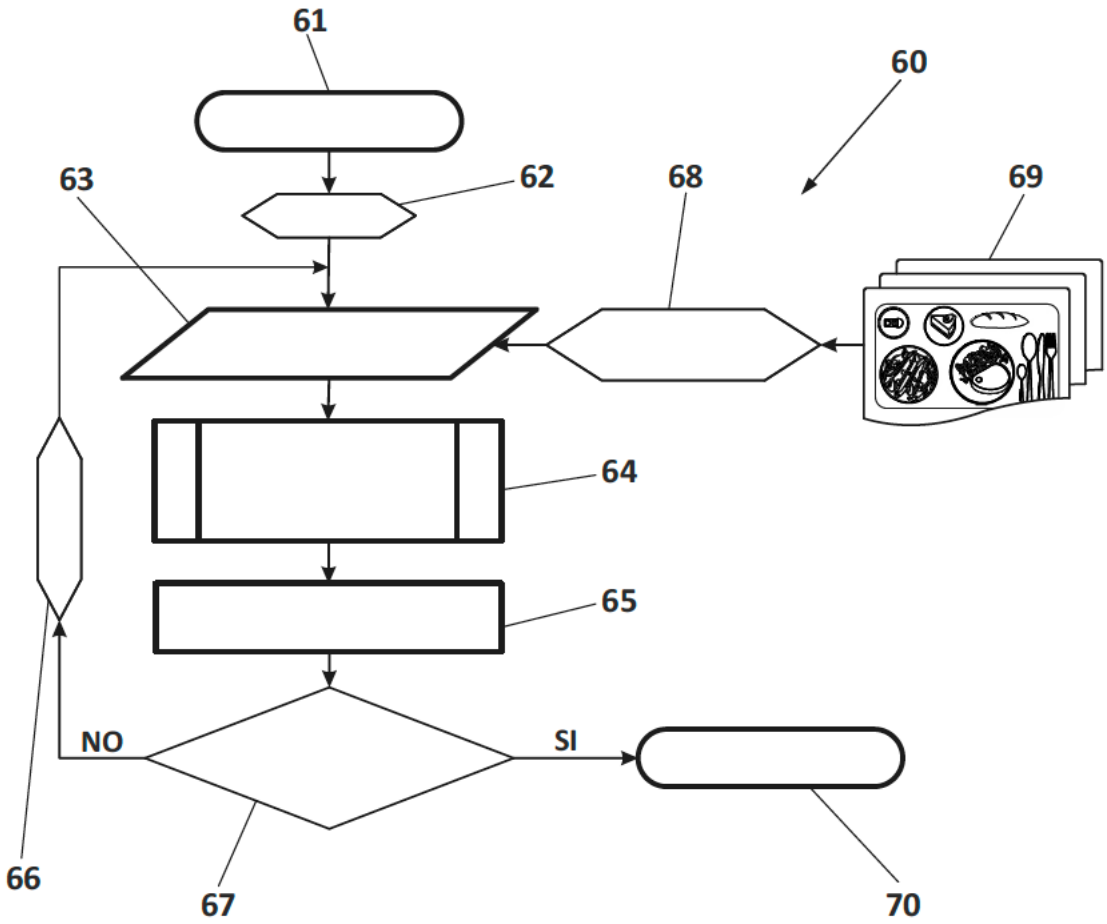

FIG. 5

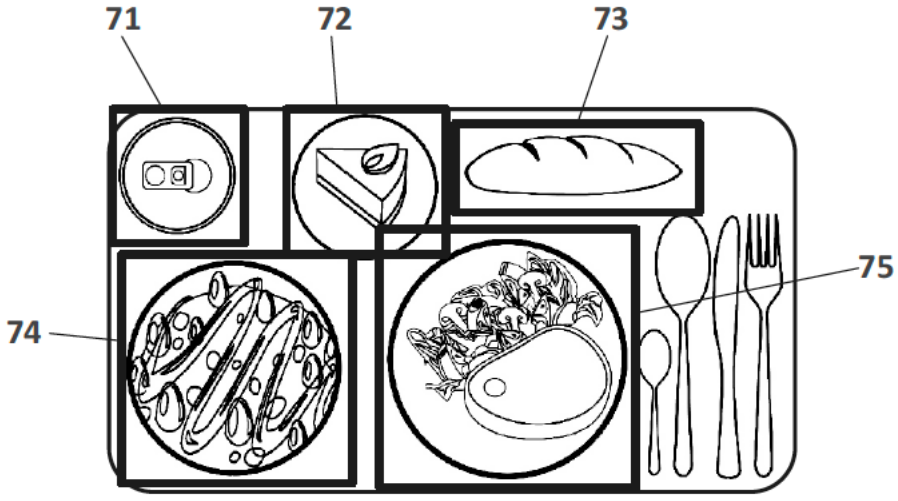

FIG. 6