ScienceDirect
Disponible en www.sciencedirect.com
Revista Iberoamericana de Automática e Informática industrial 11 (2014) 348–357
www.elsevier.es/RIAI

Conteo de personas con un sensor RGBD comercial

M. Castrillón-Santana*, J. Lorenzo-Navarro, D. Hernández-Sosa

SIANI, Universidad de Las Palmas de Gran Canaria (ULPGC), Spain.

Resumen

En este trabajo se demuestra que la información de profundidad proporcionada por una cámara RGBD comercial de bajo coste, es una fuente fiable de datos para realizar de forma robusta el conteo automático de personas. La adopción de una configuración de vista cenital reduce la complejidad del problema, al mismo tiempo que permite preservar la privacidad de las personas monitorizadas. Para llevar a cabo el estudio experimental se han considerado dos técnicas propias del campo de análisis de imágenes 2D trasladadas al contexto de imágenes de profundidad. Las pruebas evaluaron su rendimiento con vídeos reales sin restricciones de iluminación, incluyendo episodios de iluminación cambiante o muy baja. En este conjunto experimental se realizó la detección, seguimiento y análisis de patrones de comportamiento de las personas que cruzaban el campo de visión. Los resultados obtenidos alcanzan una tasa de acierto próxima al 95 %, superando los obtenidos con técnicas actuales basadas exclusivamente en información visual. Estos resultados sugieren la utilidad del uso de información de profundidad en esta tarea particular.

Palabras Clave:
Conteo de personas, cámaras de profundidad, detección de eventos, detección de objetos

1. Introducción

El conteo de personas es una característica deseable para un sistema automático con aplicaciones potenciales en múltiples escenarios. Por ilustrar con algunos ejemplos, es necesario conocer el número de pasajeros que entran y salen de un medio de transporte público para llevar a cabo su control y gestión. En pubs y discotecas los protocolos de evacuación están diseñados de acuerdo con la capacidad del local, no debiendo sobrepasarse para evitar situaciones peligrosas. El control de presencia es también esencial para la implantación de políticas de ahorro energético. En otro contexto, la obtención automática de datos de audiencia, así como la medida de tiempos de atención, es también una capacidad de interés para empresarios y anunciantes. Debido a las diversas aplicaciones, y su creciente interés, la literatura reciente describe distintas soluciones para conocer lo más exactamente posible, el número de personas que accede y sale de un espacio delimitado. Las dos principales tecnologías que los investigadores han utilizado hasta la fecha para resolver este problema son: 1) imágenes de los escenarios analizadas con técnicas de Visión por Computador, y 2) el análisis de las firmas que ofrecen diferentes haces de luz en presencia de personas.

1.1. Estado actual

Como se indicaba en el apartado anterior, previo a la aparición de las cámaras RGBD, los sistemas basados en información visual y los basados en haces de luz han sido los más utilizados para el conteo de personas.

Los sistemas basados en emisión de haces de luz tienen la ventaja de preservar la privacidad, al no capturar imágenes de las personas. Sin embargo, el rango de detección efectiva de estos sistemas es limitado. Algunos ejemplos relevantes de estos métodos son Nakamura et al. (2006); Fod et al. (2002); Katabira et al. (2004); Mathews y Poigné (2009). Nakamura et al. (2006) proponen un método para el seguimiento de peatones en áreas abiertas utilizando un conjunto de láseres. Tras la eliminación del fondo, las diferentes lecturas se integran evitando problemas de oclusión que podrían afectar a un sistema basado en un único emisor láser. Un enfoque similar fue presentado por Fod et al. (2002) donde los modelos de fondo y primer plano se obtienen a partir de las lecturas de un láser que proporciona la distancia. Katabira et al. (2004) presentaron un sistema basado en un sensor colocado en el techo de un pasillo. Las formas humanas se extraen de la lectura del sensor de distancia mediante la transformación de los datos al plano *xz*, asociando la detección de

*Autor en correspondencia
Correos electrónicos: mcastrillon@iusiani.ulpgc.es (M. Castrillón-Santana), jlorenzo@iusiani.ulpgc.ess (J. Lorenzo-Navarro), dhernandez@iusiani.ulpgc.es (D. Hernández-Sosa)

http://dx.doi.org/10.1016/j.riai.2014.05.006

personas a la localización de objetos prominentes. Mathews y Poigné (2009) introdujeron un sistema basado en un conjunto de emisores de haces en infrarrojo. La detección de personas se realiza con una red neuronal entrenada previamente con patrones de movimientos obtenidos en un simulador.

A diferencia de los sistemas basados en emisión de haces de luz, los sistemas basados en visión pueden aplicarse con menores restricciones también en espacios amplios y abiertos, como estaciones o zonas comerciales. En el caso concreto de los espacios abiertos, las soluciones hacen uso de técnicas de análisis de multitudes, como en los trabajos de Zhan et al. (2008); Moore et al. (2011), para estimar el número de individuos y su comportamiento. Otra solución, como la presentada por Hou y Pang (2011), utiliza la cantidad de píxeles del primer plano para procesar el número de personas en escenas muy concurridas. Proponen tres métodos de estimación basados en un aprendizaje previo considerando la relación entre los píxeles de primer plano y el número de personas presentes. Como inconveniente, las condiciones de luz tienen una importante influencia en el rendimiento de estos sistemas, además de sólo proporcionar una estimación que puede no ser suficiente en algunas aplicaciones.

Dependiendo de la ubicación de la cámara, se distingue entre sistemas cenitales y no cenitales. Indicar que los sistemas basados en visión con cámaras no cenitales no se pueden utilizar en aplicaciones donde la privacidad sea un requerimiento a cumplir. Al contrario, la configuración cenital es capaz, en principio, de preservar la privacidad porque los rostros de los individuos no son capturados.

Entre las propuestas no cenitales, Lee et al. (2008) hacen uso de un láser como fuente de luz estructurada. De esta manera, la estimación se realiza por medio de la integración de imágenes consecutivas. Cuando las personas cruzan la zona monitorizada, se obtiene un patrón que permite tanto contar el número de apariciones de individuos, como la dirección de movimiento. Un trabajo más reciente de Zeng y Ma (2010), presenta una extensión del detector del patrón cabeza-hombros basado en HOG-LBP, introduciendo PCA y el procesamiento a varios niveles para lograr un 95 % de exactitud en la detección y el seguimiento. Zhao et al. (2009) utilizan detección de rostros y seguimiento para contar el número de personas. Las caras se detectan haciendo uso de histogramas de color, y el seguimiento se basa en un filtro de Kalman. Tras generar una trayectoria, un clasificador de k vecinos más cercanos determina las trayectorias reales y las cuenta. Un inconveniente de este sistema es que requiere varias cámaras para detectar a las personas en ambas direcciones.

Considerando por otro lado las propuestas con cámaras cenitales, Chan et al. (2008); Chan y Vasconcelos (2012) proponen un método basado en el análisis de una multitud, que hace uso de una mezcla de texturas dinámicas para segmentar la multitud en diferentes direcciones, es decir, evitando características individuales. Tras una posterior corrección de perspectiva, algunas de las características se calculan en cada segmento detectado como primer plano para finalmente obtener el número de personas empleando un proceso de Gauss. Kim et al. (2002) proponen un método basado en la sustracción de fondo y seguimiento con una configuración cenital. La sustracción de fondo se realiza de forma adaptativa para gestionar posibles cambios de las condiciones de luz. El proceso de conteo se lleva a cabo observando las entidades que cruzan el área definida. Septian et al. (2006) presentan un método similar, pero para imágenes en color, utilizando heurísticas basadas en áreas para detectar cuando hay más de una persona en las zonas consideradas de primer plano. El método descrito por Albiol et al. (2001) determina el número de personas que entran y salen de un vagón de tren. La cámara cenital colocada en el marco de la puerta, hace uso de tres líneas para obtener una imagen de la integración de sus muestras en el tiempo, produciendo diferentes patrones de acuerdo con la densidad de personas. Barandiaran et al. (2008) proponen un método muy similar con líneas de conteo. Bozzoli et al. (2007) introducen un método de conteo de personas en ambientes concurridos como puertas de autobús o tren. La propuesta se basa en el cálculo de un modelo de fondo promedio sobre imágenes de contornos con el fin de evitar la influencia de los cambios bruscos de las condiciones de iluminación. Los bordes del primer plano se seleccionan calculando el flujo óptico. Por último, cada vector de movimiento se asigna a un segmento y a continuación, se agrupan y se combinan para producir una estimación de la gente que cruza en cada dirección. Otro sistema con cámaras cenitales es el propuesto por Antic et al. (2010), donde las personas se detectan por medio de la segmentación de la imagen con un algoritmo de agrupación k-media y el seguimiento de los grupos resultantes en la secuencia. Otro enfoque indirecto para contar el número de las personas en ambientes de multitudes es el propuesto por Albiol y Silla (2010). En su propuesta, se estima el número de pasajeros que suben y bajan de un tren estableciendo una estimación a partir del número de esquinas detectadas, asumiendo que sin pasajeros hay un bajo número de esquinas, y el número promedio de esquinas por persona.

La emisión de haces de luz proporciona información de profundidad, que también puede ser suministrada por los sistemas estéreo, con el objetivo de reducir los problemas de iluminación inherentes al canal visual García et al. (2012). Efectivamente, como argumenta Harville (2004), la profundidad es una fuente de información importante para la segmentación, siendo prácticamente insensible a los cambios de iluminación. Además, una configuración cenital aumenta sus beneficios como se revela en Cohen et al. (2000) y Harville (2004). Así, Beymer (2000) emplea un par estéreo calibrado para obtener la trayectoria de una persona en un mapa de ocupación calculado a partir de la transformación de perspectiva de la imagen de disparidad. Modelos de mezcla de gaussianas combinados con un filtro de Kalman proporcionan las trayectorias que se clasifican en cuatro categorías. van Oosterhout et al. (2011) presentan un enfoque similar con el uso de una máscara circular para detectar las cabezas en la zonas segmentadas tras una detección mediante la sustracción de fondo. El proceso de detección de la cabeza permite a los autores distinguir los segmentos con más de una cabeza mediante la proyección de los píxeles de la cabeza en el plano de tierra. Una sola cámara calibrada proporciona la señal visual empleada por Velipasalar et al. (2006), estando, por lo tanto, expuestos a los artefactos de iluminación. Yu et al. (2007) afirman que el uso de la información de profundidad calculada

a partir de un par estéreo mejora el rendimiento en comparación con un sistema monocular. El sistema presentado en Qiuyu et al. (2010) demuestra los beneficios de la vista cenital para el conteo de personas. Los autores se centran en el tiempo real, integrando un DSP, sin embargo, la reducción de la resolución de la imagen simplificaría ese aspecto. Yahiaoui et al. (2010) utilizan una cámara estéreo cenital para controlar el número de pasajeros que entran y salen de un autobús, reduciendo por lo tanto el escenario de aplicación y simplificando el problema.

Más recientemente, la aparición de cámaras RGBD asequibles hace posible su aplicación no sólo para el conteo automático de personas como proponen Hernández et al. (2011), sino también en tareas de re-identificación Albiol et al. (2012); Barbosa et al. (2012); Oliver et al. (2012); Satta et al. (2013); Lorenzo-Navarro et al. (2013). En particular en nuestros trabajos previos, se ha hecho uso de un emisor láser para el conteo de personas Hernández-Sosa et al. (2011), obtenido resultados preliminares de conteo y re-identificación con un modelo simple del fondo sobre información de profundidad en Hernández et al. (2011), o hecho uso de características de biometría blanda como la altura, la constitución del cuerpo y volumen en un escenario de vista cenital en Lorenzo-Navarro et al. (2013) para re-identificación. En estos trabajos, la adición de la información de profundidad ayuda a resolver ambigüedades, primero durante la detección, y posteriormente al modelar las diferentes identidades. El trabajo aquí descrito, presenta un estudio detallado de técnicas de sustracción de fondo basadas en información de profundidad en condiciones de iluminación exigentes, para el problema específico de conteo de personas.

Otros trabajos han propuesto la fusión de información de pronfundidad y visual para el conteo y seguimiento. Cui et al. (2007, 2008) describen un método que fusiona datos de un láser y un método de seguimiento visual. El seguimiento con láser se basa en la integración de varias lecturas para detectar pares de piernas y posteriormente realizar el seguimiento con un filtro de Kalman para estimar la posición, velocidad y aceleración de dichas piernas. Una cámara calibrada permite realizar seguimiento visual con información de color que alimenta un sistema de seguimiento *mean-shift*. Por último, los resultados de ambos procesos de seguimiento se fusionan con un enfoque bayesiano. Bellotto y Hu (2009) siguen un enfoque similar, teniendo en cuenta el patrón característico de las piernas de una persona, combinado con un detector visual de rostros. Otros autores han propuesto enfoques multisensoriales, véanse por ejemplo Blanco et al. (2003); Scheutz et al. (2004).

2. Escenario y propuesta

El objetivo del sistema descrito en este trabajo consiste en detectar y contar las personas que cruzan una puerta, como por ejemplo en escenarios de transporte público, tiendas, locales nocturnos, etc., sin restringir las condiciones de iluminación. De acuerdo con la literatura anterior, el uso de haces de luz proporciona precisión a un coste mayor, mientras que los enfoques basados en visión, incluso siendo más rentables, no son adecuados para escenarios oscuros o de iluminación cambiante, ya que requieren técnicas más elaboradas de sustracción de fondo para resolver los problemas intrínsecos introducidos por los cambios de las condiciones iluminación.

Llegando a la conclusión de que, si bien la información proporcionada por un sensor visual es una valiosa fuente para resolver diferentes problemas de visión por ordenador, también incorpora un cierto grado de ambigüedad que puede entorpecer el proceso que nos ocupa. En este sentido, para diseñar un sistema lo suficientemente flexible para adaptarse a cualquier situación, y observando las dificultades inherentes a la iluminación no controlada, se ha realizado un análisis comparativo de las posibilidades que ofrece un sistema exclusivamente basado en información de profundidad, frente a los basados en información visual. Un objetivo posterior trataría de obtener lo mejor de ambos enfoques, basando el procesamiento en todos los canales de información alineados proporcionados por los sensores RGBD de bajo coste actuales Shotton et al. (2011).

La información de profundidad se relaciona con la información de la distancia proporcionada por un haz de luz, y sus beneficios para el problema han sido ya sugeridos por la literatura, como por ejemplo Fanelli et al. (2011a). Sin embargo, la profundidad no se ha utilizado comúnmente para la tarea del conteo automático, probablemente por el coste de los sensores disponibles previamente, con la excepción de la implementación de los sistemas de estéreo descritos en la sección anterior, cuyo rendimiento se puede ver negativamente afectado en condiciones de iluminación no controladas.

Las cámaras RGBD de consumo actuales son capaces de proporcionar información de profundidad añadida al color. Esta información adicional, obtenida de forma económica y compacta, ha comenzado a ser utilizada en un amplio número de escenarios de interacción hombre-máquina. La información proporcionada por sensores tales como Microsoft Kinect se utiliza para la detección de personas Xia et al. (2011); Albiol et al. (2012), inferir la pose de la cabeza Fanelli et al. (2011b), o del cuerpo Shotton et al. (2011), o d escribir la actividad que realiza Marcos et al. (2013), pero rara vez en configuraciones cenitales como la descrita por Hernández et al. (2011). Como se ha mencionado anteriormente, este tipo de sensor se puede aplicar en entornos donde otros enfoques disminuyen su rendimiento debido a las malas condiciones de iluminaciones como es el caso de los pares estéreo.

El enfoque propuesto no se basa en la información visual, como la requerida por un par estéreo, para obtener los datos de profundidad, por lo que puede emplearse en entornos con nivel de iluminación variable e incluso muy baja como cines y pubs. Por lo tanto, con la solución propuesta la gama de aplicaciones va más allá de los que se pueden abordar mediante visión estéreo. Además, se adopta la configuración cenital argumentando que preserva la privacidad, y facilita, en particular en datos de profundidad, la detección y el seguimiento de objetos que cruzan escenarios afectados por condiciones de iluminación cambiantes. Para ello se ha montado una cámara Kinect fija, proporcionando una vista cenital que cubre nuestro escenario, como se ilustra en la Figura 1, capturando imágenes como las que se presentan en la Figura 2. Las imágenes tienen una resolución de 640x480 píxeles con 24 bits por píxel para la información visual (RGB), y 8 para la de profundidad. En la

imagen de fondo, los objetos más cercanos se representan con tonos más oscuros, mientras que los píxeles blancos representan los puntos cuya información de profundidad no ha podido ser obtenida por la cámara. Con el fin de extraer la mayor cantidad de información posible de cada imagen, la sustracción de fondo se aplica para detectar objetos salientes. Para dichos objetos detectados se realiza un seguimiento en el tiempo, siendo asignados a las trayectorias que finalmente son etiquetadas. La Figura 3 describe de forma esquemática el procesamiento aplicado a cada fotograma capturado.

La geometría del escenario definido se describe en la Figura 1, analizando a continuación el contexto de aplicación de la solución propuesta. Haciendo uso de las especificaciones de campo de visión vertical de 43° proporcionado por Microsoft para la cámara Kinect, la longitud de r se puede calcular según la siguiente expresión,

$$r = 2\tan\left(\frac{43^\circ}{2}\right)(T-h) \quad (1)$$

donde T es la distancia de la cámara al suelo, y h la altura de la persona. Para una persona de altura promedio, $h = 1,75m$ y la cámara colocada a $T = 3m$ del suelo, se obtiene $r = 0,98m$. Para una velocidad de paso estándar de $1,4m/s$, la cámara captura un promedio de 18 imágenes de la persona atravesando el campo de visión, asumiendo una frecuencia de captura en el rango de 15 – 25 imágenes por segundo, debe ser información suficiente para decidir la etiqueta a asignar a la acción realizada por el individuo.

De acuerdo con (1) la modificación de altura de la cámara sólo afecta el campo de visión. En el estudio empírico de Spinello y Arras (2011) sobre el alcance efectivo de la cámara Kinect, se demuestra que una distancia máxima de 10 metros es admisible a costa de una menor resolución en las mediciones de profundidad.

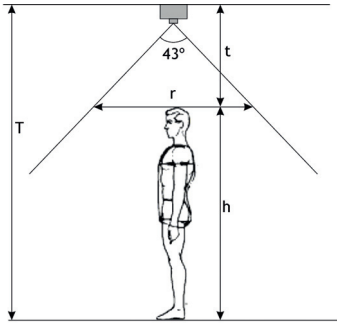


Figura 1: Geometría de la configuración experimental

Una vez descrita la visión general del sistema, continuaremos proporcionando más detalles de los diferentes módulos en las siguientes secciones. La sección 3 introduce brevemente las técnicas de modelado de fondo analizadas. La sección 4 resume la detección basada en profundidad, y el seguimiento. Finalmente las secciones 5 y 6 presentan respectivamente la configuración experimental y las conclusiones del trabajo.

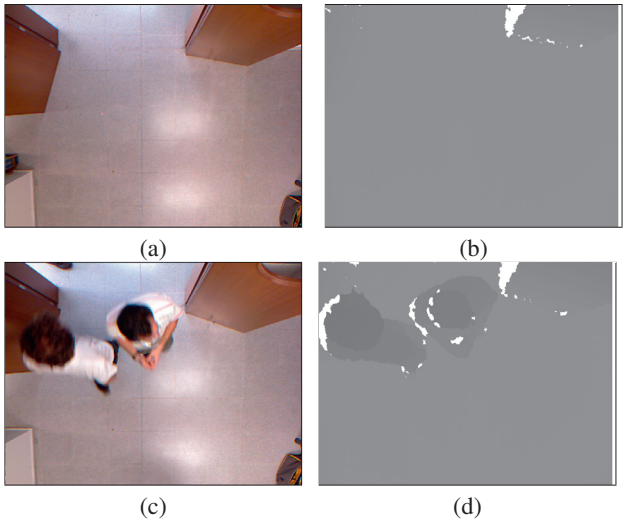


Figura 2: (a, c) Imágenes RGB (b, d) y de profundidad de un escenario con vista cenital en dos situaciones diferentes.

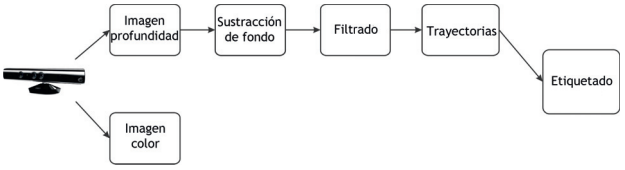


Figura 3: Ilustración del proceso de conteo

3. Técnicas de modelado de fondo

Esta sección describe las técnicas empleadas para modelar el fondo con el fin de realizar la detección de objetos en imágenes. El modelado de fondo tiene una notable tradición en la literatura de visión por computador Brutzer et al. (2011); Piccardi (2004), presentando diferentes desafíos en relación con los cambios repentinos de iluminación, sombras, oclusión, etc. En nuestro escenario, de forma oportunista, se adaptan técnicas habitualmente utlizadas con imágenes RGB, a la información de profundidad para reducir las incertidumbres y dificultades de este problema de cara a realizar la detección y segmentación de individuos.

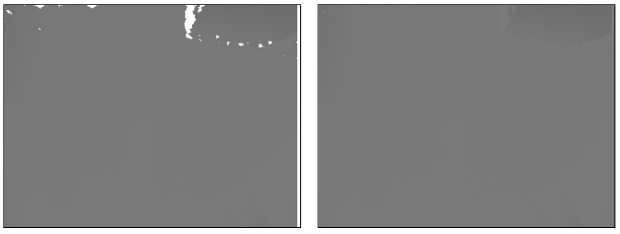
El objetivo es ofrecer al lector evidencias acerca de la simplificación que el uso de la información de profundidad ofrece en la tarea de sustracción de fondo en un escenario de configuración cenital. La selección de la configuración de la vista cenital se adopta en primer término para reducir la complejidad del problema a resolver, siendo además, como ya se ha indicado, adecuada para aplicaciones con requisitos de preservar la privacidad. Esta configuración de cámara, tanto con cámara de profundidad o par estéreo, ya ha demostrado ventajas en el contexto de procesamiento visual cuando se utilizan múltiples cámaras. Entre sus beneficios, se puede mencionar que esta configuración reduce la oclusión y los problemas de calibración de cámaras como señalan Gollan et al. (2011), en concreto para la detección de personas, como sugieren Englebienne y Krose.

(2010); Englebienne et al. (2009).

Evidentemente, una vez obtenido el modelo de fondo, las personas que transitan de forma normal por el escenario aparecerán especialmente salientes en relación con el fondo, dada la configuración de la cámara. Para ilustrarlo, como se observa en las figuras 2b y 2d, las cabezas destacan en la mayoría de los casos. Ante esta evidencia, una técnica de sustracción de fondo será capaz de segmentar de forma sencilla y robusta las zonas correspondientes a la presencia de personas. En esta sección se describen dos técnicas, aplicables tanto sobre la información visual o sobre la de profundidad: 1) modelo basado en umbral simple de Heikkila y Silven (1999), y 2) modelo basado en combinación de gaussianas de Zivkovic y der Heijden (2006). Reiterar que como se demuestra en los experimentos, el modelo de fondo construido con las imágenes de profundidad reduce la influencia de los cambios bruscos de iluminación, o los problemas de color exhibidos por las soluciones basadas exclusivamente en información visual. Si bien debemos mencionar que la comunidad ya comienza a describir técnicas que realizan el modelado fusionando ambas fuentes de datos como Camplani et al. (2014).

3.1. Modelado simple

Esta sección describe un modelado básico del fondo, que tras su cálculo en base a las imágenes iniciales, permite con un coste muy reducido, obtener información lo suficientemente robusta como para realizar la detección y segmentación de personas. Esta solución aprovecha la configuración estática de la cámara, y la escasa influencia de los cambios de iluminación en la imagen de profundidad.



(a) (b)

Figura 4: (a) Ejemplo imagen de profundidad de imagen y (b) modelo de fondo estimado para el escenario experimental.

Para definir el modelo de fondo de la escena, $fondo$, se calcula previamente la imagen promedio de profundidad y el umbral de saturación de profundidad. La imagen promedio de profundidad, $\overline{prof}$, se calcula como el promedio de las primeras k imágenes de profundidad (se asume que no habrá personas estáticas en todas ellas) como:

$$\overline{prof}(i,j) = \frac{\sum_{l=1}^{k} prof^{l}(i,j)}{k} \tag{2}$$

donde $prof^{l}(i,j)$ es el valor del píxel (i,j) de la l-ésima imagen de profundidad de la secuencia.

En la imagen promedio resultante, pueden aparecer diferentes puntos singulares, debido a las sombras, oclusiones, etc., representados en blanco en la figura 4a. Para evitar su influencia, se modifican haciendo uso de un umbral de saturación. La definición de este umbral tiene en cuenta la configuración cenital de la cámara, que contempla la escena desde una vista superior. Por este motivo se puede considerar que la mayor parte del escenario visible es el suelo de la escena, una superficie plana. Siguiendo esta idea, se calcula el valor de píxel promedio de la imagen promedio de profundidad como:

$$prof_{th} = \frac{\sum_{i=1}^{alto} \sum_{j=1}^{ancho} \overline{prof}(i,j)}{ancho \times alto} \tag{3}$$

Correspondiendo *ancho* y *alto* respectivamente al ancho y alto de la imagen en píxeles. Asumiendo que los píxeles más cercanos se presentan como más oscuros, para calcular el modelo de fondo, $fondo$, el umbral de saturación, $prof_{th}$, se utiliza para forzar el valor de cualquier píxel con valores de profundidad mayores, es decir, más claros

$$fondo(i,j) = \begin{cases} \overline{prof}(i,j) & \text{if } \overline{prof}(i,j) < prof_{th} \\ prof_{th} & \text{en otro caso} \end{cases} \tag{4}$$

La Figura 4b ilustra el modelo de fondo calculado para el escenario presentado en la figura 2, junto a una imagen de profundidad utilizada para su cálculo, Figura 4a, permitiendo al lector observar las diferencias.

Una vez que el modelo de fondo está disponible, se aplica una técnica de sustracción de fondo sencilla y poco costosa, siguiendo un enfoque similar al propuesto por Heikkila y Silven (1999). El primer plano se calcula umbralizando la imagen de profundidad, considerando los píxeles extraídos como en primer plano, es decir, la región de interés en el problema de la detección. Para un píxel de una determinada imagen de profundidad, $prof(i,j)$, su correspondiente píxel de la imagen en primer plano, pp, se calcula como:

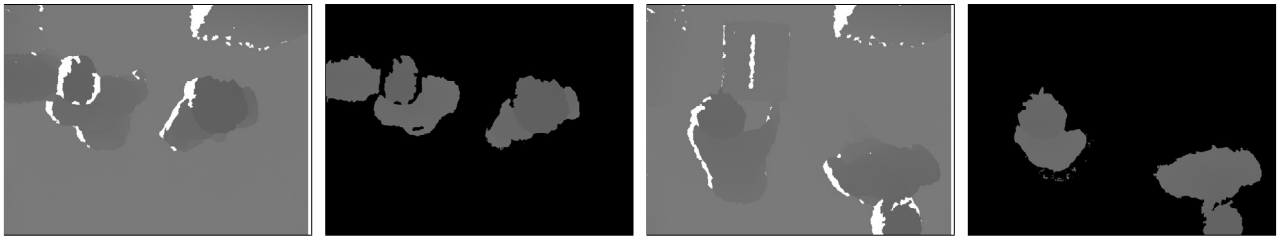
$$pp(i,j) = \begin{cases} prof(i,j) & \text{if } prof(i,j) < fondo(i,j) \times \tau \\ 0 & \text{en otro caso} \end{cases} \tag{5}$$

A diferencia de Heikkila y Silven (1999), el lector debe observar que el valor de los píxeles de primer plano se mantiene para su posterior análisis. De esta manera, la información de profundidad no sólo proporciona la ubicación, sino también información de la altura del píxel, es decir, del individuo. La Figura 5 muestra dos ejemplos de segmentación, presentando a la izquierda la imagen de entrada, y la segmentación resultante del primer plano a su derecha.

La definición del umbral en la ecuación 5 se determina según el escenario de aplicación. Para nuestro propósito, estamos interesados en detectar personas que cruzan una entrada, objetos muy salientes con respecto al fondo (principalmente suelo) en términos de distancia. En los resultados presentados en este documento, el umbral se ha fijado en $\tau = 0{,}9$.

3.2. Modelado basado en combinación de gaussianas

El segundo método de sustracción de fondo analizado ha sido propuesto por Zivkovic y der Heijden (2006). La inclusión



(a) (b) (c) (d)

Figura 5: (a, c) Imágenes de profundidad de entrada y (b, d) imágenes de segmentación resultantes.

de este método en la comparación es debido a su buen rendimiento en imágenes de color, y al hecho de que un esquema muy similar ha sido aplicado previamente en un escenario de conteo de personas por van Oosterhout et al. (2011). Este método realiza un modelo de fondo a nivel de píxeles basado en una combinación de gaussianas (GMM siglas en inglés), que extiende el método propuesto por Stauffer y Grimson (1999). El modelo de fondo se describe como:

$$p(\vec{x}|\mathcal{X}_T, fondo) \approx \sum_{m=1}^{C} \hat{\pi}_m \mathcal{N}(\vec{x}, \hat{\vec{\mu}}_m, \hat{\sigma}_m^2 I) \quad (6)$$

donde $\mathcal{X}_T = \{\vec{x}^{(t)}, \ldots, \vec{x}^{(tT)}\}$ es el conjunto de entrenamiento, $p(\vec{x}|\mathcal{X}_T, fondo)$ es la probabilidad de que el píxel $\vec{x}$ pertenezca al segundo plano o fondo, dado el conjunto de entrenamiento $\mathcal{X}_T$. $\hat{\vec{\mu}}_1, \ldots, \hat{\vec{\mu}}_C$ son las estimaciones de las medias, mientras que $\hat{\sigma}_1, \ldots, \hat{\sigma}_C$ de las varianzas, siendo I la matriz identidad. El peso de cada componente en (6) está definido por $\hat{\pi}_m$. Si se ordenan en orden descendente, el número de componentes C se puede obtener como

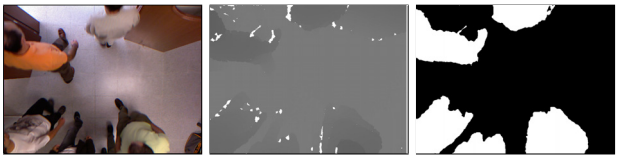
$$C = \arg\min_{c} \left(\sum_{m=1}^{c} \hat{\pi}_m > (1 - c_f) \right) \quad (7)$$

donde c se corresponde con el total de componentes considerado y c_f controla la cantidad de los datos que pueden pertenecer a objetos de primer plano sin influir en el modelo de fondo, siendo el número de componentes en el GMM no fijo a diferencia de otros métodos basados en GMM, como por ejemplo Stauffer y Grimson (1999); Han et al. (2004). El parámetro c_f controla cuando un objeto se considera como parte del primer plano. Cuanto menor sea el valor de c_f y por tanto mayor el valor de $(1 - c_f)$, mayor será el tiempo requerido para que un objeto pase a no ser considerado fondo. En nuestros experimentos, se probaron varios valores entre 0,05 y 0,9, y al final se concluyó que un valor de $c_f = 0,2$, exhibe un buen equilibrio entre la adaptación a los cambios del escenario y la eliminación de espúreos en la imagen de profundidad.

De acuerdo con (6) y considerando que una muestra $\vec{x}$, como la profundidad de un píxel en la posición (i, j), es decir $prof(i, j)$, un píxel se clasifica como primer plano, $pp(i, j)$, si

$$pp(i,j) = \begin{cases} prof(i,j) & \text{si } p(prof(i,j)|fondo) > c_{th} \\ 0 & \text{en otro caso} \end{cases} \quad (8)$$

donde c_{th} es el valor umbral que define si un píxel en una determinada posición pertenece al fondo en función de la distancia de Mahalanobis entre el valor del píxel y los centroides de las gaussianas que componen el modelo para dicha posición. Este valor se establece en tres veces la desviación típica de la gaussiana de acuerdo a Zivkovic y der Heijden (2006).



(a) (b) (c)

Figura 6: (a) Imagen RGB, (b) Imagen de profundidad y (c) Máscara de resultado de segmentación del primer plano utilizando GMM

4. Detección y seguimiento

Una vez que se ha obtenido la imagen del primer plano de la imagen actual, pp, se analizan las componentes conectadas en esa imagen. Después de filtrar las zonas menores por tamaño de forma similar a Brutzer et al. (2011), asumiendo que pp contiene un conjunto de m componentes válidas $B = \{b_1, b_2, ..., b_m\}$. De forma ilustrativa, esta operación de detección de objetos salientes aplicada a las imágenes de entrada presentadas en las figuras 5a, 5c y 6b producen respectivamente los resultados de las Figuras 5b, 5d y 6c.

Las aproximaciones de seguimiento por detección han mostrado un buen rendimiento en diferentes escenarios restringidos Andriluka et al. (2008); Leibe et al. (2008). por este motivo, se ha adoptado este enfoque para conectar las sucesivas detecciones en términos de trayectorias en el tiempo.

Para realizar este conexión, se realiza un encaje entre fotogramas consecutivos. Dadas las componentes detectadas en la imagen l, $B^l = \left\{b_1^l, b_2^l, ..., b_{m_l}^l\right\}$, se relacionan con las detectadas en la imagen anterior $B^{l-1} = \left\{b_1^{l-1}, b_2^{l-1}, ..., b_{m_{l-1}}^{l-1}\right\}$ por medio de un test de superposición. Dada la componente b_p^l, en la imagen actual, se localiza la componente de la imagen anterior con mayor superposición:

$$mb_p^l = \arg \max_{k=1,\ldots,m_{l-1}} \left(b_p^l \cap b_k^{l-1}\right) \quad (9)$$

El test de superposición en este escenario es válido en la mayoría de los casos porque con personas caminando a paso normal, la superposición de componentes es lo suficientemente alta entre fotogramas consecutivos. Además, la interacción entre los sujetos no es frecuente en un lugar de paso como una puerta, y las oclusiones son particularmente raras.

Un resultado positivo establece una conexión temporal entre las dos componentes, asignándose a la misma trayectoria. Una trayectoria se define entonces como una lista de componentes coincidentes y relacionadas en imágenes sucesivas, $t_i = \left\{b_t^1, b_t^2, ..., b_t^l\right\}$, donde la primera componente de la trayectoria se define como b_t^1, y la última como b_t^l. Una componente del fotograma actual sin encaje desencadena una nueva hipótesis de trayectoria. Trayectorias demasiado pequeñas se consideran ruido.

Se calcula un vector de características para cada componente p de la imagen actual l, conteniendo el área de blob, $área_p^l$, las dimensiones de su caja límite, sx_p^l, sy_p^l, y la localización de su punto más alto, px_p^l, py_p^l.

$$\mathbf{vb}_p^l = \left\{área_p^l, sx_p^l, sy_p^l, px_p^l, py_p^l\right\} \tag{10}$$

El conjunto de vectores de características correspondientes a las componentes que forman a una trayectoria permite describir dicha trayectoria de forma temporal. Para nuestro sistema de conteo de personas, cuando una trayectoria cesa será etiquetada como ENTRADA/SALIDA/TRÁNSITO.

El esquema aplicado en el procesamiento de imágenes se resume en el algoritmo 1. En nuestra configuración experimental, el etiquetado se realiza en base en la información proporcionada por el desplazamiento en el eje y sobre la línea de paso definida por la puerta.

Algoritmo 1 Esquema de procesamiento

```
Captura de imagen de profundidad
if fondo no está definido then
    Calcula fondo
else
    Sustracción de fondo
    Extracción de componentes
    for blob = 1 → m do
        Encaje de componentes con trayectorias activas
        if Concidencia then
            Actualización trayectoria
        else
            Crear una nueva trayectoria activa
        end if
    end for
    for trayectorias activas do
        if La trayectoria no se encajó then
            Etiquetar la trayectoria como ENTRADA/SALIDA/TRÁNSITO
            Desactivar trayectoria
        end if
    end for
end if
```

5. Resultados experimentales

Para probar la validez del enfoque, se han capturado imágenes con una configuración cenital en la puerta de entrada a una sala como se ha mostrado en las figuras previas. Esta grabación se ha llevado a cabo en dos sesiones diferentes, capturando imágenes para las que posteriormente se ha anotado de forma manual la acción de cada individuo como ENTRADA/SALIDA/TRÁNSITO. La cámara se encuentra en una entrada, por lo tanto se considera una acción como ENTRADA cuando una persona entra en la habitación y se mantiene en el interior desapareciendo de la vista, SALIDA cuando la persona deja la habitación y no es visible, y TRÁNSITO cuando la persona se muestra en el campo de visión sin cruzar la entrada o si el número de pasos o cruces por la zona de la puerta fuese par. Los parámetros utilizados para esta validación han sido de $k = 500$ para calcular el modelo de fondo simple (ecuación 2).

La primera secuencia de vídeo, de unos 15 minutos, contiene alrededor de 15000 imágenes. El número total de personas que cruzan, es decir, de eventos o acciones es 258, presentando diversidad de las condiciones de iluminación del escenario, como se puede observar en las Figuras 2 y 7. Para este conjunto de datos se realiza una comparativa aplicando las técnicas de modelado de fondo descritas de forma independiente tanto sobre la imagen visual (sólo la más potente basada en gaussianas), como la de profundidad.

La segunda secuencia de vídeo es más corta en duración, presentando 70 eventos. Sin embargo, la mayor parte de la misma se adquirió en condiciones de muy baja iluminación como se evidencia en la Figura 8. Los enfoques basados en información visual, tanto monocular como estéreo, no pueden aportar resultados dignos, es por ello que se incluyen sólo los resultados basados en la información de profundidad.

En primer lugar, se analizan los resultados alcanzados al etiquetar automáticamente las trayectorias de las acciones de ENTRADA/SALIDA/TRÁNSITO sobre la primera secuencia. La Tabla 1 presenta los resultados obtenidos indicando la exhaustividad o ratio de etiquetado correcto de las trayectorias, $TPR = \frac{TP}{TP+FN}$, y la precisión o valor predictivo positivo, $PPV = \frac{TP}{TP+FP}$, logrado por los distintos enfoques. Siendo TP, FP y FN respectivamente el número de verdaderos positivos, falsos positivos, y falsos negativos.

Observando los resultados de la Tabla 1, se puede concluir que incluso un modelo de fondo simple aplicado a la información de profundidad, se comporta mejor que uno más potente aplicado sobre la información visual. Este comportamiento es más evidente en un escenario cuyos cambios de iluminación son notables debido al autoajuste de la cámara, y la presencia de un piso reflectante, como se aprecia en la Figura 7. Ambos esquemas de modelado basados en información de profundidad proporcionan mejor exhaustividad y precisión. Por contra, al usar exclusivamente información visual, los cambios de iluminación y los reflejos del suelo afectan al rendimiento del enfoque visual resultando en la disminución en más de 15 puntos del TPR, y exhibiendo alrededor de 9 puntos menos de precisión. Se puede concluir que la información de profundidad proporcionada por una cámara comercial de consumo estándar es capaz de dar información válida y fiable para resolver el problema de conteo de personas con buena viabilidad.

Centrándonos en los resultados proporcionados exclusivamente por la información de profundidad, ambos modelados

Figura 7: Fotogramas del primer vídeo mostrando cambios de iluminación.

Tabla 1: Resumen de resultados de conteo de personas para la primera secuencia de vídeo. BM se refiere a modelado de fondo (BM sus siglas en inglés).

Anotación manual	Etiquetado automático (profundidad) simple BM		GMM BM		Etiquetado automático (RGB) GMM BM	
	TPR	PPV	TPR	PPV	TPR	PPV
ENTRADA	0,98	0,99	1,0	0,99	0,82	0,94
SALIDA	0,98	0,98	0,99	0,99	0,78	0,92
TRÁNSITO	0,73	1,0	0,81	1,0	0,57	0,57
Total	0,96	0,98	**0,98**	**0,99**	0,80	0,91

muestran un rendimiento bastante similar en nuestra configuración experimental. Debemos recordar al lector que no se hace uso de información de apariencia, por lo tanto, los resultados son notoriamente insensibles a los cambios de iluminación presentes en las imágenes.

Sin embargo, el modelado simple de fondo no es capaz de detectar objetos pequeños (con una altura menor a 1,20 metros), tales como sillas de ruedas, niños, etc. Por otro lado, esta aproximación, detecta sólo unos pocos falsos positivos. Se deben a fallos de seguimiento, que produjeron una escisión de la trayectoria, considerándose dos veces. Como se indicó en la sección 2, una trayectoria típica tiene alrededor de 18 cuadros, por lo que se ha optado por considerar las trayectorias breves como ruido, como se producen típicamente cuando el individuo está parcialmente fuera del campo de vista.

Para el enfoque de modelado basado en gaussianas, el detector de objetos es más sensible a los objetos pequeños. En cuanto a los errores, en ciertas situaciones con seguimiento de varias componentes, las trayectorias pueden duplicarse apareciendo detecciones espúreas que dan lugar a etiquetados incorrectos. Con este método la máscara de primer plano fue típicamente mayor que con el modelo simple, provocando en ocasiones dificultades para no agrupar a varios individuos en una componente. En cualquier caso, se observó que evidentemente para ambos enfoques, no es posible detectar múltiples sujetos u objetos cruzando ocultos por un paraguas o similar. De hecho, incluso un observador humano erraría.

Un experimento adicional fue realizado con la segunda secuencia de vídeo, donde la iluminación exhibe en la mayor parte del vídeo condiciones muy duras como se muestra en la Figura 8. La exhaustividad o tasa de verdaderos positivos se resume en la Tabla 2. Los resultados globales son ligeramente peores, si bien hay que hacer notar el menor tamaño de la secuencia y un agrupamiento de personas afecta en mayor medida a la tasa final. Se ha observado que esto ocurre para algunos eventos de SALIDA en los que un grupo de personas se mueve de forma compacta y en la misma dirección. Circunstancia que el sistema actual sólo es capaz de etiquetar como demasiado grande para una persona promedio, sin identificar la presencia de varios individuos en una componente.

6. Conclusiones

En este trabajo se ha hecho uso de los datos proporcionados por una cámara RGBD de consumo para detectar, seguir y

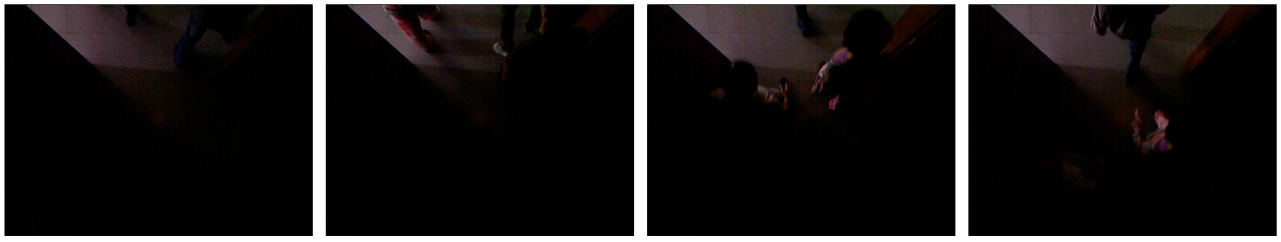


Figura 8: Fotogramas del segundo vídeo mostrando situaciones de escasa iluminación.

contar las trayectorias de las personas que cruzan un espacio. Nuestra principal conclusión es que la información de profundidad es una señal válida y sólida para resolver el problema de conteo automático de personas, proporcionando ventajas adicionales en términos de facilitar la segmentación de personas.

Para argumentar esta conclusión, se han comparado los resultados obtenidos con los dos esquemas de modelado de fondo para la posterior aplicación de la sustracción de fondo. El enfoque basado en gaussianas proporciona resultados ligeramente mejores.

El método propuesto es particularmente robusto en escenarios de baja o media densidad de personas incluso si se producen grandes o repentinos cambios de iluminación, o prácticamente ausencia de la misma como pudieran ser locales nocturnos o cines. Para escenarios con multitudes y aglomeraciones de personas, estimamos que sería necesario hacer cooperar tanto técnicas que integren información visual, como mejorar la identificación de personas en las componentes salientes.

Destacar que las condiciones de iluminación no juegan un papel clave en la solución propuesta. Sin embargo, el campo de visión está limitado por el sensor, de manera similar a los escenarios donde los sensores láser han demostrado hasta ahora ser más eficaces que los enfoques basados en información exclusivamente visual.

Tabla 2: Resumen de resultados de conteo de personas para la segunda secuencia de vídeo.

Anotado	TPR	PPV
ENTRADA	0,95	1,0
SALIDA	0,79	1,0
TRÁNSITO	1,0	0,75
Total	0,88	0,98

English Summary

People counting using a consumer RGBD camera

Abstract

In this paper, we prove that depth information provided by a consumer depth camera is a reliable data source to perform robust people counting. The adoption of a top view configuration reduces the space problem complexity for this task, while preserving privacy. Two different background subtraction approaches for color images are transferred to this context and tested in real video to perform detection, tracking, and behavioral patterns analysis of subjects crossing the field of view. The results achieved in an experimental setup with real video reported a TPR over 95 %, beating robust GMM background subtraction based only on the visual cue. The results suggest the benefits of the depth cue for this particular task.

Keywords:

People counting, Consumer depth cameras, Event detection, Object detection

Agradecimientos

Trabajo parcialmente apoyado por el Departamento de Informática y Sistemas de la ULPGC.

Referencias

Albiol, A., Albiol, A., Oliver, J., ., J. M., September 2012. Who is who at different cameras: people re-identification using depth cameras. IET Computer Vision 6 (5), 378–387.

Albiol, A., Mora, I., Naranjo, V., December 2001. Real-time high density people counter using morphological tools. IEEE Transactions on Intelligent Transportation Systems 2 (4), 204–218.

Albiol, A., Silla, J., 2010. Statistical video analysis for crowds counting. En: Proceedings of the 16th IEEE international conference on Image Processing (ICIP). pp. 2569–2572.

Andriluka, M., Roth, S., Schiele, B., 2008. People-tracking-by-detection and people-detection-by-tracking. En: IEEE Conf. on Computer Vision and Pattern Recognition.

Antic, B., Letic, D., D. Culibrk, V. C., 2010. K-means based segmentation for real-time zenithal people counting. En: Proceedings of the 16th IEEE International Conference on the Image Processing (ICIP). pp. 2565–2568.

Barandiaran, J., Murguia, B., Boto, F., 2008. Real-time people counting using multiple lines. En: Ninth International Workshop on Image Analysis for Multimedia Interactive Services. pp. 159–162.

Barbosa, B. I., Cristani, M., Bue, A. D., Bazzani, L., Murino, V., 2012. Re-identification with RGB-D sensors. En: 1st International Workshop on Re-Identification.

Bellotto, N., Hu, H., Feb- 2009. Multisensor-based human detection and tracking for mobile service robots. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 39 (1), 167–181.

Beymer, D., 2000. Person counting using stereo. En: Workshop on Human Motion. pp. 127–133.

Blanco, J., Burgard, W., Sanz, R., Fernandez, J., 2003. Fast face detection for mobile robots by integrating laser range data with vision. En: Proc. of the International Conference on Advanced Robotics (ICAR). pp. 953–958.

Bozzoli, M., Cinque, L., Sangineto, E., 2007. A statistical method for people counting in crowded environments. En: 14th International Conference on Image Analysis and Processing.

Brutzer, S., Hoferlin, B., Heidemann, G., 2011. Evaluation of background subtraction techniques for video surveillance. En: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1937–1944.

Camplani, M., del Blanco, C. R., Salgado, L., Jaureguizar, F., Garcí, N., January 2014. Advanced background modeling with RGB-D sensors through classifiers combination and inter-frame foreground prediction. Machine Vision and Applications 25 (1), 122–136.

Chan, A. B., Liang, Z.-S. J., Vasconcelos, N., 2008. Privacy preserving crowd monitoring: Counting people without people models or tracking. En: Computer Vision and Pattern Recognition. pp. 1 – 7.

Chan, A. B., Vasconcelos, N., April 2012. Counting people with low-level features and bayesian regression. IEEE TRANSACTIONS ON IMAGE PROCESSING 21 (4), 2160–2177.

Cohen, I., Garg, A., Huang, T., 2000. Vision-based overhead view person recognition. En: 15th International Conference on Pattern Recognition.

Cui, J., Zha, H., Zhao, H., Shibasaki, R., 2007. Laser-based detection and tracking of multiple people in crowds. Computer Vision and Image Understanding 106, 300–312.

Cui, J., Zha, H., Zhao, H., Shibasaki, R., 2008. Multi-modal tracking of people using laser scanners and video camera. Image and Vision Computing 26 (2), 240–252.

Englebienne, G., Krose., B., 2010. Fast bayesian people detection. En: 22nd Benelux Conference on Artificial intelligence.

Englebienne, G., van Oosterhout, T., Krose, B., 2009. Tracking in sparse multi-camera setups using stereo vision. En: Third ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC).

Fanelli, G., Gall, J., Gool, L. V., 2011a. Real time head pose estimation with random regression forests. En: Computer Vision and Patter Recognition (CVPR).

Fanelli, G., Weise, T., Gall, J., Gool, L. V., 2011b. Real time head pose estimation from consumer depth cameras. En: 33rd Annual Symposium of the German Association for Pattern Recognition (DAGM).

Fod, A., Howard, A., Mataric, M. J., May 2002. Laser-based people tracking. En: IEEE International Conference on Robotics and Automation (ICRA). Washington D.C., pp. 3024–3029.

García, J., Gardel, A., Bravo, I., Lázaro, J. L., Martínez, M., Rodríguez, D., Octubre-Diciembre 2012. Detección y seguimiento de personas basado en estereovisión y filtro de kalman. Revista Iberoamericana de Automática e Informática Industrial 9 (4).

Gollan, B., Wally, B., Ferscha, A., 2011. Id management strategies for interactive systems in multi-camera scenarios. En: 4th Conference on Context Awareness for Proactive Systems (CAPS). Budapest.

Han, B., Comaniciu, D., Zhu, Y., Davis, L., 2004. Incremental density approximation and kernel-based bayesian filtering for object tracking. En: IEEE Conf. on Computer Vision and Pattern Recognition. pp. 638–644.

Harville, M., 2004. Stereo person tracking with adaptive plan-view templates of height and occupancy statistics. Image and Vision Computing 22 (2), 127–142.

Heikkila, J., Silven, O., June 1999. A real-time system for monitoring of cyclists and pedestrians. En: IEEE Workshop on Visual Surveillance. Fort Collins, Colorado, pp. 82–90.

Hernández, D., Castrillón, M., Lorenzo, J., 2011. People counting with re-identification using depth cameras. En: 4th International Conference on Imaging for Crime Detection and Prevention (ICDP).

Hernández-Sosa, D., Castrillón-Santana, M., Lorenzo-Navarro, J., 2011. Multi-sensor people counting. En: IbPRIA. pp. 321–328.

Hou, Y., Pang, G., 2011. People counting and human detection in a challenging situation. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 41, 24–33.

Katabira, K., Nakamura, K., Zhao, H., Shibasaki, R., November 22 - 26 2004. A method for counting pedestrians using a laser range scanner. En: 25th Asian Conference on Remote Sensing (ACRS 2004). Thailand.

Kim, J. W., Choi, K. S., Park, W.-S., Lee, J.-Y., Ko, S. J., September 2002. Robust real-time people tracking system for security. Intelligent Building Society (IBS) 2 (3), 184 – 190.

Lee, G.-G., ki Kim, H., Yoon, J.-Y., Kim, J.-J., Kim, W.-Y., 2008. Pedestrian counting using an IR line laser. En: International Conference on Convergence and Hybrid Information Technology 2008.

Leibe, B., Schindler, K., Cornelis, N., Gool:, L. J. V., 2008. Coupled object detection and tracking from static cameras and moving vehicles. IEEE Trans. Pattern Anal. Mach. Intell. 30 (10), 1683–1698.

Lorenzo-Navarro, J., Castrillón-Santana, M., Hernández-Sosa, D., 2013. On the use of simple geometric descriptors provided by RGB-D sensors for re-identification. Sensors 13 (7), 8222–8238.

Marcos, A., Pizarro, D., Marrón, M., Mazo, M., Abril 2013. Captura de movimiento y reconocimiento de actividades para múltiples personas mediante un enfoque bayesiano. Revista Iberoamericana de Automática e Informática Industrial 10 (2).

Mathews, E., Poigné, A., 2009. Evaluation of a "smart"pedestrian counting system based on echo state networks. EURASIP Journal on Embedded Systems 2009, 1–9.
DOI: http://dx.doi.org/10.1155/2009/352172

Moore, B. E., Ali, S., Mehran, R., Shah, M., December 2011. Visual crowd surveillance through a hydrodynamics lens. Communications of the ACM 54 (12), 64–73.

Nakamura, K., Zhao, H., Shibasaki, R., K.S., Ohga, T., Suzukawa, N., 2006. Tracking pedestrians using multiple single-row laser range scanners and its reliability evaluation,. Systems and Computers in Japan 37, 1–11.

Oliver, J., Albiol, A., Albiol, A., 2012. 3d descriptor for people re-identification. En: 21st International Conference on Pattern Recognition (ICPR).

Piccardi, M., 2004. Background subtraction techniques: a review. En: IEEE International Conference on Systems, Man and Cybernetics. pp. 3099–3104.

Qiuyu, Z., Li;, T., Yiping, J., Wei-jun, D., 2010. A novel approach of counting people based on stereovision and dsp. En: The 2nd International Conference on Computer and Automation Engineering (ICCAE).

Satta, R., Pala, F., Fumera, G., Roli, F., 2013. Real-time appearance-based person re-identification over multiple kinect cameras. En: 8th International Conference on Computer Vision Theory and Applications (VISAPP). Barcelona, Spain.

Scheutz, M., McRaven, J., Cserey, G., 2004. Fast, reliable, adaptive, bimodal people tracking for indoor environments. En: Proceedings. 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2004. (IROS 2004). Vol. 2. pp. 1347–1352.

Septian, H., Tao, J., Tan, Y.-P., 5-8 Dec. 2006 2006. People counting by video segmentation and tracking. En: 9th International Conference on Control, Automation, Robotics and Vision, 2006. ICARCV '06. Singapore, pp. 1–4.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipma, A., Blake, A., June 2011. Real-time human pose recognition in parts from a single depth image. En: Computer Vision and Pattern Recognition.

Spinello, L., Arras, K. O., 2011. People detection in RGB-D data. En: Proc. of The International Conference on Intelligent Robots and Systems (IROS).

Stauffer, Grimson, 1999. Adaptive background mixture models for real-time tracking. En: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 246–252.

van Oosterhout, T., Bakkes, S., Kröse, B., 2011. Head detection in stereo data for people counting and segmentation. En: International Conference on Computer Vision Theory and Applications (VISAPP). pp. 620–625.

Velipasalar, S., li Tian, Y., Hampapur, A., July 2006. Automatic counting of interacting people by using a single uncalibrated camera. En: IEEE International Conference on Multimedia and Expo. Toronto, ON, Canada.

Xia, L., Chen, C.-C., Aggarwal, J. K., June 2011. Human detection using depth information by kinect. En: International Workshop on Human Activity Understanding from 3D Data in conjunction with CVPR (HAU3D). Colorado Springs, CO.

Yahiaoui, T., Khoudour, L., Meurie, C., July 2010. Real-time passenger counting in buses using dense stereovision. J. Electron. Imaging 20.

Yu, H., Liu, J., Liu, J., 2007. 3d feature extraction of head based on target region matching. En: Proceedings of the International Conference on Computational Intelligence and Security. pp. 366–370.

Zeng, C., Ma, H., 2010. Robust head-shoulder detection by pca-based multilevel HOG-LBP detector for people counting. En: 20th International Conference on Pattern Recognition (ICPR). Istambul, pp. 2069–2072.

Zhan, B., Monekosso, D. N., Remagnino, P., Velastin, S. A., Xu, L.-Q., 2008. Crowd analysis: a survey. Machine Vision and Applications 19, 345–357.

Zhao, X., Dellandréa, E., Chen, L., 2009. A people counting system based on face detection and tracking in a video. En: Advanced Video and Signal Based Surveillance.

Zivkovic, Z., der Heijden, F., 2006. Effcient adaptive density estimation per image pixel for the task of background subtraction. Pattern Recognition Letters 27, 773–780.