Memorias del XVI Congreso Latinoamericano de Control Automático, CLCA 2014
Octubre 14-17, 2014. Cancún, Quintana Roo, México

Detección y Seguimiento de Personas con Análisis de Color en Datos RGB-D ⋆

Karla Lourdes Luna Gallegos *
Elvia Ruth Palacios Hernández **
Antonio Marín Hernández ***

* *CIEP FI - UASLP, Av. Manuel Nava s/n 78290, San Luis Potosí, SLP, México (e-mail: karla.luna@alumnos.uaslp.edu.mx).*
** *Facultad de Ciencias - UASLP, Av. Salvador Nava s/n 78290, San Luis Potosí, SLP, México (e-mail: epalacios@fciencias.uaslp.mx).*
*** *DIA - UV, Sebastián Camacho No. 5, 91000, Xalapa, Veracruz México (e-mail: anmarin@uv.mx).*

Abstract: La detección y localización de personas es una tarea útil en múltiples contextos y aplicaciones. Algunas de estas requieren, debido a restricciones particulares, que el equipo sea ligero y a un costo accesible; lo cual limita los recursos que pueden usarse para el procesamiento. En este trabajo se presenta un algoritmo ligero para detección y seguimiento de personas mediante el análisis de datos RGB-D proporcionados por un sensor Kinect. Este sensor cuenta con una cámara RGB y un sensor/emisor de luz infrarroja que mediante triangulación obtiene los datos de profundidad (Depth) de la escena, a un costo accessible. El algoritmo propuesto procesa en tiempo real la información adquirida. Comienza por detectar los rostros de las personas que se encuentren en el campo visual del sensor aplicando un filtrado de color de piel. La persona más cercana es identificada como usuario y se genera para esta un perfil de color en el espacio HSV en función del color de su ropa. Una vez realizado esto se genera un proceso de segmentación realizado por el método de crecimiento de regiones en el espacio 3D, teniendo como punto semilla la profundidad a la que se encuentra la persona con respecto al marco referencial del sensor. El algoritmo es capaz de ubicar y seguir a una persona dentro del área de trabajo, y se implementará en un robot móvil para el seguimiento de personas adaptando la velocidad del robot a la del usuario.

Keywords: Datos RGB-D, Seguimiento de Personas, Detección de Rostros.

1. INTRODUCCIÓN

Los problemas de detección y seguimiento de personas son unos de los más estudiados en el área de la visión por computadora. Los desarrollos en esta área han generado múltiples aplicaciones tales como los sistemas de seguridad y las cámaras fotográficas donde utilizan la detección de rostros y sonrisas, solo por mencionar algunas. En el área de la robótica llamada Interacción Humano-Robot o HRI (por sus siglas en ingles), esta es una de las tareas principales que debe realizar un robot.

La detección de personas es un caso especial dentro de los varios grupo de algoritmos existentes en la comunidad de visión por computadora. Estos algoritmos extraen características tales como los contornos, tonos de piel, algunos puntos de interés, etc., con el fin de hacer la detección. En Viola and Jones (2001) se propuso un método de detección extremadamente rápido, que se basa en la extracción de características tipo Haar para el reconocimiento de objetos donde algunos grupos de variaciones de contraste forman una característica Haar. También utiliza Adaboost para la clasificación, el cual utiliza el valor de cambio en el contraste para determinar las áreas de luz y oscuridad. Los clasificadores Haar son de los métodos más utilizados, ya que permiten realizar una segmentación en tiempo real como en Talele et al. (2012), Viola and Jones (2004) y Wilson and Fernandez (2006). Otros trabajos relacionados a este problema utilizan las máquinas de soporte vectorial (SVMs) Sahbi and Boujemaa (2002) y los histogramas de gradientes orientados (HOG) en Do and Kijak (2012). La información necesaria para la detección de rostros y el seguimiento de personas es provista por los sensores, siendo la cámara RGB-D la más utilizada. En Pamplona et al. (2013) se presenta un sistema de autenticación de un rostro en 3D donde se utiliza una cámara RGB-D. Dos cámaras inteligentes de protocolo de Internet (IP) son utilizadas para obtener y localizar rostros de personas en Manap et al. (2010).

En Kristou et al. (2011) el seguimiento de personas se soluciona con un telémetro láser (LRF) y con una cámara omni-direccional. Otra solución para el seguimiento de personas es con un sensor identificador de Radio-Frecuencias (RFID) y una cámara montada en una unidad Pan-Tilt, ver Germa et al. (2010).

En 2010, la compañía Microsoft lanzó el sensor Kinect, originalmente fue fabricado para la consola Xbox360, pero

⋆ Esta investigación contó con el apoyo del Consejo Nacional de Ciencia y Tecnología CONACYT (CVU-372360).

ha sido ampliamente utilizado en la investigación debido a las posibilidades que ofrece y su bajo costo. El dispositivo permite el manejo de la consola mediante gestos, objetos, voz e imágenes sin necesidad de utilizar un dispositivo adicional de comunicación.

Algunos artículos utilizan el sensor Kinect para el reconocimiento de rostros por ejemplo Y.L. Li (2013), Hg et al. (2012) y otros para el seguimiento de personas como Luber et al. (2011) y Hoshino et al. (2011). En este trabajo se propone utilizar un sensor Kinect para la obtención de imágenes donde se quiere detectar el rostro de una persona en tiempo real. Para ello se desarrolla e implementa un algoritmo de detección del rostro, clasificación y de segmentación de personas.

La organización de este artículo se presenta de la siguiente manera. Las características del sensor Kinect se presentan en la Sección II. Una descripción del algoritmo de detección de rostro, de clasificación y segmentación son descritos en la Sección III. La Sección IV muestra los detalles de implementación final de los algoritmos. Finalmente la Sección V presenta los resultados experimentales de esta aplicación y las conclusiones del trabajo se describen en la Sección VI.

2. SENSOR KINECT

El sensor Microsoft Kinect consta de una cámara RGB, un emisor/sensor de luz infrarroja (IR) que por medio de triangulación funciona como un sensor de profundidad. Además de ello este sensor tiene un arreglo múltiple de micrófonos y un acelerómetro de 3 ejes en un motor de inclinación, ver Figura 1.

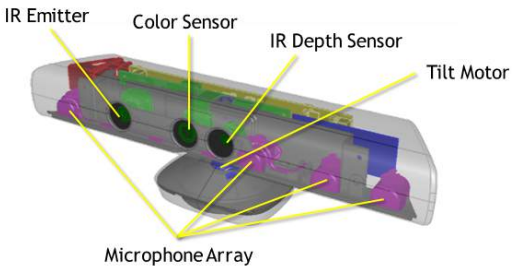

Fig. 1. Componentes del Sensor Kinect, Microsoft (2014).

De manera predeterminada la cámara RGB del Kinect, obtiene imágenes de 640 x 480 píxeles, la secuencia de vídeo en RGB utiliza 8-bits por canal y 11-bit para medir la profundidad a una velocidad de 30 cuadros/s, con un ángulo de apertura de 43° en vertical y 57° en horizontal.

Este proyecto utiliza la cámara RGB y el sensor de profundidad, esto provee la información necesaria para el procesamiento de imágenes. Para más información consultar Microsoft (2014).

3. ALGORITMO DE DETECCIÓN DE PERSONAS

El objetivo en la detección de rostros es básicamente reconocer si una persona se encuentra en el área de trabajo y localizarla con respecto a la posición 3D del robot. Existen muchos métodos utilizados para la detección de rostros, reconocimiento y seguimiento. El algoritmo para el seguimiento de rostros utilizado en este proyecto esta basado en Viola and Jones (2001), en donde el clasificador Haar extrae características y el AdaBoost los clasifica.

El algoritmo propuesto, se divide en tres etapas; la primera realiza el reconocimiento de rostro con el algoritmo de Viola and Jones (2001), el cual es validado por un umbral de color de piel y el porcentaje de dicho color en la región identificada como rostro. Después de detectar un rostro, se generá un perfil de color del usuario mediante una estadística del color de ropa usada por este. El perfil de color selecciona de una región determinada (unos centímetros abajo del rostro) el color predominante en el espacio de color Matiz, Saturación, Valor, conocido cómo HSV (por sus siglas en ingles Hue, Saturation, Value). Finalmente se realiza una segmentación de los puntos correspondientes al usuario, con los datos de profundidad proporcionados por sensor Kinect. Con estos datos se ubica a la persona en un espacio de 3D.

3.1 Detección de Rostros

El algoritmo para la detección de rostros realiza una primera transformación de la imagen mediante la generación de una nueva imagen llamada Imagen Integral (II). Se realiza una extracción de características usando filtros Haar, y finalmente se utiliza el Boosting para construir clasificadores en cascada. El esquema de la metodología utilizada se muestra en la Figura 2.

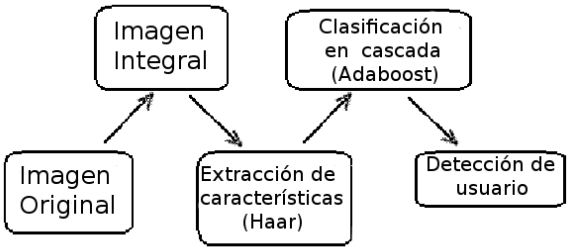

Fig. 2. Esquema de la metodología Viola and Jones (2001).

La imagen integral fue introducida a la visión por computadora por Viola and Jones (2001). La imagen integral permite una forma rápida de calcular la suma de los valores de los píxeles en una imagen dada y extraer características a diferentes escalas.

La imagen integral en un punto (x, y) contiene la suma de los píxeles de arriba y a la izquierda con (x, y) incluido:

$$II(x,y) = \sum_{x'<x, y'<y} i(x', y'), \qquad (1)$$

donde $II(x, y)$ es la imagen integral y $i(x, y)$ es la imagen original.

Las características de cada objeto son extraídas con la implementación de ciertos patrones, esto se realiza mediante la aplicaciíon de filtros de tipo Haar los cuales realizan una detección de la diferencia de intensidad, que generan curvas de nivel, puntos y líneas, mediante la adopción de zonas de contrastes. La percepción se procesa para encontrar lineamientos que se pueden utilizar para clasificar un objeto en particular.

La clasificación asigna una clase de características similares, de acuerdo con un modelo obtenido mediante un entrenamiento. El método de *boosting* presentado en Freund and Schapire (1997), consiste de varios clasificadores esenciales en un clasificador más complejo, siempre que tengan un número suficiente de muestras (en este caso 6000). Este algoritmo utilizado por Viola and Jones (2001) es conocido como AdaBoost y su procedimiento se muestra a continuación.

- Dadas unas muestras de imágenes $(x_1, y_1), ..., (x_n, y_n)$ donde $y_i = 0, 1$ para muestras negativas y positivas respectivamente
- Inicializa los pesos $w_{1,i} = \frac{1}{2m}, \frac{1}{2l}$ para $y_i = 0, 1$ respectivamente, donde m y l son los números de muestras negativas y positivas respectivamente.
- Para $t = 1, ..., T$:
  (1) Normaliza los pesos $w_{t,i} \leftarrow \frac{w_{t,i}}{\sum_{j=1}^{n} w_{t,j}}$
  (2) Para cada característica, j, entrenar un clasificador h_j el cual está restringido al uso de una sola característica. El error es evaluado respecto a $w_t, \epsilon_j = \sum_i w_i |h_j(x_i) - y_i|$
  (3) Escoge el clasificador, h_t, con el más pequeño ϵ_t
  (4) Actualiza los pesos:
  $$w_{t+1,i} = w_{t,i}\beta_t^{1-e_i} \tag{2}$$
  donde $e_i = 0$ si la muestra x_i es clasificada correctamente, $e_i = 1$ en otro caso, y $\beta_t = \frac{\epsilon_t}{1-\epsilon_t}$
- El clasificador robusto final es:
$$h(x) = \begin{cases} 1 & \sum_{t=1}^{T} \alpha_t h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} \alpha_t \\ 0 & en\ otro\ caso \end{cases} \tag{3}$$
donde $\alpha_t = log\frac{1}{\beta_t}$

Por último se realiza una validación del algoritmo para asegurar que se encontró a una persona, con un umbral de predefinido del color de tez de piel y con un porcentaje mínimo de 40 % de dicho color de piel en el la zona del rostro. El algoritmo ubica el rostro de hasta tres usuarios, tomando como referencia el punto central del rostro.

3.2 Clasificación del usuario

Después de detectar el rostro se realiza una clasificación del color de ropa del usuario en el espacio de color HSV para la localización de la persona de una forma más robusta. El primer paso es convertir la imagen obtenida por el sensor (RGB) al espacio HSV. En el espacio de color (HSV) el matiz proporciona el valor de cromaticidad, la saturación representa la pureza e intensidad de un color y el valor o luminancia corresponde a la claridad u oscuridad. Este espacio ofrece una forma más sencilla e intuitiva de segmentación de la que ofrece el espacio RGB como se explica en Gil and F. Torres (2004), por ejemplo, las zonas de sombra y brillo pueden provocar segmentaciones incorrectas en el espacio de RGB.

El algoritmo de detección de rostros es capaz de localizar a tres personas en el área de visión, siendo el más cercano el objetivo a clasificar y segmentar del resto de la imagen. En el caso de que el sensor pierde de vista al usuario objetivo, entra en operación el algoritmo de clasificación para la localización del usuario por medio del color de su ropa. La metodología que utiliza se muestra a continuación.

(1) Adquiere la información de la imagen en RGB y la convierte al formato de HSV
(2) Partiendo de la ubicación del rostro encontrado con el algoritmo anterior se obtiene una región de interés, la zona debajo del rostro del usuario.
(3) Se detecta el color predominante de esta región mediante el calculo del histograma en el espacio de color HSV.
(4) Se segmenta este color en el campo de visión del sensor.
(5) Finalmente, se ubica al usuario en un espacio de 2D (u, v), es decir, sobre los píxeles de la imagen $(640, 480)$ dando prioridad al punto de referencia del rostro.

3.3 Segmentación y Localización del Usuario

Finalmente se segmenta al usuario con el método de crecimiento de regiones, pero en esta ocasión en el espacio 3D correspondiente a los datos de profundidad proporcionados por el sensor. El método de crecimiento de regiones, trata de la selección de puntos de semilla inicial, examina los píxeles vecinos de los "puntos de semilla" iniciales y determina si los píxeles vecinos deben añadirse a la región. Esta adición depende de un criterio de región de pertenencia que podría ser, intensidad de los píxeles, la textura de nivel de gris, el color, la profundidad del objeto, etc.

El pseudocódigo del algoritmo se presenta a continuación:
Entradas:

- Lista de Regiones= $\{R_i\}$
- Puntos disponibles= $\{P\}$
- Punto semilla= $\{S\}$
- Región actual $\{R_c\} \rightarrow \emptyset$

Algoritmo:

- Mientras $\{P\}$ no sea cero:
  - Leer $\{S\}$
  - Encontrar vecinos más cercanos del punto semilla actual $\{S_c\} \rightarrow \{S_c\} \cup \{P\}$
  - Si $\{S_c\}$ cumple con criterio de pertenencia agregar a $\{R_c\} \rightarrow \{R_c\} \cup \{P\}$
  - si no $\{R_{c2}\} \rightarrow \{R_{c2}\} \cup \{P\}$

Donde R_i debe cumplir con:

- La segmentación debe ser completa, es decir, cada píxel debe pertenecer a una región.
- Los puntos de una región deben estar conectados de algún modo predefinido.
- Las regiones deben ser disjuntas.

El punto de semilla inicial esta dado por los algoritmos anteriores, el cual corresponde al punto central del rostro o de la persona. El criterio para la adición de los píxeles vecinos sera la distancia a la que se encuentra el usuario que es proporcionada por el sensor Kinect. Además, el sensor brinda la ubicación de los objetos en su área de visión en un espacio de 3D (x, y, z), haciendo posible la reconstrucción del lugar de trabajo y la localización del usuario en éste.

Para la localización en 3D se utiliza las librerías de nubes de puntos (PCL), estas librerías son independientes y de código abierto PointClouds (2014).

4. IMPLEMENTACIÓN

Los algoritmos son implementados en el middleware Robot Operating System (ROS). ROS es un conjunto de software de código abierto que incluye librerías y herramientas para los desarrolladores de software para robots. Éste funciona como sistema operativo, proporciona controladores de dispositivo, acceso a hardware, permite paso de mensajes entre procesos y da soporte a gran variedad de robots. ROS es un software distribuido de procesos (también conocido como nodos) que permite a los ejecutables ser diseñados de forma individual y acoplados en tiempo real. Estos procesos se pueden agrupar en paquetes y pilas, que puede ser fácilmente compartidos y distribuidos. ROS también es compatible con un sistema de repositorios que permiten la colaboración y distribución de código, ver ROSIntroduction (2012).

OpenNI y OpenCV son proyectos enfocados en la integración de sensores PrimeSense con ROS. El controlador ROS es compatible con los dispositivos de PrimeSense (PSDK5.0) y con el Kinect ROSIntroduction (2012). Este paquete permite obtener información del sensor y procesarla.

Con los datos que proporciona el sensor (RGB y depth) es posible determinar, cuando alguien se encuentra dentro del campo de visión de la cámara (2D) donde serán detectados y localizados en un espacio de tres dimensiones.

En la primer etapa del algoritmo, el usuario es encontrado por su rostro, la clasificación entra como auxiliar para la ubicación de la persona en caso de perder la detección del rostro y despues se realiza una segmentación para la localización del usuario en el espacio de trabajo.

El esquema final del algoritmo de seguimiento de personas se muestra en la Figura 3.

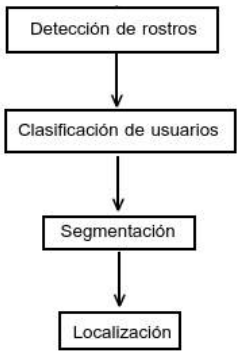

Fig. 3. Esquema del algoritmo de seguimiento de personas.

5. RESULTADOS

Esta sección presenta los resultados experimentales en cada parte del algoritmo. Las imágenes son obtenidas con el sensor Kinect descrito en la sección II.

La Figura 4 muestra como el algoritmo distingue la cara de la persona y la señala con un recuadro verde, incluso si la persona gira un poco el rostro, el algoritmo sigue detectando.

Fig. 4. Detección de rostro.

Este algoritmo es capaz de detectar hasta 4 rostros en el área de visión (Figura 5) y reconocer distintos tonos de piel (Figura 6)

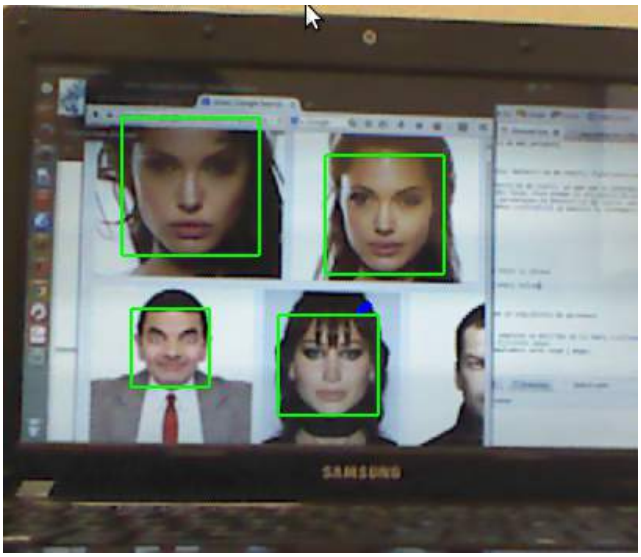

Fig. 5. Detección de múltiples rostros.

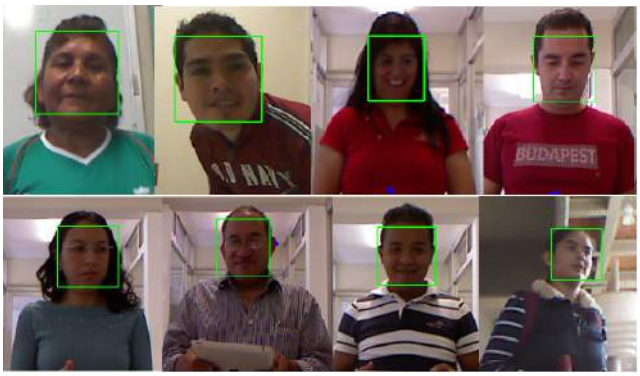

Fig. 6. Detección de rostros distintos tonos de piel.

Después de detectar el rostro se busca una región de importancia (zona del pecho) por medio de una máscara en la imagen, se aplica un histograma en esta región para obtener el color (en el espacio de HSV) de mayor predominancia y así localizar al usuario incluso si el rostro no es ubicado (Figura 7). Las características deseadas en la ropa que debe usar el usuario para una eficiente clasificación por color son: ropa con color distinto al fondo y con uno o dos colores dominantes, las texturas en la ropa pueden ser pequeñas.

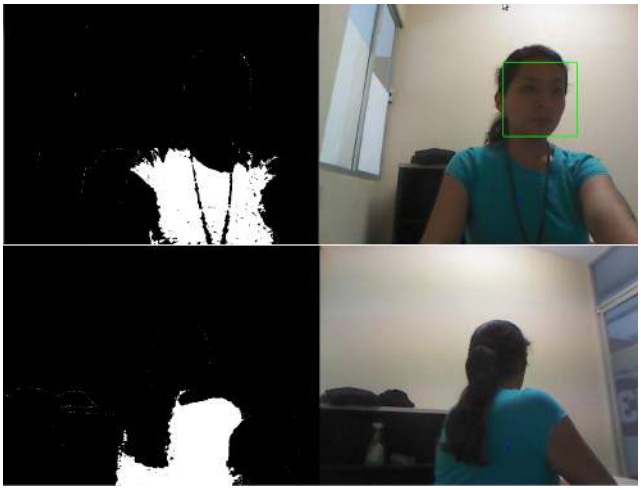

Fig. 7. Clasificación por color.

En caso de existir otra persona en el campo de trabajo será distinguido el usuario siempre que no porten el mismo color de ropa. En la fase de segmentación, se toma la posición del usuario, en un espacio de dos dimensiones, obtenido con el algoritmo de detección de rostro junto a la clasificación de color, entonces con el método de región de crecimiento es posible hacer la reconstrucción del cuerpo completo y separarlo del fondo. El usuario será ubicado siempre que el rostro se mantenga en el área de visión, para distinguir al usuario al encontrar más personas en el espacio de trabajo se realiza la segmentación por distancia. Cuando el rostro del usuario no es localizado el algoritmo de segmentación pierde la ubicación del usuario, para resolver este problema se implemento el clasificador por color en el espacio de HSV. Las tres partes del algoritmo mejoran la localización del usuario (Figura 8)

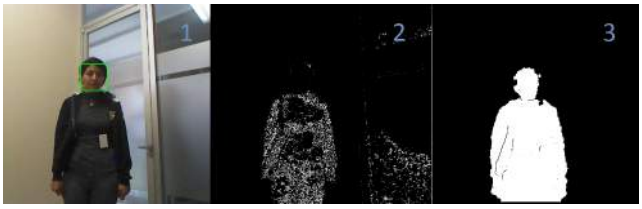

Fig. 8. Ubicacion del usuario.

Con la información de profundidad que el sensor ofrece, podemos conocer la posición de la persona en un espacio de tres dimensiones. La Figura 9 señala con un cuadro color magenta donde se localiza el usuario, la línea color cyan es la referencia (ortogonal al plano de visión) y el cubo rojo representa el sensor.

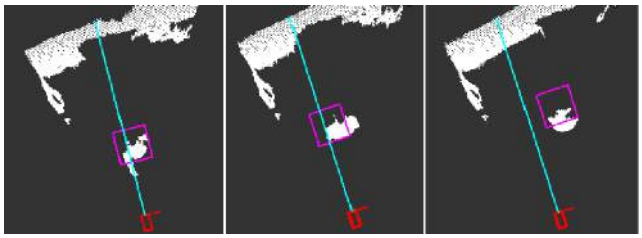

Fig. 9. Localización de la persona en un espacio de 3D

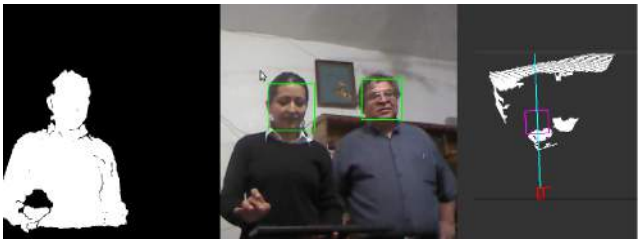

Fig. 10. Ubicación del usuario en existencia de más personas

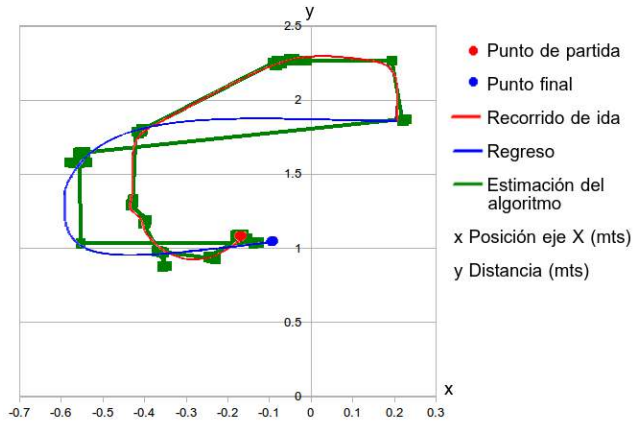

Fig. 11. Ubicación del usuario contra estimación del algoritmo

Como se puede observar en la Figura 9 el algoritmo es capaz de seguir a la persona en el campo de visión.

El algoritmo de detección de rostros localiza a los usuarios en el área de visión, distinguiendo al primer usuario detectado. Una vez que se tiene la información de posición (2D) y de color (HSV) del usuario, éste se puede separar del resto de personas y del fondo para ser ubicado en un espacio tridimensional (Figura 10).

Para mostrar el seguimiento que realiza el algoritmo se presenta la gráfica de la Figura 11 donde se muestra el recorrido que realizo el usuario y el que es detectado por el algoritmo.

Entonces el algoritmo se integra de 3 elementos: detección de rostro, clasificación y segmentación. La fase principal del algoritmo es la de detección de rostro, ya que con la información que está proporciona se inicializan las otras dos fases. Para probar la eficiencia de esta primer etapa se realizaron pruebas para obtener los porcentajes de detección de rostro, así como los falsos negativos y falsos positivos. En la Tabla 1 se muestra la información obtenida.

Muestras			
Positivo	Falso Positivo	Falso Negativo	Total
467	14	19	500
Porcentajes de detección			
93.4	2.8	3.8	100

Table 1. Porcentajes de detección de rostros

6. CONCLUSIONES

Este artículo propone un método para la detección y seguimiento de personas. Este método se realiza en dos partes, un algoritmo detecta el rostro y clasifica al usuario por su color de ropa y otro algoritmo lo segmenta y ubica en un espacio tridimensional. Además se presentan varios casos en los que el algoritmo es sujeto a condiciones reales, por ejemplo, cuando el usuario se encuentra en constante movimiento.

El algoritmo de detección de personas en combinación con el algoritmo de segmentación 3D, son capaces de seguir a una persona dentro del espacio visual del sensor. En comparación con otros sensores, el sensor Kinect mejora la robustez del problema pues permite adicionar condiciones de detección y clasificación con la información obtenida de ambos sensores que lo integran (cámara RGB, sensor de profundidad). Este sensor es de tecnología última generación y es interesante estudiar más de lo que éste dispositivo puede ofrecer y explorar sus capacidades.

Como trabajo futuro, implementaremos el algoritmo a un robot móvil para realizar un control de velocidad dependiente del usuario.

REFERENCES

Do, T. and Kijak, E. (2012). Face recognition using co-ocurrence histogram of oriented gradients.

Freund, Y. and Schapire, R.E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55, 119–139.

Germa, T., F.Lerasle, N.Ouadah, and V.Cadenat (2010). Vision and rfid data fusion for tracking people in crowds by a mobile robot. *Computer Vision and Image Understanding*, 114, 641–651.

Gil, P. and F. Torres, F.O. (2004). Detección de objetos por segmentación multinivel combinada de espacios de color.

Hg, R., Jasek, P., Rofidal, C., Nasrollahi, K., Moeslund, T., and Tranchet, G. (2012). An rgb-d database using microsoft's kinect for windows for face detection.

Hoshino, F., Luber, M., Spinello, L., and Arras, K. (2011). Human following robot based on control of particle distribution with integrated range sensors.

Kristou, M., Ohya, A., and Yuta, S. (2011). Target person identification and following based on omnidirectional camera and lrf data fusion.

Luber, M., Spinello, L., and Arras, K. (2011). People tracking in rgb-d data with on-line boosted target models.

Manap, N., Caterina, G.D., Soraghan, J., Sidharth, V., and Yao, H. (2010). Face detection and stereo matching algorithms for smart surveillance system with ip cameras.

Microsoft (2014). http://msdn.microsoft.com/en-us/library/jj131033.aspx.

Pamplona, M., Sarkar, S., Goldgof, D., Silva, L., and Bellon, O. (2013). Continuous 3d face authentication using rgb-d cameras. *IEEE Conference on Computer Vision and Pattern Recognition*.

PointClouds (2014). http://pointclouds.org/.

ROSIntroduction (2012). http://wiki.ros.org/ros/introduction.

Sahbi, H. and Boujemaa, N. (2002). Coarse-to-fine support vector classifiers for face detection. *IEEE Proceedings International Conference on Pattern Recognition*, 3, 359–362.

Talele, K., Kadam, S., and Tikare, A. (2012). Efficient face detection using adaboost.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Conference on Computer Vision and Pattern Recognition*, 1, 511–518.

Viola, P. and Jones, M. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57, 137–154.

Wilson, P.I. and Fernandez, J. (2006). Facial feature detection using haar classifiers. *Circuits, Systems and Computers (JCSC)*, 21, 127–133.

Y.L. Li, A.S. Mian, W.L.A.K. (2013). Using kinect for face recognition under varying poses, expressions, illumination and disguise.