Congreso Internacional en Ingeniería Electrónica. Mem. Electro 2015, Vol. 37, pp 216-221, Chihuahua, Chih. México
http://depi.itchihuahua.edu.mx/display/memorias_electro/MemoriaElectro2015
ISSN 1405-2172

USO DE PUNTOS DE ARTICULACIONES PARA LA SEGMENTACIÓN DE IMÁGENES DE PROFUNDIDAD OBTENIDAS DEL SENSOR KINECT Y ESCALAMIENTO DE ARTICULACIONES

Chavez-Montes Oscar Alejandro, Chacon-Murguia Mario I., Ramirez-Quintana Juan A.
Laboratorio de Percepción Visual con Aplicaciones en Robótica, Instituto Tecnológico de Chihuahua
oachavezm@outlook.com, mchacon@ieee.org, jaramirez@itchihuahua.edu.mx

RESUMEN.

La detección de personas convencional mediante sistemas de visión se realiza en la mayoría de los casos en imágenes tomadas por cámaras con sensor en el rango de luz visible. Estos métodos imitan el proceso de detección que los humanos utilizamos. Se utilizan propiedades basadas en gradientes, como histogramas de gradientes orientados HOG por sus siglas en inglés, o extraen puntos de interés como la transformada SIFT. En este artículo se propone utilizar las imágenes de profundidad obtenidas con el sensor Kinect para la detección de personas. Se utilizará el algoritmo de Jamie Shotton para la extracción de las articulaciones del cuerpo humano, y técnicas de procesamiento de imágenes para la manipulación de estas. Se utilizaran los puntos de articulaciones obtenidos, para determinar métricas para la calibración del sistema de terapia de brazos, así como el escalamiento de estos puntos para lograr un rango de extensión de brazos más amplio y así poder manipular cómodamente los objetos dentro de un ambiente de terapia virtual. Finalmente se presentan la implementación y los resultados obtenidos en un escenario virtual orientado a la terapia física.

Palabras clave— Segmentación, Procesamiento en profundidad, Factor de Escala, Rehabilitación Virtual.

ABSTRACT.

Conventional people detection algorithms are mostly given in images taken by visible light cameras. These methods mimic the detection process that humans use. They often use properties based on gradients, such as histograms of oriented gradients (HOG), or feature extraction like SIFT transform, etc. This paper aims to use the depth images obtained with the Kinect sensor for detecting people. Jaimie Shotton algorithm is used for extracting human body joints, and image processing techniques for manipulating the depth image. Joints will be used to find metrics used to calibrate the arms therapy system, also this joints will be escalated to achieve a larger extension range so objects in the virtual environment can be manipulated comfortably inside the virtual environment. Finally the implementation and the results obtained in a virtual rehabilitation scenario oriented to therapy are presented.

Keywords: Segmentation, Depth processing, Scale Factor, Virtual rehabilitation.

1. INTRODUCCIÓN.

El campo del procesamiento digital de imágenes se refiere a la interpretación de una imagen digital por medio de una computadora, esta imagen tiene un número finito de elementos que tienen una posición particular y un valor que corresponde a la intensidad del nivel de gris o color [1]. En el caso de las imágenes de profundidad este valor corresponderá a la distancia entre el objeto y la cámara. Los sensores 3-D nos proporcionan información tanto de color o niveles de gris, como también información de profundidad, teniendo así información en 3 dimensiones (X, Y, Z).

El creciente desarrollo en la tecnología utilizada en la fabricación de sensores 3-D así como en la forma en la que estos operan, ha logrado que se puedan obtener sensores de bajo costo accesibles para todo el público como el Kinect [2], abriendo así un sin número de posibilidades para el desarrollo de aplicaciones basadas en esta tecnología. La información de profundidad [3] proporcionada por estos sensores, en especial la información de profundidad obtenida con el Kinect, contiene una gran cantidad de ruido, ya que muchos de los pixeles en esta imagen pueden no tener valor de profundidad, como consecuencia de la transparencia o la mala reflexión del patrón de luz utilizado por el sensor en los objetos presentes en la escena. Esta información poco confiable o faltante (huecos) dependiendo de la exigencia de la aplicación, tiene que ser recuperada antes de ser usada.

La capacidad de trabajar con información de profundidad brindada por los sensores 3-D permite analizar el movimiento de las personas de una manera no invasiva, es decir; sin necesidad de colocar marcadores o sensores en puntos estratégicos del cuerpo humano para analizar sus movimientos.

El análisis de movimiento humano ha sido de gran interés desde los inicios de la visión por computadora debido a su relevancia en una gran variedad de aplicaciones. Con el desarrollo de nuevos sensores y algoritmos de estimación de pose han surgido grandes oportunidades en este campo [3, 4]. El análisis del movimiento no solo consta de obtener datos del esqueleto humano para estimar su pose, sino también en entender el movimiento que se realiza.

El estudio del reconocimiento de actividades mediante el uso de partes del esqueleto ha demostrado que se pueden reconocer una gran variedad de acciones utilizando información de articulaciones que representen el esqueleto de una persona [5]. La idea de analizar el movimiento utilizando articulaciones del esqueleto ha sido muy explorada, se tienen principalmente 3 maneras de obtener información del esqueleto: sistemas de

captura de movimiento (MoCap), imágenes de color de una sola vista o de múltiples vistas y mapas de profundidad. La gran diferencia entre estos métodos es la cantidad de ruido embebido que contienen, en resumen el MoCap es el más libre de ruido [6].

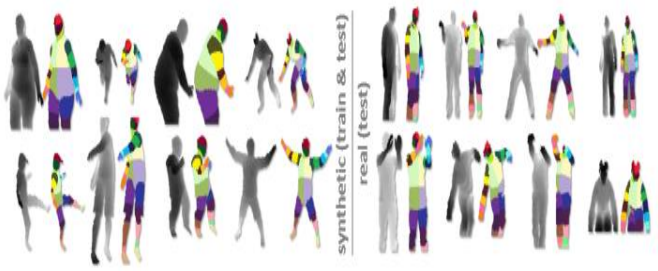
El propósito del trabajo reportado en este artículo es desarrollar un entorno virtual en el que se realice el análisis de movimiento del usuario en tiempo real, empleando un sensor de profundidad (Kinect), como la interfaz entre el usuario y el ordenador. Se emplearon técnicas de procesamiento de imágenes con la finalidad de analizar el movimiento de los miembros superiores del usuario, lo cual será utilizado en un ambiente de terapia virtual.

En las siguientes secciones se hablará sobre el método utilizado para la extracción de articulaciones, para más adelante hablar sobre cómo estos puntos son utilizados para realizar la segmentación de la imagen y el cálculo de las métricas necesarias para realizar la calibración de alcance de brazos, y una vez obtenida esta información se hablará de la influencia que tiene para el escalamiento de los puntos de articulaciones de manos para lograr un rango de movimiento más grande en relación a la distancia recorrida por la mano del usuario. Finalmente se hablará sobre los resultados obtenidos y las conclusiones a las que se llegaron al implementar lo realizado en este artículo en un ambiente de terapia virtual para brazos.

2. EXTRACCIÓN DE ARTICULACIONES.

Para extraer las posiciones de interés en el esqueleto se utiliza el método presentado por Shotton *et al.* [8] que consiste en el reconocimiento de articulaciones del cuerpo en 3-D en una imagen de profundidad sin utilizar información temporal. En este método se diseña una representación intermedia de las partes del cuerpo mapeando la dificultad de la estimación de pose en una clasificación por pixel más simple, también se entrena un clasificador aleatorio de decisiones, el cual funciona cuadro por cuadro a través de diferentes tipos de cuerpos (forma, tamaño, vestimenta, etc.), y el aprendizaje discriminativo elimina naturalmente las oclusiones y el recortado de la persona por el cuadro de la imagen. Aun sin utilizar información temporal, la estimación de la posición de las articulaciones es estable y precisa, siendo que este clasificador no utiliza información temporal, solo trabaja con las poses estáticas y no con el movimiento, por lo que se descartan poses redundantes de la captura de movimiento inicial utilizando el vecino más lejano.

La aportación clave del método de Shotton *et al.* es la representación de la parte intermedia del cuerpo. Se define una serie de etiquetas codificadas en color para representar partes del cuerpo y cubrir robustamente el mismo, lo cual es ilustrado en la Figura 1. Algunas de estas partes son definidas para localizar las articulaciones de interés, mientras las demás se utilizan para rellenar huecos o para servir de ayuda en la estimación de otras partes del cuerpo. Las definiciones de estas partes pueden ser cambiadas para utilizarse en aplicaciones particulares.



(a) (b)

Figura 1. (a) Imágenes de profundidad utilizadas para el entrenamiento y prueba. (b) Imágenes de profundidad utilizadas para la prueba del método.

Para la comparación de características se emplea un simple algoritmo inspirado en [7], en un pixel dado *x*. Las características pueden calcularse mediante:

$$f_\theta(I,x) = d_I\left(x+\frac{u}{d_I(x)}\right) - d_I\left(x+\frac{v}{d_i(x)}\right) \qquad (1)$$

Donde $d_I(x)$ es la profundidad del pixel *x* en la imagen *I*, y los parámetros *u* y *v* son los desplazamientos. La normalización de estos desplazamientos asegura que estas características sean invariantes en profundidad. Con la ecuación 1 se puede calcular la característica f_{θ_1} la cual dará una respuesta positiva para los pixeles *x* cerca de la parte superior del cuerpo, pero valores cercanos a cero para los pixeles *x* cerca de la parte baja del cuerpo. La característica f_θ ayudará a encontrar información de estructuras delgadas como lo es el brazo. Individualmente estas características proveen una señal débil sobre la parte del cuerpo a la cual pertenece el pixel, pero en combinación con un árbol de decisiones son suficientes para eliminar la ambigüedad de las todas las partes entrenadas.

Un árbol de decisiones consiste de un nodo de separación y un nodo de hojas. Cada nodo de separación consiste de una característica f_θ y un umbral τ para clasificar un pixel *x* en la imagen *I*. Se empieza por la raíz y se evalúa repetidamente la Ecuación 1, yendo por la rama izquierda o derecha dependiendo de la comparación con el umbral τ. Al alcanzar el nodo izquierdo del árbol *t*, una distribución aprendida $P_t(c \mid I,x)$ sobre una parte del cuerpo se guarda en una etiqueta *c*. las distribuciones son promediadas junto a todos los arboles de decisiones para resultar en una clasificación final, al conjunto de *T* arboles de decisiones se le conoce como bosque.

$$P(c \mid I, x) = \frac{1}{T}\sum_{t=1}^{T} P_t(c \mid I, x) \quad (2)$$

Para el entrenamiento de cada árbol se utilizó un conjunto aleatorio distinto de imágenes sintetizadas. Se tomó un subconjunto aleatorio de 2000 pixeles de cada imagen como ejemplos, para asegurar una distribución uniforme sobre todas las partes del cuerpo. Cada árbol es entrenado utilizando el siguiente algoritmo [8].

- Se propone un conjunto el cual se divide en candidatos $\phi = (\theta, \tau)$

Partiendo del conjunto de ejemplos $Q = \{(I, x)\}$ hacia la derecha e izquierda de cada subconjunto ϕ :

$$Q_1(\phi) = \{(I, x) \mid f_\theta(I, x) < \tau\} \quad (3)$$

$$Q_r(\phi) = Q \setminus Q_1(\phi) \quad (4)$$

- Calcular ϕ dada la ganancia más grande en información:

$$\phi^* = \arg\max G(\phi) \quad (5)$$

$$G(\phi) = H(Q) - \sum_{s \in \{l,r\}} \frac{|Q_s(\phi)|}{|Q|} H(Q_s(\phi)) \quad (6)$$

Donde la entropía de Shanon $H(Q)$ es calculada en el histograma normalizado de las partes del cuerpo etiquetadas como $l_I(x)$ para todo $(I, x) \in Q$.

Si la ganancia $G(\phi^*)$ es suficiente, y la profundidad del árbol está por debajo del máximo, entonces se hace una recursión para los subconjuntos de la derecha e izquierda $Q_r(\phi^*)$ y $Q_1(\phi^*)$.

El reconocimiento de las partes del cuerpo infiere información por pixel, esta información se agrupa a través de los pixeles para generar proposiciones confiables sobre las posiciones de las articulaciones del esqueleto en 3-D. Estas proposiciones son la salida final del algoritmo.

Para acumular los centros globales 3-D de la masa de probabilidad para cada parte, se utiliza un enfoque basado en el cambio de la media [9] con un kernel gaussiano ponderado.
El estimador de densidad para las partes del cuerpo se define como:

$$f_c(\hat{x}) \alpha \sum_{i=1}^{N} w_{ic} \exp\left(\left\| \frac{\hat{x} - \hat{x}_i}{b_c} \right\|^2\right) \quad (7)$$

Donde $\hat{x}$ es el espacio de coordenadas en 3-D, N es el número de pixeles en la imagen, w_{ic} es la ponderación del pixel, x_i es la reproyección (cambio de sistema de coordenadas con otra proyección) del pixel dentro del espacio de coordenadas dado por la profundidad $d_I(x_i)$ y b_c es el ancho de banda aprendido por parte del cuerpo. La ponderación del pixel w_{ic} considera ambas probabilidades de las partes del cuerpo inferidas en el pixel y en la superficie del área del pixel:

$$w_{ic} = P(c \mid I, x_i) \cdot d_I(x_i)^2 \quad (8)$$

Esto asegura que las densidades estimadas son invariantes a la profundidad. El desplazamiento de la media de estas densidades es utilizado para encontrar eficientemente las modas en esta densidad. Todos los pixeles que se encuentren sobre un umbral de probabilidad λ_c serán utilizados como puntos de inicio por parte del cuerpo *c*. Se provee una estimación de confianza final, siendo esta la suma de los pesos de los pixeles que se acerquen a cada moda. Estas modas corresponden a información de pixeles que se mapean a la superficie del cuerpo, cada una de las cuales es integrado a la escena por un offset *z* aprendido ζ_c para producir una estimación de articulaciones final.

3. SEGMENTACIÓN DEL CUERPO HUMANO Y CALIBRACIÓN DE ALCANCE.

Para la segmentación del cuerpo humano se utilizó la información de las articulaciones obtenida con el método anteriormente mencionado. Para esto se utiliza un umbral de profundidad (T_D) en el cual lo que se encuentre detrás de la persona detectada será eliminado, para determinar el umbral se utiliza el punto de cadera central, de no existir o no encontrarse este punto se utiliza el de la cabeza, teniendo así:

$$I_D(x, y) = \begin{cases} I(x, y) & si \quad I(x, y) < T_D \\ 0 & de\,otra\ forma \end{cases} \quad (8)$$

Este umbral se encuentra tomando los puntos de los hombros, codos y cabeza, teniendo que el máximo entre estos 3 puntos será el umbral de profundidad.

$$T_D = Sp \vee (Hp \vee Ep) \quad (9)$$

En la Figura 2 se puede observar la información sin procesar obtenida con el sensor Kinect, en la cual se puede apreciar la información de profundidad no medida con color negro. En la Figura 3 se muestra la segmentación de la información cruda

del Kinect obtenida utilizando la Ecuación 9, con el umbral calculado con la Ecuación 10.

Figura 2. Imagen obtenida con Kinect

Figura 3. Imagen segmentada

Una vez se tiene la imagen segmentada, se hace el cálculo de la distancia euclidiana entre las coordenadas del hombro derecho (SD_x, SD_y) y mano derecha (HD_x, HD_y) así como del hombro izquierdo (SL_x, SL_y) y mano izquierda (HL_x, HL_y), para así obtener la máxima extensión de brazos posible y poder determinar el límite para la colocación de los objetos que servirán para la terapia física.

$$DL = \sqrt{(SL_x - HL_x)^2 + (SL_y - HL_y)^2} \quad (10)$$

$$DR = \sqrt{(SD_x - HD_x)^2 + (SD_y - HD_y)^2} \quad (11)$$

Estas distancias se estarán actualizando a lo largo de la terapia, si se sobrepasa un umbral de distancia, es decir, si el paciente se mueve hacia enfrente o hacia atrás y sobrepasa el umbral, también solo se tomara el máximo de cada una de estas distancias, para que el límite de colocación de objetos sea la máxima apertura de los brazos. En la Figura 4 se puede apreciar en azul el limite en el cual se colocarán los objetos para su manipulacion.

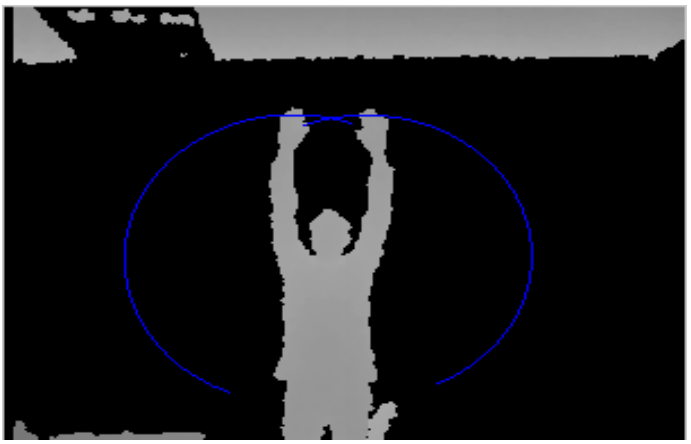


Figura 4. Círculos para la colocación de objetos

Para colocar los objetos dentro de la escena se utilizan círculos, cuyo radio será la distancia euclidiana mencionada anteriormente, de estos círculos solo se utilizaran 8 octantes del círculo unitario con centro en cada uno de los hombros como se muestra en la Figura 4.

Para retirar estos objetos con coordenadas (O_x, O_y), se utilizará de igual manera la distancia euclidiana entre las manos y los objetos.

$$DL_{OI} = \sqrt{(HL_x - O_x)^2 + (HL_y - O_y)^2} \quad (12)$$

$$DR_O = \sqrt{(HD_x - O_x)^2 + (HD_y - O_y)^2} \quad (13)$$

Si esta distancia euclidiana es menor a un umbral, entonces el objeto se retira, de no ser así el objeto permanece en su sitio. La terapia concluye al remover exitosamente todos los objetos de la escena.

4. ESCALAMIENTO DE PUNTOS.

El paso de calibración, comentado en la sección anterior, nos sirve para encontrar información referente al usuario, es

decir, en una primera etapa de la terapia se debe realizar una calibración para conocer los parámetros del usuario para así dar la posibilidad de interacción con los objetos del escenario, la información obtenida de esta función será:

- La distancia entre la mano y el hombro en ambos hemisferios del cuerpo: esto para determinar que tanto se puede extender un brazo y que diferencia existe entre un hemisferio del cuerpo y el otro, lo cual sirve para determinar si un brazo tiene un rango de extensión más pequeño y realizar las correcciones necesarias dentro del ambiente virtual.
- Factor de escala global: este factor de escala es utilizado para modificar el alcance de la mano del paciente, con el objetivo de lograr una fácil interacción con todos los objetos del escenario.

La distancia entre las manos y los hombros se calcula por medio de las Ecuaciones 11 y 12, en las cuales se utiliza la distancia euclidiana para conocer un valor numérico que representara la extensión que se obtiene de ambos brazos.

Una vez calculada la extensión de los brazos, se debe de almacenar y reportar la máxima extensión de los brazos por separado (MDL y MDR) para esto se utiliza la Ecuación 14.

$$\begin{aligned} MDL &= DL \vee MDL(t-1) \\ MDR &= DR \vee MDR(t-1) \end{aligned} \tag{14}$$

Esta distancia máxima nos indica cual es la máxima extensión lateral de brazos que logró el paciente en la etapa de calibración. Al conocer la máxima extensión de brazos podemos determinar si existe una extensión mayor o menor en cualquiera de los brazos. De existir un problema, es necesario corregirlo. Para esto se utiliza un factor de escala (sF), el cual se calcula con la Ecuación 15.

$$\begin{aligned} sFL &= \frac{MDL}{MDR} \\ sFR &= \frac{MDR}{MDL} \end{aligned} \tag{15}$$

La influencia de este factor se muestra en la Figura 6, donde en la Figura 6(a) observamos los puntos de las manos determinados por Kinect representados por guantes negros, los cuales se colocan sobre las manos para indicar su posición, mientras que en la Figura 6(b) se observa cómo se escala el punto de la mano con menor capacidad de extensión para lograr el mismo alcance que la mano con mayor extensión, teniendo así una nueva coordenada para el punto que corresponde a la mano afectada.

Existe otro problema a considerar, dado que se trabaja con un ambiente virtual, debemos asegurarnos de que el paciente sea capaz de interactuar con todos los objetos que se encuentran dentro del ambiente virtual. La interacción con los objetos terapéuticos dentro del entorno virtual puede verse afectada por el tamaño de los brazos y la distancia a la que se encuentre el paciente del sensor. Para lograr una interacción satisfactoria con el ambiente se hace uso de otro factor de escala (factor de escala global).

(a) (b)

Figura 6. Influencia del factor de escala, (a) Punto sin factor de escala, (b) Punto de interés escalado

El factor de escala global (gsF) se calcula utilizando la distancia máxima encontrada con la Ecuación 14 y la distancia entre el hombro y las coordenadas de objetos más alejados del usuario, las coordenadas utilizadas no necesariamente pertenecen al mismo objeto ya que se busca encontrar la combinación de coordenadas que se encuentren más retiradas del usuario. Para este cálculo también utilizamos la distancia euclidiana como se muestra en la Ecuación 16.

$$\begin{aligned} DL_{SO} &= \sqrt{\left(SL_x - mO_x\right)^2 + \left(SD_y - mO_y\right)^2} \\ DR_{SO} &= \sqrt{\left(SD_x - MO_x\right)^2 + \left(SD_y - MO_y\right)^2} \end{aligned} \tag{16}$$

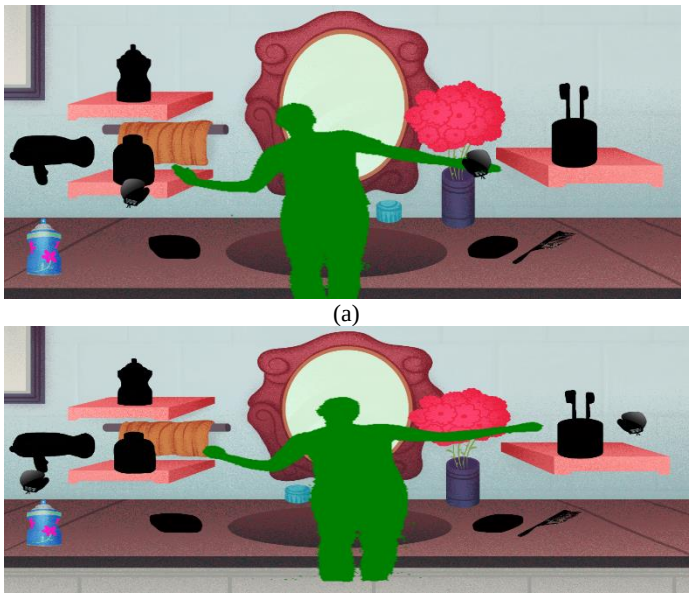
Donde (mO_x, mO_y) es la combinación de coordenadas de objetos más cercanos a la esquina superior izquierda, y (MO_x, MO_y) es el par de coordenadas más cerca de la esquina inferior derecha de la imagen. De igual manera se utiliza la máxima distancia entre hombro y objeto para cada brazo (MDL_{SO}, MDR_{SO}) para calcular el factor de escala, el máximo de estas distancias se calcula con la Ecuación 17.

$$\begin{aligned} MDL_{SO} &= DL_{SO} \vee MDL_{SO}(t-1) \\ MDR_{SO} &= DR_{SO} \vee MDR_{SO}(t-1) \end{aligned} \tag{17}$$

Por último se calcula el factor de escala global, el cual nos muestra la relación que existe entre la extensión del brazo y la distancia entre el hombro y el objeto. El factor de escala se calcula como se muestra en la Ecuación 18.

$$\begin{aligned} gsFL &= MSL_{SO} / MDL \\ gsFR &= MDR_{SO} / MDR \end{aligned} \tag{18}$$

En la Figura 7 se puede apreciar la influencia del factor de escala global, en la Figura 7(a) se observan los puntos de las manos (guantes negros) corregidos cuando existe una diferencia de extensión de brazos, pero también se aprecia que la máxima extensión de los brazos es insuficiente para la manipulación de los objetos sin necesidad de cambiar de posición, es para esto que se aplica el *gsF*, cuyo impacto se puede observar en la Figura 7(b), en la cual se presentan los puntos de las manos (guantes negros) con una distancia mayor a la posición de las manos. Esto nos permite manipular todos los objetos de la escena sin la necesidad de cambiar de posición.



(a)

(b)

Figura 7. Influencia del factor de escala global, (a) punto sin *gsF*, (b) punto con *gsF*.

5. RESULTADOS Y CONCLUSIONES

En este trabajo se logró obtener exitosamente las articulaciones de interés mediante el algoritmo de Shotton *et al*. Así como también una segmentación con buena velocidad de procesamiento, ya que se procesa la información a la misma velocidad de adquisición del Kinect la cual es de 30 cuadros por segundo, permitiendo realizar la terapia virtual en tiempo real. La posición de los objetos se determinó exitosamente mediante la distancia euclidiana y la colocación de estos se encuentra dentro del rango de alcance de la mano de la persona. Para determinar si se tocó un objeto también se utilizó la distancia euclidiana con cierto umbral dando esto buenos resultados. También se logró implementar esto en un ambiente virtual con un fondo y objetos como imágenes. Para la manipulación correcta de estos se utilizaron satisfactoriamente los factores de escala, de forma que se logra extender el alcance de los puntos de la mano encontrados, para virtualmente observar un movimiento más grande con un movimiento físico más pequeño.

6. AGRADECIMIENTOS

Esta investigación fue apoyada por el Fondo Mixto de Fomento a la Investigación Científica y Tecnológica CONACYT-Gobierno del Estado de Chihuahua bajo el proyecto CHIH-2012-C03-193760, and CHI-MCIET-2013-230.

7. REFERENCIAS

[1] R. C. Gonzales and R. E. Woods.Digital Image Processing, Second Edition ed. Upper Saddle River, New Jersey, U.S.A. Prentice Hall, 2002.

[2] Microsoft Corp. Redmond WA. Kinect for Xbox 360.

[3] J. Han, L. Shao, D. Xu, J. Shotton. "Enhanced Computer Vision with Microsoft Kinect Sensor: A Review," Cybernetics, IEEE Transaction,. vol. 43, 2013, pp. 1318-1334.

[4] Y. Berdnikov and D. Vatolin. "Real-time depth map occlusion filling and scene background restoration for projected-pattern-based depth camera," in Proc. Int. Conf. Comput. Graph. Vision, 2011, pp. 1-4.

[5] S. Milani and G. Calvangno. "Joint denoising and interpolation of depth maps for MS Kinect sensors," in Proc. ICASSP, 2012, pp. 797-800.

[6] M. Schmeing and X. Jiang. "Color Segmentation Based Depth Image Filtering," in Proc. Int. Workshop Depth Image Anal., 2012.

[7] M. Camplani and L. Salgado, "Efficient spatio-temporal hole filling strategy for Kinect depth maps," in Proc. SPIE Int. Conf. 3-D Image Process. Appl., vol. 8290, 2012, pp. 1–10.

[8] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "*Real-Time Human Pose Recognition in Parts from a Single Depth Image*," in proc *CVPR, IEEE*, 2011.

[9] X. Lu, C. Chiah-Chih, J.K. Aggarwal, "Human deteccition using depth information by Kinect," in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops(CVPRW), 2011, pp. 2160-7508.