Congreso Internacional en Ingeniería Electrónica. Mem. Electro 2015, Vol. 37, pp 211-215, Chihuahua, Chih. México
http://depi.itchihuahua.edu.mx/display/memorias_electro/MemoriaElectro2015
ISSN 1405-2172

MÉTODO PARA SEGMENTACIÓN DE MOVIMIENTO EN ANÁLISIS DE MARCHA HUMANA EN TIEMPO REAL CON VIDEOS DE ALTA VELOCIDAD

Corral S. Alma D., Ramírez Q. Juan A., Chacón M. Mario I.
Instituto Tecnológico de Chihuahua
División de Estudios de Posgrado e Investigación
Ave. Tecnológico No. 2909, Chihuahua, Chih., México, C.P. 31310
Tel. 52 (614) 2012000 Ext. 112
adcorral@itch.edu.mx, jaramirez@itch.edu.mx, mchacon@itch.edu.mx

RESUMEN.

Análisis de marcha humana mediante video es un tema que ha generado gran interés por sus aplicaciones en robótica, seguridad y terapia médica. Uno de los mayores retos en análisis de marcha es detectar en tiempo real el movimiento de la silueta humana utilizando videos adquiridos con cámaras de alta velocidad. En este artículo se propone SOM Retinotópica Simplificada (SRESOM), un método con capacidad de realizar segmentación en tiempo real de la silueta humana en movimiento. SRESOM utiliza la técnica de substracción de fondo basándose en la red neuronal SOM Retinotópica. SRESOM fue implementada en C++ y tuvo resultados satisfactorios en la segmentación de la silueta humana en movimiento en escenarios controlados y con velocidades de procesamiento que son factibles para aplicaciones en tiempo real con videos adquiridos con cámaras de alta velocidad.

Palabras Clave: segmentación de video, análisis de marcha humana, redes neuronales.

ABSTRACT.

Human gait analysis has been a paramount issue in applications related to robotics, security and medical diagnosis. One of the most important challenges in gait analysis is to detect movement of the human silhouette in videos acquired with high speed cameras. This paper proposes a real-time method to motion segmentation of human silhouette that we called Simplified RESOM (SRESOM). SRESOM uses background subtraction based on Artificial Neural Network (ANN) Retinotopic SOM (RESOM). SRESOM was implemented in C++ achieving accurate results in the segmentation of human motion in controlled scenarios. Furthermore, the frame rate of SRESOM is feasible to real-time applications using videos acquired with high speed cameras.

Keywords: video segmentation, human gait analysis, neural network.

1. INTRODUCCIÓN

Análisis de marcha humana a partir de video, es la detección de los marcadores que describen los movimientos de cadera, rodilla y tobillo durante el ciclo de marcha [1] de una persona. Dicho análisis tiene múltiples aplicaciones, entre las que destacan sistemas de seguridad, robótica y rehabilitación médica [1] [2] y se ha venido realizando en forma no invasiva con una o varias cámaras [1] [4] [5] o con el sensor kinect [2]. La resolución en el tiempo de los marcadores que describen la marcha hace necesaria la detección de movimiento de una persona con cámaras de alta velocidad [6]. Sin embargo, aunque existen una gran cantidad de métodos para el análisis de marcha como flujo óptico [7], substracción de fondo [8] [9] [1] o tracking [10], el problema de realizar una segmentación coherente de movimiento en tiempo real con cámaras de alta velocidad sigue siendo un tema de interés en la literatura.

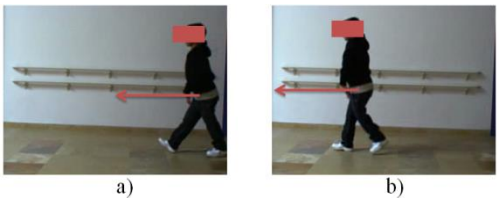
En aplicaciones de tiempo real, las Redes Neuronales Artificiales (RNA), han sido uno de los paradigmas de procesamiento más utilizados por su factibilidad para implementaciones en Hardware [11] y una de las aplicaciones más comunes de las RNA en análisis de video es detección de movimiento en tiempo real [12] [13] [14]. Por lo tanto, para contribuir en el desarrollo de nuevas metodologías para análisis de marcha humana, en este trabajo se propone el método SOM Retinotópica Simplificada (SRESOM), un método para realizar la segmentación de movimiento de personas basado en la SOM Retinotópica (RESOM) [13] [14], una RNA para análisis de video en tiempo real. SRESOM debe ser coherente en la detección de la silueta de la persona en movimiento en un escenario, y lo debe hacer en tiempo real aun con videos adquiridos con cámaras de alta velocidad, dejando tiempo de procesamiento para algoritmos de búsqueda de marcadores para aplicaciones medicas. Adicionalmente, SRESOM forma parte del proyecto "Determinación de movimiento humano mediante visión artificial enfocado al análisis y terapia médica" para apoyar la parte de diagnóstico médico de pacientes con problemas de marcha al caminar.

La organización del artículo es la siguiente: en la sección 2 se describe el método SRESOM, en la sección 3 se muestran los resultados y en la sección 4 se presentan las conclusiones.

2. ESQUEMA DEL MÉTODO SRESOM

El objetivo del método SRESOM es segmentar la silueta humana en movimiento de escenarios controlados, obtenidos a partir de videos utilizados para analizar el ciclo de marcha [1]. En los videos el paciente camina horizontalmente por el

escenario como se muestra en la Figura 1. El paciente puede atravesar una y otra vez el escenario en cualquier dirección.



a) b)

Figura 1. Video de ciclo de marcha en video *Em1.avi*. a). Cuadro 50 de la secuencia de video. b) Cuadro 150 de la secuencia de video.

SRESOM consta de tres módulos: entrada, substracción de fondo y segmentación, tal como se muestra en la Figura 2.

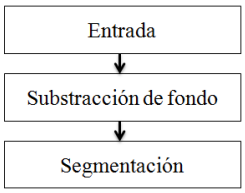


Figura 2. Esquema general de SRESOM.

El módulo de entrada adquiere un cuadro del espacio de color RGB de una secuencia de video. El módulo de substracción de fondo obtiene los pixeles candidatos a pertenecer a objetos en movimiento mediante un modelado de fondo y la diferencia entre éste y el video de entrada. En el módulo de segmentación, se realiza una binarización de pixeles para distinguir los objetos en movimiento identificados. A continuación se describen cada uno de los módulos.

2.1 Módulo de entrada.

En este módulo se adquiere un cuadro $I(x,y,z)^t$ de una secuencia de video obtenida con cámaras estacionarias, donde t es el índice de cuadro, (x,y) representa la posición del píxel y z representa los canales de color. Los cuadros son adquiridos en el espacio de color RGB ($z=\{R,G,B\}$). Cada cuadro $I(x,y,z)^t$ es transformado a $I(k,z)^t$, donde $k = y + Y(x-1)$, Y es el número máximo de columnas. Los videos utilizados fueron adquiridos en escenarios controlados en donde no existen objetos en movimiento en los primeros cuadros y a una taza de 15, 25, 30 y 75 cuadros por segundo (*frames per second, fps*).

2.2 Módulo de substracción de fondo.

En este módulo se realiza una segmentación de la silueta humana en movimiento mediante la técnica substracción de fondo [14], la cual consiste en obtener la diferencia entre el modelado de fondo (objetos del escenario que no están en movimiento) y el video de entrada. Lo que resulta de tal procesamiento es la identificación de los pixeles que contienen información de un objeto en movimiento. Existen muchos métodos para realizar el modelado de fondo [14] dentro de los cuales las redes neuronales artificiales (RNA) han resultado de gran utilidad [11] [13] [15]. El modelado de fondo de SRESOM emplea una RNA llamada RESOM propuesta en [12], que en este trabajo fue modificada eliminando ciertos parámetros para obtener mayor eficiencia en los tiempos de procesamiento en el universo de casos. La arquitectura de la RESOM se muestra en la Figura 3, donde se aprecia que cada canal de color RGB es procesado por una red RESOM, la cual está definida por un conjunto de pesos dados por $\omega(k,m)^{z,t}$, donde m es el índice de neurona. Cada neurona tiene k pesos conectados punto a punto con cada elemento del cuadro de entrada [12]. A partir de reglas de aprendizaje que se definirán en la subsección 2.2.1, las neuronas de RESOM aprenden correctamente la información del fondo y de forma parcial e incompleta los objetos en movimiento.

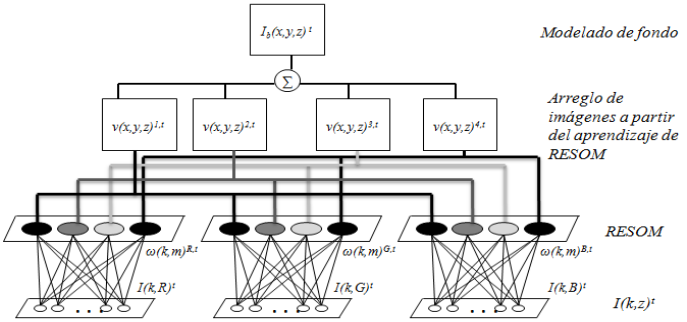
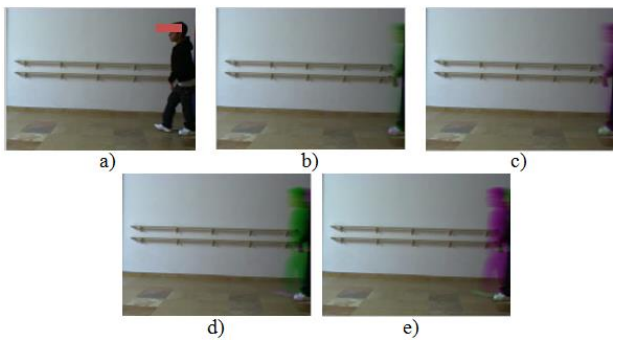


Figura 3. Arquitectura del módulo de modelado de fondo.

La información aprendida por las tres redes RESOM se utiliza para generar un conjunto de imágenes $v(x,y,z)^{m,t}$ donde los pesos $\omega(k,m)^{z,t}$ se transforman a $\omega(x,y,m)^{z,t}$ y posteriormente $v(x,y,z)^{m,t}=\omega(x,y,m)^{z,t}$. Finalmente, el modelado de fondo es promedio dado por:

$$I_b(x,y,z)^t = \frac{1}{M}\sum_{m=1}^{M} v(x,y,z)^{m,t} \qquad (1)$$

Con base en [13] SRESOM tiene cuatro neuronas por red $(m=1,\ldots,4)$. En la Figura 4, se muestra un ejemplo de $v(x,y,z)^{m,t}$ donde se refleja el aprendizaje de las neuronas, las cuales aprenden correctamente el fondo, pero la silueta en movimiento se aprende de manera muy distinta en cada neurona.



a) b) c) d) e)

Figura 4. Respuesta de la RESOM modificada en video *Em1*. a) $I(x,y,z)^{40}$. b) $v(x,y,z)^{1,t}$. c) $v(x,y,z)^{2,t}$. d) $v(x,y,z)^{2,t}$. e) $v(x,y,z)^{4,t}$.

2.2.1 Aprendizaje de la RESOM

El aprendizaje de la RESOM [12] [13], está definido por la regla hebbiana dada por:

$$\omega(k,m)^{z,t+1} = \omega(k,m)^{z,t} + \alpha\left[I(k,z)^t - \omega(k,m)^t \right]\beta(m)^z \qquad (2)$$

El modelo reportado en este trabajo tiene modificaciones en la cantidad de iteraciones por cuadro (SRESOM solo considera una iteración por cada cuadro) y los parámetros hebbianos α y $\beta(m)^z$. α es la función de vecindario asociada al aprendizaje de la red dado por:

$$\alpha^t = \exp\left(-\frac{6t_a}{T_f} \right) \qquad (3)$$

t_a=0 en $t=1$ y para cada cuadro $t_a=t_a+1$, T_f caracteriza la caída exponencial y con base en [12] se determinó que $T_f=30$. El parámetro $\beta(m)^z$ es la función de vecindario asociada a la capacidad de aprendizaje de cada neurona dentro de la RESOM y está dada por:

$$\beta(m)^z = \exp\left(-\frac{\left(m_w^z - m\right)^2}{\left(\sigma_\beta^t\right)^2} \right) \qquad (4)$$

donde σ_β^t es el radio de vecindario dado por:

$$\sigma_\beta^t = 13.5\exp\left(-\frac{5t_a}{T_f} \right) \qquad (5)$$

y m_w^z es el índice de la neurona ganadora de cada canal z de color que se obtiene a partir de la menor magnitud de $\eta(m)^z$ dado por:

$$\eta(m)^z = \sum_k \left| I(k,z)^t - \omega(k,m)^{z,t} \right| \qquad (6)$$

donde m_w^z es el valor del índice m de min($\eta(m)^z$) de cada red. En el modelo original en [12][13], α^t y σ_β^t realizan una serie de cálculos para definir el aprendizaje de la RESOM con base en condiciones de escenarios complejos y videos de tamaños indefinidos. Sin embargo, los escenarios de este trabajo tienen un fondo que cambia muy poco al trascurrir los cuadros y los videos tienen una duración finita. Por lo tanto, se modificó el modelo original simplificando α^t y σ_β^t con un comportamiento exponencial decreciente que de acuerdo al análisis basado en [13], en $t_a=0$, $I_b(x,y,z)^t$ va a aprender el primer cuadro, luego con la caída exponencial de α^t y σ_β^t las neuronas de la RESOM van a aprender los cambios que suceden en el fondo pero sin que $I_b(x,y,z)^t$ aprenda los objetos en movimiento.

2.2.2 Distancia euclidiana

Luego de modelar el fondo, se obtiene la diferencia entre éste y el video de entrada mediante la distancia euclidiana definida por:

$$I_e(x,y)^t = \left\| I(x,y,z)^t - I_b(x,y,z)^t \right\|_2 \qquad (7)$$

En la Figura 5 se puede observar el resultado del modelado de fondo basado en RESOM y el resultado de $I_e(x,y)^t$.

a) b) c)

Figura 5. Resultados del módulo de substracción de fondo de video *Em1.avi*. a) $I(x,y,z)^t$ en $t=40$. b) $I_b(x,y,z)^t$. c) $I_e(x,y)^t$.

Otra función de la distancia euclidiana es determinar si existen cambios considerables en el fondo, los cuales pueden ser causados por diversos motivos como cambios repentinos de iluminación. De acuerdo a las condiciones establecidas en las terapias, la silueta humana no ocupa más del 15% del área de la imagen, por lo tanto se establece la siguiente regla:

$$t_a = \begin{cases} 0 & MI_e > 0.15 \\ t_a + 1 & otra \quad forma \end{cases} \qquad (8)$$

donde MI_e es el promedio de los niveles de grises de $I_e(x,y)^t$. Mediante la ecuación (8), se reinicia el valor de t_a que define el comportamiento de α^t y σ_β^t .

2.3 Módulo de segmentación.

En este módulo se procesa la substracción de fondo $I_e(x,y)^t$ con el objetivo de realizar la segmentación del objeto en movimiento de la silueta humana. Para ello, se realiza la separación de la información de $I_e(x,y)^t$ en dos clases: fondo y objeto dinámico (silueta humana en movimiento). Para esto, se realiza una segmentación por umbral dada por:

$$I_{din}(x,y)^t = \begin{cases} 1 & si \quad I_e(x,y)^t > Th_1 \\ 0 & otra \quad forma \end{cases} \qquad (9)$$

donde Th_1 es el umbral que separa el objeto dinámico del fondo. La selección de Th_1 depende de la cantidad de ruido presente en $I_e(x,y)^t$. De acuerdo a las diversas pruebas de videos con escenarios de aulas para terapia que se encuentran en condiciones controladas, se estableció que cualquier umbral Th_1 con valor entre 0.1 y 0.2 genera en $I_{din}(x,y)^t$ una segmentación coherente de la silueta humana. Sin embargo, con estos valores de Th_1 se incrementa la probabilidad de tener ruido en los resultados de la segmentación por lo que el umbral Th_1 será adaptivo pretendiendo obtener la mejor definición de la silueta, con el mínimo ruido posible. Para definir el valor adaptivo de Th_1, primero se realiza una separación preliminar fondo-objeto dinámico con el mínimo valor de umbral obtenido en los videos de los escenarios de aulas, dada por:

$$I_u(x,y)^t = \begin{cases} 1 & si \quad I_e(x,y)^t > 0.1 \\ 0 & otra \quad forma \end{cases} \qquad (10)$$

Luego, se calcula la media μI_u de $I_u(x,y)^t$ para estimar la cantidad de pixeles con valor de uno que son los que corresponden al ruido y al objeto dinámico, entre el total de píxeles del cuadro. Debido a que la silueta de los pacientes no genera un área mayor al 15% de la imagen, se asume que si μI_u

es mayor 0.15, entonces existe ruido en los resultados de $I_u(x,y)^t$. Por lo tanto, μI_u puede ser utilizado como métrica para el ruido que existe en $I_e(x,y)^t$ y para definir Th_1 con μI_u cuando $\mu I_u > 0.15$, de la siguiente manera:

$$Th_1 = \begin{cases} 0.15 & si \quad \mu I_u < 0.15 \\ \mu I_u & otra \quad forma \end{cases} \tag{11}$$

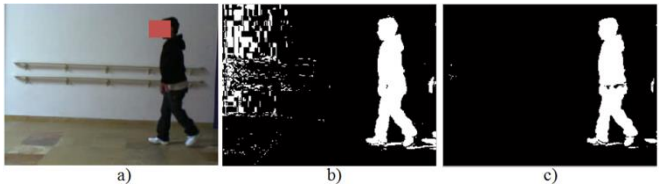


Figura 6. Resultados del módulo de segmentación en *t=67 en video Em1.avi*. a) $I(x,y,z)^t$. b) $I_u(x,y)^t$. c) $I_{din}(x,y)^t$.

3. RESULTADOS

La implementación de SRESOM se realizó en C++ utilizando la plataforma de desarrollo Visual C++ 2012 con el paquete de OpenCV 3.0 para Windows. Se realizaron pruebas con 22 videos para analizar el ciclo de marcha adquiridos a 15, 25, 30 y 75 *fps*. SRESOM generó resultados coherentes respecto a la silueta humana en movimiento, ya que son comparables con otros métodos de segmentación aplicados al análisis de marcha [16]. En la Figura 7 se pueden ver cuatro ejemplos con distintos tipos de escenarios: condiciones normales, sombras, problemas de camuflaje y con fondo distinto.

Las pruebas de velocidad de SRESOM se realizaron en una computadora con procesador Intel® Core™ i5 de 2.5 GHz, con 4 GB de RAM y Windows 7 de 64 bits. La Tabla I reporta los resultados de tiempo y velocidad de procesamiento de SRESOM con videos de 268 cuadros y con diferentes resoluciones. Para cada resolución se realizaron ocho mediciones de tiempos por cada video y en todas ellas se obtuvieron resultados factibles para aplicaciones en tiempo real. Con base en los datos de la tabla I, se puede asumir que en resoluciones menores o iguales a 480x640 se puede utilizar una cámara que adquiera a más de 30 *fps*. De acuerdo a [16], una resolución de 240x320 es suficiente para el análisis de marcha y como se puede ver en la tabla I, a esta resolución se pueden utilizar cámaras que adquieren a 500 *fps*.

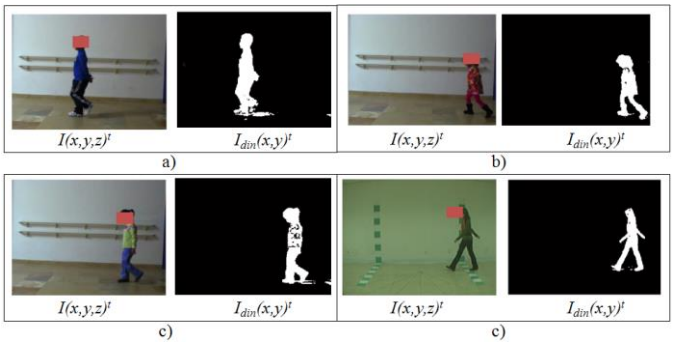


Figura 7. Resultados de SRESOM. a) Video *Em5.avi*, el cual tiene problemas de sombras. b) Video *Em10.avi* que se encuentra en condiciones normales. c) Video *Em15.avi* con problemas de camuflaje entre la ropa del paciente y el fondo. c) Video *107-nm-06-090.avi* con de la base de datos CASIA.

En la tabla I se puede observar que las velocidades de procesamiento en resoluciones menores 240x320 son mayores a 1000 *fps*, lo cual define a SRESOM como el método más rápido para detectar movimiento en la literatura analizada, ya que algunos métodos exitosos como SOBS [11], SOM-CNN [16], BNN [17] o el método propuesto en [1] tienen velocidades de procesamiento menores a 1000 *fps* en videos con resoluciones de 120x160. Se puede observar en la tabla I que al aumentar la resolución espacial de $I(x,y,z)^t$ la velocidad de procesamiento va cayendo de tal manera que existe un punto de inflexión entre las resoluciones de 240x320 y 360x480. En pruebas realizadas con resoluciones distintas a las de las reportadas en la tabla I, se pudo observar que existe un punto de inflexión entre las resoluciones de 255x340 y 270x360, ya que las velocidades de procesamiento varían de 1500 a 80 *fps* en dichas resoluciones.

4. CONCLUSIONES

En este artículo se presentó SRESOM, un método de segmentación de movimiento de la silueta humana para aplicaciones de análisis de marcha. Este método se basa en una variante de la red neuronal RESOM, la cual fue diseñada para análisis de video en tiempo real. Los resultados de SRESOM mostraron un desempeño satisfactorio en sus tareas de segmentación en escenarios controlados. Además, de acuerdo a las pruebas, SRESOM tiene velocidades de procesamiento que lo hacen uno de los métodos más rápidos en la literatura, ya que

Tabla I. Tiempos y velocidades de procesamiento de SRESOM.

Resolución	**Velocidad de procesamiento en *fps***								**Promedio *fps***	**Tiempo promedio**
240x320	2436	2851	3435	2851	2677	2284	2008	2851	2674.125	373 µs
360x480	61	62	62	62	62	61	62	61	61.625	16 ms
480x640	30	31	31	31	31	31	31	31	30.875	32 ms
600x800	15	19	20	20	16	16	20	20	18.25	54 ms

para resoluciones de 240x320 llega a tener velocidades mayores a 1000 *fps*, mientras que otros métodos tienen velocidades de procesamiento menores a 1000 *fps* con resoluciones de 120x160. En conclusión, SRESOM es un método que genera buenos resultados en la segmentación de movimiento de la silueta humana en escenarios controlados y con videos adquiridos con cámaras de alta velocidad.

Como trabajo futuro, se pretende implementar SRESOM en un prototipo de terapia para análisis de marcha humana que forma parte del proyecto FOMIX "Determinación de movimiento humano mediante visión artificial enfocado al análisis y terapia médica". Dicho prototipo será utilizado por los Centros de Atención Múltiple (CAM) del estado de Chihuahua.

Agradecimientos

Los autores agradecen al Fondo Mixto de Fomento a la Investigación Científica y Tecnológica CONACYT-Gobierno del Estado de Chihuahua y al Tecnológico Nacional de México bajo los apoyos -2012-C03-193760 y CHI-MCIET-2013-230.

Referencias.

[1] Mario I Chacon-Murguia, Rafael Sandoval-Rodriguez y Omar Arias-Enriquez, "Human Gait Feature Extraction Including a Kinematic Analysis Toward Robotic Power Assistance," International journal of advanced robotic systems, vol. 9, no. 68, 2012, pp. 1-9,

[2] Baiqing Sun, Xiaogang Liu, Xuetang Wu y Haiyang Wang, "Human Gait Modeling and Gait Analysis Based on Kinect," Int conf on robotics & automation, Hong Kong, 2014, pp. 3173-3178.

[3] Vartiainen P, Bragge T, Arokoski JP y Karjalainen PA, "Nonlinear State-Space Modeling of Human Motion Using 2-D Marker Observations," IEEE Transactions on Biomedical Engineering, vol. 61, no. 7, julio 2014, pp. 2167-2178.

[4] Haitao Liu, Yang Cao y Zengfu Wang, "A novel algorithm of gait recognition," Int conf on wireless communications & signal processing, Nanjing, noviembre 2009, pp. 1-5.

[5] Tao Liu, Inoue Y, Shibata K y Morioka H, "Development of wearable sensor combinations for human lower extremity motion analysis," Int conf on robotics and automation, Orlando, 2006, pp. 1655-1660.

[6] Maodi Hu, Yunhong Wang, Zhaoxiang Zhang, De Zhang y Little JJ, "Incremental Learning for Video-Based Gait Recognition With LBP Flow," IEEE Transactions on Cybernetics, vol. 43, no. 1, junio 2012, pp. 77-89.

[7] Shelke PB y Deshmukh PR, "Gait Based Gender Identification Approach," fifth Int conf on advanced computing & communication technologies , Haryana, 2015, pp. 121-124.

[8] Gaba I y Ahuja SP, "Gait analysis for identification by using BPNN with LDA and MDA techniques," Int conf on Innovation and Technology in Education, Patiala, 2014, pp. 122-127.

[9] Cheng Yang et al., "Multiple marker tracking in a single-camera system for gait analysis," Int conf on image processing , Melbourne, 2013, pp. 3128-3131.

[10] Janardan Misra and Indranil Saha, "Artificial neural networks in hardware: A survey of two decades of progress," Neurocomputing, vol. 74, no. 1-3, diciembre 2010, pp. 239-255.

[11] Lucia Maddalena y Alfredo Petrosino, "A Self-Organizing Approach to Background Subtraction for Visual Surveillance Applications," IEEE Transactions on Image Processing, vol. 17, no. 7, julio 2008, pp. 1168-1177.

[12] Juan A Ramirez-Quintana and Mario I Chacon-Murguia, "Self-Organizing Retinotopic Maps Applied to Background Modeling for Dynamic Object Segmentation in Video Sequences," Int joint conference on neural networks, Dallas, 2013, pp. 1-8.

[13] Juan A Ramirez-Quintana and Mario I Chacon-Murguia, "An Adaptive Unsupervised Neural Network based on Perceptual Mechanism for Dynamic Object Detection in Videos with Real Scenarios," Neural Processing Letters, vol. Publicado en línea, agosto 2014, pp. 1-25.

[14] Sebastian Brutzer, Benjamin Höferlin, and Gunther Heidemann, "Evaluation of Background Subtraction Techniques for Video Surveillance," Int conf on Computer Vision and Pattern Recognition, 2011, pp. 1937-1944.

[15] Graciela Ramirez-Alonso and Mario I Chacon-Murguia, "Object detection in video sequences by a temporal modular self-adaptive SOM," Neural Computing and Applications, marzo 2015.

[16] Omar Arias Enriquez, "Análisis de la marcha humana basada en percepción visual 2D/Kinect y su diagnostico utilizando sistemas difusos," Instituto Tecnológico de Chihuahua, Chihuahua, Tesis de maestría 2012.

[17] Juan Alberto Ramirez-Quintana and Mario Ignacio Chacon-Murguia, "Self-adaptive SOM-CNN neural system for dynamic object detection in normal and complex scenarios," Pattern Recognition, vol. 48, no. 1, abril 2015, pp. 1137–1149.

[18] Dubravko Culibrk, Oge Marques, Daniel Socek, Hari Kalva, and Borko Furht, "Neural Network Approach to Background Modeling for Video Object Segmentation," IEEE Transactions on Neural Networks, vol. 18, no. 6, Nov 2007, pp. 1614-1627.