SISTEMA DE RECONOCIMIENTO DE VOCALES DE LA LENGUA DE SEÑAS MEXICANA

Eliúh Cuecuecha Hernández

Benemérita Universidad Autónoma de Puebla

eliuhcueh@gmail.com

José Javier Martínez Orozco

Benemérita Universidad Autónoma de Puebla

javier.35.93.01@gmail.com

Daniel Méndez Lozada

Benemérita Universidad Autónoma de Puebla

daniel.mendezl@alumno.buap.mx

Adán Zambrano Saucedo

Benemérita Universidad Autónoma de Puebla

zambranos.adan@gmail.com

Aldrin Barreto Flores

Benemérita Universidad Autónoma de Puebla

aldrin.barreto@correo.buap.mx

Verónica Edith Bautista López

Benemérita Universidad Autónoma de Puebla

vbautista@cs.buap.mx

Salvador Eugenio Ayala Raggi

Benemérita Universidad Autónoma de Puebla

saraggi@ece.buap.mx

Resumen

La lengua de señas es un medio de comunicación tan importante como la lengua hablada para el desarrollo social del ser humano. Tras la aparición de

sensores de reconocimiento de gestos, como *Kinect*, surge especial interés por utilizarlos para interpretar la lengua de señas.

La finalidad del presente trabajo consistió en interpretar las vocales de la lengua de señas mexicana identificadas por gestos estáticos con la mano. Para ello se utilizó el sensor *Leap Motion Controller*, ideal para esta aplicación al detectar y seguir las manos con tal precisión sin necesidad de entrar en contacto con el usuario. Para lograr el reconocimiento de los gestos correspondientes a las vocales se utilizó el modelo de perceptrón multicapa junto a una interfaz visual en tiempo real. La red fue entrenada y calibrada por un experto en lenguaje de señas, logrando así una razón de reconocimiento de hasta 100%.

Palabras Claves: Leap Motion Controller, Lengua de Señas Mexicana (LSM), perceptrón multicapa, reconocimiento de Imágenes.

Abstract

The Sign Language is a communication mean as important as the speaking language, for the social development of the human being. Thus, after gesture recognition sensors emerged such as Kinect, interest arises for utilize them to interpret the Sign language.

The present work aims to interpret the Mexican Sign Language vowels identified by one-hand static gestures. In order to do so, the Leap Motion Controller was used, ideal for this application while detecting and tracking hand position with such accuracy, without the need to physically interact with the user.

To achieve the vowels' gesture recognition, the multilayer perceptron model was used along with a real time visual interface. The network was trained and calibrated by a sign language expert, achieving a recognition ratio up to 100%.

Keywords**:** *Image Recognition, leap motion controller, Mexican sign language, multilayer perceptron model.*

1. Introducción

El lenguaje de señas es un medio de comunicación para gente con problemas de audición y de habla. Este lenguaje usa información visual mediante

movimientos con los dedos, la mano y el brazo. El lenguaje de señas mexicano tiene el alcance de representar las 27 letras del alfabeto con una sola mano y así proveer un medio de comunicación de ayuda a personas con discapacidad del habla o audición [Serafín, 2011].

Sin embargo, interpretar el lenguaje de señas resulta complicado para la gran mayoría de personas no relacionadas con esta lengua. Por lo tanto, ante la imposibilidad de que las personas con discapacidad auditiva utilicen la palabra hablada, se plantea desarrollar un sistema confiable y fácil de usar para el reconocimiento de la Lengua de Señas Mexicana, y así proveer una plataforma de interfaz natural que facilite la comunicación. Más aún, esta plataforma también sirve para facilitar el proceso enseñanza-aprendizaje de la lengua de señas.

Múltiples trabajos de investigación se han llevado a cabo con el *Leap Motion Controller* para desarrollar aplicaciones de traducción del lenguaje de señas. En particular, [Naglot, 2015] presenta un notable trabajo con razón de reconocimiento de 96.15%. Aquí se plantea un sistema de reconocimiento en tiempo real del lenguaje de señas americano (ASL). Para ello, se procesan 520 muestras de letras del alfabeto mediante una red neuronal. Los datos de entrada comprenden distancias euclidianas entre diferentes partes de la mano, y a la salida se obtienen activaciones correspondientes a las 26 letras del alfabeto a identificar.

En [Fok et al., 2015] se muestra que utilizando más sensores es posible mejorar el reconocimiento de la Lengua de Señas. En este trabajo se usan dos sensores que en conjunto determinan la posición estática de la mano correspondiente a los 9 dígitos del ASL. La información se procesa mediante el Modelo Oculto de Markov (HMM) y muestra un mínimo de 68.78% de reconocimiento para un sólo sensor y un mínimo de 84.68% utilizando dos sensores. El experimento demuestra que es posible desarrollar un sistema de reconocimiento robusto utilizando más sensores.

Por otra parte, en [Chuan, 2014] se muestra el uso de Machine Learning para el reconocimiento del alfabeto del ASL. No obstante, en este estudio la máxima razón de reconocimiento es de 72.78% y la aplicación del sistema está pensada en auxiliar a instructores en la enseñanza de la Lengua de Señas. Aquí se plantea, además, incorporar una cámara web para mejorar el reconocimiento de gestos.

De los trabajos anteriores resulta que la mayor razón de reconocimiento se obtiene con una red neuronal de perceptrón multicapa. Por esta razón, en el presente trabajo se decidió utilizar este modelo en conjunto con el *Leap Motion Controller* para el reconocimiento de vocales de la Lengua de Señas Mexicana (LSM), en vez del ASL como la mayoría de trabajos plantean.

2. Métodos

El Leap Motion Controller es un pequeño dispositivo USB diseñado para que el usuario pueda manejar una computadora mediante gestos manuales. El dispositivo, mostrado en la figura 1, consiste en un sensor de movimiento en 3D capaz de detectar la trayectoria de las manos, dedos y objetos similares, reportando su posición y movimiento.

Figura 1 Detección del Leap Motion Controller [Leap Motion, 2017].

La figura 2 muestra las principales características que se pueden obtener a través del ambiente de programación de *Leap Motion Controller*. Entre estas se encuentran:

- Mano: Este modelo entrega información sobre el tipo (derecha, izquierda, ambas manos), posición (la posición central de la palma en milímetros), velocidad de movimiento, etc.
- Dedos: Características relacionadas con dirección (descrita por un vector unitario), longitud del dedo (en milímetros), ancho, posición de la punta, posición de la falange proximal (primera falange), posición de la falange distal (última falange) y la posición del metacarpo.

- Gestos: Algunos patrones de movimiento son reconocidos por el sensor, por ejemplo, el trazado de un círculo (*circle*) con un dedo, o el movimiento lineal que se interpreta como un deslizamiento (*swipe*), o el gesto de teclear (*key tap*).

Figura 2 Partes de la mano identificadas por Leap Motion Controller [Davis, 2017].

Gracias a las múltiples variables de reconocimiento del *Leap Motion*, varios trabajos relacionados a identificación y clasificación se han llevado a cabo con este dispositivo [Erdogan et al., 2016], [Funasaka, 2015] y [Nájera et al., 2016].

De especial interés son las posiciones de los dedos (*tip position*), pues no importando la orientación de la mano, la representación de vocales en el LSM depende únicamente de la configuración espacial de los dedos. Por esta razón se decidió tomar como datos a ciertas distancias clave, pues resultan ser suficientes para poder discernir entre las distintas vocales. El *Leap Motion* calcula con gran precisión las puntas de los dedos gracias a los algoritmos de reconocimiento implementados en este dispositivo. Algunos algoritmos remarcables se pueden consultar en [Langford et al., 2016] y [Priego, 2012].

La información proporcionada a la red consiste entonces en los siguientes datos:

- Las distancias euclidianas entre la posición de la punta de cada dedo y la posición de la palma de la mano (5 datos).
- Las distancias euclidianas entre la posición de la punta de cada dedo, de manera consecutiva (4 datos).

El conjunto de datos recabados consta de 100 muestras (20 para cada vocal), tomadas de 4 autores del proyecto y un intérprete traductor experto en lenguaje de señas, trabajador del DIF municipal de la ciudad de Puebla. La mitad de estas muestras sirve como conjunto de aprendizaje y la otra mitad como conjunto de validación. El conjunto de datos es normalizado antes de ser utilizado para el entrenamiento utilizando ecuación 1.

$$N(x) = \frac{1}{max_x - min_x}(x - min_x) \quad (1)$$

La tabla 1 muestra los datos obtenidos para el gesto correspondiente a la vocal *A*. Estos datos corresponden a las distancias clave correspondientes de la mano, expresadas en milímetros. Este conjunto de datos junto con el del resto de las vocales se deben normalizar para poder ingresar a la red mediante la función dada por la ecuación 1. Los valores máximos y mínimos se obtienen del conjunto de datos de todas las vocales.

Tabla 1 Datos de posición para la vocal *A* obtenidos con el Leap Motion Controller.

Patrón	**Palma Pulgar**	**Palma Índice**	**Palma Medio**	**Palma Anular**	**Palma Menique**	**Pulgar Índice**	**Índice Medio**	**Medio Anular**	**Anular Menique**
1	166.09	49.90	49.34	51.70	56.16	153.66	18.23	29.44	22.97
2	178.47	49.08	48.56	50.98	50.27	160.31	12.06	27.47	23.64
3	158.74	51.75	50.48	55.92	57.39	144.52	17.14	30.19	25.18
4	168.79	48.01	45.13	46.14	56.65	149.69	18.30	31.94	26.78
5	185.49	56.70	46.65	43.74	44.66	149.92	24.88	27.88	22.25
6	168.83	48.31	48.15	50.57	54.36	154.29	15.29	31.16	24.12
7	173.08	45.66	41.89	42.90	51.40	151.38	16.72	30.20	26.52
8	176.20	46.25	44.54	48.98	52.84	156.47	14.68	29.65	26.39
9	179.37	47.98	43.58	39.42	46.78	150.09	13.96	30.42	28.41
10	182.66	49.65	50.69	49.17	49.90	167.71	13.53	27.67	22.65
11	177.26	53.06	56.44	57.40	56.08	172.62	13.41	26.08	22.24
12	174.21	46.17	44.75	46.54	50.00	155.42	15.88	27.58	24.43
13	183.74	48.48	49.08	48.58	49.24	167.35	14.39	27.88	22.73
14	184.56	50.75	51.90	50.07	49.81	166.95	14.87	28.02	23.19
15	183.15	59.02	58.12	55.29	50.43	162.69	15.58	26.24	24.12
16	175.37	57.01	53.00	47.51	46.96	146.45	17.36	29.12	25.11
17	168.78	56.17	50.80	45.62	43.56	140.81	18.88	25.47	25.57
18	169.64	58.02	49.56	41.05	40.71	133.41	19.41	27.32	27.39
19	173.22	63.04	56.32	47.71	44.56	137.15	19.95	29.86	25.77
20	149.15	53.75	34.61	27.31	44.42	103.92	25.29	32.25	31.15

Para la clasificación se usa una red neuronal de perceptrón multicapa cuyo aprendizaje es supervisado y cuyo modelo consiste en prealimentar la red (*feedforward*). Esto es, se usan las características del conjunto de entrada y se obtiene una salida particular para la vocal que ha sido introducida. Para el entrenamiento se recurrió al uso del algoritmo de retropropagación de los errores, observando la evolución del error variando la razón de aprendizaje. Este consiste en introducir un conjunto inicial a la red y asignar valores aleatorios a cada uno de los pesos. Con base en estos pesos, se obtiene cada una de las salidas en las diferentes capas hasta llegar a la capa de salida, calculando así el error obtenido a la salida mediante la ecuación 2 y propagando este valor hacia cada una de las capas anteriores (ocultas y de entrada), modificando así los valores de los pesos de cada una de las interconexiones correspondientes mediante las ecuaciones 3, 4, 5 y 6 utilizando la función de activación sigmoidal.

Cálculo del error a la salida para un patrón n siendo $y(n)$ las salidas de la red y $s(n)$ las salidas deseadas, ecuación 2.

$$e(n) = \frac{1}{2}\sum_{i=1}^{n_c}(s_i(n) - y_i(n))^2 \quad (2)$$

Cálculo de los pesos capa oculta $C-1$ a la capa de salida C, ecuación 3.

$$w_{ji}^{C-1}(n) = w_{ji}^{C-1}(n-1) + \alpha\delta_i^C(n)\alpha_j^{C-1}(n) \quad (3)$$

Para $j = 1,2,\dots,n_{C-1}$ y $i = 1,2,\dots,n_C$

Cálculo de umbrales de la capa de salida, ecuación 4.

$$u_i^C(n) = u_i^C(n-1) + \alpha\delta_i^C(n) \quad (4)$$

Para $i = 1,2,\dots,n_C$

Con:

$$\delta_i^C(n) = -(s_i(n) - y_i(n))y_i(n)(1 - y_i(n))$$

Cálculo de los pesos de la capa c a la capa $c+1$, ecuación 5.

$$w_{kj}^c(n) = w_{kj}^c(n-1) + \alpha\delta_j^{c+1}(n)\alpha_k^c(n) \quad (5)$$

Para $k = 1,2,\dots,n_c$, $j = 1,2,\dots,n_{c+1}$ y $c = 1,2,\dots,C-2$

Cálculo de umbrales de las neuronas de la capa $c+1$ para $c = 1,2,,\ldots,C-2$, ecuación 6.

$$u_j^{c+1}(n) = u_j^{c+1}(n-1) + \alpha\delta_j^{c+1}(n) \quad (6)$$

Para $j = 1,2,\ldots,n_{c+1}$ y $c = 1,2,\ldots,C-2$

Con:

$$\delta_j^{c+1}(n) = a_j^c(n)(1-a_j^c(n))\sum_{i=1}^{n_{c+1}}\delta_i^{c+2}(n)\omega_{ji}^c$$

La red neuronal artificial en cuestión consta de 3 capas: una capa de entrada, una capa oculta y una capa de salida, i.e. $C = 3$. Las neuronas de la capa de entrada se conectan a las neuronas de la capa oculta y éstas a la capa de salida, lo que lleva a la interconexión de los pesos en la red.

El tamaño de la capa de entrada depende de los datos que serán introducidos a la red, en este caso 9, es decir 9 neuronas en la capa de entrada ($n_1 = 9$). De manera similar para la capa de salida se tienen 5 posibilidades (5 vocales) por lo tanto 5 neuronas en la capa de salida ($n_3 = 9$). Para la capa oculta no existe una regla establecida que calcule el número de neuronas a usar, sin embargo, debido a la gran tasa de reconocimiento alcanzada en [Naglot, 2015] parece prudente utilizar la regla empírica expuesta allí: se suma el número de neuronas de la capa de entrada más el número de neuronas de la capa de salida y este resultado se divide entre dos. Es decir, $n_2 = \frac{5+9}{2} = 7$. La arquitectura de la red neuronal queda entonces cómo se muestra en la figura 3.

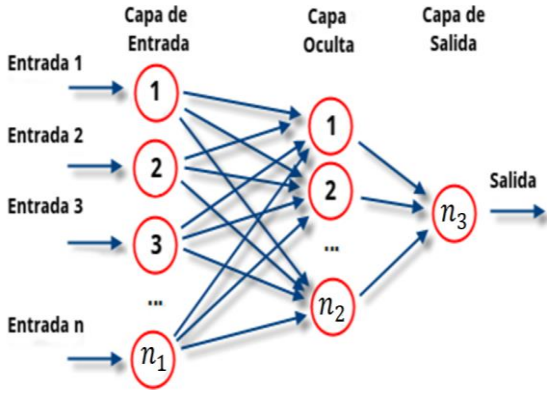
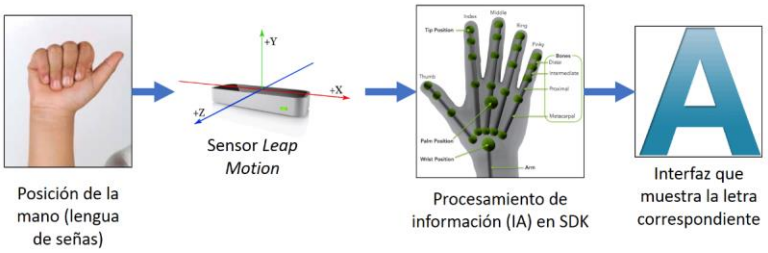


Figura 3 Modelo del perceptrón multicapa utilizado

3. Resultados

El sistema propuesto fue implementado en lenguaje Java. La figura 4 muestra un diagrama con el proceso utilizado para reconocer la Lengua de Señas.



[Chuan, 2014], [Langford et al., 2016], [Serafín, 2011]

Figura 4 Diagrama a bloques del flujo de información.

En particular, el programa de computadora desarrollado pretende ser fácil de usar. Para ello se muestra en pantalla una representación espacial de la mano vista por el *Leap Motion* (figura 5), con la finalidad de que el usuario evalúe si el reconocimiento del sensor es adecuado. Luego, para auxiliar en la realización del gesto, se muestra una imagen con la posición de la mano a realizar para cada vocal. Finalmente, a un lado se muestran las vocales que cambian de color azul a rojo, en dónde la letra más roja es la vocal identificada por el programa. El reconocimiento de la vocal se complementa con la reproducción del sonido correspondiente.

Figura 5 Interfaz del programa.

Cómo se mencionó anteriormente, del conjunto de datos obtenidos, se clasificaron las muestras en dos subconjuntos: entrenamiento y prueba. Durante el proceso de

entrenamiento, se realiza un super ajuste sobre dicho conjunto, concluyendo el proceso cuando se alcanza el primer mínimo de la función del error de validación. Cabe mencionar que para la estimación del error no se utilizó ni un método como validación cruzada o hold-out. El conjunto de entrenamiento se utilizó para determinar los parámetros del clasificador neuronal, mientras que el conjunto de prueba fue utilizado para estimar el error de generalización.

Una herramienta que permite visualizar el desempeño de un algoritmo que usa aprendizaje supervisado, como la red propuesta en este caso, es una matriz de confusión. La tabla 2 muestra una matriz de confusión para dos clases donde VP significa verdadero positivo, FP para falso positivo, VN verdadero negativo y FN falso negativo.

Tabla 2 Matriz de confusión para una red de dos clases.

		Valor predicho		
	Valor	**Positivo**	**Negativo**	**TOTAL**
Valor real	**Positivo**	VP	FN	P
	Negativo	FP	VN	N
	TOTAL	P’	N’	P+N

En esta matriz cada columna representa el número de predicciones de cada clase, mientras que la fila representa las clases reales. El principal beneficio de esta herramienta es que permite ver si la red está confundiendo dos clases.

La exactitud del sistema es el porcentaje de muestras que fueron correctamente clasificadas por la red. La precisión mide el grado de certitud del sistema, es decir, el error presente al momento de clasificar muestras de una sola clase, ecuación 7 y ecuación 8.

$$exactitud = \frac{VP+VN}{P+N} 100\% \quad (7)$$

$$precisión = \frac{VP}{VP+FP} 100\% \quad (8)$$

Para este caso en particular, las señas que representan a las vocales el LSM, no presentan un parecido notable entre ellas. Por esta razón es que para exactitud y precisión se tienen valores de 100%. Sin embargo, para mayor número de letras a identificar, estas razones disminuirán notablemente, tabla 3.

Tabla 3 Matriz de confusión del sistema propuesto.

		Valor predicho								
		A	**E**	**I**	**O**	**U**	**TOTAL**	**VP**	**FP**	**Precisión**
Valor real	**A**	**10**	0	0	0	0	10	10	0	100%
	E	0	**10**	0	0	0	10	10	0	100%
	I	0	0	**10**	0	0	10	10	0	100%
	O	0	0	0	**10**	0	10	10	0	100%
	U	0	0	0	0	**10**	10	10	0	100%
	TOTAL	10	10	10	10	10	**50**	50	0	100%
	Razón de Reconocimiento = 100%									

4. Discusión

Como se estableció anteriormente, la exactitud y precisión del sistema resulta ser de 100% para ambos casos. Esto es, con base en la arquitectura diseñada del sistema, la cual busca identificar las vocales del LSM, la identificación de cada vocal presenta un error prácticamente nulo, lo cual dice que no existe confusión del sistema para clasificar las entradas.

Con base en trabajos similares sobre reconocimiento de lenguaje de señas, se observó que el caso de mayor razón de reconocimiento se da mediante el uso de una metodología similar: una red neuronal de arquitectura perceptrón multicapa. Cabe mencionar que dicha razón es muy alta, y por ende el sistema es muy eficiente.

Finalmente, el sistema tratado en el presente trabajo obtuvo una razón de reconocimiento de hasta el 100%. Esto se debió a diferentes factores como lo son: un número preciso y calculado de capas ocultas, validación y calibración adecuada de los patrones de entrada, capacidad de generalización alta, entre otros. Además, la representación en lenguaje de señas de las vocales en el LSM disminuye la malinterpretación de estas ya que no presentan similitudes mayores entre ellas, a diferencia de otras letras del alfabeto. En conjunto, estos factores propiciaron una eficacia casi absoluta del sistema, concluyendo que el uso de redes neuronales, en específico una arquitectura perceptrón multicapa, da resultados más precisos y exactos.

A diferencia de trabajos anteriores donde se utilizan otros métodos cuyos resultados, si bien son completamente funcionales, no se asemejan a los obtenidos en este trabajo. Además, la importancia de este trabajo radica en que el

reconocimiento se basa en el lenguaje de señas mexicano, el cual no ha sido tratado en gran detalle en otros trabajos, ya que la gran mayoría se basan en el lenguaje de señas americano.

Los resultados obtenidos dan pie a que se utilicen redes neuronales para la solución de problemas que requieran de un aprendizaje continuo del sistema basado en ciertos resultados deseados, y que sean aplicables en diferentes áreas.

5. Conclusiones

El reconocimiento de lenguaje de señas mexicano mediante el sensor *Leap Motion* brinda una interfaz amigable y práctica para poder interpretar el lenguaje de personas con discapacidad auditiva y verbal, y de esta manera darles un método de comunicación que simule una conversación hablada.

En conjunto con el sensor, la arquitectura de perceptrón multicapa fue esencial para la obtención de resultados. La eficacia de este método radica en el cálculo preciso de número de capas ocultas, así como en el uso del método *backpropagation* para su entrenamiento, propiciando que el sistema obtenga una capacidad de generalización alta. Esto se refleja en la facilidad de relacionar la entrada con la salida adecuada, a pesar de algunas similitudes que puedan presentar distintas entradas.

Gracias a la razón de reconocimiento obtenida, se comprueba que el *sensor Leap Motion* resulta ideal para el tipo de aplicación, debido a su facilidad de uso y precisión de reconocimiento 3D de la mano.

Algunas aplicaciones en las que se puede implementar dicho trabajo son: reconocimiento de enfermedades en la mano, fisioterapia personalizada de la mano [Santos et al., 2015], plataforma de voz para personas sordomudas, corrección de posturas, enseñanza de lenguaje de señas, entre otras.

6. Bibliografía y Referencias

[1] Chuan, C. H. y Guardino, C., American Sign Language Recognition Using Leap Motion Sensor, International Conference on Machine Learning and Applications, 2014.

[2] Davis A., Getting Started with the Leap Motion SDK. Leap Motion. http://blog.leapmotion.com/getting-started-leap-motion-sdk/.

[3] Erdogan, K., Durdu, A., y Yilmaz, N., Intention Recognition Using Leap Motion Controller and Artificial Neural Networks, CoDIT, Malta, 2016.

[4] Fok, K. Y., Ganganath, N., Cheng, C. T., y Tse, C. K. A Real-Time Recognition System Using Leap Motion Sensors. Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, 2015.

[5] Funasaka, M., Ishikawa, Y., Takata, M., y Kazuki, J., Sign Language Recognition using Leap Motion Controller, PDTA, 2015.

[6] Langford Cervantes, J., Alencastre Miranda, M., Muñoz Gómez, L., Navarro Hinojosa, O., Echeverría Furio, G. Manrique Juan, C., y Maqueo, M. Detección y seguimiento de palmas y puntas de los dedos en tiempo real basado en imágenes de profundidad para aplicaciones interactivas, Research in Computing Science, pp. 137-149, 21 Marzo 2016.

[7] Leap Motion. API Overview: https://developer.leapmotion.com/ documentation /csharp/devguide/Leap_Overview.html.

[8] Naglot, D., y Kulkarni, M., Real Time Sign Language Recognition Using the Leap Motion Controller, Vishwakarma Institute of Technology.Pune, India, 2015.

[9] Nájera Romero, L. O., López Sánchez, M., González Serna, J. G., Pineda, T. R., y Arana Llanes, J. Y. Recognition of Mexican Sign Language through the Leap Motion Controller, CSC, 2016.

[10] Priego Pérez, F. P. Reconocimiento de Imágenes del Lenguaje de Señas Mexicano, Centro de Investigación en Computación IPN, México, 2012.

[11] Santos, A., Guimaraes, V., Matos, N., Cevada, J., Ferreira, C., y Sousa, I., Multi-sensor Excercise-based Interactive Games for Fall Prevention and Rehabilitation, 9th International Conference on Pervasive Computing Technologies for Healthcare, 2015.

[12] Serafín de Fleischmann, M. E., González Pérez, R., Manos con Voz, CONAPRED, México, 2011.