INTERACCIÓN HOMBRE-MÁQUINA USANDO GESTOS MANUALES EN TEXTO REAL*

*Nelson Balsero, Diego Botero, Juan Zuluaga***
*Carlos Alberto Parra Rodríguez****

Resumen: en este artículo se presenta el desarrollo de un sistema electrónico capaz de reconocer, en tiempo real, doce gestos manuales realizados por un interlocutor con una de sus manos en una escena con iluminación y con fondo controlados. El sistema implementado es robusto a rotaciones, translaciones y cambios de escala de la mano del interlocutor en el plano de la cámara. El sistema funciona tanto en una computadora personal, como en la tarjeta de evaluación *ADSP BF-533 EzKit Lite* de *Analog Devices*. Como etapa final se propone en la tarjeta de desarrollo la visualización en un *display* de la letra asociada al gesto reconocido. Por otra parte, en el computador personal se tiene una herramienta visual que presenta el procesamiento realizado en cada una de las etapas del algoritmo propuesto. Se tiene un sistema eficiente para la comunicación entre un hombre y una máquina y se vislumbran futuras aplicaciones orientadas a posibilitar la interacción de personas sordomudas con la población, en general.

Palabras clave: reconocimiento e interpretación de imágenes, interacción hombre-máquina en tiempo real, procesador Blackfin 533.

Abstract: in this paper we present the development of an electronic system able to recognize, in real time, a set of twelve manual gestures carried out by a person with one of his hands in a controlled illumination and background scene. The implemented system shows rotational, translational and scale change robustness. The system is intended to

* *Fecha de recepción: 5 de septiembre de 2005. Fecha de aceptación para publicación: 29 de noviembre de 2005. Este artículo se deriva de un trabajo con el mismo nombre presentado en el IEEE Colombian Workshop on Robotics and Automation,* (*CWRA*) *en agosto de 2005, organizado por el IEEE-Colombia y cuyas memorias fueron publicadas con el ISBN 958-695-182-0. Se publica con autorización del coordinador del comité evaluador del evento.*

** *Ingenieros electrónicos, Pontificia Universidad Javeriana, Bogotá, Colombia.*

*** *Ingeniero electrónico, Pontificia Universidad Javeriana. Magíster en Ingeniería Eléctrica, Universidad de los Andes. Magíster en DEA y Doctorado en Automatique et Informatique Industrielle, Universite de Toulouse III. Profesor asociado, Departamento de Electrónica, Pontificia Universidad Javeriana, Bogotá, Colombia. Correo electrónico: carlos.parra@javeriana.edu.co.*

be performed in a personal computer as well as in the Analog devices' evaluation platform *ADSP Blackfin 533 Ez Kit Lite*. As a final step, in the Blackfin's platform, we propose a view option in a display of the associated letter to the recognized gesture. In the personal computer we present an illustration tool intended to show the results in different steps of the proposed algorithm. We obtained an efficient system for human machine interaction and future applications intended to enable the interaction of deaf and mute people with the population in general.

Key words: recognition and interpretation of images, human-machine interaction in real time, Blackfin 533 processor.

1. INTRODUCCIÓN

Actualmente, para tratar el problema de reconocimiento se dispone de un conjunto de técnicas muy potentes; no obstante, su costo computacional normalmente es muy alto, lo cual las hace poco viables para implementarlas en tiempo real usando un procesador embebido.

Actualmente se destaca la técnica propuesta por Viola y Jones [2001] y, posteriormente, mejorada por Lienhart y Maydt [2002], que está basada en las *Wavelets de Haar*, las cuales hacen un análisis multi-resolución de la imagen. Además, la librería *OpenCV* [Intel, 2004] incluye funciones que permiten encontrar la mano y la cara mediante clasificadores paralelos tipo *Adaboost* [Lienhart, Maydt, 2002] los cuales arrojan excelentes resultados. Otros autores usan técnicas morfológicas como la esqueletización, para identificar órganos en el cuerpo humano [Di Ruberto, Rodríguez, Casta, 200?].

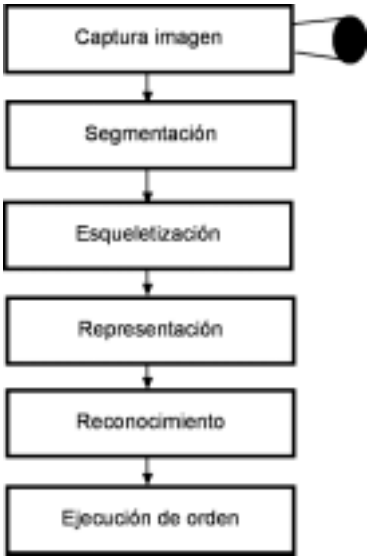
Con el objetivo de buscar ciertas características que clasifiquen cada uno de los gestos para diferentes individuos, se propone como solución realizar un análisis morfológico de la imagen. De esta forma, se establece un nuevo alfabeto, en el cual cada dedo de la mano representa un *bit*, estableciéndose un conjunto de gestos altamente diferenciables y un problema de naturaleza binaria, el cual se aborda mediante un análisis morfológico de la imagen. Finalmente, se demuestra la viabilidad y la eficiencia del algoritmo desarrollado y se obtienen tiempos de procesamiento bastante buenos en el procesador *Blackfin* 533 [ADSP-BF533, 200?], usando la tarjeta de evaluación *EZ-kit lite* de *Analog Devices*

2. DESCRIPCIÓN GENERAL DEL SISTEMA

El sistema implementado ejecuta el procesamiento de una secuencia de video, en la cual se encuentra una persona usando camisa oscura de manga larga gesticulando en primer plano, y su imagen es capturada por una cámara bajo condiciones de iluminación y de fondo controlados.

En la Figura 1 se presenta el diagrama en bloques del sistema. En la primera etapa se realiza la captura de la imagen a partir de una secuencia de video; posteriormente, se tiene una etapa de segmentación en la cual se determina la región en la que se encuentra la mano del interlocutor gestual (región de la imagen que va a ser procesada). En la región de interés se realiza un proceso de adelgazamiento para limitar la cantidad de información que se va a procesar y para permitir el éxito de la etapa de reconocimiento. Una vez la imagen es adelgazada, se identifican puntos de interés, cuya posición respecto al centro de masa de la mano, permitan su representación en un vector, cuya dimensión es igual al número de dedos. Cada uno de sus componentes tiene información sobre la inclinación del dedo, respecto a la inclinación del antebrazo. Finalmente, usando el error cuadrático medio, se decide si el vector es lo suficientemente parecido a alguno de los vectores base, los cuales son establecidos en una etapa de entrenamiento previa al funcionamiento del sistema.

Figura 1. Diagrama en bloques del sistema



Fuente: presentación propia de los autores.

2.1. ALFABETO

En el marco del presente trabajo se propone un nuevo alfabeto, basado en el número de dedos y la ubicación de estos en la mano; éste otorga la flexibilidad de obtener un conjunto bastante amplio de gestos, con diferencias estructurales bastante definidas. Asimismo, el objetivo es identificar la presencia del dedo y su ubicación con respecto al centro de masa de la mano, es decir, este nuevo alfabeto está basado en la modelación de los dedos de la mano como entradas binarias al sistema de reconocimiento y se considera el dedo pulgar como el *bit* más significativo. Así, como se ilustra en la Figura 2, en la cual se presentan los

símbolos generados en este trabajo, la letra A se representa en términos binarios por 10000, debido a que el gesto presenta sólo el dedo pulgar. Al representar de forma binaria una mano se pueden obtener treinta y dos gestos, con la posibilidad futura de ampliar el alfabeto, aproximadamente a sesenta y cuatro gestos (usando las dos caras de la mano) y más de dos mil gestos con la utilización de las dos manos.

2.2. Entrenamiento del sistema

La fase de entrenamiento es una etapa del diseño del sistema para establecer los vectores base; estos contienen información de cada uno de los gestos para los cuales el sistema va a responder, siendo una pequeña base de datos de la vectorización de imágenes correspondientes a gestos válidos. De igual forma, de los vectores base depende el éxito del reconocimiento, razón por la cual éstos se establecen mediante una serie de pruebas y de análisis estadísticos de los resultados obtenidos al aplicar el algoritmo desarrollado en diferentes interlocutores. Si se promedian los datos correspondientes a varios individuos se pueden definir los vectores base y cuando se toman muestras de una población significativa se logra desarrollar un sistema funcional para la población, en general.

2.3. Segmentación

Mediante el empleo de fondo y de iluminación controladas, la región correspondiente a la piel resulta ser la más brillante. De este modo, si se usa la información de luminancia, aquellos píxeles que superen un umbral establecido son considerados como piel y, por tanto, deben ser tenidos en cuenta para un posterior análisis. Si se garantiza una buena iluminación y un fondo suficientemente opaco, se puede establecer un umbral con el cual se puede lograr una buena segmentación.

2.4. Región de interés

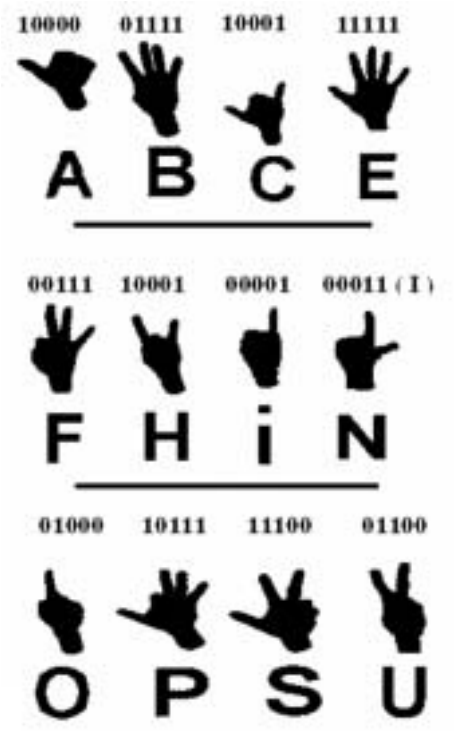
En esta etapa del proceso se establece una región que debe contener únicamente la mano segmentada para garantizar un posterior reconocimiento. Dentro de la escena en la cual se encuentra el interlocutor, es posible que aparezcan otros objetos; el sistema debe ser capaz de ubicar estos objetos, que resultan ser ruido para la aplicación, y filtrarlos. Teniendo esto en cuenta se debe procesar la imagen para establecer una región de interés (ROI). Mediante el establecimiento de la ROI, se reduce el área de la imagen sobre la cual se debe buscar el objetivo, con lo cual se optimiza el proceso.

2.5. Esqueletización

Para disminuir la cantidad de información por procesar, conservando la distribución topológica de las manos, se puede realizar una operación morfológica sobre la región de interés como es el adelgazamiento [González, 1993].

El adelgazamiento remueve información redundante, lo cual produce una imagen más simple y reduce el espacio y el tiempo de acceso a memoria y se facilita la extracción de características topológicas de la región de interés. El resultado del proceso de adelgazamiento de la imagen segmentada debe mantener ciertas propiedades, para posibilitar una correcta conservación de las características topológicas de un gesto determinado y permitir un correcto reconocimiento futuro. El resultado de la operación morfológica en cuestión debe garantizar que la imagen resultante sea de un píxel de ancho [González, 1993]; de esta forma se pueden encontrar fácilmente las ramificaciones que corresponden a píxeles con más de dos vecinos y los puntos terminales que son los píxeles con un sólo vecino de interés.

Figura 2. Alfabeto generado.



Fuente: presentación propia de los autores.

En este trabajo se evaluaron dos algoritmos de adelgazamiento para determinar el método apropiado para el cumplimiento del reconocimiento en tiempo real. Finalmente, se implementó el algoritmo *Zhang Suen* [Forsyth, 2003] y el *Medial Axis Transform* (MAT) [Haralick, 1993].

2.6. Filtrado de imagen esqueletizada

Antes de la obtención de los puntos finales se hace necesario realizar un proceso de limpieza de la imagen adelgazada resultante. Se determinan las distancias de los puntos finales obtenidos de la imagen adelgazada

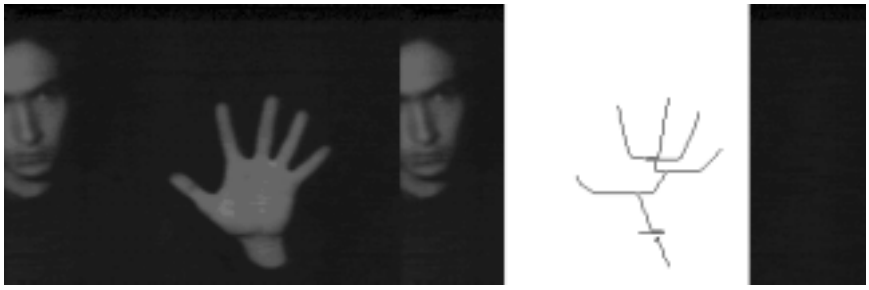
respecto al centro de masa de la imagen segmentada y se establece un umbral. En efecto, los puntos finales correspondientes a las distancias que se muestren menores que dicho umbral son descartados.

Para mantener la robustez del sistema a la distancia cámara–usuario, el proceso de limpieza de la imagen adelgazada establece un umbral, el cual es una proporción de la distancia más grande entre uno de los puntos finales y el centro de masa de la mano. Por ello, se establece unumbral adaptativo.

2.7. Representación y reconocimiento

Para reconocer en una imagen un gesto, se hace un análisis morfológico de la imagen en busca de una vectorización apropiada de la mano que permita un posterior reconocimiento.

Figura 3. Puntos finales del esqueleto.*



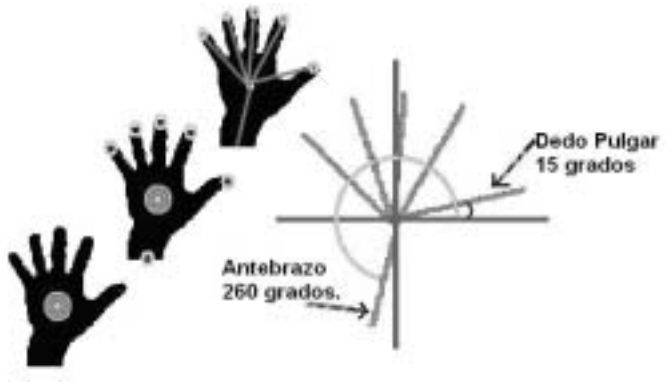
Fuente: presentación propia de los autores.

*En el mejor de los casos corresponden a los dedos o al punto final del antebrazo.

Una vez la imagen ha sido adelgazada –basados en un soporte teórico formal– se trata de justificar la escogencia de los puntos de control, los cuales van a ser preponderantes en la intención de extraer las características topológicas de la mano, a partir de una imagen adelgazada. Se quiere, a partir de una imagen adelgazada, hallar los puntos de control más adecuados para la representación de la curva resultante del proceso de adelgazamiento, teniendo en cuenta que limitar su número es importante para el cumplimento del objetivo de operaciones, en el tiempo real del procesamiento que se quiere realizar. Así pues, se escogen los puntos finales de la imagen adelgazada, dado que estos proveen información clave de la estructura topológica de la mano, además del centro de masa de la mano segmentada, para representar la idea de la geometría que ésta contiene con los puntos de control que se van a analizar. El proceso de reconocimiento se basa en encontrar los ángulos de los dedos y del antebrazo (puntos finales) con respecto a un punto de referencia. Para calcular cada uno de estos ángulos se usa como referencia el centro de masa de la mano segmentada del interlocutor (ver Figura 4).

Cuando la mano se encuentra completamente vertical, el ángulo del antebrazo respecto al origen es idealmente 270 grados, y el ángulo entre el dedo pulgar y el antebrazo es un poco mayor a 90 grados, como se muestra en la Figura 4.

Figura 4. Ángulos relativos al antebrazo.



Fuente: presentación propia de los autores.

Cuando se rota la mano en el plano, el ángulo del antebrazo con respecto al origen cambia, tal y como era de esperarse. Sin embargo, la diferencia de ángulos entre el dedo pulgar y el antebrazo sigue siendo un poco mayor a 90 grados, de la misma forma la diferencia de ángulos entre el antebrazo y cada uno de los dedos, permanece constante.

Se genera un vector con las diferencias de ángulos, el cual, posteriormente, usando el criterio del error cuadrático con respecto a unos vectores base establecidos en la fase de entrenamiento, se decide si el vector calculado se parece lo suficiente a alguno de los vectores almacenados en memoria y se toma como base esta decisión para reconocer cada uno de los gestos. Estos vectores serán arreglos con longitud igual al número de dedos en cada gesto, así la longitud máxima de uno de estos vectores base es cinco (correspondientes a los cinco ángulos entre el antebrazo y cada uno de los cinco dedos). Para cada uno de los gestos, para los cuales el sistema fue entrenado, existe al menos un vector. Se trabaja con los ángulos para que el sistema sea robusto a translaciones y a cambios de escala, dado que los ángulos dependen de las relaciones entre longitudes, las cuales serán constantes, siempre y cuando el objetivo se encuentre en el plano de la cámara.

Para encontrar cada uno de los ángulos se usa como punto de referencia el centro de masa, que puede variar a medida que el interlocutor gesticula. Para lograr un centro de masa más estático cabe la posibilidad de descubrir más el antebrazo del interlocutor, lo cual trae como consecuencia que los ángulos entre los dedos sean más parecidos, con una mayor probabilidad de error. Después de una serie de pruebas se estableció que el punto óptimo en el cual debe dejarse la manga es aproximadamente 3 cm debajo de la mano del interlocutor.

3. DESARROLLO EN EL PROCESADOR *ANALOG DEVICES ADSP BF-533*

La implementación del sistema es orientada a tiempo real, usando un procesador dedicado, lo cual permite aplicaciones portátiles. En el desarrollo del sistema se usaron, principalmente, el codificador de video ADV7183, la interfaz paralela de periféricos (PPI), el controlador de DMA y la memoria asíncrona SDRAM. El PPI junto al DMA permiten implementar un sub-muestreo de la imagen exclusivamente con hardware. Este sub-muestreo no afecta de manera significativa el desempeño de la aplicación, pero sí optimiza la velocidad de proceso, pues se disminuyen los accesos a memoria. Se configura el DMA para que genere una interrupción, una vez la imagen completa haya sido almacenada en memoria e interrumpa la transferencia de datos. De este modo, se tiene almacenada en memoria una imagen blanco y negro correspondiente a la escena capturada. La rutina de interrupción del DMA corresponde al proceso de la imagen. Una vez se ha procesado la imagen se habilita, nuevamente, el DMA para que transfiera otra imagen y se repita el proceso.

4. EVALUACIÓN DE RESULTADOS

Para evaluar el algoritmo de reconocimiento se analizaron un total de diecinueve mil doscientas imágenes correspondientes a diferentes individuos gesticulando. Se evaluaron cuatro aspectos de eficiencia del sistema, cada uno con cuatro mil ochocientas imágenes independientes y los resultados son los siguientes:

- Verdaderos aciertos 79,27%
- Verdaderos rechazos 99,50%
- Falsos aciertos 0,27%
- Falsos rechazos 38,39%

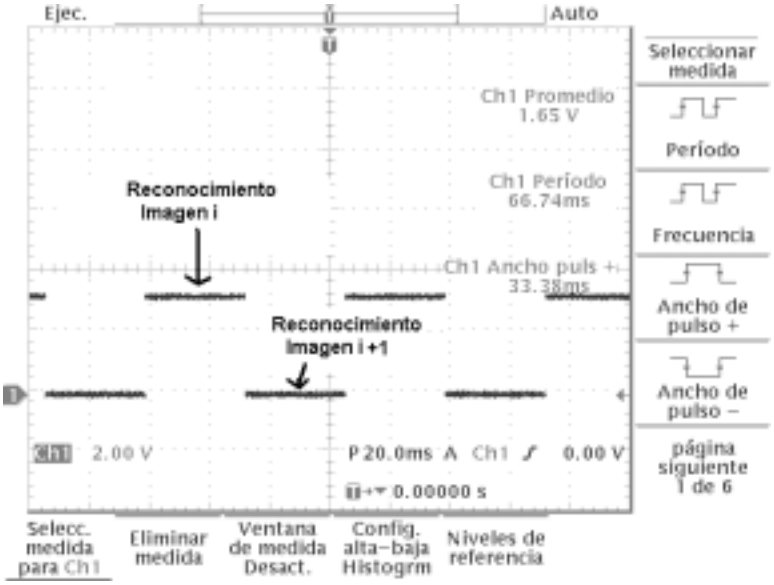
Cabe recalcar que el sistema reconoce el 79% de los cuadros analizados, lo cual es muy alto si se tiene en cuenta que en un segundo se procesan veinticinco imágenes. El procesamiento de la imagen consta de dos etapas; una primera en la cual se localiza la mano dentro de la imagen y una segunda etapa, en la cual se reconoce el objeto. La localización de la mano en la imagen, que corresponde al proceso de fijación de la ROI, es un proceso mucho más costoso computacionalmente que el proceso de reconocimiento, tal y como se aprecia en las Figuras 5 y 6, en las cuales se visualiza una señal correspondiente a la bandera programable del procesador de *Analog Devices*, con la cual se determina el tiempo de proceso (medio periodo de la señal corresponde a un procesamiento de la imagen).

De este modo, se toma la decisión de implementar una localización de la ROI cada cien procesos. En efecto, se establece la región de interés y sobre esta región se realizan los cien reconocimientos posteriores. Finalmente se refresca la ROI, como se aprecia en la Figura 7. Como se puede observar, se logra obtener un mayor número de reconocimientos

en un determinado tiempo. El algoritmo de adelgazamiento implementado en la DSP fue el MAT debido a que el tiempo de proceso (35 ms) resulta ser entre cinco y seis veces menor que el tiempo logrado con el algoritmo de adelgazamiento *Shang Zuen* (150 ms). En el entorno de programación Visual C++ se desarrolla un sistema más robusto, en comparación con el implementado en la tarjeta de desarrollo, dado que es posible determinar constantemente la región de interés de forma recursiva, sin afectar su funcionamiento en tiempo real. La función de ubicación de la mano con algoritmos recursivos resulta ser óptima en relación con su versión no recursiva en cuanto a tiempo; sin embargo, en la tarjeta de evaluación, por el gran número de iteraciones implicadas, se implementa la versión no recursiva del algoritmo de ubicación.

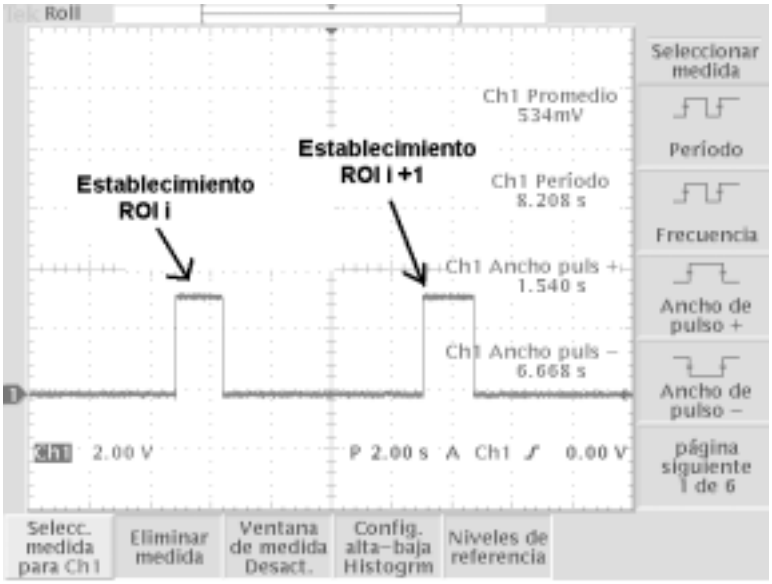
Al implementar una función recursiva, el procesador debe guardar el contexto en cada iteración y considerando que el número de iteraciones es proporcional al número de píxeles analizados en la imagen, resulta ser un inconveniente por las limitaciones del espacio, en memoria rápida, en la cual se pueda almacenar el contexto.

Figura 5. Tiempo promedio de proceso de reconocimiento



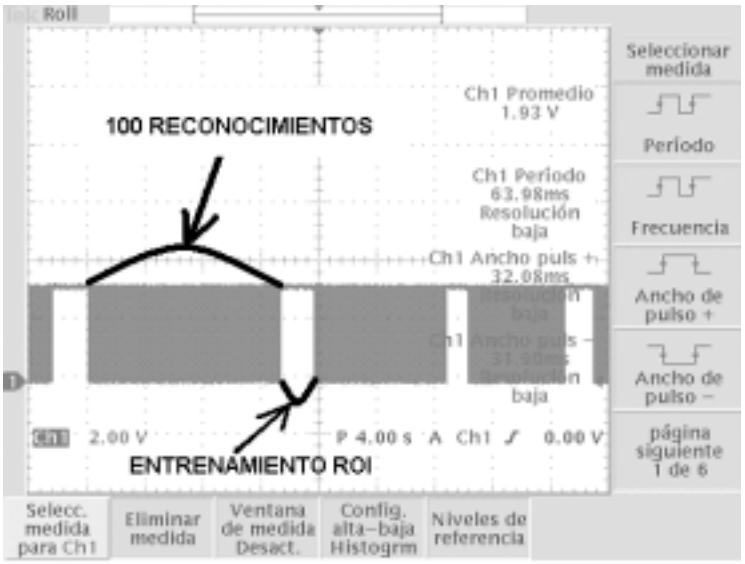
Fuente: presentación propia de los autores.

Figura 6. Tiempo de establecimiento de la ROI



Fuente: presentación propia de los autores.

Figura 7. Proceso de localización de la mano



Fuente: presentación propia de los autores.

5. CONCLUSIONES

Se obtuvo una herramienta eficiente que permite la comunicación entre un usuario y una máquina, abriéndole a éste la posibilidad de controlarla a distancia y en tiempo real. De la misma manera abre la posibilidad de manejar puertos y demás periféricos de la computadora personal, permitiendo desarrollos futuros enfocados a posibilitar teleconferencias guiadas por personas sordomudas. Se vislumbra entonces, la posibilidad de permitir a dicha población limitar su aislamiento, para poder interactuar con personas que son ajenas al lenguaje implementado, a través de una máquina que sintetiza un sonido o genera un texto. En una futura mejora del sistema se propone trabajar con espacios de color con lo cual se podría llegar a tener con cualquier clase de fondo.

REFERENCIAS

ADSP-BF533 Blackfin Processor Hardware Reference, Rev. 3.0. Analog Devices, s.d.

Di Ruberto, C., Rodriguez, G., Casta, L. *Recognition of Shapes by Morphological Attributed Relational Graphs*. s.d.: Department of Mathematics, University of Cagliari, s.d.

Forsyth, D. *Computer Vision: A Modern Approach*. New Jersey: Prentice Hall, 2003.

González, R. "Digital Image Processing". *Reading: Addison-Wesley*, 1993.

Haralick, R. M. "Computer and Robot Vision". Vol. 2. *Reading: Addison-Wesley*, 1993.

Intel® Software Products Open Source *Intel Open Source Computer Vision Library Reference Manual* Diciembre 2004. Disponible en la dirección electrónica http://www.intel.com/research/mrl/research/opencv.

Lienhart, R., Maydt, J. "An Extended Set of Haar-like Features for Rapid Object Detection". IEEE ICIP, 2002.

Viola, P., Jones, M.J. "Rapid Object Detection using a Boosted Cascade of Simple Features". IEEE CVPR, 2001.