CONVERSIÓN DEL LENGUAJE DE SEÑAS A VOZ

CONVERSION OF SIGN LANGUAGE TO VOICE

Jorge Del Carpio Salinas[1], Amilcar Mescco Mizares[2]

RESUMEN

El presente trabajo usa los métodos de procesamiento de imagen y redes neuronales. Para clasificar los patrones del lenguaje de señas, se ha implementado en Matlab un conjunto de funciones para tal fin. Nuestro vocabulario del lenguaje de señas esta compuesto de 24 letras pertenecientes al abecedario, este trabajo considera solo aquellos gestos que no presentan movimiento, nuestro hardware son una cámara digital y una PC.

Palabras clave.- Lenguaje de señas, Método skin color segmentation, Redes neuronales, Transformada de Karhunen-Lòeve.

ABSTRACT

This work uses the image processing methods and neural network to classify the sign language patterns, a set of function had been implemented in Matlab to achieve our goal. Our vocabulary of sign language are composed of 24 letters belong to the alphabet, this work considers only those gestures without movement, our hardware are a digital camera and a PC.

Key words.- Sign language, Skin color segmentation method, Neural network, Karhunen-Lòeve transform.

INTRODUCCIÓN

El estudio de la adquisición de imagen, extracción de características, clasificación y reconocimiento de patrones, constituyen factores importantes en el procesamiento de imágenes, de igual manera es importante para la construcción de sistemas prácticos.

En el presente trabajo titulado Conversión del Lenguaje de Señas a Voz se utilizan las etapas mencionadas con un propósito aplicativo que es reconocer las señas del lenguaje de las personas sordomudas para llevarlos a voz.

Para el desarrollo del presente trabajo se han seguido una serie de etapas, el primero de los cuales es la adquisición de imágenes. La realidad no es óptima para la toma de imágenes que posteriormente se procesará, ya que ésta presenta dificultades siendo la más importante la iluminación y la variación de la iluminación de un lugar a otro la que dificulta el procesado de las imágenes y siendo ésta la primera etapa es de vital importancia obtener imágenes óptimas. Para tratar de atenuar este problema se tomó en cuenta la iluminación del medio donde se tomó las fotografías tratando de evitar las sombras y brillo siempre presentes.

La segunda etapa, es la ubicación de la región de la mano para la identificación de las señas, en esa etapa se eliminó todo lo que no pertenece a la región de la mano obteniendo como resultado una imagen normalizada donde se muestra la seña que se realiza. La técnica usada fue la de localización por el color de la piel. Los problemas que se presentaron en la etapa de ubicación de la región

[1]Dr. Docente investigador de la Facultad de Ingeniería Eléctrica y Electrónica de la Universidad Nacional de Ingeniería, [2]Ingeniero, egresado de la Facultad de Ingeniería Eléctrica y Electrónica de la Universidad Nacional de Ingeniería.

de la mano, es que parte del fondo de la imagen puede confundirse con el color de la piel o pequeños puntos que no pertenecen a la región de la mano, estos problemas se superaron variando los parámetros Y, y I de una imagen que representan la luminancia y la información del color, también se utilizó los conceptos de erosión y dilatación para la eliminación de ciertas zonas no deseables, la salida para esta etapa es la imagen de entrada recortada y normalizada para un tratamiento homogéneo de todas las imágenes. Luego todas las imágenes normalizadas disponibles se representan como un vector fila y se forma una nueva matriz cuyas filas son imágenes representadas como vector.

Con la matriz disponible se procedió a la extracción de las características de la imagen, esta es una etapa que ayuda también a la reducción de la matriz, para el entrenamiento de la red neuronal se logró reducir la matriz que representa la imagen a un vector con relativamente pocos elementos. La técnica empleada fue la Transformada Karhunen-Lòeve (KLT).

Otro problema que se presentó fué que el tiempo de procesamiento de las imágenes era considerable, por ello se guardó como archivos de datos en Matlab los resultados previos, de esa manera cada vez que se necesitaba utilizar esos datos lo único que se hace es cargarlo a la función que lo solicite.

Se utilizó una red neuronal para el reconocimiento de los patrones del lenguaje de señas, el tipo de entrenamiento utilizado fue el de Levenberg-Marquardt con la que se obtuvo un entrenamiento rápido. Las etapas mencionadas se presentan a continuación en un diagrama de bloques, Fig. 1.

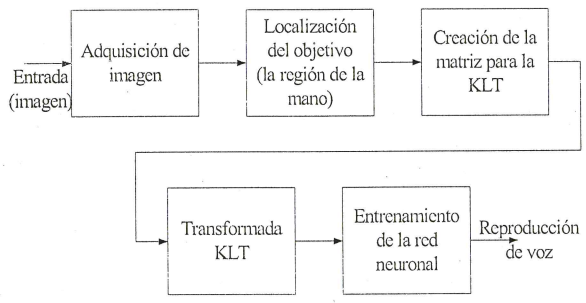

Fig. 1 *Diagrama de bloques del sistema "Conversión del Lenguaje de Señas a Voz".*

La entrada al sistema son imágenes ya sea adquiridas previamente o capturadas de una secuencia de video.

ADQUISICIÓN Y CARACTERISTICAS DE LAS IMÁGENES DIGITALES

Para esta primera etapa se considero especificaciones en cuanto a la ubicación de la cámara para la obtención de imágenes, región de enfoque de la cámara, distancia de la cámara con respecto a la persona que realiza las señas, la iluminación en el espacio de trabajo, la cantidad de imágenes y las variaciones de las señas realizadas.

La cámara debe de estar ubicada al frente de la persona que realiza las señas a una distancia de aproximadamente de 0.4 metros, (ver Fig. 2) en la región de enfoque de la cámara se debe evitar de tener colores semejantes al de la piel humana ya que esto confundiría a la etapa de localización del objetivo (esta es la segunda etapa), la iluminación en esta etapa juega un papel muy importante, el espacio de trabajo no debe presentar excesiva sombra y brillo en la región de enfoque de la cámara, entonces se debe ubicar a la persona de acuerdo a la iluminación del medio o arreglar un escenario óptimo a base de lámparas, en cuanto a la cantidad de imágenes se ha considerado 20 imágenes por cada seña entre las cuales se tendrá la seña realizada de manera correcta y pequeñas variaciones de la misma.

Teniendo en cuenta las consideraciones mencionadas se ha construido una base de datos que esta constituida por un total de 960 imágenes con una resolución de 640x480 píxeles (40 por cada seña de las letras del abecedario para cuya representación no se requiere de movimiento) adquiridas con una cámara digital.

Estas imágenes digitales tienen la característica de estar conformado por tres campos de colores las cuales son un campo de color rojo, un campo de color verde y otro de color azul los cuales se conocen como RGB (red, green y blue), mediante software se pueden extraer cada uno de estos campos, para el procesamiento de imagen se trabaja con las imágenes a escala de grises.

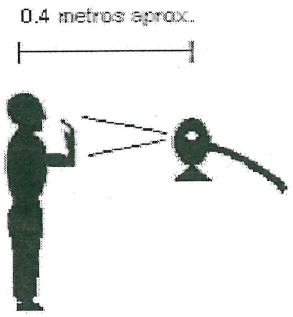

Fig. 2 Adquisición de imágenes.

UBICACIÓN DE LA REGIÓN DE LA MANO

Para la localización de la región de la mano se ha utilizado el método de "skin color segmentation", la cual usa los parámetros Y, I y Q. La relación entre los parámetros YIQ y los parámetros RGB es la siguiente:

$$\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ 0.596 & -0.275 & -0.320 \\ 0.212 & -0.523 & 0.311 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (1)$$

Y es el parámetro de la luminancia, I y Q representan la crominancia, para la localización de la región de la mano se han utilizado los parámetros Y e I. Experimentalmente se halló que el rango del color de la piel esta entre: $60<Y<300$ y $2<I<70$. Este rango varía de un lugar a otro por la diferencia en la iluminación, esto se nota principalmente cuando se trabajan con imágenes adquiridas de una secuencia de video. Es por eso que el sistema cuenta con un ajuste para los parámetros Y e I. En la Fig. 3 se observa el funcionamiento del método de ubicación de la región de la mano con los valores de Y e I hallados experimentalmente.

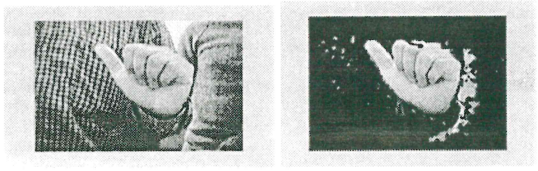

Fig. 3 Ubicación de la región de la mano con los parámetros Y e I.

Como se observa el exceso de brillo dificulta la completa localización de la región de la mano pero en este caso no afectó de manera considerable, hay que tener presente de adecuar el espacio de trabajo con una buena iluminación con eso nos referimos que se tiene que evitar las sombras y exceso de brillo, otro problema que se puede presentar como en este caso es la ubicación del fondo de la imagen como parte la mano, lo cual, se manifiesta como las manchas alrededor de la mano, este problema se solucionó con la teoría de erosión y dilatación para una imagen y si aún así quedan imperfecciones en la localización de la mano, esta se eliminará en el momento del recorte de la imagen en esta misma etapa. Se realiza el recorte de la imagen con el fin de eliminar todo lo que no sea dato útil refiriéndonos con eso al fondo de la imagen, una vez que se tiene la imagen recortada se hace uso de la transformada proyectiva para normalizar las imágenes, se requiere de imágenes de iguales dimensiones para un tratamiento homogéneo de ellas, Fig. 4.

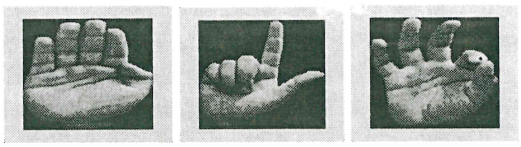

Fig. 4 Resultado del recorte y normalización de la imagen.

ORDENACIÓN DE LAS IMÁGENES

Las imágenes normalizadas que se han conseguido con el procedimiento anterior serán utilizadas para realizar la Transformada Karhunen-Lòeve (KLT) que será descrito posteriormente. Para usar estas imágenes normalizadas se lleva cada una de estas imágenes a un vector fila, ya que en el presente trabajo las imágenes se normalizaron a la dimensión de 60x40, se tiene que el vector fila obtenido es de 1x2400, con este conjunto de imágenes representados como vectores fila se forma una matriz que será la entrada para la KLT, para nuestro caso esta matriz tiene dimensiones de 480x2400 ya que se cuenta con 480 imágenes para el entrenamiento de la red de los 960 de la base de datos (el resto de imágenes fué usado en las pruebas del sistema).

EXTRACCIÓN DE CARACTERÍSTICAS

Para la extracción de características se ha usado la Transformada Karhunen-Lòeve (KLT) la que tiene por entrada una matriz cuyas filas representan determinado gesto del lenguaje de señas, lo que se quiere conseguir en esta parte del trabajo es un

conjunto de coeficientes (la menor cantidad posible) que represente a una imagen, esto se logra hallando la covarianza de la matriz de entrada para que de este resultado se hallen los valores propios (λ_i) y vectores propios (V_i). Los vectores propios hallados constituyen las bases ortonormales (V) requerida para representar una imagen con un conjunto de coeficientes. Los coeficientes buscados se pueden hallar matricialmente de la siguiente manera:

$$\vec{\alpha} = \left(V^{t}\right)^{*} \vec{J} \qquad (2)$$

O como un producto escalar por medio de:

$$\alpha_i = < v_i, J > \qquad (3)$$

Donde J es cada una de las imágenes representadas como vector fila, de esta manera cada imagen se puede expresar en función de los vectores propios y los coeficientes hallados por medio de:

$$J = \alpha_1 v_1 + \alpha_2 v_2 + \ldots + \alpha_N v_N \qquad (4)$$

Donde el máximo valor de N es 2400 ya que contamos con 2400 vectores propios con igual número de coeficientes, lo que se hizo para cumplir el objetivo de esta etapa es tomar una pequeña porción del total es decir N<2400, el criterio de elección de la cantidad de coeficientes se basó en el gráfico de los valores propios los cuales se ordenaron de mayor a menor, se tomaron los valores propios cuyas magnitudes sean los mayores de entre el total de los valores propios, según la grafica que se muestra se puede tomar los 200 primeros valores propios con sus respectivos vectores propios, ver Fig. 5.

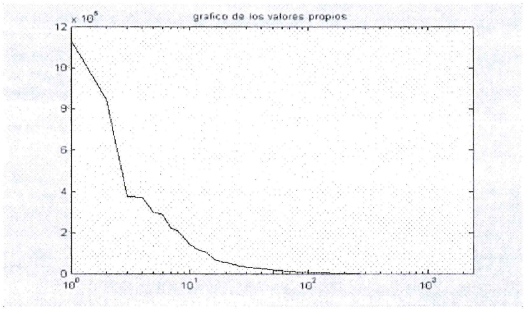

Fig. 5 *Gráfico de los valores propios.*

De esta manera se tiene cada una de las imágenes representadas por un conjunto pequeño de coeficientes, se puede recuperar la imagen con una cantidad N de coeficientes mediante:

$$\vec{J} = V\vec{\alpha} \qquad (5)$$

También se puede medir el grado de exactitud de la aproximación N-ésima con el cociente:

$$\rho = \frac{\sum_{i=1}^{N} \lambda_i}{\sum_{i=1}^{2400} \lambda_i} \qquad (6)$$

En el presente trabajo se podría haber tomado 200 coeficientes que según la tabla 1 representaría un grado de exactitud de 99.27%, en nuestro primer intento se tomó los 200 coeficientes pero luego en la simulación del sistema se obtuvo un pobre resultado, luego de varios intentos se llegó a la decisión de tomar 51 coeficientes, con lo cual el rendimiento del sistema en el reconocimiento del lenguaje de señas, es óptimo. En la Tabla 1 se muestra el número de coeficientes y el grado de exactitud que representa.

Número de Coeficientes tomados	Grado de exactitud %
51	85.8
60	87.4
70	89.31
80	90.91
90	92.28
100	93.43
200	99.27
2400	100

Tabla 1*. Coeficientes y grado de exactitud.*

Se puede tomar valores mayores que 51 pero lo que se haría es incrementar el cálculo computacional y de acuerdo con los resultados del entrenamiento de la red neuronal, 51 coeficientes es el adecuado. A continuación vemos en la Fig. 6, 7 y 8, la misma imagen que ha sido recuperado con diferente número de coeficientes.

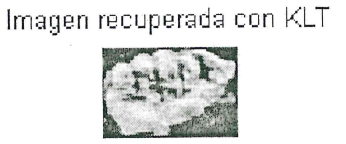

Fig. 6 *Recuperación de las imágenes con 51 coeficientes.*

Fig. 7 *Recuperación de las imágenes con 70 coeficientes.*

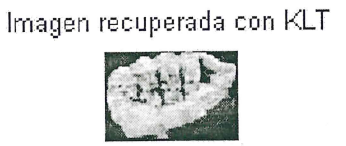

Fig. 8 *Recuperación de las imágenes con 200 coeficientes.*

El esquema de cómo se realiza la Transformada Karhunen-Lòeve (KLT) para las imágenes es la siguiente

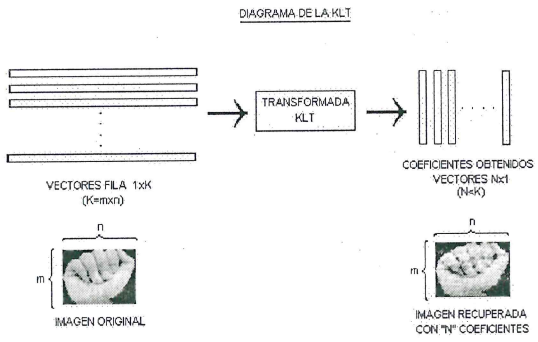

Fig. 9 *Esquema de la transformada Karhunen-Lòeve (KLT).*

RECONOCIMIENTO DE IMÁGENES

Para la clasificación de las imágenes se usó una red neuronal, específicamente una red Backpropagation, la red creada presenta la capa de entrada, una capa oculta y la capa de salida; como se mencionó la capa de entrada es un vector columna de 51x1, el número de neuronas para la capa oculta se halló de manera experimental, para un adecuado resultado se necesitó de 18 neuronas, dado que nuestro vocabulario esta constituido de 24 letras se tomó 24 neuronas para la capa de salida, el método de entrenamiento usado fue el método de Levenberg-Marquardt.

Con todos estos parámetros se logró que el sistema clasifique las imágenes que muestran el lenguaje de señas con una eficiencia de 98% considerando las imágenes con las que no ha sido entrenada la red, la salida del sistema es la reproducción de un archivo de audio previamente grabado.

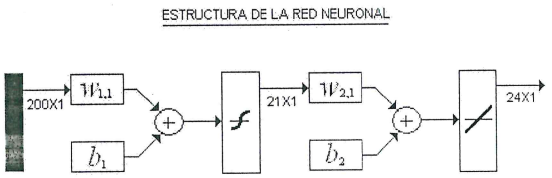

Fig. 10 *Esquema de la red neuronal los valores mostrados corresponden al primer entrenamiento.*

La red neuronal es entrenado con imágenes que presentan una determinada iluminación, al cambiar de lugar el parámetro Y varia con lo que se requiere un ajuste de los valores lo que ocasiona que la red neuronal disminuya su eficiencia, es por eso que es recomendable entrenar la red neuronal con la iluminación del nuevo ambiente de trabajo, esto es más crítico cuando se extraen imágenes de un secuencia de video.

La curva de entrenamiento de la red neuronal es el siguiente.

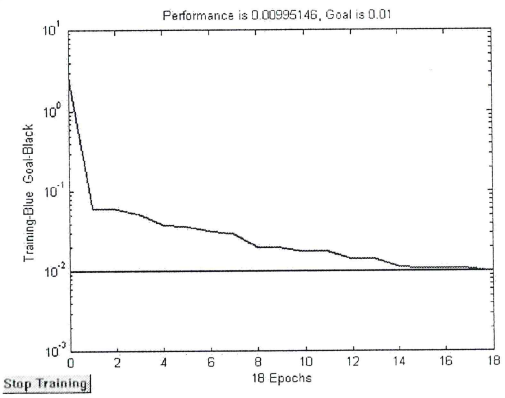

Fig. 11 *Curva de entrenamiento de la red neuronal.*

Como se observa en la Fig. 11, se llega al rendimiento especificado en el entrenamiento, hay

casos en que no se logra llegar a la meta lo cual se evidencia en el momento de la simulación ya que el rendimiento se ve disminuido.

RESULTADOS

Se ha logrado el reconocimiento del lenguaje de señas de las letras del alfabeto que no presentan movimiento, que son en número 24, en un 98%, las entradas para el sistema provienen de imágenes adquiridas previamente o adquiridas de una secuencia de video.

La interfase para el usuario presenta las imágenes de entrada ya sea de una secuencia de video o imágenes adquiridas previamente y la señal de voz muestreada.

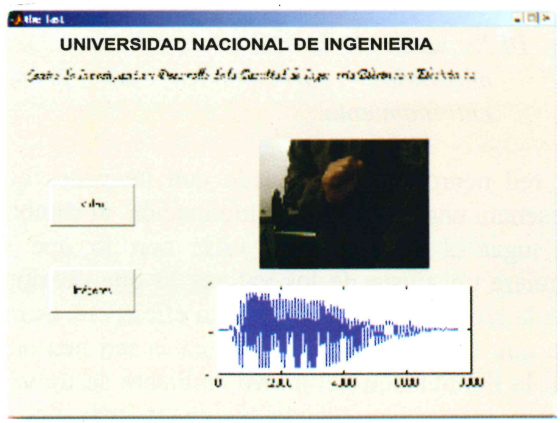

Fig. 12 Interfase para el usuario.

CONCLUSIONES

La iluminación es el factor determinante para una buena recolección de datos es por eso que en el lugar de trabajo se debe buscar iluminación homogénea sobre nuestro objetivo que es la región de la mano.

Los parámetros Y e I utilizados para la localización de la mano pueden ser ajustados, los valores tomados en el presente trabajo son experimentales.

El tiempo que demora para el entrenamiento es considerable por lo que una vez procesado se guarda como un archivo de datos para cargarlos a la función que lo necesite. Se ha implementado una función de simulación para las nuevas imágenes que se quiera clasificar, estas nuevas imágenes también usan las base ortonormales hallados con anterioridad.

La red neuronal se entrenó con 480 imágenes de la base de datos, para lo cual tenemos que la eficiencia de la red con el total de imágenes que se cuenta en la base de datos (960 imágenes) es de 98%.

En el caso de la adquisición de imágenes de una secuencia de video es crítica la ubicación de la región de la mano, ya que en esto interviene también la resolución y la rapidez de toma de imágenes de la cámara

La conversión del lenguaje de señas a voz de imágenes adquiridas de una secuencia de video, presenta un pequeño retardo por el procesamiento que se realiza, en este retardo también influyen el tipo de cámara utilizado y el entorno el que ha sido implementado los algoritmos utilizados.

Se ha intentado usar la Transformada Discreta de Coseno en dos dimensiones para la extracción de características, pero los resultados obtenidos no fueron satisfactorios.

REFERENCIAS

1. **Forsyth, D. A., Ponce, J.**, "Computer Vision a Modern Approach", Editorial Prentice Hall, 2003.
2. **Jain, A.K.,** "Fundamentals of Digital Image Processing", Editorial Prentice Hall, 1989.
3. **Gonzáles, R., Wood, R.,** "Tratamiento Digital de Imágenes", Editorial Wesley, 1992.
4. **Bhuiyan, A., Ampornaramveth, V., Ueno, H., "**Detection and Facial Feature Localization for Human-machine interface", 2003, http://www.nii.ac.jp/hrd/HTML/Journal/pdf/05/05-03.pdf.
5. **Demuth, H., Beale, M.,** "Image Processing Toolbox for Use with MATLAB® User's Guide", Versión 3, 2002.
6. **Domingo, M.,** "Visión Artificial" Departamento de Ingeniería Informática Universidad de Santiago de Chile, 2003, http://www2.ing.puc.cl/~dmery/vision/guia1_2003.pdf.
7. **Martín, F.,** "Transformadas de Imagen" Universidad de Vigo. http://www.gts.tsc.uvigo.es/pi/Transformadas.pdf.
8. **Demuth, H., Beale, M.,** "Neural Network Toolbox for Use with MATLAB®", versión 4, 2004.

9. **Freeman, J., Skapura, D.**,"Redes Neuronales Aplicaciones y Técnicas de Programación", Editorial Addison-Wesley, 1991.
10. **Lagunas, M. A.,** "Procesado en dos dimensiones", 2003. www.cttc.es/docs/CapVI.pdf.
11. **Chen, Y.,** "Chinese Sign Language Recognition and Synthesis", Institute of Computing Technology Chinese Academy of Sciences, 2003. http://brigade.umiacs.umd.edu/iccv2003/yqchen_demo.pdf.
12. **http://www.fotodigital.tk/Tecnologia.html.**
13. **The MathWorks,** "Creating Graphical User Interfaces", Version 7, 2004.

Correspondencia: amescco@yahho.es

Recepción de originales: enero 2007
Aceptación de originales: abril 2007