Estudio comparativo de técnicas para el reconocimiento de gestos por visión artificial

Comparative study of techniques for gesture recognition by artificial vision

Sandra E. Nope R. Esp., Humberto Loaiza C. PhD, Eduardo Caicedo B., PhD
Grupo de Percepción y Sistemas Inteligentes(PSI) – E.I.E.E., Universidad del Valle, Cali-Colombia
{sandrano,ecaicedo,hloaiza}@univalle.edu.co

Recibido para revisión: 2 de Septiembre de 2008, Aceptado: 28 de Noviembre de 2008, Versión final: 10 de Diciembre de 2008

Resumen—Se aborda el problema del reconocimiento visual de gestos realizados con las manos mediante diferentes técnicas de reconocimiento de patrones. Los gestos son capturados a través de una cámara Web, y, se extraen primitivas de movimiento inspiradas del procesamiento de la información de movimiento que ocurre en el cerebro de macacos. Los resultados alcanzados en el reconocimiento indican que características usadas en el reconocimiento son bastante discriminantes, por lo que, incluso con técnicas simples de reconocimiento de patrones, se obtuvieron buenos resultados. Sin embargo, es a través de un estudio que abarque diferentes técnicas de reconocimiento de patrones que los resultados en el reconocimiento pueden ser optimizados. En el presente trabajo, se evaluó el desempeño de seis técnicas de reconocimiento estándar para distinguir entre cuatro gestos diferentes, en donde el porcentaje de reconocimiento correcto osciló entre 87.88% y 97.14%.

Palabras Clave—Codificación y primitivas de movimiento, Integración temporal, Reconocimiento de gestos, Robótica.

Abstract—We address the problem of visual hand gesture recognition performed using different techniques of pattern recognition. Gestures are captured from a Webcam, and, motion primitives are extracted inspired on the motion perception process in macaque's brain. The recognition results obtained show the discriminate power of the features; for this reason, a simple recognition algorithm would provide satisfactory results. However, the recognition results can be improved by a study about different techniques of pattern recognition. Thus, we evaluate the performance of six standard techniques to distinguish between 4 different gestures, obtaining recall rates amongst 87.88% and 97.14%.

Keywords—Motion codification and primitives, Temporal integration, Gesture recognition, Robotics.

I. INTRODUCCIÓN

En la vida cotidiana, la gente usa los gestos para comunicarse con otras personas y/o para enriquecer la comunicación verbal. Un aspecto físico que pueda ser controlado, como posiciones físicas o movimientos de los dedos, manos, brazos o cuerpo de una persona; puede ser la base para un conjunto de gestos con diversidad de aplicaciones [1].

El reconocimiento de gestos ha sido un área bastante estudiada en los últimos tiempos con el fin de usarlos para transmitir información o para controlar dispositivos o aplicaciones. Los gestos realizados con manos y brazos han recibido más atención; en el primer caso, esto se debe a que las manos con 30 grados de libertad (incluyendo la muñeca) [2], son extremadamente hábiles y expresivas.

Los gestos pueden clasificarse en estáticos o dinámicos [3]. En los gestos dinámicos, los movimientos realizados corresponden al gesto en sí mismo; mientras que en los gestos estáticos son atribuidos a una cierta pose o configuración. En cualquier caso, para poder usarlos es necesario definir como deben ser interpretados en el contexto de tarea. Las interfaces hombre-máquina basadas en gestos pueden llevarse a cabo mediante visión artificial o guantes especiales; sin embargo, el primer método es más desafiante y natural que un dispositivo invasivo dedicado a la adquisición.

En cuanto al reconocimiento de gestos usando visión, [4] usaron modelos ocultos de Markov – HMM, para ello, ya que el problema de representar patrones de no gestos es difícil tanto para redes neuronales como para HMMs, lo resolvieron creando un modelo umbral que consiste en una copia del estado de todos los modelos de gestos de entrenamiento en el sistema; el cual es usado como un umbral de clasificación adaptativo. [5] propusieron el uso de Máquinas de Estado Finito – FSM semi-

automáticas para el reconocimiento de gestos en 2D. La estructura del modelo de las FSM se realiza inicialmente de forma manual, basándose en la observación de la topología espacial de los datos, y, posteriormente se refina de acuerdo con los datos de entrenamiento. La ventaja de esta aproximación reside en la eficiencia computacional de las FSM lo que permite que la clasificación se haga en tiempo real. Adicionalmente, no requieren grandes cantidades de datos de entrenamiento para obtener un buen modelo.

Por otro lado, los gestos son altamente variables de una persona a otra, e incluso de un ejemplo a otro durante la ejecución por una misma persona [6]. Por ello, es importante usar propiedades invariantes que representen el gesto.

Un aspecto importante en la construcción de un sistema de reconocimiento de gestos, además de escoger una representación en si misma, es la creación de una base de datos actualizada de los gestos conocidos y el criterio utilizado para reconocerlos. Es en este último aspecto en que se concentrará este trabajo. El objetivo que se desea alcanzar a futuro es programar un robot mediante aprendizaje por demostración; es decir, lograr que los robots adquieran nuevas habilidades a través de la observación y de esta forma, aprendan comportamientos complejos e interactúen inteligentemente con el ambiente. Para ello, se ha identificado como primera fase el reconocimiento de gestos.

En contraposición a los abordajes tradicionales que suelen hacerse en la literatura para abordar la extracción de características discriminantes que faciliten el proceso de reconocimiento, en este trabajo se extraen primitivas de movimiento bio-inspiradas. La ventaja de este enfoque, es que se tiene la certeza de esta utilizando una estrategia exitosa, en este caso, empleada por macacos, una especie de monos que posee un sistema visual muy similar al humano [7].

El artículo inicia, en la sección 2, con una descripción general del sistema, y, dada la importancia de la representación ,y aunque no es el objetivo principal de este trabajo, se describe brevemente en la sección 3 la forma en la que se extrajeron las primitivas de movimiento bio-inspiradas usadas para el reconocimiento. La sección 4 describe rápidamente las seis técnicas usadas en el reconocimiento de patrones en este trabajo, mientras que en la sección 5 se presentan los resultados obtenidos con cada una de ellas. Las conclusiones y trabajo futuro se presentan en la sección 6.

II. DESCRIPCIÓN DEL SISTEMA

La Figura 1 presenta el diagrama de bloques del sistema de reconocimiento de gestos que se utilizó. Para identificar y localizar el objeto de interés en las imágenes (la mano), se usó la información de color. El bloque de "Representación del Movimiento" usa esta información, junto con las derivadas espacio-temporales de las imágenes para estimar el movimiento en forma de vectores de flujo óptico. La salida de este bloque corresponde a un conjunto de respuestas neuronales que codifican el movimiento instantáneo, en donde la variable t corresponde al número de imágenes del vídeo en el que se realiza un gesto. A continuación está el bloque de "Integración Temporal" que recopila la información de movimiento instantáneo provista por los bloques precedentes, y cuya salida corresponde a un conjunto de imágenes que representan la evolución temporal de las respuestas neuronales. Debido a la alta dimensionalidad de la salida de este bloque, las respuestas neuronales se procesan de tal forma que se reduzca la dimensionalidad de los datos y se facilite la tarea de reconocimiento de gestos realizada por el siguiente bloque.

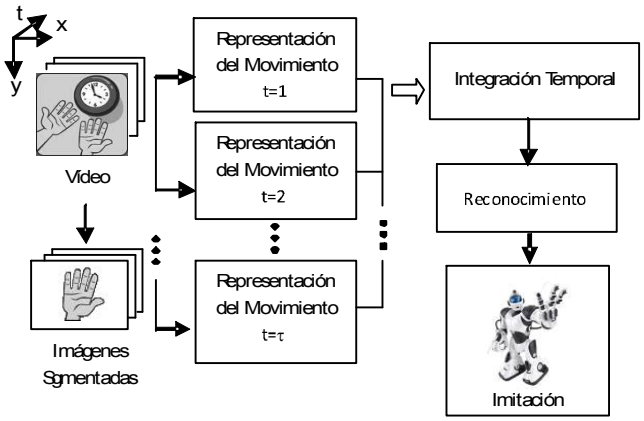

Figura 1. Diagrama de Bloques del Sistema de Reconocimiento de Gestos.

III. REPRESENTACIÓN DE LOS GESTOS

El proceso de obtención de las primitivas de movimiento usadas en el reconocimiento se gestos dinámicos se dividió en cuatro partes: segmentación de la mano en las imágenes, representación instantánea del movimiento, integración temporal y reducción de la dimensionalidad.

A. Segmentación de la Mano

Se utiliza la técnica de detección del color piel por píxel propuesta por [8]. Esta técnica es rápida, simple, y produjo los mejores resultados dentro de las técnicas píxel a píxel estudiadas. Consiste en usar un umbral el plano I del espacio de color YIQ por encima del cual los puntos se consideran de color piel. Los resultados reportados en la literatura alcanzaron un porcentaje de verdaderos positivos de 94.7% y de 30.2% falsos positivos. En el presente trabajo se adicionó un umbral inferior y superior en el plano Q con el fin de disminuir los falsos positivos. Los umbrales usados en las pruebas de laboratorio fueron determinados de manera heurística y corresponden a un valor de 13.7 en el plano I, y de -10 y 22 como límites inferior y superior en el plano Q.

En aras de eliminar puntos ruidosos y objetos indeseados del fondo de la imagen que no correspondan a la mano, sólo se escogen aquellos puntos conexos; correspondiéndole a la mano la mayor región dentro de la primera imagen. En imágenes consecutivas, la mano corresponde a la mayor región de color piel dentro de una ventana de búsqueda.

B. Representación del Movimiento

Los puntos identificados como pertenecientes a la mano en cada instante de tiempo, se usan para estimar el flujo óptico afín [9]. Esta técnica combina la ecuación de restricción del flujo óptico y las ecuaciones correspondientes al modelo de formación de imágenes afín. A través de la estimación de flujo óptico se obtiene un conjunto de vectores que representan el desplazamiento de los píxeles en las imágenes.

Los vectores de flujo óptico son la representación más simple del movimiento. Sin embargo, para percibir el movimiento es necesario procesar la información que contienen dichos vectores. Para ello, se usaron las ideas principales de [10] para simular en computador el procesamiento de movimiento en cerebro de macacos.

La codificación de movimiento realizada en el presente trabajo se dividió en dos partes: Procesamiento a bajo nivel y la codificación del movimiento. La primera reduce la resolución sin pérdida significativa de información, mientras que la segunda permite identificar la velocidad, dirección del movimiento e identificar la clase de movimiento complejo (rotación, compresión o expansión) que se está realizando.

1) Codificación a Alto Nivel

Para ahorrar tiempo de cómputo sin pérdida de información relevante y agilizar los procesos subsecuentes, se utilizó otra idea de la biología: Los Campos Receptores (*Receptive Fields* - RF). Los campos receptores se simularon calculando la media de todos los puntos dentro de círculos fijos solapados de diámetro D píxeles. La entrada a los RFs corresponde a la matriz de magnitud del flujo óptico o a la de su ángulo. Matemáticamente, dada la matriz de entrada a los campos receptivos $I_{in}(x,y)$, la matriz de salida de los campos receptivos $I_{out}(i,j)$, está definida por (1):

$$I_{out}(i,j) = \sum_x \sum_y k(i,j) * I_{in}(x,y) \tag{1}$$

donde,

$$k(i,j) = \begin{cases} 1/n & if \sqrt{(x-i)^2 + (y-j)^2} \le D/2 \\ 0 & \text{en otro caso} \end{cases} \tag{2}$$

2) Codificación del Movimiento

La selectividad de las neuronas a una velocidad y dirección particular, se simuló mediante la respuesta de un filtro Gausiano sintonizados a una determinada velocidad y dirección de movimiento. Esta respuesta se aproximó mediante la multiplicación de la respuesta separada de dos filtros Gausianos diferentes; uno selectivo a una velocidad particular y otro selectivo a una dirección particular. Matemáticamente, la respuesta de un filtro bidimensional $G(s_k,\theta_p)$ sintonizado a una velocidad s_k y a una dirección de movimiento θ_k está determinada por (3). En donde I_s y I_θ son la respuesta de los RFs ante la magnitud del flujo óptico y ante el ángulo respectivamente y, σ_s y σ_θ la desviación estándar de los mismos.

$$G(s_k,\theta_p) = e^{\frac{-(I_s - s_k)^2}{\sigma_s}} e^{\frac{-(I_\theta - \theta_p)^2}{\sigma_\theta}} \tag{3}$$

Para obtener un conjunto completo de respuestas neuronales, se realizan diferentes combinaciones (multiplicaciones) de los diferentes filtros Gausianos sintonizados a velocidad con los diferentes filtros Gausianos sintonizados a dirección. Así, si se relacionan las respuestas neuronales con imágenes, el resultado corresponderá a imágenes cuyos píxeles tienen una mayor intensidad de brillo en aquellos puntos con valores cercanos a los de sintonización de los filtros.

Es posible obtener una representación robusta a cambios del punto de vista, y que permite identificar la clase de movimiento que se está ejecutando. Para ello, en lugar de usar la dirección de movimiento, se utiliza el ángulo entre los vectores de flujo óptico y el gradiente de la magnitud del flujo óptico (*a*). Puede verificarse que, en el caso de un movimiento complejo como la rotación de la mano en el sentido de las manecillas del reloj, el ángulo *a* entre estos vectores tomando como referencia los vectores de flujo óptico corresponde a 90°; mientras que para una rotación en el sentido inverso es de 270°; para un movimiento de expansión es de 0° y de 180° para un movimiento de compresión. Valores angulares diferentes corresponden a una mezcla entre estos movimientos básicos.

C. Integración Temporal

Este bloque se encarga de reunir la información de movimiento instantánea provista por el bloque anterior, lo que es necesario para reconocimiento de gestos dinámicos. En este trabajo, la integración se realizó incluyendo la ejecución completa de los gestos, lo que le da mayor robustez al reconocimiento.

Con este objeto, [11] propusieron en su trabajo dos plantillas temporales que posteriormente procesaron para el reconocimiento de diferentes ejercicios aeróbicos: la Imagen de la Energía del Movimiento (*Motion Energy Image* – MEI) y la Imagen de la Historia del Movimiento (*Motion History Image* – MHI). En este trabajo se utilizó la MEI y una

adaptación de la MHI propuesta por Bobick y Davis.

La construcción de dichas plantillas requiere, inicialmente, estimar una imagen binaria $D(s_k,\theta_p,t)$. Esta imagen binaria, para este caso, indica las regiones en las que las respuestas neuronales son más fuertes de acuerdo con su sintonización en cuanto a velocidad y dirección del movimiento $D_\theta(s_k,\theta_p,t)$, lo que se refleja en (4), o en (5) para las neuronas sintonizadas a velocidad y detección de la clase de movimiento complejo $D_\alpha(s_k,\alpha_p,t)$.

$$D_\theta(s_k,\theta_p,t)=\begin{cases}1 & si\ G(s_k,\theta_p,t)>th_\theta\\ 0 & en\ \ otro\ \ caso\end{cases} \tag{4}$$

$$D_\alpha(s_k,\alpha_p,t)=\begin{cases}1 & si\ G(s_k,\alpha_p,t)>th_\alpha\\ 0 & en\ \ otro\ \ caso\end{cases} \tag{5}$$

1) Imagen de la Energía del Movimiento

Las imágenes de energía del movimiento $E_\theta(s_k,\theta_p,t)$ y $E_\alpha(s_k,\alpha_p,t)$ para $G(s_k,\theta_p,t)$ y $G(s_k,\alpha_p,t)$, están definidas por las ecuaciones (6) y (7) respectivamente. Las MEIs resultantes son brillantes en los puntos en los que las neuronas se han activado en cualquier instante de tiempo.

$$E_\theta(s_k,\theta_p,t)=\bigcup_{\tau=0}^{t}D_\theta(s_k,\theta_p,\tau) \tag{6}$$

$$E_\alpha(s_k,\alpha_p,t)=\bigcup_{\tau=0}^{t}D_\alpha(s_k,\alpha_p,\tau) \tag{7}$$

2) Imagen de la Historia del Movimiento

Sea $H_\theta(s_k,\theta_p,t)$ la imagen de la historia del movimiento para $G(s_k,\theta_p,t)$ y $H_\alpha(s_k,\alpha_p,t)$ para $G(s_k,\alpha_p,t)$, definidos por las ecuaciones (8) y (9) respectivamente. En esta representación, los puntos más brillantes corresponden a aquellos puntos que se activaron reiterativamente en cualquiera de los instantes de tiempo.

$$H_\theta(s_k,\theta_p,t)=\mathrm{H}_\theta(s_k,\theta_p,\mathrm{t}\text{-}1)+1 \quad si\ D_\theta(s_k,\theta_p,t)=1 \tag{8}$$

$$H_\alpha(s_k,\alpha_p,t)=\mathrm{H}_\alpha(s_k,\alpha_p,\mathrm{t}\text{-}1)+1 \quad si\ D_\alpha(s_k,\alpha_p,t)=1 \tag{9}$$

$H_\theta(s_k,\theta_p,t)$ y $H_\alpha(s_k,\alpha_p,t)$ son una representación del movimiento durante toda su ejecución, contienen mucha información redundante y presentan una alta dimensionalidad. Para reducir la carga computacional y facilitar el reconocimiento, se disminuyó la dimensión de los datos tal como se explica a continuación.

D. Disminución de la Dimensionalidad

Se puede preservar la información más relevante de las diferentes respuestas neuronales y, al mismo tiempo obtener una codificación de baja resolución, calculando un histograma conformado por el promedio de las respuestas de cada una de las neuronas. Estos valores dependen de la actividad de las neuronas individuales en el tiempo, e intuitivamente, pueden relacionarse con las probabilidades de ocurrencia de un movimiento a una velocidad y dirección dada, y de un movimiento complejo determinado a una determinada dirección.

Matemáticamente, sea h_θ el histograma para $H_\theta(s_k,\theta_p,t)$, y h_α el histograma para $H_\alpha(s_k,\alpha_p,t)$ de acuerdo con las ecuaciones (10) y (11).

$$h_\theta(s_{k=i},\theta_{p=j},t)=\frac{1}{M}\sum H_\theta(s_i,\theta_j,t) \tag{10}$$

$$h_\alpha(s_{k=i},\alpha_{p=j},t)=\frac{1}{M}\sum H_\alpha(s_i,\alpha_j,t) \tag{11}$$

Donde M corresponde al tamaño de cada respuesta neuronal integrada en el tiempo $H_\theta(s_k,\theta_p,t)$ o $H_\alpha(s_k,\alpha_p,t)$. El mismo procedimiento se realiza sobre las imágenes MEI para obtener una representación de menor dimensión.

Esta representación puede disminuirse aún más aplicando una técnica como el Análisis de Componentes Principales [12] – PCA sobre los histogramas. Sin embargo, es buena idea normalizar los datos para que tengan media de cero y desviación estándar uno.

En su trabajo, [12] muestra que un vector r_m puede representarse de forma exacta como la suma ponderada de todos los eigen-valores de un conjunto de vectores.

Para aplicar PCA en este trabajo, se usan 36 vectores de entrenamiento para crear una matriz f, en donde cada columna f_m corresponde a los histogramas normalizados h_θ y h_α extraídos de los vídeos de entrenamiento. Ya que los eigen-vectores corresponden a las variaciones más significativas dentro del conjunto de entrenamiento; r_m puede ser aproximado por los l primeros términos en la suma, tal como se expresa en (12).

$$r_m\approx\sum_{i=1}^{l}f_{m_i}e_i+c \tag{12}$$

donde c es el promedio del conjunto de vectores (cero si los vectores están normalizados).

El tamaño inicial de los vectores usados para el reconocimiento es de 72x1 (36x1 para MHI y 36x1 para MEI) y se puede usar el porcentaje de información que se desea conservar como parámetro para elegir el valor de l. Así, por

ejemplo, si se mantienen seis componentes principales se conserva 71.15% de la información, mientras que con 19 se conserva el 90.05%. Estos vectores reducidos, son los que finalmente se utilizaron en el reconocimiento de gestos.

IV. APLICACIÓN AL RECONOCIMIENTO DE GESTOS

La codificación de movimiento anteriormente descrita fue aplicada al reconocimiento de 4 gestos: El gesto 1 corresponde rotar la mano en sentido inverso a las manecillas del reloj y devolverse (saludar), el gesto 2 corresponde a bajar y subir la mano (abanicar), el gesto 3 corresponde a rotar la mano en sentido inverso a las manecillas del reloj, y el gesto 4 corresponde a acercar y alejar la mano respecto a la cámara.

Para el reconocimiento de gestos, se implementaron las siguientes técnicas estándar de reconocimiento de patrones, en donde se trató de probar al menos con una técnica dentro de cada clase de clasificador, así: Clasificadores paramétricos (clasificador bayesiano), clasificadores no paramétricos (el vecino más próximo, los *k*-vecinos y distancia mínima al centroide) y redes neuronales artificiales (redes neuronales probabilísticas y redes neuronales perceptron multicapa).

1) Clasificador Bayesiano

Este clasificador paramétrico supone que los datos poseen una distribución de probabilidad Gausiana. Para determinar los parámetros del modelo normal, se usó el principio de máxima verosimilitud que pretende encontrar los valores óptimos que maximicen una función de verosimilitud derivada de los datos de entrenamiento. A través de estas ecuaciones es posible estimar la probabilidad a priori $P(x|C_i)$, esto es la probabilidad de que un dato de entrada $\boldsymbol{x}$ se presente dado que pertenece a la clase C_i. La probabilidad de mala clasificación es minimizada seleccionando la clase C_i que tiene la mayor probabilidad *a-posteriori* $P(C_i|x)$, tal que el dato de entrada $\boldsymbol{x}$ es asignado a la clase C_i si,

$$P(C_i|x) > P(C_j|x) \quad \forall_{i\neq j} \tag{13}$$

(13) es equivalente a (14)

$$P(x|C_i)P(C_i) > P(x|C_j)P(C_j) \quad \forall_{i\neq j} \tag{14}$$

2) Clasificadores No Paramétricos

Las reglas de decisión no paramétricas son atractivas por no requerir algún tipo de conocimiento *a-priori* sobre la distribución de los datos. En su lugar, se requiere un conjunto de entrenamiento representativo con datos etiquetados con la clase a la que pertenecen.

a) El vecino más próximo:

Consiste en comparar un vector de entrada desconocido con todos los vectores de entrenamiento que corresponden a diferentes gestos, y escoger aquel con la distancia más pequeña $d^{(i)}$. Este método es un caso particular del método de *k*-vecinos que se explicará más adelante. Si se usa como métrica la distancia euclidiana, la regla de decisión queda definida matemáticamente por (15).

$$d^{(i)} = \arg\min \left\| x - f^{(i)} \right\| \tag{15}$$

donde $\boldsymbol{x}$ es el vector de entrada proyectado en el eigen-espacio, y $f^{(i)}$ es el conjunto de muestras de entrenamiento también proyectadas.

b) k-vecinos:

Este método es un caso particular de la clasificación MAP (máximo *a-posteriori*) con funciones de densidad de probabilidad estimadas por el método de Parzen con ventana adaptativa [13]. Los *k*-vecinos es un modelo basado en la memoria definida por un conjunto de ejemplos debidamente etiquetados. La regla de los *k*-vecinos se basa en la suposición de que los prototipos más cercanos tienen una probabilidad *a-posteriori* similar. Las desventajas de este método radican en que la estimación resultante no es una función de densidad de probabilidad verdadera y que los datos de entrenamiento deben mantenerse almacenados.

Para clasificar un dato de entrada $\boldsymbol{x}$, deben encontrarse los *k*-patrones de entrenamiento más próximos a dicho dato y determinar la clase más votada; esta será la clase que se le asignará a dicho dato. La proximidad puede ser medida entre otras, con la distancia euclidiana o la distancia mahalanobis.

La elección de *k* es esencial en este método pues determina la calidad de las predicciones. Una forma apropiada de mirar el número *k* de vecinos cercanos, es verlo como un parámetro de suavizado. Un valor pequeño permitirá una gran varianza en la predicción, mientras que un valor grande conducirá a un modelo muy sesgado. Así, *k* debe ser un valor lo suficientemente grande para minimizar la probabilidad de mala clasificación y lo suficientemente pequeño para que los *k* puntos más próximos estén cerca del punto en cuestión. Por otro lado, la elección de un número grande de vecinos cuando se cuenta con pocos patrones de entrenamiento lleva a que la decisión no se tome con base en la concentración local de patrones de entrenamiento sino en propiedades globales con poco significado para la decisión.

c) Mínima distacia al centroide:

Este método es una simplificación de la clasificación bayesiana con funciones Gausianas de probabilidad, donde se asume que todas las matrices de covarianza de clase son diagonales e iguales y que las probabilidades *a priori* de clase también lo son. Los resultados obtenidos por este algoritmo

son buenos cuando las clases forman nubes de datos poco dispersas y bien separadas.

El método es relativamente rápido y uno de los más simples. La regla de decisión consiste en asignar al dato de entrada a la clase cuyo centroide sea el más cercano. Al igual que en casos anteriores, hay varias formas de medir dicha distancia, pero la más empleada es la distancia euclidiana. Así, la función de costo que debe ser minimizada está dada por (16).

$$d^{(i)} = \arg\min_i \left\| x - \hat{x}_i \right\| \tag{16}$$

en donde $\hat{X}_i$ corresponde a los centroides de cada una de las clases *i*.

3)Redes Neuronales

a) Redes neuronales probabilísticas:

La red neuronal probabilística (*Probabilistic Neural Network* – PNN) [14] se basa en el cálculo de la probabilidad *a-posteriori* (método de máxima verosimilitud). Para ello, se realiza un modelado Gausiano de los vectores de entrada, centrados en las muestras de entrenamiento. La PNN no tiene una verdadera etapa de entrenamiento sino que recorre todo el conjunto clasificación en cada ejecución, lo que da lugar a un método lento. La ventaja de estas redes es que producen altos porcentajes de clasificación usando un conjunto de datos mucho menor que los que requerirían otras redes neuronales como el perceptron.

Las funciones discriminantes de la clase *i-ésima* para una PNN están definidas por (17). Dado un patrón de entrada, se calculan dichas funciones discriminantes (una por clase). La que resulte máxima determinará la clase a la cual se asigna el dato.

$$D_i(x) = \frac{p(C_i)}{N_i} \sum_{j=1}^{N_i} \exp\left[-(1/2\sigma^2) d^2(x, x_j^{(i)}) \right] \tag{17}$$

donde $d^2(x, x_j^{(i)})$ es la distancia euclidiana entre las características del patrón de entrada y la de la muestra de entrenamiento *j-ésima* para la clase *i-ésima*. N_i es el número de muestras para cada clase *i* y $p(C_i)$ es la probabilidad a priori de clase *i*. s es un parámetro a determinar de manera heurística.

b) Red Perceptron Multicapa (MLP)

El perceptron multicapa es un aproximador universal y quizás la red neuronal supervisada más usada. Esta es una red de varias capas, usualmente tres: la capa de entrada, la capa oculta y la capa de salida. La capa de entrada está constituida por aquellas neuronas que introducen los patrones de entrada a la red; en estas neuronas no se produce procesamiento. La capa oculta utiliza como función de transferencia funciones sigmoidales, mientras que las funciones de la capa de salida pueden ser lineales o sigmoidales. La inclusión de varias capas le permite resolver problemas que no son linealmente separables.

La principal característica de este tipo de redes es el uso de la función de aprendizaje de Retropropagación hacia atrás o Regla delta generalizada. La regla delta trabaja de dos maneras: aprendizaje por lotes o aprendizaje en serie. El aprendizaje por lotes acumula las variaciones de los pesos y al final de cada ciclo, actualiza a la vez todos los pesos. El aprendizaje en serie va actualizando los pesos cada vez que se presenta un nuevo dato, lo que le da mayor velocidad, pero respetando el orden de presentación de las entradas.

La codificación de movimiento anteriormente descrita fue aplicada al reconocimiento de 4 gestos: El gesto 1 corresponde rotar la mano en sentido inverso a las manecillas del reloj y devolverse (saludar). El gesto 2 corresponde a bajar y subir la mano (abanicar). El gesto 3 corresponde a rotar la mano en sentido inverso a las manecillas del reloj. El gesto 4 corresponde a acercar y alejar la mano respecto a la cámara.

La base de datos empleada contiene 70 secuencias de vídeo grabadas en el laboratorio para cada uno de los cuatro gestos, de las cuales, 35 fueron usadas para el entrenamiento y las 35 restantes para la validación.

Para el reconocimiento de gestos se usaron diferentes clasificadores estándar: Paramétricos (clasificador Bayesiano), No-paramétricos (el vecino más próximo, los *k-vecinos* y distancia al centroide), y Redes neuronales (probabilísticas y perceptron multicapa). La Tabla 1 resume los resultados de clasificación con dichas técnicas de aprendizaje si se conserva el 90% de la información; esto es, si se usan 19 componentes principales.

V. RESULTADOS EN EL RECONOCIMIENTO DE GESTOS

Los resultados del reconocimiento se presentan en forma resumida a través de la Tabla 1 para cada una de las técnicas de reconocimiento de patrones presentadas en la sección anterior. La base de datos usada se conformó a partir de 280 vídeos, 70 para cada uno de los cuatro gestos; la mitad de esta base de datos fue usada durante la fase de entrenamiento y las 140 restantes para validación.

Los resultados para el caso de la red MLP se obtuvieron con una configuración de 32 neuronas en la capa de entrada con funciones de activación 'tansig', 4 neuronas en la capa de salida con funciones de activación 'logsig' y mediante el algoritmo de entrenamiento de propagación hacia atrás de gradiente conjugado.

El menor porcentaje de clasificación (88.57%) se obtuvo usando 6 componentes principales junto con la técnica de clasificación de distancia mínima al centroide; que es un buen porcentaje de reconocimiento, especialmente si se tienen en cuenta las suposiciones restrictivas iniciales respecto a las clases. Mientras que el mayor porcentaje de clasificación (95%) se obtuvo usando los mismos 6 componentes principales junto con las redes neuronales bayesianas.

Por otro lado, y coherente con lo esperado, la técnica de los *k* vecinos definió mejor la distribución de los datos que la del vecino más próximo. En cuanto a las redes neuronales, se obtuvo un mejor reconocimiento con la red neuronal probabilística que con la red neuronal perceptron multicapa.

Los buenos porcentajes de reconocimiento obtenidos en todos los casos, permiten afirmar que las características extraídas de las imágenes son bastante discriminantes.

Usando 19 componentes, los mejores resultados se obtuvieron con las redes MLP (97.14%), y los peores, al igual que en el caso anterior, con la mínima distancia al centroide (87.88%). Con las redes PNN y el algoritmo del vecino más próximo se obtuvieron porcentajes de reconocimiento iguales (95%); valor que corresponde al porcentaje más alto obtenido usando sólo 6 componentes principales. Mientras que la técnica de los *k*-vecinos mantuvo el mismo porcentaje de buena clasificación respecto a la obtenida con el uso de 6 componentes.

Tabla 1. Resumen de los porcentajes de clasificación para las seis técnicas de reconocimiento de patrones usadas

Técnica \ % de aciertos	6 PCA	19 PCA
Bayesianas	95.00	92.14
Vecino más próximo	91.42	95.00
k-vecinos (*k*=5)	94.29	94.29
Mínima distancia	88.57	87.88
PNN (σ=0.3)	93.57	95.00
MLP	92.14	97.14

Las técnicas de Bayes y Mínima Distancia presentaron una leve disminución en sus porcentajes, que pudo ser ocasionada por el ruido introducido por el mayor número de componentes en el vector.

En general, al usar más componentes principales, y por lo tanto, conservar un mayor porcentaje de información, los porcentajes de reconocimiento mejoraron, aunque sin mantener el mismo orden.

En la Tabla 2 y 3 permiten detallar los resultados en el reconocimiento empleando 6 y 19 componentes respectivamente, mediante matrices de confusión promedio. Lo ideal, es que la diagonal de las tablas tenga valores iguales a 1 y el resto de sus celdas sean cero. El valor de 0.152 en la Tabla 2 corresponde al mayor error de reconocimiento para el caso de seis componentes, e indica que el Gesto 4 (fila) ha sido confundido en promedio el 15.22% de las veces con el Gesto 2 (columna). En contraposición, el gesto que mejor reconoce el sistema es el Gesto1 (96.2% de las veces).

Tabla 2. Matriz de confusión promedio para el caso de 6 PCA

	Gesto 1	Gesto 2	Gesto 3	Gesto 4
Gesto 1	0.962	0.010	0.014	0.014
Gesto 2	0.010	0.952	0.000	0.038
Gesto 3	0.057	0.000	0.943	0.000
Gesto 4	0.000	0.152	0.005	0.843

Tabla 3. Matriz de confusión promedio para el caso de 19 PCA

	Gesto 1	Gesto 2	Gesto 3	Gesto 4
Gesto 1	0.957	0.019	0.024	0.000
Gesto 2	0.019	0.924	0.000	0.057
Gesto 3	0.024	0.000	0.976	0.000
Gesto 4	0.000	0.114	0.000	0.886

Cuando se usó 19 componentes, el gesto que mejor reconoce el sistema corresponde al Gesto3, y, al igual que en el caso anterior, la mayor error ocurrió el 11.4% de las veces, en donde el Gesto 4 fue erróneamente identificado como Gesto 2. Esta confusión se debe al hecho a que el ángulo entre los vectores de flujo óptico y el gradiente de la magnitud es similar para ambos gestos, y se diferencian más por la dirección del movimiento, la cual no siempre es suficientemente discriminante.

VI. CONCLUSIONES Y TRABAJO FUTURO

Se presentó un estudio sobre el desempeño de seis técnicas de clasificación de patrones para el reconocimiento de cuatro gestos realizados por manos. Las técnicas estudiadas abarcan un amplio espectro: paramétricas, no-paramétricas y de inteligencia artificial.

El enfoque de utilizar la bio-inspiración para extraer primitivas de movimiento que se utilizan como características para discriminar entre gestos, es novedoso. Para los casos analizados, se alcanzó un porcentaje de reconocimiento que oscila entre 87.88% (distancia mínima al centroide) y 97.14% (MLP), en ambos casos conservando 19 componentes principales.

Los porcentajes de reconocimiento alcanzado son sensibles al cambio en el número de características utilizadas, sin embargo se encontró que existen algunos clasificadores que son más robustos a dicho cambio, como las técnicas de k-vecinos y Mínima Distancia.

Al conservar un mayor número de componentes principales, y, por lo tanto, de la información; las técnicas de MLP y el Vecino Más Próximo alcanzaron una mejora notable en los porcentajes de éxito en el reconocimiento alcanzados.

Los altos porcentajes de reconocimiento alcanzados en todos los casos indican que las características escogidas son bastante discriminantes, lo que le da al sistema robustez independientemente de la técnica de reconocimiento de patrones empleada para el reconocimiento de gestos. Esta

característica permite que la selección definitiva de la técnica y el número de componentes atienda a un compromiso entre la complejidad computacional y el desempeño deseado.

El siguiente desarrollo en esta línea de investigación se enfocará en la imitación de los gestos percibidos por el sistema de visión artificial y su acople a un brazo robótico. El trabajo futuro se orientará al reconocimiento de gestos compuestos generados por la mezcla de varios gestos básicos.

AGRADECIMIENTOS

Agradecemos al Programa de Apoyo a Doctorados Nacionales de Colciencias, a la Universidad del Valle y al Instituto Técnico Superior (IST) – Portugal, por el soporte a este trabajo. Un reconocimiento especial al profesor José Santos-Victor del IST – Portugal por su orientación, consejo y apoyo durante el desarrollo de este proyecto.

REFERENCIAS

[1] Lenman, S., Bretzner, L.y Thuresson, B., 2002. Computer Vision Based Recognition of Hand Gestures for Human-Computer Interaction. *Technical Report TRITA-NA-DO2209, CID-report*,

[2] Ling, J., Wu Y. y Huang, T.S., 2002. Modeling the Constraints of Human Han Motion. *Proceeding of the Wórkshop o Human Motion*, pp. 121-126.

[3] Chang, C-C., Chen, J-J., Tai W-K. y Han, C-C., 2006. New approach for Static Gesture Recognition. *Journal of Information Science and Engineering*, vol. 22, pp. 1047-1057.

[4] Hyeon-Kyu L. y Kim, J.H., 1999. An hmm-based threshold model approach for gesture recognition. *Pattern Analysis and Machine Intelligence*, vol. 21, pp. 961-973.

[5] Hong, P., Turk M. y Huang, T.S., 2000. Constructing finite state machines for fast gesture recognition. *Pattern Recognition*, vol. 3, pp. 391-694.

[6] Sandberg, A., s.a., Gesture recognition using neural networks. Master's Thesis, Stockholm University.

[7] DeVanois, R.L. Morgan, M.C. and Snodderly, D.M., 1974. Psychophysical studies of monkey vision. III. Spatial luminance contrast sensitivity test of macaque and human observers. *Vsion Research*, vol. 14, pp. 53-67.

[8] Wang, C. and Brandstein, M., 1999. Multi-source face tracking with audio and visual data. *IEEE MMSP*, pp. 169-174.

[9] Nope, S., Loaiza, H. y Caicedo, E., 2006. Review of Techniques for Motion Estimation in Artificial Vision. En *Revista Colombiana de Tecnologías de Avanzada*, vol. 2, pp. 102-108.

[10] Pomplun, M., Martinez-Trujillo, J., Simine, E., Liu, Y., Treue, S. y Tsotsos, J. K., 202. A Neurally-Inspired Model for Detecting and Localizing Simple Motion Patterns in Image Sequences. *Presented at Wórkshop on Dynamic Perceptio*n, Bochum, Germany.

[11] Bobick, A. F. and Davis, J. W., 2001. The Recognition of Human Movement using Temporal Templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 257-267.

[12] Oja, E., s.a. Subspace Methods of Pattern Recognition. *Research Studies Press*, Hertfordshire.

[13] Marques, J., 1999. Reconhecimrnto de *pa*trões: métodos estatísticos e neuronais. IST Press, Lisboa -Portugal, ISNN: 972-8469-08X.

[14] Blue, J.L., Candel, G.L., Grother, P.J. y Wilson, C.L., s.a. Evaluation of Pattern Classifiers for fingerprint and OCR applications. *Pattern Recognition*, vol. 27, pp. 485-501.

Este trabajo hace parte de la tesis doctoral de Sandra Esperanza Nope Rodríguez sobre una arquitectura de control basada en el aprendizaje por imitación de gestos aplicada en robótica, becaria Colciencias en el programa de apoyo a doctorados nacionales.

Sandra E. Nope R.: Ingeniera en Electrónica y Telecomunicaciones de la Universidad del Cauca. Estudiante de doctorado en Ingeniería, becaria Colciencias, sandrano@univalle.edu.co

Eduardo Caicedo Bravo: Doctor en Informática Industrial de la Universidad Politécnica de Madrid. Profesor de la Universidad del Valle, ecaicedo@univalle.edu.co

Humberto Loaiza Correa: Doctor en Robótica de la Université d'Evry, Francia. Profesor de la Universidad del Valle, hloaiza@univalle.edu.co