MANIPULACIÓN DE ROBOTS CON BASE EN POSTURAS LABIALES

LIP GESTURE-BASED ROBOT MANIPULATION

JUAN B. GÓMEZ

Universidad Nacional de Colombia Sede Manizales, jbgomezm@unal.edu.co

JORGE E. HERNÁNDEZ

Universidad Nacional de Colombia Sede Manizales, jehernandezl@unal.edu.co

FLAVIO PRIETO

Universidad Nacional de Colombia Sede Manizales, faprietoo@unal.edu.co

Recibido para revisar Abril 17 de 2007, aceptado Agosto 30 de 2007, versión final Septiembre 09 de 2007

RESUMEN: En este artículo se propone un nuevo método para el comando de tres grados de libertad de un robot manipulador, por medio de gestos de la boca. Las diferentes posiciones son registradas en secuencias de video, las cuales son procesadas y clasificadas en tiempo real. Se utilizan diferentes técnicas de procesamiento de imágenes a cada cuadro, para lograr la adecuada segmentación y caracterización del área de la boca. Posteriormente, se utiliza la información de clasificación en una máquina de estados que estabiliza la detección del gesto e interactúa con la interfaz de comando del robot, indicándole la operación a realizar. Los resultados obtenidos muestran que la metodología propuesta es altamente efectiva para aplicaciones en tiempo real, siendo lo suficientemente rápida y adecuada para la detección de las posturas seleccionadas.

PALABRAS CLAVE: Interfaz hombre-máquina, segmentación de la boca, detección de gestos.

ABSTRACT: In this paper, a novel approach for the mouth-gestures based command of three degrees of freedom of a robot is proposed. The different selected gestures are recorded in video sequences, which are processed and classified in real time. Several image processing techniques are applied in each frame, in order to achieve an appropriate feature extraction and classification of gestures. After that, the output of the classifier is used as the input of a state machine which stabilizes the command selection and sends the selected operation to the robot's command interface. The method shows to be very effective for real time applications, giving both enough speed and good gesture detection.

KEYWORDS: Human-machine interface, mouth segmentation, gesture detection.

1. INTRODUCCIÓN

La cirugía tradicional en laparoscopia requiere la ayuda de una persona para manipular el endoscopio, según las instrucciones del cirujano. Esta técnica de operación no es óptima porque el endoscopio se mueve constantemente, debido a los temblores de la mano del operador. Las órdenes del cirujano pueden ser mal interpretadas por el operador y, por lo tanto, mal ejecutadas.

Este problema puede ser resuelto desarrollando un Sistema de Posicionamiento del Laparoscopio por un brazo robotizado (SPRL). Es decir, un robot controlado directamente por el cirujano, manipula el laparoscopio [1]. Utilizando una interfaz cirujano-robot de alto nivel, el cirujano puede controlar por sí mismo el laparoscopio mediante la voz, una palanca de mando o mediante los movimientos de la cabeza.

Los SPRL que utilizan una interfaz basada en palancas de mando o en pedales requieren de la mano o del pie del cirujano para controlar la cámara. Estos tipos de interfaz no son de uso fácil, porque el cirujano ya tiene ocupadas sus manos y pies para controlar una gran variedad de instrumental quirúrgico. Algunos investigadores intentaron utilizar la voz, para desarrollar un sistema de posicionamiento para endoscopia robotizada [2]; estos sistemas presentan como inconveniente el ruido de fondo, el cual puede ser interpretado por el robot como órdenes. Por lo tanto, parece ser que la mejor manera de controlar un SPRL es mediante la utilización de los gestos de la cara. El sistema FAce MOUSe [3], es una interfaz basada en los movimientos del rostro, en el que una cámara fotográfica normal observa la cabeza del cirujano quien con movimientos intencionales de su cabeza, controla la posición y orientación del laparoscopio. De esta manera, el cirujano puede controlar un SPRL mediante los movimientos de su cabeza, sin ningún dispositivo especial. Sin embargo, parece más natural controlar el movimiento de un robot sólo con el movimiento de los labios.

Los movimientos del laparoscopio son restringidos a cuatro grados de libertad (GDL). Los primeros dos GDL son movimientos de inclinación perpendicular (*pan* y *tilt*) alrededor del punto de inserción del trocar, que permite la introducción del laparoscopio. El tercer GDL es el de inserción y retracción del laparoscopio, el cual corresponde al *zoom* de las imágenes. El último GDL, el de rotación del laparoscopio, siempre es evitado durante la operación quirúrgica, porque la observación de estas imágenes rotadas demanda esfuerzo mental adicional (muy importante) por parte de los cirujanos [4]. De esta manera, el Sistema Robotizado de Posicionamiento para Laparoscopio sólo requiere tres GDL. Los movimientos normales de la cabeza y de los labios, permiten reproducir estos tres GDL. Por supuesto, para este sistema, los labios del cirujano deben ser visibles por la cámara.

La consola de operación del Sistema Quirúrgico DaVinci [5] generalmente está ubicada a 3 metros del paciente; en esta consola, el cirujano no requiere tapabocas y por tanto puede utilizar sus labios para controlar la cámara del laparoscopio. Este control se realiza mediante una cámara de video normal que sigue el movimiento de los labios del cirujano. El movimiento del laparoscopio pude ser modelado por una máquina de estados, a partir de unas entradas definidas por la posición de los labios.

Los sistemas de visión artificial están compuestos por diversas etapas, desde el momento de la captura de las imágenes o secuencias, hasta la interpretación de los resultados. En general, se puede decir que dichos sistemas están compuestos por las siguientes etapas:

- Adquisición de las imágenes o secuencias de video.
- Pre-procesamiento de las secuencias.
- Segmentación de los objetos de interés.
- Caracterización de los objetos segmentados.
- Clasificación de los objetos.
- Interpretación de la información de la escena.

Existen en la literatura diversas técnicas de adquisición y de pre-procesamiento de la información; por este motivo, el objetivo del trabajo, en cuanto a visión artificial se refiere, se concentra en determinar estrategias adecuadas de segmentación, caracterización y clasificación de gestos bucales en tiempo real. Adicionalmente, la interpretación de los resultados se utiliza en una máquina de estados que genera la secuencia de comandos de operación de tres grados de libertad de un robot manipulador.

La estructura del documento es como se sigue: en la Sección 2 se presenta el trabajo en el campo de la manipulación de sistemas robotizados de cirugía utilizando diversos mecanismos, entre ellos la visión artificial. En la Sección 3 se describe el método utilizado para la segmentación del área de la boca en secuencias de video en tiempo real. En la Sección 4 se expone la estrategia de caracterización de la región segmentada de la boca. En la Sección 5 se muestra la máquina de estados utilizada en la clasificación de gestos bucales y la interacción con el robot. En la Sección 6 se muestran los resultados de las pruebas de operación del sistema. En la Sección 7 se concluye el trabajo, y se proponen los lineamientos de la continuación del mismo.

2. TRABAJO RELACIONADO

La etapa más crítica en el proceso de detección y clasificación de los gestos bucales en secuencias de video es la segmentación de la región de la boca. Una manera de acercarse al problema de segmentación de labios, es encontrando una transformación de espacio de color apropiada que refuerce la diferencia entre el área de los labios y la región del rostro. En este campo, se han desarrollado varios trabajos. En [6], se afirma que la componente roja es predominante en el área de la cara, en el espacio de color RGB y la separación entre la piel y los labios es más fácil de ver en la relación entre las componentes G y B. En [7], se presenta un nuevo conjunto de transformaciones no lineales compuestas desde el espacio de color YC_bC_r. Ellos muestran que la transformación no lineal puede mejorar significativamente el contraste entre el área de la boca y el resto de la cara.

En [8], se define una nueva transformación basada en el espacio de color RGB llamada el mapa de la curva cromática. Esa transformación refuerza la diferencia entre los labios y la piel, y permite la detección robusta del labio bajo condiciones de iluminación no uniformes y sin el uso de cualquier maquillaje en particular. La transformación se basa en el hecho de que la cantidad de verde en el área de la piel, comparada con la componente azul es más grande que en el área de los labios. En [9], se presenta un sistema automático para la lectura de los labios y la reproducción sintética de gestos y sonido. En este trabajo se utilizó una nueva transformación de espacio de color logarítmica HSV, y un análisis de vecindarios espacio-temporales para segmentar el área de los labios apropiada en las secuencias de video. En [10] se definen los umbrales de la componente H del espacio HSV que discrimina los labios del área de la piel.

Una fuente reiterativa en la segmentación de imágenes se fundamenta en las medidas de conectividad difusa. Estas técnicas permiten radios de detección muy altos en escenas en las cuales los bordes son poco definidos. Sin embargo, dichas técnicas suelen ser costosas computacionalmente, y por tanto se descartan para aplicaciones en tiempo real. En [11] se propone un método de segmentación difusa de los labios, basada en un multi-fondo y un esquema del objeto. Ellos utilizan una función de distancia dual que tenga una parte euclídea y una parte elíptica. Presentaron una función de costo que se deriva del algoritmo de conectividad difuso (FCM). Otro trabajo que utiliza FCM es el presentado en [12]; en éste, se utiliza un segmentador FCM basado en una representación en los espacios de color de CIELAB y de CIELUV. Una estimación iterativa del parámetro para las funciones de la calidad de miembro del proceso de FCM que utilizó, demuestran una buena convergencia en tres iteraciones.

En [13] se propone un método basado en un modelo estadístico de forma, con descriptores Gaussianos de apariencia local. Se muestra que, en algunos casos, la respuesta de los descriptores locales puede predecir la forma. Esta predicción se logra por medio de una red neuronal artificial no lineal.

En la interfaz Hombre-Robot [14], se presentó un sistema basado en el reconocimiento de la cara y los gestos. Se usó la información de la cara y de las manos (movimientos de los dedos) como las entradas en una regla de decisión. La segmentación de la piel se realizó usando la representación de color YIQ. Los comandos al robot se enviaron por la red TCP/IP. En [15] y [16] se propuso un sistema robotizado para la interacción de un operador humano. La interfaz del hombre-robot es un sistema basado en visón para lograr una interacción natural entre el operador y el robot. El sistema de visión encuentra y sigue el rostro de los operadores, reconoce los gestos faciales y determina la mirada fijamente del usuario. En ambos trabajos, el tiempo real no se tuvo en cuenta para el desarrollo de los algoritmos.

3. SEGMENTACIÓN DE GESTOS BUCALES EN SECUENCIAS DE VIDEO

La primera etapa antes del proceso de extracción de características en secuencias de video faciales es la segmentación de los labios y la boca. El proceso se muestra en la Figura 1.

A pesar de la cantidad de trabajos en el área de reconocimiento automático de la región de la boca en imágenes y video, el compromiso entre la

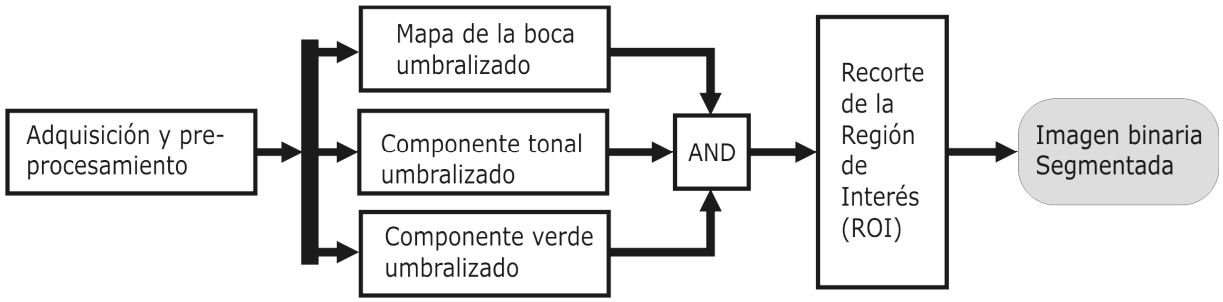


Figura 1. Proceso de Segmentación de los Labios.
Figure 1. Lips segmentation process.

calidad y la velocidad de operación alcanzada, no suple las necesidades de sistemas de misión crítica en tiempo real. Además, presentan inconvenientes cuando el sujeto en las imágenes posee vello facial y/o es de tez oscura. Por éste motivo, en el presente artículo se plantea una metodología diferente, que probó ser robusta en la presencia de vello facial, y cuyo rendimiento es adecuado para el uso en sistemas en tiempo real.

La primera fase consiste en la descomposición de la información de la imagen en tres componentes de color: el tono (H), la componente verde (G) y la transformación de color llamada *mapa de la boca,* propuesta en [7]. En una segunda fase, las tres componentes son binarizadas y mezcladas por medio del operador de conjunción AND. El resultado es la exclusión (o segmentación) adecuada del área de la boca, y unas pocas regiones residuales de tamaño mucho menor al de la boca.

Finalmente, se realiza un recorte de la región de interés (ROI, por sus siglas en inglés), delimitada por medio de una elipse que envuelve el área de la boca. Esta elipse es calculada de forma dinámica mediante el uso de la información proveniente de iteraciones anteriores del algoritmo. En la primera iteración, la búsqueda de la región de interés se realiza en toda la imagen. A continuación se explican en detalle el método propuesto para la segmentación de la boca.

3.1 Extracción del área de la boca

La primera componente de color utilizada para resaltar la información del área de la boca es la componente verde (G) del espacio de color *RGB*.

En las imágenes que poseen información de piel y labios, la información de la componente verde es una característica discriminante entre ellas. Para mejorar el contraste entre las regiones se realizó una expansión dinámica de la componente verde.

La siguiente componente es el *mapa de la boca,* presentada en [7]. Esta componente nace de una transformación no lineal del espacio de color YC_bC_r. La expresión que resalta la componente de la boca está descrita por la Ecuación (1).

$$f\left(C_b, C_r\right) = C_r^2\left(C_r^2 - \eta\frac{C_r}{C_b}\right)^2 \qquad (1)$$

Donde: C_r^2 y C_r / C_b son normalizados en el rango de [0, 255], y η es la relación promedio entre C_r^2 y C_r / C_b. Una vez que la componente es calculada, ésta se normaliza en el rango [0,255].

La tercera componente es el tono (*H*) del espacio de color *HSV*. En el trabajo de Eckert [10], se proponen ciertos umbrales fijos en la componente de tono que resaltan de forma adecuada el área de la boca en imágenes faciales, representadas en la componente tonal. Estos umbrales mostraron ser adecuados para ubicar la boca, pero en algunos casos demasiado exclusivos y con tendencia a eliminar el borde externo de los labios. Por lo tanto, dichos valores fueron tomados como base para la binarización de la componente tonal, adicionando un desplazamiento de 2 unidades en cada sentido, y dando como resultado una ampliación en la banda de selección.

En la componente verde se usó una binarización adaptativa basada en la información estadística de

la media $\left(\mu_{g\exp}\right)$ y la varianza $\left(\sigma_{g\exp}\right)$ de la imagen. Resultados experimentales mostraron que, en imágenes en las cuales no aparecen los dientes ni exceso de barba, el umbral definido por

$$\mu_{g\exp} - 1.5\sigma_{g\exp} \le mouth \le \mu_{g\exp} + 1.7\sigma_{g\exp} \quad (2)$$

resalta de forma adecuada la región de la boca. Cuando aparecen dientes o barba, la región segmentada en la componente verde resulta ser laxa y por tanto la exclusión queda en manos de las otras dos componentes.

En la componente de mapa de la boca se utilizó un umbral con la información de media $\left(\mu_{f(C_r,C_b)}\right)$. Finalmente, el rango dinámico de la binarización del mapa de boca está definido por

$$mouth \le \mu_{f(C_b,C_r)} \quad (3)$$

La conjunción de las tres imágenes binarias, produce una imagen que resalta la zona de los labios y el interior de la boca.

3.2 Recorte de la región de interés (ROI)

Con base en la imagen binaria generada mediante la binarización de las tres componentes y su conjunción, se aplicó una condición de recorte elíptico, la cual restringe la región de interés para la búsqueda de la boca en la próxima iteración. Además reduce la presencia de regiones residuales en el resto de la imagen. Cualquier píxel que se encuentre fuera de la zona elíptica hallada es descartado y no pertenece a la región de interés.

El cálculo iterativo de la zona de recorte elíptico, se realiza con base en las características detectadas de la región de la boca, y se explica en la Subsección 4.2.

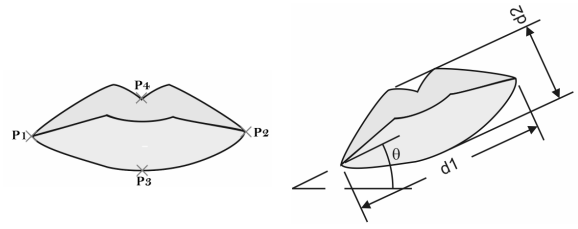
3.3 Resultados de la segmentación

La Figura 2 muestra algunos resultados de la segmentación para distintos sujetos. Los mejores resultados fueron obtenidos para rostros pálidos y sin barba. El filtro elíptico reduce el problema del ruido, que fue introducido por el componente de tono, pero hace que el sistema se torne inestable y pierda la boca con facilidad. La inestabilidad es debida a las condiciones de expansión-compresión usadas para adaptar la elipse sobre el tiempo, y el ruido que aparece cerca del área de la boca. Los primeros dos sujetos muestran problemas de segmentación en el labio superior, debido principalmente a la proyección de sombras por parte de la nariz sobre dicha zona.

Figura 2. Ejemplos de segmentación.
Figure 2. Snapshots of the segmentation process.

4. EXTRACCIÓN DE CARACTERÍSTICAS

Para la caracterización de la boca se implementó un algoritmo que realiza la búsqueda de los cuatro puntos característicos: las dos esquinas horizontales de la boca (izquierda y derecha) y las dos esquinas verticales de la boca (superior e inferior), como se muestra en la Figura 3(a).



(a) Esquinas de la boca.

(b) Medidas de labio tomadas en cuenta para el proceso de clasificación

Figura 3. Características y métricas de la Boca.
Figure 3. Mouth landmarks and measurements.

La búsqueda de los puntos se realiza dentro de la caja que limita la boca. Esta caja se encuentra sumando los píxeles blancos en cada uno de los ejes (X y Y), comenzando desde los límites de la imagen. En cada caso la selección se realiza sobre un umbral de la suma de los píxeles, en la cual la primera ocurrencia se tiene como referencia de la fila y columna correspondiente. Finalmente, las referencias son movidas proporcionalmente según los rangos verticales y horizontales, con el fin de cubrir el área completa de la boca.

El cálculo de la boca se realiza solamente en la primera iteración. En las siguientes iteraciones la búsqueda de los puntos se hace en el vecindario de los puntos encontrados en las iteraciones anteriores y que cumplan la condición elíptica. La búsqueda de los puntos se efectuó utilizando la información de la mejor recta que caracteriza la boca. Los parámetros: pendiente y punto de corte, de la línea recta son calculados utilizando una regresión lineal con todos los puntos que forman el área de la boca segmentada. Sin embargo, el valor de la pendiente calculada por la regresión es promediado con el valor obtenido de la pendiente entre los puntos p_1, p_2. Después del cálculo de la nueva recta, se determinó la línea recta perpendicular que pasa por el punto medio entre p_1 y p_2. En el siguiente paso, los puntos son calculados de forma convencional y se verifica la proximidad de los puntos p_1 y p_2 con la línea recta horizontal y de los puntos p_3 y p_4 con la línea recta vertical, respectivamente. Cuando la distancia de alguno de los puntos sobrepasa un umbral, el punto es proyectado sobre su respectiva recta y se mueve sobre ella hasta el corte con la boca. Utilizando las comisuras encontradas, se procede con la extracción de las características. En este trabajo se propuso el uso de dos índices de apertura de la boca, y otro para la rotación angular de la cara.

4.1 Cálculo de los índices de apertura y cierre mediante lógica borrosa

Para el cálculo de los índices, se usó un sistema de inferencia borroso (FIS) con una variable de entrada con dos funciones de membresía (μ_{Th} y μ_{Op}), y dos variables de salida (i_{Th} y i_{Op}). La variable de entrada $\gamma(k)$ es escogida de la relación $\gamma(k) = d_2(k)/(d_1(k)+d_2(k))$ (ver Figura 3(b)), y las salidas son el grado de *delgadez* (i_{Th}) y *apertura* (i_{Op}) del gesto de la boca. La Figura 4 muestra los conjuntos difusos seleccionados para la variable borrosa de entrada $\gamma(k)$. La forma de las funciones de membresía de los conjuntos difusos de entrada está descrita en la Ecuación 4.

Las posiciones Γ_{Op} y Γ_{Th}, en la variable de apertura, son puntos que definen el valor medio de apertura y delgadez para los gestos de boca abierta y boca delgada respectivamente. Dichos valores pueden ser obtenidos a partir de un análisis estadístico de la media en un conjunto de

$$\mu_{Th}(\gamma) = \begin{cases} 1, & si \quad 0 \le \gamma < \Gamma_{Th} - \frac{\Delta_{Th}}{2} \\ 1 - \frac{2}{\Delta_{Th}^2}\left(\gamma - \Gamma_{Th} + \frac{\Delta_{Th}}{2}\right)^2, & si \quad \Gamma_{Th} - \frac{\Delta_{Th}}{2} \le \gamma < \Gamma_{Th} \\ \frac{2}{\Delta_{Th}^2}\left(\gamma - \Gamma_{Th} - \frac{\Delta_{Th}}{2}\right)^2, & si \quad \Gamma_{Th} \le \gamma < \Gamma_{Th} + \frac{\Delta_{Th}}{2} \\ 0, & si \quad \Gamma_{Th} + \frac{\Delta_{Th}}{2} \le \gamma \end{cases}$$

$$\mu_{Op}(\gamma) = \begin{cases} 0, & si \quad 0 \le \gamma < \Gamma_{Op} - \frac{\Delta_{Op}}{2} \\ \frac{2}{\Delta_{Op}^2}\left(\gamma - \Gamma_{Op} + \frac{\Delta_{Th}}{2}\right)^2, & si \quad \Gamma_{Op} - \frac{\Delta_{Op}}{2} \le \gamma < \Gamma_{Op} \\ 1 - \frac{2}{\Delta_{Op}^2}\left(\gamma - \Gamma_{Op} - \frac{\Delta_{Op}}{2}\right)^2, & si \quad \Gamma_{Op} \le \gamma < \Gamma_{Op} + \frac{\Delta_{Op}}{2} \\ 1, & si \quad \Gamma_{Op} + \frac{\Delta_{Op}}{2} \le \gamma \end{cases} \tag{4}$$

estudio. El valor de $\Gamma \pm \Delta/2$ (*a* y *b* en la Figura 4), define el intervalo de incertidumbre con respecto a las posiciones de apertura y delgadez. Nótese que, para el primer o segundo caso, si el valor Δ tiende a 0, los conjuntos difusos asociados tienden a ser conjuntos clásicos. Por otro lado, si los valores difusos son grandes tendrán un área de superposición grande. El conjunto difuso es complementario si los dos valores de caída (Δ_{Th} y Δ_{Op}) tienen un valor igual a $(\Gamma_{Op} - \Gamma_{Th})$. Seleccionando el método de inferencia por centro de masa, el valor de los

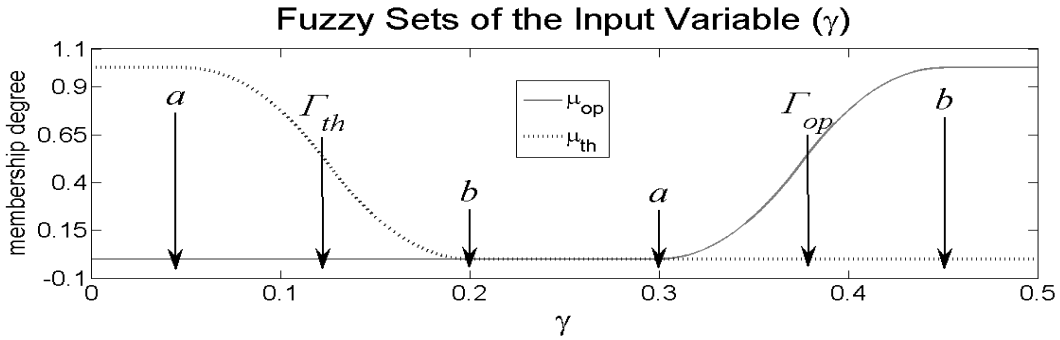


Figura 4. Comportamiento del Sistema de Inferencia Borrosa en términos de γ.
Figure 4. Fuzzy inference system behavior shown in terms of γ.

índices de apertura y de delgadez puede ser calculado como se muestra en la Ecuación 5.

$$i_{Th}(\gamma)=\frac{\mu_{Th}(\gamma)}{\mu_{Th}(\gamma)+\mu_{Op}(\gamma)}$$
$$i_{Op}(\gamma)=\frac{\mu_{Op}(\gamma)}{\mu_{Th}(\gamma)+\mu_{Op}(\gamma)} \quad (5)$$

i_{Th} es igual a 1 si el gesto de la boca está con "labios escondidos", y decrece rápidamente, hasta que alcanza el valor de 0. i_{Op} es igual a 1 si la boca se abre considerablemente, y decrece rápidamente hacia 0 cuando el gesto tiende a un estado normal de la boca o estado de reposo. Nótese que los valores de Δ_{Th} y Δ_{Op} tienen un control indirecto sobre las pendientes de caída y crecimiento de los índices, como se muestra en la Figura 4.

Otra característica que se usó, es el ángulo entre el eje horizontal de la imagen y el eje principal de la boca (identificado como θ en la Figura 3b). Esta característica, ayuda a determinar la rotación de la boca.

4.2 Cálculo iterativo de la región de interés por medio de recorte elíptico

La elipse de recorte que rodea la región de interés se puede parametrizar a través de su centro p_{CM} y la diagonal mayor y menor (r_1, r_2); dos vectores ortonormales $(\vec{k},\vec{l})$ sirven a su vez como bases del sistema rotado de referencia de la elipse. Estos parámetros son calculados gracias a la información proveniente de la inicialización del algoritmo, en la cual se realiza la detección de los puntos base sin tener en cuenta el recorte. El centro de la elipse es calculado como el centro de masa de la boca en la imagen actual. Los vectores normales y la distancia de cada eje se calculan a partir de las comisuras $\vec{p}_1$ y $\vec{p}_2$, como muestra en la Ecuación (6) y en la Ecuación (7).

$$\vec{k}=\frac{\vec{p}_2-\vec{p}_1}{\|\vec{p}_2-\vec{p}_1\|};\quad \vec{l}=\begin{bmatrix}0 & 1\\ -1 & 0\end{bmatrix}\vec{k} \quad (6)$$

$$\begin{cases} r_1 = 1.2\times\max\left(\|p_1-p_{CM}\|,\|p_2-p_{CM}\|\right)\\ r_2 = 1.2\times\min\left(\max\left(\|p_4-p_{CM}\|,\|p_3-p_{CM}\|\right),r_1\right)\end{cases} \quad (7)$$

Los valores calculados son utilizados para determinar la ROI, la cual está definida por la Ecuación (8).

$$ROI:\left\{x,y\middle|r_1^2<u^2+(\phi v)^2\right\} \quad (8)$$

La relación elíptica ϕ define la relación de aspecto de la elipse, y una matriz de transformación bidimensional traslada las coordenadas de la imagen (x y y) a las

coordenadas de la elipse (u y v). Esta transformación ajusta la rotación y la traslación apropiada, tal que el centro de la elipse localizado en ($u = 0$, $v = 0$), corresponda al centro de masa del área de la boca en la iteración anterior, y el eje principal de la elipse tenga la misma pendiente del eje principal de la boca en la iteración actual.

4.3 Resultados de la extracción de características y la detección inicial de los gestos

La Tabla 1 muestra las medidas de las características obtenidas en las imágenes de la Figura 2.

Con la información obtenida de los puntos se pueden calcular los índices de apertura y de cierre, que sirven para determinar de forma inicial el gesto que se observa. Para efectos interpretativos, se toman como gestos iniciales la boca delgada, que corresponde a i_{th}=1, la boca abierta, que corresponde a i_{op}=1, y se considera como "otro" cualquier combinación diferente.

Tabla 1. Valores calculados de las características para la Figura 2.
Table 1. Calculated feature values for Figure 2.

Rostro	$\left\|(\vec{p}_1 - \vec{p}_2)\cdot \hat{i}_x\right\|$	$\left\|(\vec{p}_1 - \vec{p}_2)\cdot \hat{j}_y\right\|$	θ
1	124.04	22.02	181.28
2	120.20	165.25	183.34
3	149.05	6.00	178.85
4	145.88	68.01	165.26

5. CLASIFICACIÓN DE GESTOS Y CONTROL DEL ROBOT

Para controlar los movimientos del robot, se diseñó una máquina de estados, donde las entradas son los índices difusos de "apertura" y "delgadez", y la rotación de la boca. Estas entradas son filtradas usando una media temporal deslizante de orden 8. En otras palabras, la decisión que se tome en un instante de tiempo depende de las detecciones halladas en dicho cuadro y siete anteriores.

El diagrama de estados de la máquina utilizada se muestra en la Figura 5. La máquina de estados está dividida en dos partes principales: desactivado (D) y activado (A). Con el objeto de aumentar la fiabilidad del sistema, se debe asegurar que cualquier movimiento involuntario

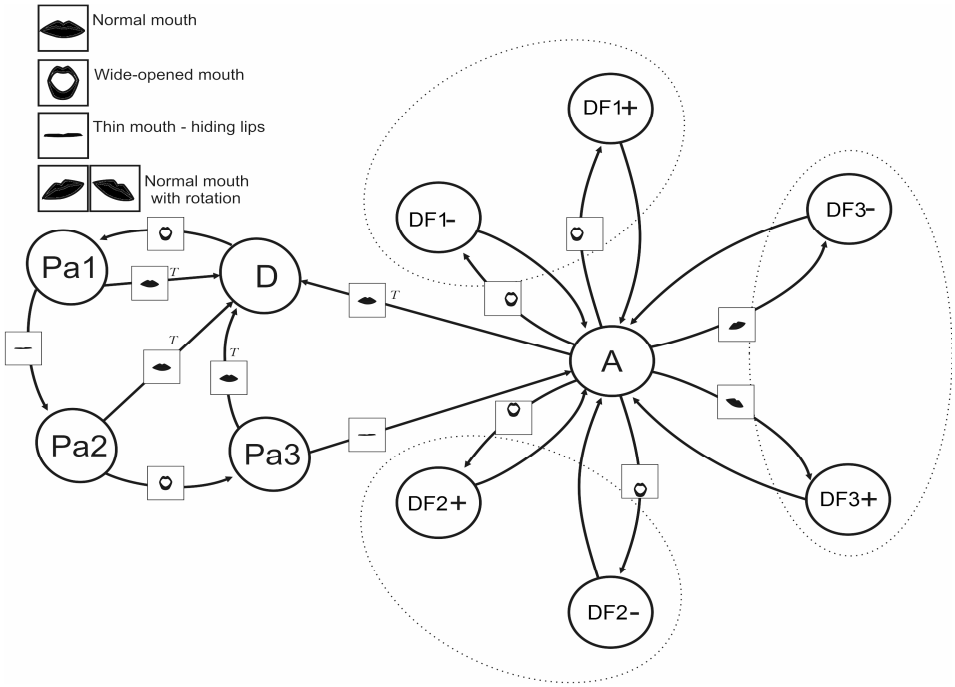


Figura 5. Máquina de Estados para el comando del robot**.**
Figure 5. Robot command state machine.

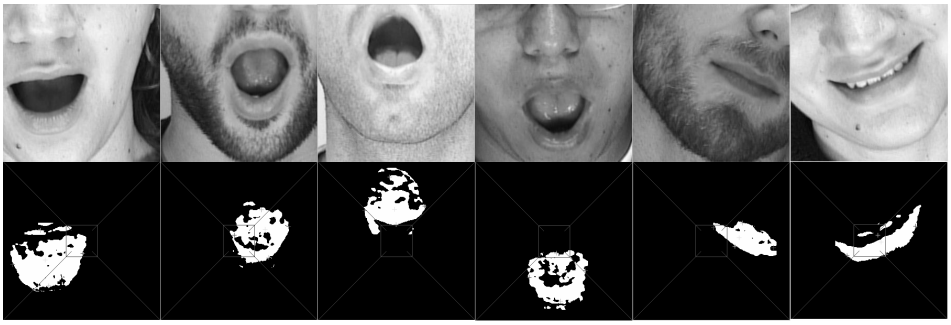


Figura 6. Seis diferentes tipos de gestos de la boca utilizados en el mando del robot**.**
Figure 6. Six different mouth gestures are used in the robot command.

de la boca no pueda causar un movimiento del robot. Por esta razón, se diseñó una secuencia de movimientos, que controla el paso entre desactivado y activado a través de tres estados intermedios (Pa1… Pa3).

Adicionalmente, hay un tiempo límite de 10 segundos entre cada transición. Si la transición supera el tiempo límite, el sistema retorna automáticamente al estado desactivado. Una vez el sistema se encuentra activado, este puede ser llevado a seis diferentes tipos de movimientos asociados a los tres grados de libertad controlados (ver Figura 6). Dos movimientos utilizan la rotación cuando la boca no se encuentra ni abierta ni cerrada. Los otros cuatro dependen de la localización del centro de masa relativo con el centro de imagen, con la condición adicional que la boca debe estar abierta.

6. PRUEBAS Y RESULTADOS

El sistema está conformado por un computador con procesador Pentium IV 3.2GHz y 1 GB de memoria RAM, una cámara de video de tele-conferencia SONY con auto compensación de la iluminación y un robot Staübli RX90 con su interfaz de mando. Para la adquisición del video se utilizó la tarjeta de digitalización IMAQ 1411, la cual se enlaza usando las librerías provistas por el fabricante. Los algoritmos fueron implementados en lenguaje C++. El sistema de video fue configurado en estándar PAL, con una resolución de 640x480 píxeles a 25 fps. Para compensar el efecto de las sombras en las inmediaciones de la región de la boca se utilizaron dos lámparas de 20 vatios. La utilización de dos focos de luz a los lados de la cara permite la adecuada manipulación de las sombras causadas por la nariz y los pómulos, de forma independiente al tipo de iluminación de techo.

Cinco secuencias de video de cuatro sujetos diferentes, con más de 5000 cuadros en total, fueron clasificadas de forma manual con el fin de medir el rendimiento del segmentador y del pre-clasificador (detección previa a la máquina de estados). El índice de detección global fue de un 82.72%. Esto indica el número de detecciones correctas en todos los cuadros de todas las secuencias. De la totalidad de los cuadros, el 35.39% corresponden al gesto denominado “boca abierta”; el 5.30% corresponden al gesto denominado “boca delgada”; y el 59.31% corresponden a los gestos diferentes y denominados “otros”.

En la Tabla 2 se presentan los resultados comparativos para cada gesto, generados por el algoritmo de detección (los valores de la diagonal corresponden al porcentaje de detecciones correctas en cada caso):

Tabla 2. Comparación de la detección inicial de gestos.
Table 2. Initial gesture detection comparisons.

Gesto detectado / Clasificación manual	Abierto	Delgado	Otro	
Abierto	**87.64%**	0.48%	11.88%	**100%**
Delgado	22.12%	**46.08%**	31.8%	**100%**
Otro	7.95%	8.98%	**83.07%**	**100%**

Tabla 3. Comparación de la detección de la rotación de la boca.
Table 3. Comparisons of the detection rate in the mouth rotation.

Rotación detectada / Clasificación manual	Normal	Rotada	
Normal	**95.13%**	4.87%	**100%**
Rotada	10.05%	**89.95%**	**100%**

Se realizó un estudio similar para los diferentes estados de rotación de la boca. Dado que la forma de la boca es aproximadamente simétrica con respecto al eje vertical central, para la prueba se unieron en una clase las inclinaciones derecha e izquierda, y en otra la postura normal. Los resultados obtenidos se muestran en la Tabla 3.
Un estudio de la media y desviación estándar de las características sobre secuencias de video pregrabadas, permitió seleccionar los valores de 0.01 para Δ_{Th} y Δ_{Op}, 0.38 para Γ_{Op} y 0.1 para Γ_{Th}. Usando estos valores, el método propuesto

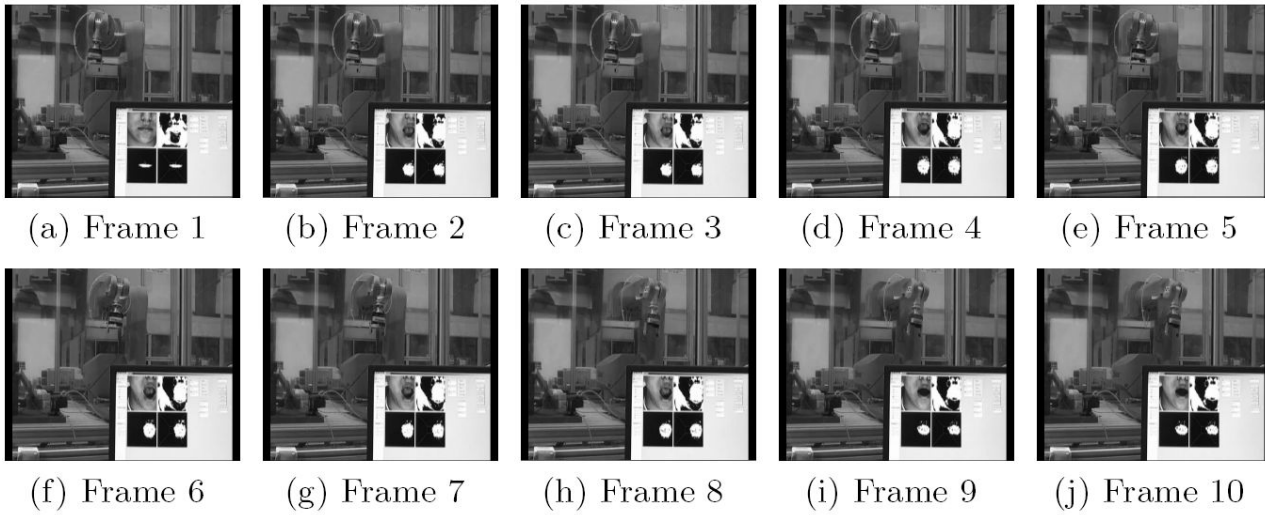


Figura 7. Secuencia de gestos para el movimiento de la articulación de la base del robot.
Figure 7. Gesture sequence for the movement of the robot's base.

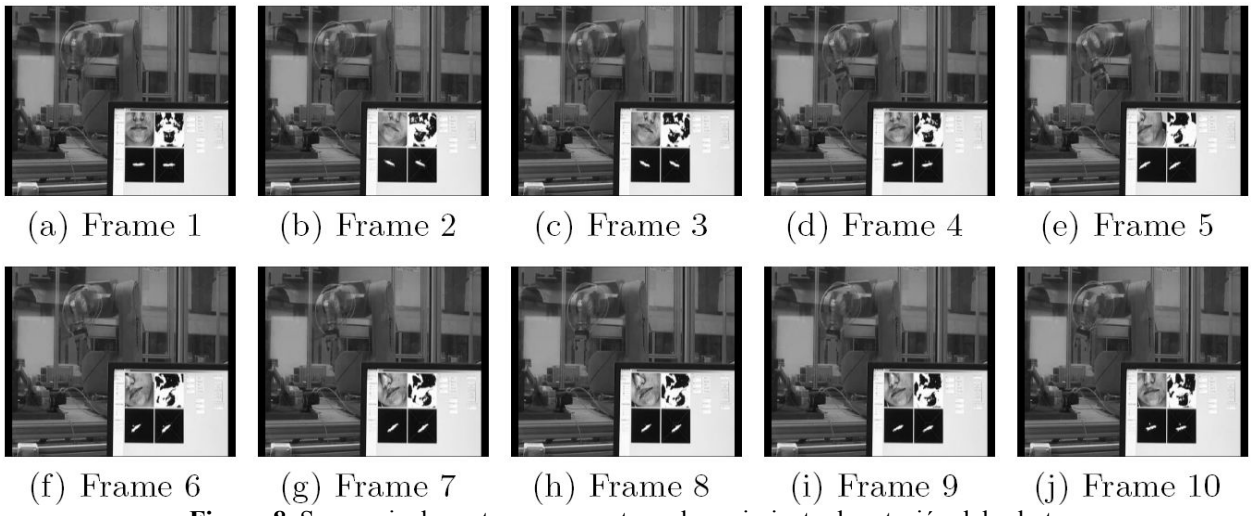


Figura 8. Secuencia de gestos que muestran el movimiento de rotación del robot.
Figure 8. Gesture sequence for the rotation of the robot's tool.

pudo detectar y clasificar los gestos de la boca con exactitud en las secuencias de video, incluso con los individuos barbados. El movimiento del robot utilizando los gestos de la boca se puede observar en las Figuras 7 y 8.

El tiempo de detección de los gestos del algoritmo se mantiene por debajo de 25ms, permitiendo su implementación para aplicaciones en tiempo real. En las pruebas realizadas, con la cámara PAL fue posible mantener su operación sin pérdida de cuadros mientras que, en secuencias pre-grabadas, la tasa de cuadros por segundo se mantenía entre 30 y 60 cuadros por segundo. Parte del costo computacional de los algoritmos dependen del tamaño de la región de interés estimada para la siguiente iteración. Así, cuando la boca es detectada como delgada, el costo de detectarla en el siguiente cuadro es mucho menor que cuando se detecta abierta.

A pesar de que el tiempo de detección de los gestos es pequeño comparado con el tiempo de muestreo del estándar PAL (~40ms), el uso de la media temporal introduce un retardo de 320ms en la toma de la decisión para el cambio en la máquina de estados. Por lo tanto, este es el tiempo mínimo de permanencia que debe tener un gesto para ser tomado en cuenta como un comando para el robot. Esto eleva de forma considerable la confiabilidad en la realización de las operaciones, y le permite al usuario sostener conversaciones sin que estas sean interpretadas como comandos. Una vez que alguna decisión sea tomada, el sistema de comunicación con el robot tarda aproximadamente un segundo en convertirla en movimiento.

7. CONCLUSIONES

Se diseñó un método capaz de detectar apropiadamente la boca en secuencias de video de diferentes individuos.

El sistema es capaz de segmentar el área de la boca en condiciones de luz variable, siempre que el efecto de las sombras generadas sobre el rostro se pueda compensar en algún grado.

El recorte elíptico utilizado ayuda en la selección de la región de interés ROI y en el descarte del ruido ocasionado por los orificios nasales. Sin embargo puede ser inestable cuando el gesto de boca delgada tiene un espesor de menos de cinco píxeles, o cuando se presentan cambios bruscos entre un cuadro y otro en la secuencia. En dichos casos, es necesario recurrir al cálculo inicial de la región de interés tomando como base la imagen completa.

Aunque el sistema tiene la capacidad de procesar las imágenes y clasificar los gestos de la boca en tiempo real, la interfaz con el robot no alcanza la misma velocidad. Además, el uso de la media temporal introduce un tiempo de retardo que depende del número de cuadros utilizados y del tiempo de muestreo fijado por el estándar que se utilice.

Los valores óptimos de $\Gamma_{Th}, \Gamma_{Op}, \Delta_{Th}$ y Δ_{Op} son medidas particulares de cada persona, y estadísticamente son objeto de estudio. Sin embargo, una selección conveniente se logra escogiendo los valores usados. Se propone un método para la estimación de los parámetros en línea, comenzando desde los valores recomendados. El método consiste en el cálculo incremental de la detección de la región de la boca delgada y abierta en cada imagen de la secuencia de video. El valor promedio de $d_2/(d_1+d_2)$ para cada caso puede ser usado como la media γ correspondiente, y los valores delta pueden ser escogidos a partir del máximo valor entre las desviaciones medidas del mismo parámetro en la secuencia de imágenes de boca abierta y boca delgada.

Aunque las pruebas se realizaron gracias a la ayuda de personas que fueron instruidas en el manejo del software, se presentan ciertas prácticas comunes en el movimiento de la boca y la generación de los gestos que afectan de forma clara los índices de error. Los casos más significativos son la visibilidad de los dientes en imágenes con la boca abierta y el exceso de presión en los labios en la boca delgada. En el primer caso, el efecto puede ser compensado con técnicas de preprocesamiento de imágenes y visión artificial; sin embargo, el segundo no es compensable y por tanto depende directamente del entrenamiento que reciba el usuario antes de operar el sistema.

El sistema se ha probado de forma exitosa en personas de tez clara y trigueña, con y sin vello facial. Dado que el método elegido para la segmentación del área de la boca se fundamenta en características de color, los resultados pueden variar para personas de tez oscura.

8. AGRADECIMIENTOS

Los autores agradecen al soporte dado por el programa ECOS Franco-Colombiano (ECOS-Nord/COLCIENCIAS/ICFES/ICETEX).

REFERENCIAS

[1] J. M. SACKIER; Y. WANG. "Robotically assisted laparoscopic surgery from concept to development," Surgical Endoscopy, vol. 8, no. 1, pp. 63–66, Jan. 1994.

[2] V. F. MURIOZ; C. VARA-THORBECK; J. G. DEGABRIEL; J. F. LOZANO; E. SANCHEZ-BADAJOZ; A. GARCIA-CEREZO; R. TOSCANO; A. JIMENEZ-GARRIDO. "A medical robotic assistant for minimally invasive surgery," in Proc. IEEE Int. Conf. Robotics and Automation, San Francisco, CA, Apr. 2000, pp. 2901–2906.

[3] A. NISHIKAWA; T. HOSOI; K. KOARA; D. NEGORO; A. HIKITA; S. ASANO; H. KAKUTANI; F. MIYAZAKI; M. SEKIMOTO; M. YASUI; Y. MIYAKE; S. TAKIGUCHI; M. MONDEN. "Face Mouse: A Novel human-machine interface for controlling the position of a laparoscope," IEEE Trans. On Robotics and Automation, vol. 19, no. 5, pp. 825-841, Oct. 2003.

[4] A. CASALS; J. AMAT; E. LAPORTE. "Automatic guidance of an assistant robot in laparoscopic surgery," in Proc. IEEE Int. Conf. Robotics and Automation, Minneapolis, MN, Apr. 1996, pp. 895–900.

[5] The DaVinci Surgical Systems. Página web, http://www.davincisurgery.com/surgery/system/index.aspx

[6] L. POWERS; D.M.W. POWERS. "Lip Feature Extraction Using Red Exclusion". Trent W. Pan-Sydney Workshop on Visual Information Processing, 2001.

[7] R.L. HSU; M. ABDEL-MOTTALEB: A.K. JAIN. "Face Detection in Color Images". IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 5, pp. 696-706, May 2002.

[8] N. EVENO; A. CAPLIER; P.Y. COULON. "A new color transformation for lips segmentation". In: IEEE Fourth Workshop on Multimedia Signal Processing. (2001).

[9] M. LIEVIN, P.DELMAS, P.Y. COULON; F. LUTHON; V. FRISTOT. "Automatic lip tracking: Bayesian segmentation and active contours in a cooperative scheme". In: ICMCS. (1999) vol. 1.

[10] M. ECKERT. "Compensación de movimiento avanzada para codificación de vídeo". PhD thesis, Universidad Politécnica de Madrid (2003).

[11] S.L.WANG; W.H. LAU; S.H. LEUNG; A.W.C. LIEW. "Lip segmentation with the presence of beards". In: IEEE International Conference on Acoustics, Speech and Signal Processing. (2004).

[12] I. ARSIC; R.VILAGUT; J.P. THIRAN. "Automatic extraction of geometric lip features with application to multi-modal speaker identification". In: IEEE International Conference on Multimedia and Expo (ICME). (2006).

[13] P. GACON; P.Y. COULON; G. BAILLY. "Non–linear active model for mouth inner and outer contours detection". In: 13th European Signal Processing Conference. (2005).

[14] M. HASANUZZAMAN; T. ZHANG; V. AMPORNARAMVETH; H. UENO. "Gesture-based human-robot interaction using a knowledge-based software platform". Industrial Robot: An International Journal. Vol. 33, 2006. pp. 37 – 49.

[15] A. ZELINSKY; J. HEINZMANN. "Human-robot interaction using facial gesture recognition". In Proceedings of the International Workshop on Robot and Human Communication. 1996. pp. 256–261.

[16] J. HEINZMANN. "Real-time human face tracking and gesture recognition". Master's thesis, Universitat Karlsruhe, Fakultat für Informatik. (1996).