SEGUIMIENTO DEL CONTORNO EXTERNO DE LA BOCA EN IMÁGENES DE VÍDEO

Alexánder Ceballos[1]
Juan Bernardo Gómez[2]
Flavio Prieto[3]

Recibido: 27/03/2009
Aceptado: 07/05/2009

RESUMEN

El seguimiento preciso de la boca de una persona, cuando está hablando, es un desafío importante en varias aplicaciones, como la identificación de la cara o la interacción con el computador. La complejidad de forma, textura y color de la boca, y los cambios de iluminación y fondos de los posibles escenarios hacen que este sea aún un problema abierto. En este artículo se propone un algoritmo para el seguimiento del contorno externo de la boca, sin utilizar marcadores o alguna clase de maquillaje para resaltar los labios, basado en apariencia y en restricciones morfológicas definidas en el estándar MPEG-4. El algoritmo es robusto ante la presencia de barba, tono de piel y calidad de la imagen.

Palabras clave: visión por computador, MPEG-4, puntos característicos del contorno externo de la boca, seguimiento de la boca.

1 Ingeniero Electrónico, Universidad Nacional de Colombia, Sede Manizales y miembro del Grupo de investigación en Percepción y Control Inteligente. Carrera 27 No. 64-60, Manizales (Caldas), Colombia. Teléfono (6) 8879300 – 55798. Correo aceballosa@unal.edu.co

2 Profesor del Departamento de Ingeniería Eléctrica, Electrónica y Computación, Universidad Nacional de Colombia, Sede Manizales. Carrera 27 No. 64-60, Manizales (Caldas), Colombia. Teléfono (6) 8879300 – 55798. Correo jbgomezm@unal.edu.co

3 Profesor del Departamento de Ingeniería Mecánica y Mecatrónica, Universidad Nacional de Colombia, Sede Bogotá. Carrera 30 No 45-03, Bogotá, Colombia. (1) 316 5000 – 14103. Correo faprietoo@unal.edu.co

LIP CONTOUR TRACKING IN VIDEO IMAGES

ABSTRACT

An accurate tracking of a person's mouth when he/she is speaking is an important challenge in several applications such as face identification or interaction with computer. Complexity of shape, texture, and color of the mouth, as well as changes in lighting and backgrounds of possible scenarios makes of it an open problem yet. This article proposed an algorithm for a tracking of the mouth external contour without using markers or any kind of make-up for highlighting lips, based on appearance and morphological restrictions defined by the MPEG-4 Standard. Algorithm is robust before the presence of beard, skin tone, and image quality.

Key words: Computer-assisted view, MPEG-4, special points of the mouth external contour, mouth tracking.

1. INTRODUCCIÓN

La extracción de características de la región de la boca ha surgido como un campo activo de visión por computador, debido al interés en aplicaciones como reconocimiento automático del habla audio-visual, reconocimiento de gestos, medición antropométrica y reconocimiento de personas.

El reconocimiento audio-visual del habla ha surgido como un campo activo, gracias a los avances en visión artificial, el procesamiento de señales y el reconocimiento de patrones (Goecke, 2005), ya que promete extender el reconocimiento de habla por computador a ambientes adversos como oficinas, aeropuertos, estaciones de trenes o automóviles en movimiento. De hecho, se ha estimado que observar al hablante equivale a una ganancia de 15 dB en la relación señal a ruido (Campbell, 2006, 2008), y los esfuerzos se han concentrado en la representación visual del habla (Aleksic y Katsaggelos, 2005), (Kratt et al., 2004), (Nefian et al., 2002), y se justifica en que ésta es invariante al ruido acústico.

Al utilizar este esquema se han obtenido buenos resultados. Por ejemplo, en Kim et al., (2006) se usa un perceptrón multicapa para combinar características de audio y visuales para compensar la pérdida de información causada por el ruido, mientras que en Salazar y Prieto, (2006), se usaron características del contorno de la boca, para el proceso de reconocimiento de 5 fonemas vocales del lenguaje español. En Potamianos, (2006) se muestran los enfoques usados para enfrentarse al problema del reconocimiento audio-visual del habla, así como algunos de los resultados más significativos.

La identificación de posturas labiales permite un estudio y seguimiento de la expresión del rostro y de la información que quiere expresar; por esta razón, en las últimas décadas, se percibe un aumento sustancial de investigaciones en reconocimiento de gestos y análisis de expresiones faciales (Yang et al., 2002). En Gómez et al., (2007) y Hernández et al., (2007), se propuso un método para la segmentación y extracción de características faciales en secuencias de vídeo en tiempo real, para ser usado en una interfaz hombre-máquina. El proceso proveyó de un pequeño conjunto de características gestuales que les permitió controlar un robot.

Uno de los motivos del interés sobre el seguimiento de la boca en secuencias de vídeo ha sido la posibilidad de usarlo en herramientas de control, diagnóstico y evaluación de procedimientos quirúrgicos (Salazar et al., 2007). En Mejía y Prieto, (2004) se presentaron diferentes algoritmos y procedimientos para la extracción automática de características faciales, con el fin de obtener las medidas de algunas de las regiones del complejo facial, las cuales permitieron al especialista desarrollar un estudio antropométrico facial.

El seguimiento de los labios es aún un problema abierto de visión artificial, debido a la complejidad de las formas, colores y texturas, y a los cambios de iluminación (Zhilin et al., 2002). Este problema ha sido exitosamente tratado para vistas laterales y con el fondo controlado (Ramos et al., 1997), pero para vistas frontales y sin marcadores de labios ha mostrado ser más complicado. Para el caso de imágenes en escala de grises, los métodos fallan en localizar los límites de la boca en áreas de contraste pobre como el labio inferior y, además, son muy sensibles a cambios de iluminación o sombras.

En el presente trabajo se propone un algoritmo asistido para el seguimiento de los puntos que definen el contorno externo de la boca en imágenes a color, según el estándar MEPG-4. El algoritmo no utiliza marcadores o alguna otra clase de maquillaje para resaltar el área de la boca. En la sección 2 se muestran los enfoques usados para enfrentarse al problema de seguimiento de los labios en la literatura, mientras que en la sección 3 se explica con

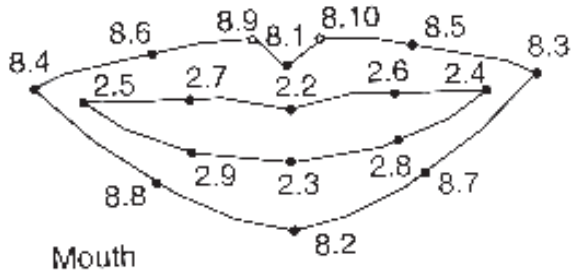
detalle el algoritmo propuesto. En la sección 4 se presentan algunos resultados sobre secuencias de vídeo y, finalmente, en la sección 5 se concluye el trabajo.

2. SEGUIMIENTO DE LOS LABIOS

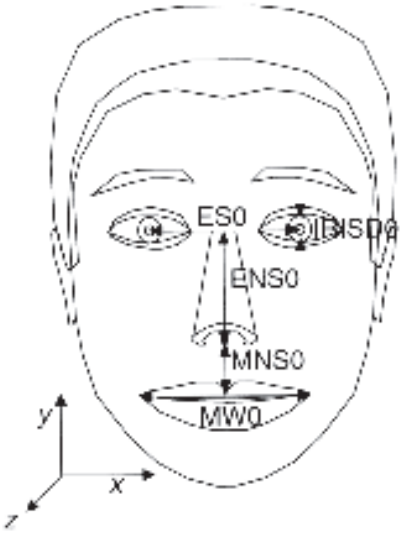
Para la extracción de las características de la región de la boca existen dos enfoques clásicos. El primero basado en la apariencia (de bajo nivel), en el cual se realizan operaciones matemáticas en píxel, sin encontrar la forma exacta de la boca. Por ejemplo, en Gómez et al., (2007) y Hernández et al., (2007) se usan diferentes algoritmos de segmentación basados en píxel y restricciones morfológicas para extraer el área de la boca, y en Zhang et al., (2002) se utilizan características basadas en el color para extraer características del habla sobresalientes en vistas frontales, mientras que el segundo enfoque se basa en la forma de la boca (de alto nivel), en el cual se hace necesaria la ubicación precisa de los contornos de los labios. En Dupont y Luettin, (2000) se extraen los contornos de la boca desde imágenes de intensidad de gris, y en Salazar y Prieto, (2006), se presentaron diferentes algoritmos y procedimientos para la extracción automática de características faciales, basadas en la región del contorno de la boca.

Para enfrentarse al problema particular de seguimiento de la boca, sobresalen los algoritmos basados en contornos con modelado polinomial estático y activo (Snakes), o la representación a trozos con polinomios (Splines). En Jiang et al., (2006) se usa una aproximación estadística llamada filtro partícula (el cual es un filtro bayesiano recursivo) y modelos de forma activa, con el fin de hacer un seguimiento determinístico estocástico de los labios sobre imágenes frontales con iluminación constante y sin marcadores. En Zhang et al., (2001) se emplean *Snakes* sobre imágenes a escala de grises para controlar labios virtuales. Primero se segmenta la región de la boca y se aplica un detector de bordes para inicializar el contorno, después se usan contornos activos para suavizar el contorno y, finalmente, se ajusta esta aproximación con un polinomio de segundo orden para el labio superior y otro para el inferior. Una idea similar fue desarrollada en Seyedarabi y Aghagolzadeh, (2006). Para la estimación inicial de la boca se usó un sistema basado en conocimiento, y consideraron *Snakes* inicializados en forma oval para modelar tanto el contorno interno como externo de la boca. En Ramos et al., (1997) se utiliza un B-Spline con forma elíptica para representar los labios. En el primer cuadro de vídeo se encuentra la boca al usar sobre el área de la cara tanto proyecciones del gradiente en gris, como un modelo estadístico basado en color. Las características extraídas fueron usadas en un sistema de identificación de personas.

MPEG-4 surgió debido a la necesidad de estandarizar los objetos virtuales de vídeo real y sintético. En él se incluyen la codificación de vídeo, la compresión de la geometría y la sincronización entre audio y vídeo. Los estándares de animación del cuerpo y de la cara definidos en MPEG-4 están basados en la estructura ósea y muscular del ser humano, y aunque no permiten que se generen todos los movimientos, debido a que algunos son propios de cada persona, es el esfuerzo más cercano hasta ahora y es el estándar que se usa en este momento en la industria cinematográfica.



a) Parámetros de definición facial de la boca



b) Modelo de cara en estado neutro

Figura 1. Parámetros definidos en el estándar MPEG-4 para la animación de boca. En a) se aprecian los grupos 2 y 8 que definen el contorno interno y externo de la boca, los cuales deben ser normalizados respecto al ancho de la boca (MW0) cuando la cara se encuentra en estado neutro (b).

Fuente: elaboración propia.

Con la finalidad de permitir la animación de rostros, en el estándar MPEG-4 se presentan dos conjuntos de parámetros que estandarizan los modelos del rostro respecto a algunas medidas antropométricas y definen su deformación. Los parámetros de animación facial, Facial Animation Parameters (FAP), son un conjunto de parámetros que permiten la animación de modelos de cara sintéticos. Estos parámetros especifican una acción particular de deformación de un modelo de cara en estado neutro. El modelo de la cara en estado neutro está definido por un conjunto de puntos característicos estandarizados denominados parámetros de definición facial (Facial Definition Parameters, o FDP). Los FDP se miden en unidades específicas, Face Animation Parameter Units (FAPU) (ISO/IEC, 1998). En la figura 1a se aprecian las medidas antropométricas normalizadas empleadas en el estándar, los cinco FAPU miden la distancia entre los ojos (ES0), el diámetro del iris (IRISD0), la separación entre los ojos y la nariz (ENS0), la separación entre la boca y la nariz (MNS0) y el ancho de la boca (MW0).

Se definen 68 FAPs divididos en 10 grupos. En reconocimiento del habla, generalmente se usan los grupos 2 y 8, que describen el movimiento del contorno interno y externo de la boca, respectivamente, mientras que para la síntesis visual del habla se usa el grupo 1, que define 14 visemas claramente distinguibles del habla inglesa (tabla 1). Un visema es el patrón visual de referencia de un fonema, y puede corresponder a varios fonemas (Pandzic y Forchheimer, 2002).

En Zhilin et al., (2002) y Zhilin y Aleksic, (2004) se propone el uso de *Snakes* para el seguimiento de los FAP, basados en el flujo vectorial del gradiente y con una plantilla parabólica como fuerza externa, con el fin de hacer funcionar un descodificador MPEG-4. El objeto animado pudo imitar a una persona real mientras los parámetros fueron extraídos satisfactoriamente. La tasa de transmisión de los parámetros sin compresión fue de 0,5 Kbps, mientras que el vídeo estándar requiere decenas de Mbps.

En la tabla 1 se presentan los 14 visemas establecidos en el grupo 1 de los parámetros de animación facial definidos dentro del estándar MPEG-4. También, se presentan los fonemas a los cuales hacen referencia los 14 visemas.

Tabla 1: Visemas y fonemas relacionados.

Visema	Fonemas	Ejemplo
0	ninguno	
1	p, b, m	put, bed, mill
2	f, v	far, voice
3	T, D	think, that
4	t, d	Tip, doll
5	k, g	call, gas
6	Ts, dZ, S	chair, join, she
7	s, z	Sir, zeal
8	n, l	Lot, not
9	R	Red
10	A	Car
11	E	bed
12	I	Tip
13	Q	top
14	U	book

Fuente: Pandzic y Forchheimer, (2002)

En Aleksic y Katsaggelos, (2005) se describe un sistema audio-visual de reconocimiento automático del habla. Como características visuales usaron los puntos que describen el contorno interno y el externo de la boca, y se empleó análisis de componentes principales (PCA) para disminuir la dimensión del vector de características. Como características de audio, se usaron los coeficientes cepstrales en frecuencia de Mel. Finalmente, en Abboud y Chollet, (2005) se hizo seguimiento de los labios, y la forma de la boca fue clonada sobre otra persona, con base en modelos de apariencia, los cuales fueron refinados usando los puntos característicos definidos en el estándar MPEG-4.

3. ALGORITMO PROPUESTO

El algoritmo que se propone para el seguimiento del contorno externo de la boca está basado en apariencia y en restricciones morfológicas definidas en el estándar MPEG-4. El algoritmo usa el grupo 8 que describe el contorno externo de los labios (Pandzic y Forchheimer, 2002), pues algunos estudios psicológicos han sugerido que es el que más influencia tiene en la lectura de los labios. Además, en Aleksic y Katsaggelos, (2005) se muestra que el uso del grupo 2, que describe el contorno interno de la boca, no aumenta significativamente el rendimiento de un sistema de reconocimiento automático de habla, y los algoritmos usados son significativamente más costosos que los del contorno externo.

En general, los pasos en un sistema de seguimiento de la boca son: detección de la región de la boca, localización de los labios (inicialización), seguimiento de los labios y la extracción de características (Zhang et al., 2001). El seguimiento es explicado con detalle en el algoritmo 1.

Paso 1, detección de la región de la boca

La región de interés es localizada de forma asistida únicamente en el primer cuadro de vídeo de la secuencia.

Algoritmo 1: Seguimiento asistido de puntos del contorno externo

Entradas: Vídeo en forma de secuencia de imágenes C_1, C_2, ..., $C_n \in C$

Salida: La secuencia de puntos del contorno externo de la boca para todos los cuadros de vídeo $S_{n\text{-}10}$.

Los siguientes pasos se realizan para todos los cuadros.

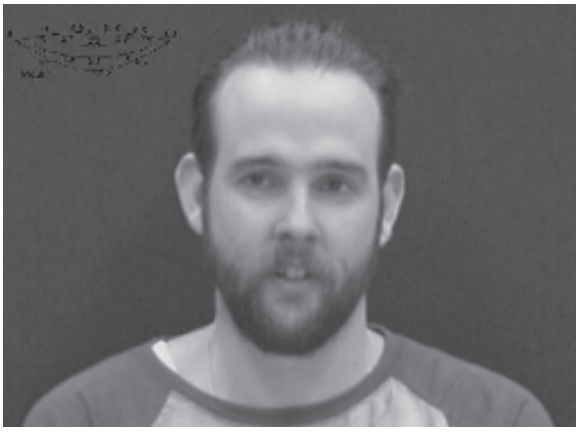
[Paso 1:] Localización de la región de interés.

[Paso 2:] Ubicación de los 10 puntos del contorno externo de la boca para el primer cuadro de vídeo. p_1, p_2, p_{10}, $\in P$

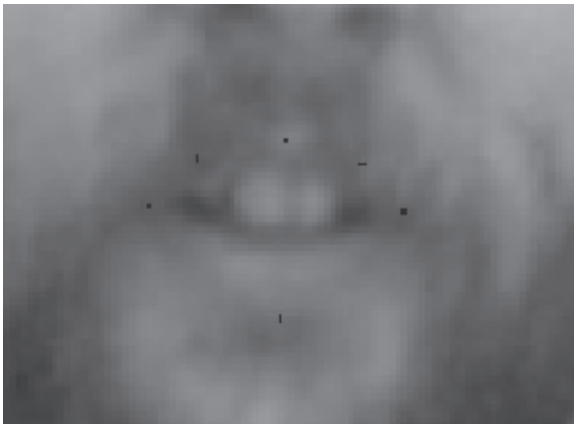
[Paso 3:]
para todos los puntos del contorno de cuadro anterior $S_{n\text{-}1}$ *hacer*
Calcular la similitud con los píxeles de la vecindad en el cuadro presente .
Escoger el candidato a punto actual $S_{n\text{-}i}$ como el píxel donde la similitud es mayor (más cercana a la unidad).
fin para
Una vez obtenidos los 10 puntos candidatos, aplicar restricciones de forma.

[Paso 4:] Cálculo de las características de la región de la boca.

En el algoritmo 1 se presenta la metodología seguida para el sistema de seguimiento de la boca.



a) Interfaz para localizar la región de interés



b) Ubicación de los 10 puntos que definen el contorno externo de la boca

Figura 2: Inicialización del algoritmo de seguimiento del contorno externo de la boca. En el primer cuadro de vídeo se ubican los 10 puntos que definen el contorno.

Fuente: elaboración propia.

Paso 2, localización de los labios

Para iniciar el algoritmo de seguimiento de los labios, se hace necesaria la ubicación exacta de los puntos que describen el contorno externo de la boca. Debido a que la segmentación robusta de la boca ante la presencia de barba, tono de piel, cambios de iluminación, presencia de lengua y calidad de la imagen aún es un problema abierto y sólo se han obtenido buenos resultados para la extracción del contorno sobre imágenes de alta definición, la inicialización en este caso se hace de forma manual, se ubican manualmente sobre el primer cuadro de vídeo los 10 puntos (figura 2).

Paso 3, seguimiento de los labios

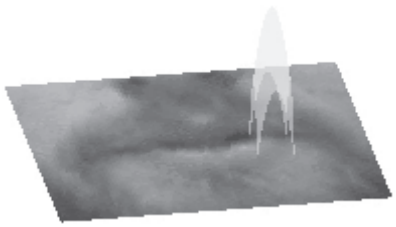
Con el fin de realizar el seguimiento de los labios se usa una medida de similitud entre cuadros de vídeo, además de algunas restricciones morfológicas dadas en el estándar MPEG-4.

La medida de similitud se hace sobre los píxeles pertenecientes a la vecindad de cada uno de los 10 puntos que definen el contorno externo. Primero, se calcula la distancia de la ventana centrada en el punto hallado en el cuadro de vídeo anterior (V) con las ventanas en el cuadro presente, centradas en cada uno de los píxeles de la vecindad de interés (V_{ij}). El cuadro presente es, además, comparado con el primer cuadro de la secuencia se vídeo, el cual posee información altamente confiable, debido a que los puntos del contorno de la boca de este cuadro no fueron calculados (ecuación 1).

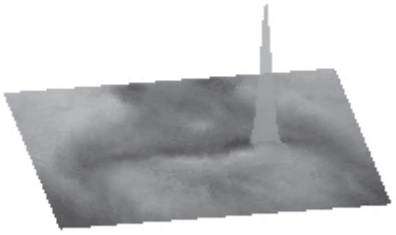
$$d_{ij} = \left\| V - V_{ij} \right\| + \left\| V_1 - V_{ij} \right\| \quad (1)$$

La distancia será mínima si la ventana del cuadro de vídeo anterior concuerda exactamente con alguna de las ventanas del cuadro de vídeo actual y su valor máximo será definido por el tamaño de las ventanas. Con el fin de normalizar la distancia y usarla como medida de similitud, la distancia es usada como el argumento de la función exponencial negativa. Siendo así, el rango se encuentra entre 1 y 0, 1 para una total concordancia y 0 para cuando las ventanas con totalmente diferentes (ecuación 2).

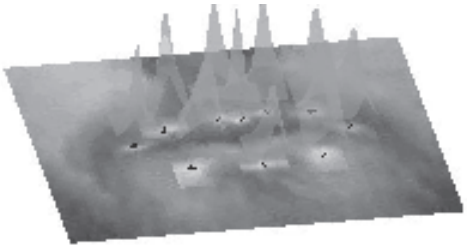
$$c_{ij} = e^{-d_{ij}} \quad (2)$$



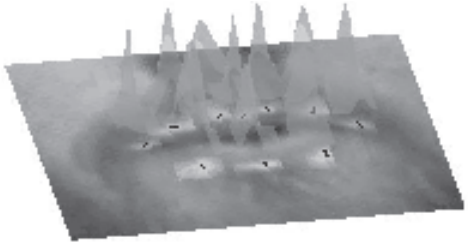
a) Función de densidad de probabilidad normal usada en la ponderación de la similitud.



b) Función de similitud ponderada calculada en la vecindad de un punto.



c) Las 10 funciones de similitud calculadas para cada uno de los puntos del contorno externo de la boca.



d) Resultado de aplicar el mismo calculo de similitud en el cuadro de vídeo siguiente.

Figura 3. La ponderación de la similitud por una función de distribución de probabilidad normal (a) se presenta en b). El resultado de aplicar el cálculo de la similitud en dos cuadros de vídeo seguidos, se muestra en c) y en d).

Fuente: elaboración propia.

La similitud es entonces ponderada usando una función de densidad de probabilidad normal con media en el punto del cuadro de vídeo anterior, y con desviación estándar igual al tamaño del vecindario (figura 3a). Así se consigue dar más peso a aquellos píxeles cercanos al punto hallado en el cuadro de vídeo anterior, debido a que es más probable que correspondan al punto en el cuadro de vídeo actual. Se escoge como candidatos a cada uno de los 10 puntos que conforman el contorno externo de la boca, aquellos cuya similitud sea más cercana a la unidad. En las figuras 3c y 3d se observa el resultado de este procedimiento en dos cuadros seguidos; los píxeles iluminados representan la probabilidad de los píxeles de convertirse en cada punto.

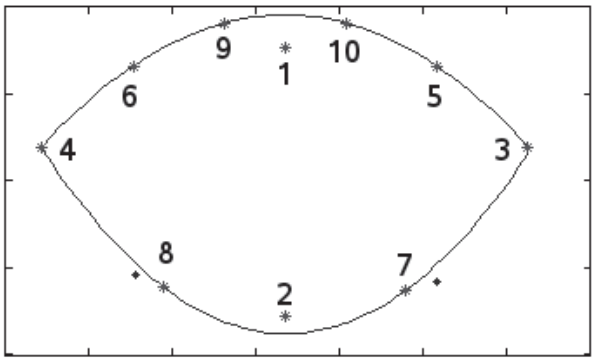


Figura 4: Restricciones morfológicas del contorno externo de los labios. Las restricciones son modeladas con dos polinomios de segundo grado.

Fuente: elaboración propia.

Con el objetivo de hacer el seguimiento de los labios más robusto, se hace que los puntos candidato, hallados con la similitud máxima cumplan las restricciones morfológicas de la boca, así como las restricciones sugeridas en el estándar MPEG-4 (tabla 2). La forma de la boca está caracterizada por ser simétrica (simetría reflexiva sobre el eje vertical). Para satisfacer las restricciones de simetría reflexiva se usan dos polinomios de segundo grado (parábolas) con eje de simetría vertical (ecuación 3) (figura 4); se deben rotar primero todos los puntos, de modo que el punto 4 y el punto 3 queden a 0°. Entonces, se ajusta el polinomio superior con los candidatos a puntos 4, 6, 9, 10, 5 y 3, y el polinomio inferior con los candidatos 4, 8, 2, 7 y 3, teniendo en cuenta las ubicaciones sugeridas por el estándar MPEG-4 en la tabla 2.

$$y=ax^2 + bx +c \quad (3)$$

En cuanto a las restricciones morfológicas sugeridas en la tabla 2, cabe recalcar que el punto 7.1, el cual corresponde al punto de rotación de la cabeza, para este caso es desconocido. Por tal motivo, las abscisas tanto del punto 1 como del punto 2, correspondientes al punto medio entre los vértices de la boca, son hechas iguales a (8.3x + 8.4x)/2, teniendo en cuenta que aunque no se pueda satisfacer la restricción dada en la tabla 2, la boca es simétrica. En el mismo orden de ideas, para los puntos 9 y 10 pertenecientes al arco de cupido (para los cuales no se definen restricciones en la tabla 2), se igualan las abscisas.

Con los puntos 9 y 10 se usan aún más restricciones, pues tampoco se permite un movimiento entre un cuadro de vídeo y otro superior al 20 % de la distancia media al punto 1, ni que alguno de los dos haga un cruce por el eje vertical.

También se ajustan las abscisas de los puntos 5, 6, 7 y 8 según la tabla 2, y se hallan las ordenadas de los puntos 6, 9, 10 y 5 al evaluar el polinomio que modela la parte superior de los labios, y de los puntos 8 y 7 al evaluar el polinomio que modela la parte inferior. Finalmente, se debe invertir la rotación hecha (figura 4).

Tabla 2: Localización recomendada para los puntos característicos del contorno externo de la boca (el punto 7.1x corresponde al punto de rotación de la cabeza).

Punto característico (FP) Localización recomendada		
Descripción		x
8.1	Punto medio del contorno externo del labio superior	701x
8.2	Punto medio del contorno externo del labio inferior	7.1x
8.3	Esquina izquierda del contorno externo de los labios	
8.4	Esquina derecha del contorno externo de los labios	
8.5	Punto medio entre FP 8.3 y 8.1 en el contorno externo del labio superior	(8.3x + 8.1x)/2
8.6	Punto medio entre FP 8.4 y 8.1 en el contorno externo del labio superior	(8.4x + 8.1x)/2
8.7	Punto medio entre FP 8.3 y 8.2 en el contorno externo del labio inferior	(8.3x + 8.2x)/2
8.8	Punto medio entre FP 8.4 y 8.2 en el contorno externo del labio inferior	(8.4x + 8.2x)/2
8.9	Punto superior derecho del arco de cupido	
8.10	Punto superior izquierdo del arco de cupido	

FP: Facial parameter

Fuente: elaboración propia.

En la tabla 2 se aprecia la ubicación recomendada para cada uno de los puntos del contorno externo de la boca definidos en el estándar MPEG-4. La localización se limita a definir las abscisas de cada uno de los puntos.

Paso 4, extracción de características

Una vez encontrados los puntos en la secuencia de vídeo, para las aplicaciones del seguimiento de los labios, las características de la forma de la boca deben ser calculadas. Teniendo los 10 puntos sobre toda la secuencia de vídeo, se puede encontrar, entre muchas otras, el área de la región dentro de los labios, la redondez, el factor de forma, la relación entre el eje horizontal y el vertical, el perímetro y diferentes relaciones geométricas entre los puntos.

El área es calculada en la forma polar según la ecuación 4, donde r_j corresponde a la distancia de cada uno de los 10 puntos hasta el centro de la boca, y $\Delta\theta$ al ángulo en radianes de separación entre un punto y otro (figura 5).

$$A = \sum_{i=1}^{10} r_i^2 \Delta\theta \tag{4}$$

Por su parte, la redondez es hallada usando la ecuación 5, en la cual A corresponde al área dentro del contorno *d* y al diámetro mayor equivalente al ancho de la boca, es decir, a la distancia entre los puntos 3 y 4 que definen el contorno externo de la boca según el estándar MPEG-4 (figura 1).

$$R = \frac{4A}{\pi d^2} \tag{5}$$

El perímetro, a su vez, se calcula al sumar las distancias entre los puntos como se indica en la ecuación 6 (figura 1). Donde p_i corresponde a las coordenadas (*x*,*y*) del punto *i*.

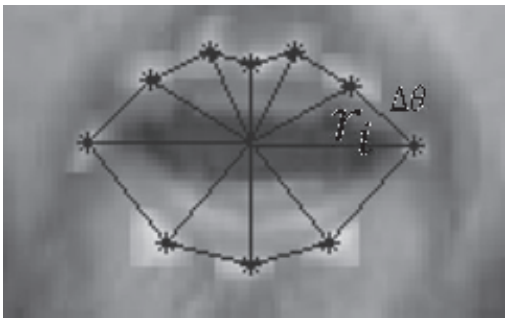


Figura 5: Cálculo del área comprendida dentro del contorno externo de la boca. El análisis se realiza usando coordenadas polares.
Fuente: elaboración propia.

Mientras que el factor de forma es encontrado al utilizar la ecuación 7. En esta ecuación es el perímetro y A el área comprendida dentro del contorno.

$$FF = \frac{P_{er}}{4\pi A} \tag{7}$$

Finalmente, la relación entre el eje vertical y horizontal de la boca es hallada al usar la ecuación 8.

$$RHV = \frac{\|p_3 - p_4\|}{\|p_1 - p_2\|} \tag{8}$$

En la figura 6 se muestra la dinámica de algunas características que pueden ser extraídas teniendo el contorno externo de la boca en una secuencia de vídeo. Se puede observar que la respuesta del factor de forma, la redondez y la relación de los diámetros (ejes vertical y horizontal) en el tiempo es similar, mientras que el área y el perímetro se comportan de manera análoga.

$$Per = \|p_4 - p_6\| - \|p_6 - p_9\| - \|p_9 - p_1\| - \|p_1 - p_{10}\| - \|p_{10} - p_5\| - \|p_5 - p_3\| - \|p_3 - p_7\| - \|p_7 - p_2\| - \|p_2 - p_8\| - \|p_8 - p_4\| \tag{6}$$

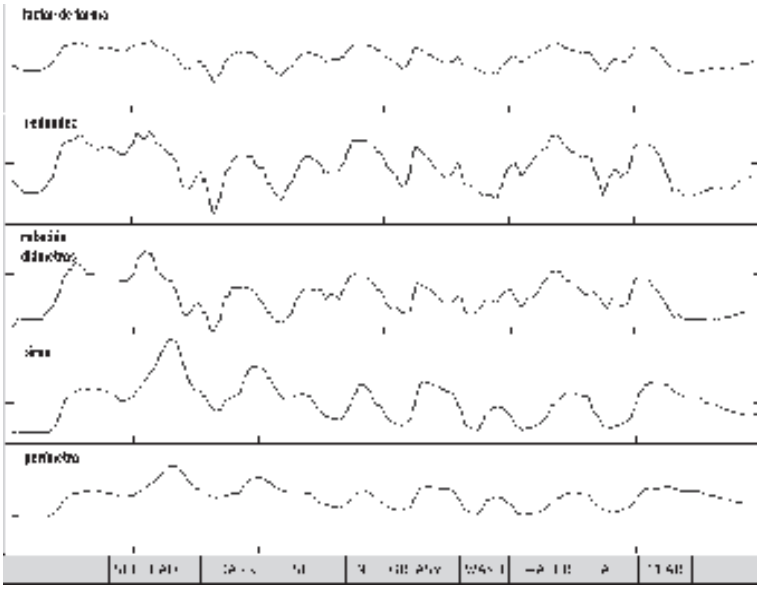


Figura 6: Algunas características de la forma de la boca que pueden ser extraídas de los 10 puntos que describen el contorno externo de la boca.

Se decidió hacer uso de los FAP que definen la deformación de los puntos característicos del contorno externo de la boca como características. Con este fin, se mide el desplazamiento de cada uno de los puntos con respecto a una boca en estado neutro (S_{neutro}), que es seleccionada de los cuadros dentro de la secuencia de vídeo (ecuación 9) (algoritmo 1). Cabe recalcar que debe haber desplazamientos, tanto positivos como negativos, para definir las deformaciones de la boca desde un estado neutro. Estos desplazamientos son, entonces, normalizados respecto al ancho de la boca, el cual es el FAPU (MW0) para los grupos 2 y 8 que describen los contornos interno y externo de la boca.

$$FAP = \frac{S_{neutro} - S}{MW0} \tag{9}$$

4. VENTAJAS DEL ENFOQUE PROPUESTO

El seguimiento automático de los labios es aún un desafío abierto. Se han conseguido buenos resultados para aplicaciones específicas, pero aún no se ha logrado establecer una metodología adecuada para realizar seguimiento preciso de la forma de los labios en tiempo real.

Por otro lado, los modelos paramétricos que minimizan una función de coste han presentado buenos resultados sólo en imágenes de alta definición y aún no resuelven el problema en tiempo real. De hecho, los algoritmos actuales dependen fuertemente de las condiciones de iluminación y son débiles ante la presencia de barba, de lengua o de los dientes, e incluso ante la diferencia entre tonos de piel.

En el trabajo presente no se utilizaron ayudas externas clásicas en el seguimiento de labios, como el uso de maquillajes o marcadores. Tampoco se controló la iluminación ni el fondo. Además, se basa en el estándar MPEG-4, reconocido internacionalmente. MPEG-4 se fundamenta en la anatomía humana y en la interacción de la estructura ósea y muscular para la animación del cuerpo y de la cara.

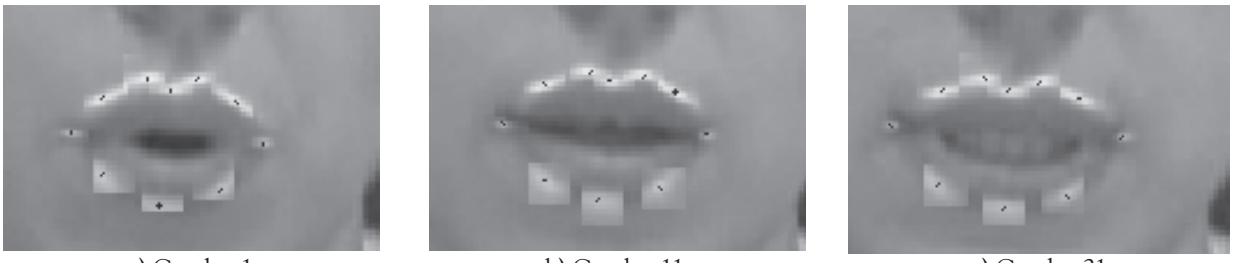
El algoritmo aún no es automático, debido a que a segmentación robusta de la boca aún es un

tema activo de investigación, y por lo tanto, debe ser inicializado de forma manual. Sin embargo, pudo seguir satisfactoriamente el contorno externo de los labios para personas con distinto tono de piel, ante la presencia de barba y sobre secuencias de vídeo con calidad de imagen pobre.

5. RESULTADOS

Con el fin de observar mejor el funcionamiento del algoritmo propuesto, en la figura 7 se presenta el resultado sobre una secuencia de vídeo cada 10 cuadros. En este caso se utilizó un vecindario de búsqueda píxeles y una ventana para el cálculo de la similitud de píxeles. Siendo así, la función de ponderación es una función de distribución normal centrada en cada uno de los 10 puntos del contorno externo del cuadro de vídeo anterior, y con desviación 11 (el tamaño del vecindario). Las zonas iluminadas alrededor de cada punto representan la probabilidad dada por la medida de similitud, de que los píxeles sean los nuevos puntos que describen el contorno externo de la boca.

El algoritmo de seguimiento asistido de los puntos que definen el contorno externo de la boca, según el estándar MPEG-4, fue utilizado tanto con la base de datos VidTIMIT (Sanderson y Paliwal, 2004) como con datos adquiridos en el laboratorio. La base de datos VidTIMIT cuenta con 103543 imágenes, cuya calidad no es muy buena, debido a que se encuentran en formato jpeg, tienen una resolución de 512x384 píxeles, son de toda la cara y la región de interés es de píxeles aproximadamente (figura 8). Por otro lado, los datos adquiridos en el laboratorio constan de 46483 imágenes y fueron tomados con un ángulo bajo, lo que no afecta la simetría de la boca (figura 9). Las imágenes son de 720x480 píxeles y se encuentran en formato png, además la región de interés es de 330x160 píxeles.

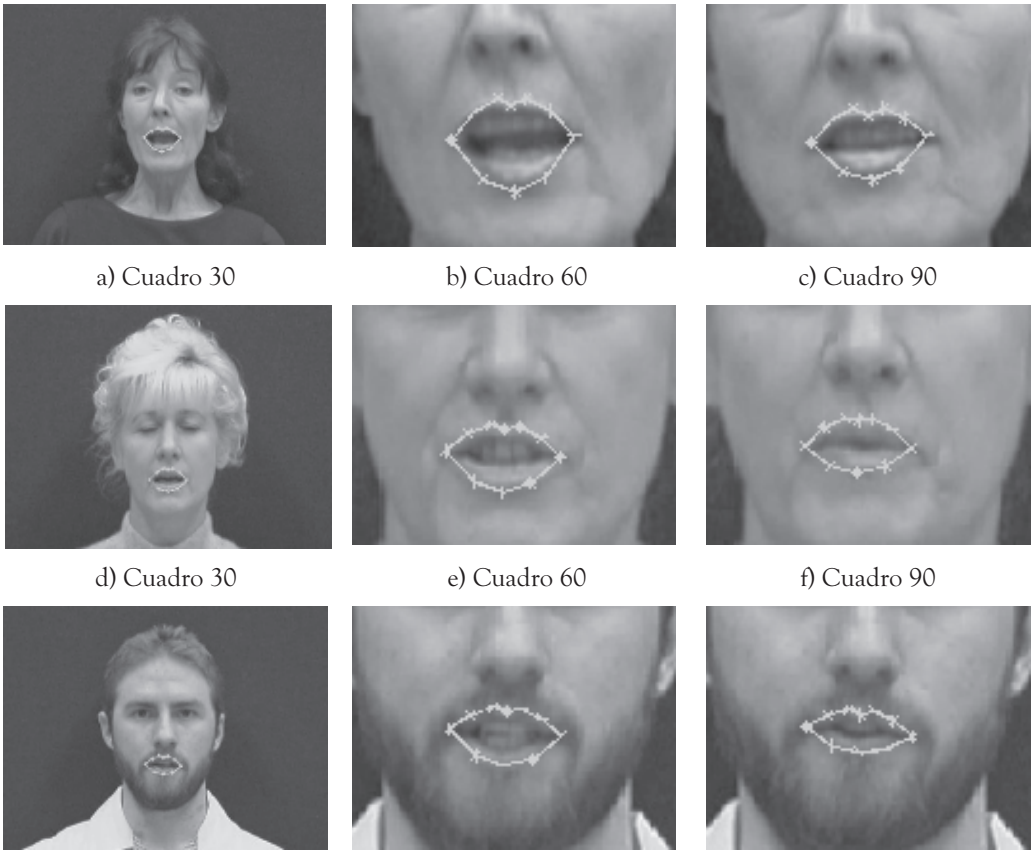


a) Cuadro 1 b) Cuadro 11 c) Cuadro 21

Figura 7: Seguimiento de los 10 puntos que conforman el contorno externo de la boca en una secuencia de vídeo cada 10 cuadros.

Fuente: elaboración propia.

En la figura 8 se aprecia el seguimiento del contorno externo de la boca sobre tres secuencias pertenecientes a la base de datos VidTIMIT. Las secuencias poseen alrededor de 100 cuadros de vídeo, y así los resultados son mostrados cada 30 cuadros. Dado que el ancho de la boca sobre la base de datos VidTIMIT tiene una media de 55 píxeles con una desviación estándar de 6,76 píxeles, se usó una ventana para el cálculo de la similitud de 21x21 y un vecindario de búsqueda de 5x5 píxeles.



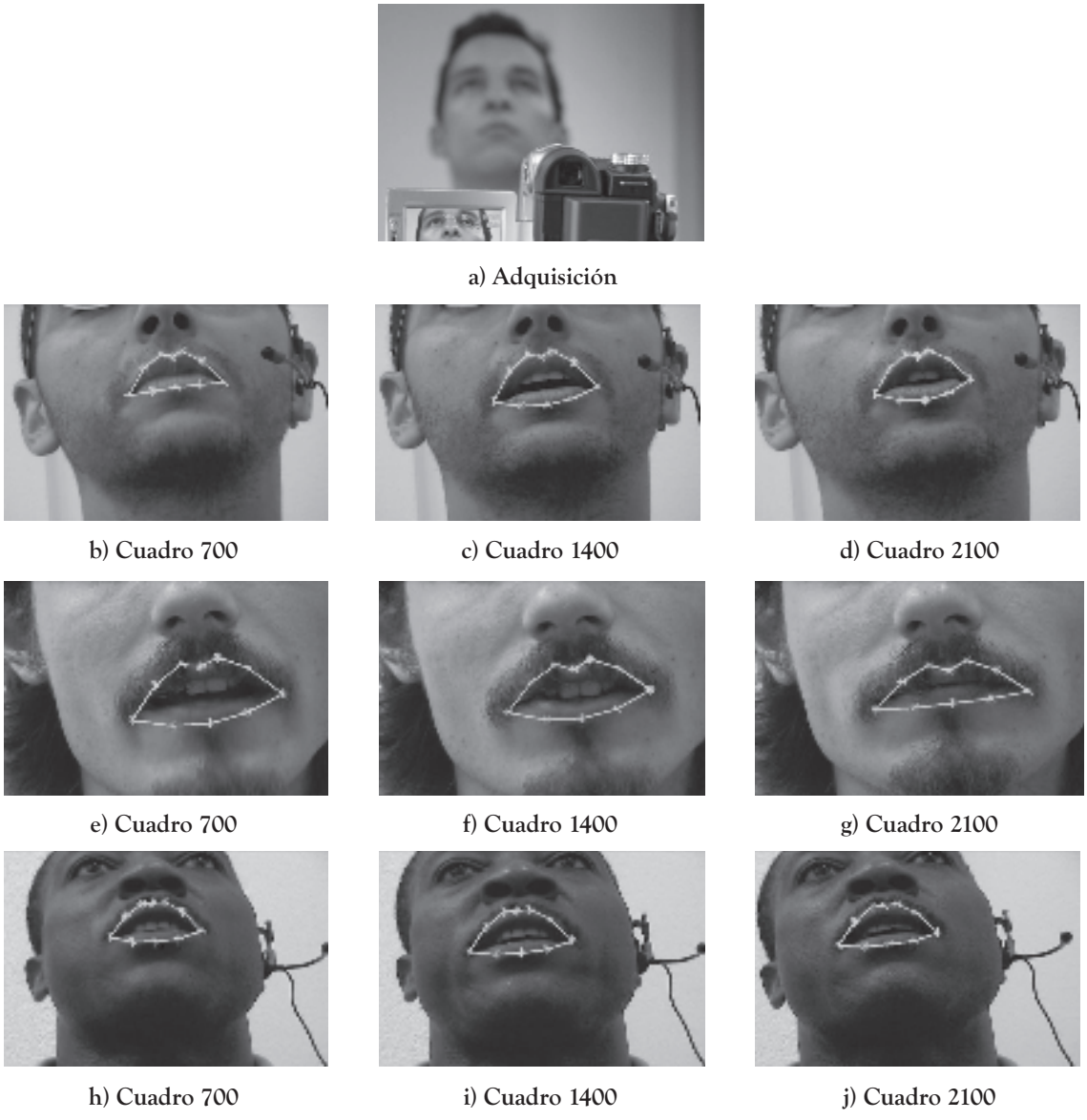
a) Cuadro 30 b) Cuadro 60 c) Cuadro 90

d) Cuadro 30 e) Cuadro 60 f) Cuadro 90

g) Cuadro 30 h) Cuadro 60 i) Cuadro 90

Figura 8: Seguimiento del contorno externo de los labios sobre tres secuencias de vídeo de la base de datos VidTIMIT. Los resultados son mostrados cada 30 cuadros.

Fuente: elaboración propia.



a) Adquisición

b) Cuadro 700 c) Cuadro 1400 d) Cuadro 2100

e) Cuadro 700 f) Cuadro 1400 g) Cuadro 2100

h) Cuadro 700 i) Cuadro 1400 j) Cuadro 2100

Figura 9: Seguimiento del contorno externo de los labios sobre tres secuencias de vídeo adquiridas en el laboratorio. Los resultados son mostrados cada 700 cuadros.

Fuente: elaboración propia.

Para el caso de los datos del laboratorio, se empleó una ventana para el cálculo de la similitud de 11x11 píxeles y un vecindario de búsqueda de 11x11 píxeles. El ancho de la boca tiene una media de 210.38 píxeles con desviación de 72,46. Los resultados de usar el algoritmo sobre estos datos se muestran en la figura 9 cada 700 cuadros, pues las secuencias de vídeo poseen 2500 cuadros aproximadamente.

En ambos casos el algoritmo siguió el contorno externo de la boca, y fue robusto a la presencia de barba, el tono de piel y calidad de la imagen. Los 10 puntos del contorno externo de la boca fueron usados para el cálculo de los FAP. En general, si se desea hacer reconocimiento de patrones dinámicos, se debe agregar información temporal al incluir derivadas en el vector de características.

6. CONCLUSIONES Y DISCUSIÓN

Se ha presentado un modelo basado en restricciones morfológicas y en una medida de similitud en píxeles para el seguimiento del contorno externo de la boca en imágenes a color. La propuesta ha mostrado ser robusta ante la presencia de barba y el tono de piel, e incluso realizó el seguimiento tanto en imágenes con buena definición adquiridas en el laboratorio, como en imágenes con menor definición presentes en la base de datos VidTI-MIT. También mostró ser fuerte ante cambios de iluminación y enfoque, pues no hubo control de iluminación y la cámara tenía autoenfoque.

El algoritmo aún no es automático, pero se ha venido trabajando en segmentación robusta de la boca usando componentes de color (Loaiza et al. 2007), y también se ha pensado en usar características de textura con este fin.

La secuencia de los 10 puntos que describen el contorno externo de la boca, según el estándar MPEG-4, obtenida al utilizar el algoritmo de seguimiento sobre una secuencia de vídeo, puede ser usada para el cálculo de características que describen la forma de la boca. Estas características, a su vez, pueden ser empleadas para hacer reconocimiento de gestos, como parte del conjunto de características para realizar identificación de personas, para realizar estudios antropométricos, y si se incluye información dinámica usando las primeras dos derivadas temporales, para reconocimiento de patrones dinámicos como el habla.

7. AGRADECIMIENTOS

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

8. REFERENCIAS

ABBOUD B. and CHOLLET G. (2005). Appearance based lip tracking and cloning on speaking faces. In Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 301 - 305.

ALEKSIC P. S. and KATSAGGELOS A. K. (2005). Comparision of MPEG-4 facial animation parameter groups with respect to audio-visual speech recognition performance. IEEE International Conference on Image Processing, 3: III- 501-504.

CAMPBELL R. (2006). Audio-visual speech processing. Elsevier, 562-569.

CAMPBELL R. (2008). The processing of audio-visual speech: empirical and neural bases. Philosophical Transactions of The Royal Society B. 1001-1010.

DUPONT S. and LUETTIN J. (2000). Audio-visual speech modeling for continuous speech recognition. IEEE transactions on multimedia, 2: 141-151.

GOECKE R. (2005). Current trends in joint audio-vídeo signal processing: a review. Proceeding of the Eighth International Symposium on Signal Processing and Its Applications, (ISSPA 2005), 1: 70 –73.

GÓMEZ J. B., PRIETO F. and REDARCE T. (2007). Lips Movement Segmentation and Features Extraction in Real Time. Innovative Algorithms and Techniques in Automation, Industrial Electronics and Telecommunications, 205 –210.

HERNÁNDEZ J. E., PRIETO F. and REDARCE T. (2007). Real-Time Robot Manipulation Using Mouth Gestures In Facial Vídeo Sequences. Advances in Brain, Vision, and Artificial Intelligence, 224-233.

ISO/IEC, (1998). Information technology-generic coding of audio-visual objects, Part 2: Visual, ISO/IEC FDIS 14496-2 (Final Drafts International Standard), ISO/IEC JTC1/SC29/WG11 N2502, Atlantic City.

JIANG M., GAN Z., HE G., and GAO W. (2006). Combining particle lter and active shape models for lip tracking. In The Sixth World Congress on Intelligent Control and Automation Proceedings (WCICA), 2: 9897- 9901.

KIM M. W., RYU J. W., and KIM E. J. (2006). Speech Recognition with Multi-modal Features Based on Neural Networks. In Lecture Notes in Computer Science. Volume 4233: 489-498.

KRATT J., METZE F., STIEFELHAGEN R., and WAIBEL A., (2004). Large vocabulary audio-visual speech recognition using the janus speech recognition toolkit. DAGM 2004, Lecture Notes in Computer Science, 3175: 488–495.

LOAIZA J., GÓMEZ J.B. and CEBALLOS A. (2007). Análisis de Discriminancia y Selección de Características de Color en Imágenes de Labios Utilizando Redes Neuronales. XII Simposio de Tratamiento de Señales, Imágenes y Visión Artificial. STSIVA 2007, 1-5

MEJÍA GÓMEZ I. M. y PRIETO ORTIZ F. A., (2004), Extracción automática de características faciales para el estudio antropométrico en niños entre 5 y 10 años de la ciudad de Manizales. En Memorias del Cuarto Encuentro de Investigación sobre Tecnologías de Información aplicadas a la solución de problemas (EITI), 171-178.

NEFIAN A. V., LIANG L., PI X., LIU X., and MURPHY K. (2002). Dynamic bayesian networks for audio-visual speech recognition. EURASIP Journal on Applied Signal Processing, 1-15.

PANDZIC I. S. y FORCHHEIMER R. 2002. (MPEG-4), Facial Animation: The Standard, Implementation and Applications, England, Wiley, 7-62.

POTAMIANOS G. (2006). Speech recognition, audio-visual. Elsevier, 800-805.

RAMOS M., MATAS J., and KITTLER J. (1997). Statistical chromaticity-based lip tracking with B-splines. In ICASSP '97: Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 4: 2973.

SALAZAR A. y PRIETO F. (2006). Extracción y Clasificación de Posturas Labiales en Niños entre 5 y 0 Años de la Ciudad de Manizales. En DYNA, Revista de la Facultad de Minas, Universidad Nacional de Colombia Sede Medellín, 73 (150): 175-188.

SALAZAR A., HERNÁNDEZ J. y PRIETO F. (2007). Automatic Quantitative Mouth Shape Analysis. Computer Analysis of Images and Patterns, 4673: 416-423.

SANDERSON C. y PALIWAL K. K. (2004). Identity verification using speech and face information. Digital Signal Processing. Elsevier, 14, Issue 5: 449-480.

SEYEDARABI H., LEE W., and AGHAGOLZADEH A. (2006). Automatic lip tracking and action units classification using two-step active contours and probabilistic neural networks. In Canadian Conference on Electrical and Computer Engineering Proceedings (CCECE), 2021-2024.

YANG M.-H., KRIEGMAN D., and AHUJA N., (2002). Detecting Faces in Images: A Survey. In IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 24(1): 34-58.

ZHANG J., KAYNAK M., CHEOK A., and KO C. C. (2001). Real-time lip tracking for virtual lip implementation in virtual environments and computer games. In The 10th IEEE International Conference on Fuzzy Systems Proceedings, 3: 1359 – 1362.

ZHANG X., MERSEREAU R. M., and CLEMENTS M. A. (2002). Audio-visual speech recognition by speechreading. The 10th IEEE Digital Signal Processing (DSP) Workshop, 1069-1072.

ZHILIN W., ALEKSIC P. S., and KATSAGGELOS A. K. (2002). Lip tracking for MPEG-4 facial animation. In Fourth IEEE International Conference on Multimodal Interfaces Processing, 293–298.

ZHILIN W. y ALEKSIC P. S. (2004). Inner lip feature extraction for MPEG-4 facial animation. In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2(iii): 633-636.