

Implementación de una Interfaz Hombre-Máquina para el Control de un Brazo Robótico Mediante Posturas Labiales

William Alfredo Castrillón Herrera



**Universidad Nacional de Colombia Sede Manizales
Facultad de Ingeniería y Arquitectura
Departamento de Ingeniería Eléctrica, Electrónica y
Computación
Grupo Percepción y Control Inteligente
Manizales
2009**

Implementación de una Interfaz Hombre-Máquina para el Control de un Brazo Robótico Mediante Posturas Labiales

William Alfredo Castrillón Herrera

Trabajo de grado como requerimiento parcial para optar al título de
M.Eng. Automatización Industrial

Director

Flavio Augusto Prieto Ortiz

**Universidad Nacional de Colombia Sede Manizales
Facultad de Ingeniería y Arquitectura
Departamento de Ingeniería Eléctrica, Electrónica y
Computación
Grupo Percepción y Control Inteligente
Manizales
2009**

Índice general

Lista de Figuras	III
Lista de Tablas	V
Prefacio	IX
Resumen	X
Abstract	XI
1. Revisión Bibliográfica	1
1.1. Detección del rostro	2
1.2. Segmentación de los labios	4
1.3. Seguimiento de características faciales	4
Introducción	1
2. Soporte Matemático	6
2.1. Color de la Piel	6
2.2. Segmentación del Rostro	7
2.3. Detección del Rostro Mediante el Algoritmo de Viola-Jones	10
2.3.1. Imagen integral	10
2.3.2. Extracción de Características	11
2.3.3. Clasificación	12
2.4. Segmentación de los Labios	14
2.4.1. Bordes Horizontales	14
2.4.2. Tono, saturación y valor (HSV)	14
2.4.3. Extracción de Características de los Labios Usando Exclusión de Rojos	15
2.4.4. Extracción automática de características de los labios para verificación “de vida” en autenticación de audio y vídeo	16
2.5. Modelos Estadísticos de Apariencia	18
2.5.1. Introducción	18
2.5.2. Modelos Estadísticos de Apariencia	19
2.5.3. Modelos de Forma Activa	20

2.5.4.	Modelos de Apariencia Activa	22
3.	Algoritmo	24
3.1.	Introducción	24
3.2.	Descripción General	24
3.3.	Detección de rostros basados en el color de la piel	25
3.4.	Detección de rostros basados en el detector de Viola-Jones	27
3.5.	Segmentación de características faciales con técnicas basadas en color	28
3.5.1.	Algoritmo de piel	28
3.5.2.	Detección de la región de la boca	29
3.5.3.	Extracción del contorno de los labios	30
3.6.	Segmentación de características faciales con modelos de apariencia activa	31
3.6.1.	Entrenamiento del modelo	32
3.6.2.	Extracción de los labios	33
3.7.	Seguimiento de los labios en la secuencia	33
3.8.	Clasificación de los gestos para determinar las órdenes	35
3.8.1.	Clasificación de los gestos	37
4.	Resultados	39
4.1.	Hardware y software utilizado	39
4.2.	Detección de rostros	39
4.3.	Detección y seguimiento de características faciales	40
4.3.1.	Entrenamiento	41
4.3.2.	Ajuste en el primer fotograma	42
4.4.	Modelo de apariencia activa como algoritmo de seguimiento	43
4.4.1.	Errores de Ajuste: No todo es “color” de rosa	44
4.5.	Clasificación de las órdenes para el control del brazo	44
5.	Conclusiones	48
A.	Adquisición	49
A.1.	Condiciones de Adquisición	49
A.1.1.	Arquitectura de la Escena	49
A.1.2.	Características de los equipos de adquisición	50
A.1.3.	Modificaciones de la Arquitectura de la Escena	50
A.1.4.	Espacio de Color Usado en la Adquisición	50
B.	Espacios de color para la segmentación del color de la piel	53
B.1.	Transformación de RGB a HSV	54
C.	Herramientas de Software	56
C.1.	OpenCV	56
C.1.1.	Rasgos de OpenCV	56

C.1.2. Inconvenientes de OpenCV	57
C.2. Herramienta de Modelado de Software - UML	57

Bibliografía	58
---------------------	-----------

Índice de figuras

2.1.	Distribución del color de la piel en el espacio H vs. S.	8
2.2.	Segmentación mediante los umbrales para H y S.	9
2.3.	Resultado de la detección del rostro.	9
2.4.	Imagen integral.	10
2.5.	Filtros Haar con traslación, rotación y cambios de escala.	11
2.6.	Característica Haar.	12
2.7.	Clasificador Boosting.	12
2.8.	Extracción de las esquinas de los labios usando valores en escala de grises. a) esquinas encontradas, b) suma de las escalas de grises, c) valores de grises para la fila de suma mínima.	15
2.9.	imágenes de la boca con puntos faciales extraídos.	17
2.10.	Etiquetas utilizadas para el algoritmo de AAMs.	19
2.11.	Buscar a lo largo de la normal a cada punto del modelo.	21
3.1.	Descripción General del Algoritmo.	25
3.2.	imágenes adquiridas bajo condiciones de iluminación no controlada (i.e.: con la iluminación artificial propia de la habitación).	26
3.3.	imágenes adquiridas bajo condiciones de iluminación semicontroladas(i.e.:iluminación artificial de la habitación + Luz frontal para eliminar las sombras).	26
3.4.	imágenes adquiridas bajo condiciones de iluminación no controladas(i.e.:luz del día que ingresa por una ventana).	26
3.5.	Rostros detectados con diferentes expresiones faciales.	27
3.6.	Respuesta del detector a imágenes adquiridas bajo condiciones de iluminación diferentes.	27
3.7.	rectángulo que encierra la región de piel.	28
3.8.	resultado de la aplicación del filtro definido por la Ecuación 3.1.	29
3.9.	Búsqueda del área de la boca.	29
3.10.	Detección de la región de la boca.	30
3.11.	Búsqueda de las esquinas de los labios y los límites superior e inferior de estos.	31
3.12.	Resultado del algoritmo de detección del contorno de la boca.	31
3.13.	Imágenes de ejemplo para el entrenamiento del modelo de apariencia activa.	32
3.14.	Modelo de forma promedio \bar{x}	33
3.15.	Variaciones de apariencia del Modelo de Apariencia Activa.	33

3.16. Ajuste del Modelo de Apariencia Activa a una imagen de entrada en donde se ven los puntos que describen el contorno de los labios.	34
3.17. Puntos del modelo que describen el contorno de los labios.	34
3.18. Resultado de seguimiento del algoritmo basado en color.	35
3.19. Resultado de seguimiento del algoritmo basado en AAMs.	36
3.20. Gestos para manejar los 3 grados de libertad del robot.	37
3.21. Diagrama de estados para controlar el brazo robótico.	38
4.1. Tiempos de respuesta en ms para el algoritmo de Viola - Jones en una imagen de 640x480 píxeles.	40
4.2. Proceso de iteración del modelo de apariencia activa, para ajustarse a la primera imagen luego de la detección del rostro por parte del algoritmo de Viola-Jones. . .	42
4.3. Ejemplos de ajuste y seguimiento del Modelo de Apariencia Activa.	43
4.4. Error en el ajuste del modelo de apariencia activa	44
4.5. Ejemplos de ajuste y seguimiento del Modelo de Apariencia Activa bajo condiciones de iluminación diferentes.	45
4.6. Ejemplos de clasificación de las órdenes para el control del brazo.	47
A.1. esquemas de adquisición inicial.	51
A.2. esquemas de adquisición mejorado.	52

Índice de cuadros

4.1.	Tiempos de respuesta en ms para el algoritmo de Viola Jones.	40
4.2.	Umbral de distancia a partir de los cuales una orden comienza a ser válida. El umbral para los ojos es una relación de aspecto de uno contra el otro.	46

Prefacio

La base de desarrollo de este trabajo se fundamenta en el procesamiento de imágenes y de video, apuntando al objetivo último de lograr una interfaz hombre máquina que sea amigable y útil en la vida diaria al usuario final. Se presenta una aproximación al alcance de este logro atacado desde dos técnicas distintas una basada en modelos de color para la detección de la región del rostro y posteriormente los labios y otra basada en modelos estadísticos de apariencia, la cual modela las variaciones de apariencia y forma que presenta un objeto determinado y que tiene la capacidad de adaptar su forma a éstas, este último es el que se utilizó para el desarrollo del prototipo del sistema. El desarrollo de este trabajo tuvo dos momentos en su línea de tiempo, el primero durante 2003, 2004 y parte de 2005, en el cual se abordó el problema desde el punto de vista de modelos de color. Estas técnicas tienen la ventaja de su facilidad de implementación y su alta velocidad de procesamiento, pero poseen la desventaja de que al estar basados en color, depende en gran medida de las condiciones de iluminación bajo las cuales se adquiere, lo cual hace que deban existir condiciones de iluminación controladas, que de una u otra forma pierde la funcionalidad en una aplicación del mundo real. El segundo momento se presenta hacia finales de 2007 y hasta la fecha, en donde se me presenta la oportunidad de trabajar en un grupo de investigación y además, pude retomar mis estudios de maestría a mediados del año 2008, en esta nueva etapa, realicé una nueva revisión bibliográfica y apoyándome también en el camino andado que tenía el grupo en lo que se refiere a procesamiento de vídeo y detección del rostro, pude entonces enfocar mi trabajo a técnicas mucho más robustas, como son el algoritmo de detección del rostro propuesto por Paul Viola y Michael Jones [1], y las basadas en modelos estadísticos de forma y apariencia para la detección, segmentación y seguimiento de características faciales, lo cual le dio un giro significativo al trabajo que se empezaba a retomar, las técnicas basadas en modelos estadísticos de apariencia más conocidos como Modelos de Forma/Apariencia Activa, ASMs y AAMs de sus siglas en inglés, presentan grandes ventajas a la hora de detectar y hacer seguimiento sobre las características del rostro. Además, de permitir una implementación en tiempo real, permite seguir todas las características faciales como son los ojos las cejas la nariz y la boca al mismo tiempo lo cual amplía el espectro sobre el cual podemos operar nuestro sistema de interacción hombre máquina ahora, no sólo concentrándonos sobre la boca exclusivamente, sino que también podemos aprovechar otras características faciales para hacer más cómodo y aproximado a la realidad nuestro sistema de interacción hombre máquina.

En este documento se presentan ambos momentos del trabajo de grado que se desarrolló, con las revisiones sobre el estado del arte que se tenían disponibles en cada uno de ellos y las técnicas que se usaron en estos, por eso de alguna forma en ciertos momento se notará la diferencia en los

enfoques y en otros se hará ambigua ya que finalmente se apuntó a resolver el mismo problema. Considero que este trabajo además de ser un ejercicio académico de investigación bastante interesante, apunta hacia el desarrollo de una aplicación del mundo real, en donde se pueda interactuar de manera natural con una computadora, y los algoritmos utilizados en este trabajo, cuentan con un nivel de generalidad y extensibilidad que pueden ser aplicados a diversos campos de la visión por computador.

Resumen

El rostro es la característica más distintiva y más usada para identificar a las personas e interactuar con ellas. La detección facial y la extracción de características faciales, es uno de los tópicos más ampliamente estudiados en la visión por computador, ya que estos apuntan al desarrollo de modos de interacción eficaz y natural con las máquinas (computadoras). El objetivo de este trabajo es desarrollar una interfaz hombre máquina que pueda ofrecer una “mano adicional” para controlar el laparoscopio a un cirujano, cuando se encuentra desarrollando una intervención quirúrgica de este tipo, lo que a menudo se hace muy necesario debido a que en la mayoría de las veces, éste tiene ambas manos e incluso ambos pies ocupados manipulando instrumentos quirúrgicos [2]. El alcance al que se pretende llegar con este desarrollo es controlar tres grados de libertad de un brazo robótico mediante gestos faciales, a través de técnicas de visión artificial que interpreten los gestos, y envíen estos comandos al robot. La interfaz desarrollada es un sistema de tiempo real, basado en visión artificial que puede seguir los gestos faciales de una persona y no requiere ningún tipo de dispositivo de contacto, tales como dispositivos para sensor puestos en el rostro. El cirujano puede de manera fácil y precisa, controlar el brazo de robot, haciendo simplemente los gestos faciales adecuados, sin tener que usar interruptores o comandos de voz para iniciar la secuencia de control. Este sistema permite una manipulación del robot, no invasiva, que da una “mano adicional” al cirujano lo cual puede ser mucho más conveniente para éste.

Abstract

The face is the most distinctive and most widely used by people feature, to identify and interact with them. Face detection and facial feature extraction is one of the most widely studied topics in computer vision, since they aim to develop ways of natural and effective interaction with machines (computers). The goal of this work is to develop a human machine interface that can offer to a surgeon an “additional hand”to control the laparoscope, when they are developing a surgery of this kind, which often becomes very necessary because in most cases it takes both hands and both feet occupied manipulating surgical instruments [2]. The scope that we pretends to reach with this development is to control three degrees of freedom of a robotic arm through facial gestures, using artificial vision techniques to interpret gestures, and send those commands to the robot. The interface developed is an artificial vision based real-time system that can track a person’s facial gestures and does not require any type of contact, such as devices to sensing in the face. The surgeon can easily and accurately control the robot arm, by making appropriate facial gestures, without having to use switches or voice commands to start the control sequence. This system enables a non-invasive robot manipulation, which gives an “ additional hand”to the surgeon which can be much more convenient for it.

Introducción

El rostro es la característica más distintiva y más usada para identificar a las personas e interactuar con ellas. La detección facial y la extracción de características faciales, es uno de los tópicos más ampliamente estudiados en la visión por computador, ya que estos apuntan al desarrollo de modos de interacción eficaz y natural con las máquinas (computadoras). El problema de detectar el rostro y las partes de éste, se ha convertido en una área popular de investigación, debido a las emergentes aplicaciones en interfaces humano-computador, humano-robot, sistemas de vigilancia, controles de acceso, videoconferencias, aplicaciones forenses, monitoreo de atención en conductores de vehículos, entre muchas otras. Este tipo de interfaces vienen ganando más espacio en los diferentes escenarios de nuestra vida: el trabajo, los hogares, los lugares públicos, etc. Y cada vez es mucho más fácil y natural la interacción con este tipo de sistemas, lo cual le brinda facilidades y herramientas adicionales a las personas, a las empresas y a la sociedad, haciendo que su trabajo, y su vida en general sea más productiva.

El objetivo de este trabajo no es más que eso: Desarrollar una interfaz hombre máquina que pueda ofrecer una “mano adicional” para controlar el laparoscopio a un cirujano, cuando se encuentra desarrollando una intervención quirúrgica de este tipo, lo que a menudo se hace muy necesario debido a que en la mayoría de las veces, éste tiene ambas manos e incluso ambos pies ocupados manipulando instrumentos quirúrgicos [2]. El alcance al que se pretende llegar con este desarrollo es controlar tres grados de libertad de un brazo robótico mediante gestos faciales, a través de técnicas de visión artificial que interpreten los gestos, y envíen estos comandos al robot.

La interfaz desarrollada es un sistema de tiempo real, basado en visión artificial que puede seguir los gestos faciales de una persona y no requiere ningún tipo de dispositivo de contacto, tales como dispositivos para sensar puestos en el rostro. El cirujano puede de manera fácil y precisa, controlar el brazo de robot, haciendo simplemente los gestos faciales adecuados, sin tener que usar interruptores o comandos de voz para iniciar la secuencia de control. Este sistema permite una manipulación del robot, no invasiva, que da una “mano adicional” al cirujano lo cual puede ser mucho más conveniente para éste.

La estructura de este trabajo se desarrolla de la siguiente manera: en el Capítulo 1, se hace una revisión sobre las diferentes técnicas de segmentación y caracterización facial, en el Capítulo 2, se presenta el soporte matemático de las técnicas utilizadas, en el Capítulo 3, se hace la descripción del algoritmo desarrollado para el control del brazo robótico, el Capítulo 4 presenta los resultados en términos de métricas de desempeño de los algoritmos, tiempos de ejecución, precisión del ajuste a las características faciales y finalmente las conclusiones y recomendaciones para trabajos futuros.

Capítulo 1

Revisión Bibliográfica

En los últimos años un nuevo campo ha ganado el interés de los investigadores de la robótica. Las técnicas de invasión mínima, como la laparoscopia, han tenido un gran crecimiento en el dominio de los sistemas robóticos [3]. En la cirugía laparoscópica, un asistente de cámara sostiene usualmente el laparoscopio para el cirujano y posiciona el objetivo de acuerdo a las instrucciones de éste. Dado que estos procedimientos pueden durar hasta dos horas (o más), la imagen de la cámara puede sufrir una pérdida significativa de estabilidad. Este método de operación es frustrante e ineficiente para el cirujano, por que las órdenes son a menudo interpretadas de manera errónea por el asistente. Las vistas pueden ser sub-óptimas e inestables, ya que algunas veces el objetivo es dirigido incorrectamente y vibra debido al temblor de la mano del asistente. La introducción de tecnologías robóticas (desarrollo de un sistema de posicionamiento laparoscópico robótico para reemplazar el asistente humano) es un paso en la solución de este problema, y el diseño de una interfaz hombre máquina amigable juega un importante papel en este paso. La mayoría de los sistemas de posicionamiento laparoscópicos robóticos tienen una interfaz hombre máquina, la cual requiere el uso de la mano o el pie del cirujano mediante una palanca o un pedal y estos tipos de interfaces son a menudo inconvenientes dado que la mayoría de las veces, el cirujano tiene ambas manos e incluso ambos pies ocupados, manipulando otro tipo de instrumentos. Este problema es solucionado utilizando la interfaz FAcE MOUSE presentada por Nishikawa et-al en [2]; en donde se introduce una técnica basada en el seguimiento de objetos dentro de una secuencia de vídeo (tracking), mediante la cual el cirujano puede controlar el laparoscopio por medio de una serie de movimientos de su cabeza. De esta forma se puede posicionar el laparoscopio, permitiéndole al cirujano moverlo de derecha a izquierda, de arriba a abajo y hacer zoom según sea la necesidad de este. Los sistemas basados en visión artificial para la interacción hombre máquina vienen siendo ampliamente utilizados en diversas aplicaciones tales como: en tareas de reconocimiento de personas [4], [5], reconocimiento de voz en ambientes ruidosos [6], [7], [8], interpretación de gestos para animación tridimensional [9], entre otras. Para llevar a cabo dicha tarea, estos se enfrentan a una serie de problemas relacionados con el procesamiento de vídeo. La segmentación, seguimiento, extracción de características, y clasificación dentro de una secuencia de imágenes son problemas que han sido abordados por diferentes autores, algunos de manera general [10], [11], y otros en aplicaciones específicas, como la detección y seguimiento de rostros [12], [4], [13], seguimiento (tracking) de labios [7], [14], [15], [5] y detección del tono de la piel [16], [13] entre otros. El procesamiento de video puede ser visto como un proceso dividido en varias etapas, a saber:

- Detección de objetos de interés dentro de una escena, lo cual es una de las tareas más com-

plejas si se quiere desarrollar un sistema completamente autónomo.

- Seguimiento (tracking), que consiste en la capacidad de seguir un objeto dentro de una secuencia de vídeo.

Uno de los problemas principales en el seguimiento de algún objeto dentro de una secuencia de vídeo, es el hecho de encontrar este objeto por primera vez dentro de la imagen o secuencia de imágenes. Para construir un sistema completamente automatizado que analice la información contenida en imágenes de rostros, se requiere de algoritmos robustos y eficientes para detección de rostros. Dada una imagen simple, el objetivo de un detector de rostros es identificar todas las regiones en la imagen las cuales contengan un rostro, independientemente de su posición tridimensional, orientación, y condiciones de luz [4]. En la aplicación que se pretende implementar, no todos los retos que se puedan presentar a un sistema de reconocimiento de rostros están presentes, ya que dentro de la imagen existirá sólo un rostro, lo cual hace que la búsqueda sea únicamente para una región. Las diversas técnicas para la detección de rostros dentro de una imagen de color o de intensidad están clasificadas de acuerdo a la técnica que usan para la detección; Yang [4] las clasifica en 4 categorías diferentes:

Métodos basados en el conocimiento. Estos métodos basados en reglas, codifican el conocimiento humano de qué constituye un rostro típico. Usualmente, las reglas capturan la relación entre las características faciales. Estos métodos son diseñados principalmente para localización de rostros.

Aproximaciones invariantes a características. Estos algoritmos pretenden encontrar características estructurales que existen aun cuando la pose, el punto de vista, o las condiciones varían, y entonces usan estas para localizar la cara. Estos métodos son diseñados principalmente para localización de rostros [13], [16].

Métodos de ajuste de plantillas. Muchos patrones estándar de un rostro son almacenados para describir la cara como un todo o como características faciales separadamente. La correlación entre una imagen de entrada y el patrón almacenado son calculados para detección. Estos métodos han sido usados para detección y localización.

Métodos basados en la apariencia. En contraste con el ajuste de plantillas, los modelos son aprendidos de un conjunto de imágenes de entrenamiento, el cual puede capturar la variabilidad representativa de la apariencia facial. Estos modelos aprendidos son usados entonces para detección. Estos métodos están diseñados principalmente para detección de rostros.

1.1. Detección del rostro

La detección de rostros y el tracking son tópicos que han sido investigados de manera exhaustiva desde hace muchas décadas. Se han propuesto diversas estrategias basadas en heurísticas y reconocimiento de patrones para obtener soluciones robustas y precisas. La mayoría de los métodos de detección de rostros que usan el color de la piel como el objetivo para la detección han ganado

mucha popularidad, dado que el color permite un rápido procesamiento y gran robustez a variaciones geométricas y condiciones de adquisición del patrón de la piel [13], [16]. La construcción de un sistema basado en el color de la piel como una característica para la detección de rostros, presenta tres problemas principales. Primero, qué espacio de color escoger, segundo, cómo debe ser modelada la distribución del color de la piel, y tercero, que tipo de procesamiento debe ser aplicado después de llevar a cabo la segmentación por color para la detección del rostro [16]. Los retos asociados con la detección de rostros pueden ser atribuidos a los siguientes factores:

Pose. Las imágenes de un rostro varían dada la pose relativa del rostro respecto de la cámara, y algunas características faciales como son un ojo o la nariz podría estar parcialmente o completamente ocultas.

Presencia o ausencia de componentes de estructura. Algunas características faciales tales como barba, bigote y anteojos pueden o no estar presentes, además de que puede existir una gran variabilidad entre estos componentes, incluyendo forma color y tamaño.

Expresión facial. La apariencia de un rostro se ve directamente afectado por la expresión facial de la persona [17].

Oclusiones. Los rostros pueden estar parcialmente ocultos por otros objetos. En una imagen con un grupo de gente, algunos rostros pueden estar parcialmente ocultos por los rostros de otras personas.

Orientación de la imagen. Las imágenes de rostros varían para diferentes rotaciones del eje óptico de la cámara.

Condiciones de adquisición. Cuando se forma una imagen, factores tales como la iluminación (espectro, distribución de la fuente y la intensidad de ésta) y características de la cámara (respuesta del sensor, lentes) afectan la apariencia de un rostro.

Existen muchos problemas relacionados con la detección de rostros. La localización de rostros trata de determinar la posición en la imagen de un solo rostro; éste es un problema simplificado en donde se asume que existe sólo un rostro en la imagen [4], [18]. El objetivo de la detección de características faciales es detectar la presencia y posición de características como los ojos, la boca, la nariz, las ventanas de la nariz, orejas, etc, asumiendo de que existe solo un rostro en la imagen [19], incluso asumiendo que sólo la característica que se desea detectar está presente en la imagen [20], [21]. En [22] se presenta una metodología capaz de reconocer rostros en imágenes y secuencias de vídeo aun cuando exista más de un rostro presente en la imagen, este método se basa en utilizar dos detectores de rostros. Un primer detector con un umbral más alto que el segundo, de tal forma que el segundo pueda detectar rostros en una región en donde el primero no. La forma de detección de estos, se utiliza también para la predicción de la posición en los siguientes fotogramas. Existen otras patentes relacionadas con el seguimiento de rostros como la KR2003073879 publicada el 19 de Septiembre de 2004 la cual presenta un “método para la detección de rostros y seguimiento de movimientos en tiempo real”, la patente KR2003062043 presenta un “Sistema para detección y seguimiento de rostros en comunicaciones de vídeo”. Dentro del marco general de la detección y reconocimiento de objetos, hoy día, se destaca sobremanera el desarrollado por Paul Viola y Michael Jones [1].

1.2. Segmentación de los labios

Como se menciona en [23], la extracción precisa de características de los labios para el reconocimiento es el primer paso y uno de los más importantes en la tarea de reconocimiento audiovisual de la voz (AVSR por su sigla en inglés). El primer conjunto de características que usualmente se extraen son las esquinas de los labios. La técnica para la extracción de las esquinas es muy similar para muchos algoritmos [7, 23, 24, 25] que se basan en la búsqueda de máximos y/o mínimos dentro de la imagen, para obtener así las esquinas de los labios. Rainer Stiefelhagen en [7] utilizan la proyección horizontal P_h para encontrar la posición vertical de la línea de los labios. En [23] se utiliza una técnica llamada exclusión de rojos, en donde se hace la segmentación de los labios utilizando los canales azul y verde de la imagen, basados en la hipótesis de que al ser los labios predominantemente rojos, son áreas casi negras en estos dos canales. En [25] se propone una técnica para extracción automática de los labios, en donde el rostro es detectado usando las componentes de color azul y roja de la imagen, combinada con restricciones geométricas y comparación con una cara promedio. La región de los labios es determinada usando derivaciones de las funciones de tono y saturación, combinado con restricciones geométricas. En [6] se presenta un método donde mediante modelos ocultos de Markov y modelos basados en forma y apariencia del contorno de los labios, pueden hacer lectura de estos, algo similar se presenta también en [14]. En [26] se hace uso de proyecciones de gradiente sobre la imagen para la segmentación de los labios, y además utiliza B-Splines para el seguimiento de estos dentro de una secuencia de video. En [5] se presenta un algoritmo el cual hace uso de un modelo estadístico 3D de los labios, el cual se entrena para hacer seguimiento de estos desde cualquier pose de la cabeza. En [19] se presenta un algoritmo que utiliza un modelo multiestado de los labios que combina información de color, forma y movimiento de los labios, con esto ellos pueden determinar 3 estados en la boca (abierta, cerrada y muy cerrada). En [20] y [21] se utiliza clustering espacial difuso para la segmentación de los labios, incorporando una función de forma elíptica, la cual les permite discriminar píxeles de color similar al de los labios pero que se encuentran fuera de estos.

la mayoría de los desarrollos en segmentación y seguimiento de los labios incorporan información estadística de forma y apariencia para sus algoritmos, lo cual les permite llevar a buen término esta tarea de una forma más robusta.

1.3. Seguimiento de características faciales

Los trabajos recientes en la detección de características faciales y seguimiento han utilizado la representación de formas y objetos deformables, donde la forma de la característica facial es representada por un sistema de puntos faciales.

Los modelos estadísticos se han empleado extensamente en análisis facial. Los modelos de apariencia activa propuesto por [27], es una aproximación estadística popular para representar objetos deformables, donde las características de las formas son representadas por un sistema de puntos. Los puntos de la característica son buscados por perfiles de niveles de gris, y se aplica análisis de componentes principales (PCA) para analizar los modos de variación de la forma de modo que la forma del objeto pueda deformarse solamente de las maneras específicas que se encuentran en los datos del entrenamiento. Los modelos de apariencia activa [28] y sus variaciones [29], [30],

se proponen posteriormente para combinar restricciones de variación de la forma y variación de la textura. En [31] se presenta una variación de los modelos de apariencia activa originales planteados por [28], en donde hace uso de la técnica de Ajuste de imágenes propuesto por [32], para ajustar la plantilla del modelo a la imagen, mediante un método que ellos llaman Modelo Inverso Composicional. En [33] se utilizan los Modelos de apariencia activa para interpretación de rostros, además incorporan un parámetro residual en el modelo a la hora del entrenamiento lo que mejora la calidad de ajuste del algoritmo. En [34] se utiliza una combinación de varios modelos de apariencia activa de un rostro visto desde diferentes puntos, para hacer seguimiento de un rostro visto desde diferentes ángulos, y además puede hacer síntesis de un punto de vista no observado del rostro. en [29] se desarrolló una implementación del algoritmo propuesto en [28], desarrollado en C++. En [35] se presenta una extensión del algoritmo original de AAMs, para seguimiento en 3D del rostro, además, presenta un nuevo algoritmo para la síntesis de la textura del modelo. En [30] se incorpora una nueva restricción al algoritmo de AAMs original, a través de un modelo 3D del rostro, lo cual restringe la variación del modelo AAM original el cual se desempeña en 2D. Los modelos de apariencia activa y las diferentes variaciones que han aparecido con el tiempo, han sido utilizados de manera exitosa en la extracción y seguimiento de características faciales en aplicaciones de tiempo real, incluso en aplicaciones en tres dimensiones, por lo tanto, se hará uso de estos modelos para la implementación de la aplicación que se desea desarrollar.

2.1. Color de la Piel

El color de la piel humana ha sido usado, y ha probado ser una característica muy efectiva en muchas aplicaciones desde detección de rostros hasta seguimiento de las manos. Aunque diferentes personas tienen diferentes colores de piel, muchos estudios han demostrado que la mayor diferencia se presenta en la intensidad y no en su cromaticidad. Muchos espacios de color han sido utilizados para etiquetar píxeles como pertenecientes a la piel, entre ellos tenemos RGB, RGB normalizado, HSV(HSL), YCrCb, YIQ, CIE XYZ y CIE LUV.

Muchos métodos han sido propuestos para construir un modelo del color de la piel. El modelo más simple es definir una región del tono de la piel usando los valores CrCb (ie. $R(Cr,Cb)$), de las muestras de píxeles de color de piel. Escogiendo umbrales $[Cr_1, Cr_2]$ y $[Cb_1, Cb_2]$, un píxel es clasificado que tiene tono de piel si su valor está entre estos valores.

Crowley y Coutaz [36] usaron un histograma $h(r, g)$ de los valores (r, g) en el espacio de color RGB normalizado. En otras palabras un píxel es clasificado como de piel si $h(r, g) \geq \tau$, donde τ es seleccionado empíricamente del histograma de las muestras. Eckert [37] propone un método basado en umbral similar al descrito para el espacio YCrCb en donde se obtienen los valores de umbral para Cr y Cb, la diferencia radica en que el espacio de color utilizado es el HSL y los umbrales obtenidos se aplican a las componentes H y S. El espacio de color HSL al igual que espacio YCrCb son los que presentan mejor desempeño a la hora de detectar los píxeles de piel. Adicionalmente, los algoritmos que utilizan estos espacios de color requieren de un costo computacional muy bajo, ya que la detección es realizada por simple umbralización.

Usualmente, el color de la piel no es una característica suficiente para la detección o seguimiento de rostros, muchos sistemas modulares usan una combinación de análisis de formas (ojos, nariz, boca), segmentación por color, e información de movimiento (parpadeo) para localizar o seguir rostros en una secuencia de video [28], [31].

Para el desarrollo de esta aplicación se asume que solo existe un rostro dentro de la imagen, por lo tanto, la información sobre el color de la piel es suficiente para detectar el rostro dentro de ella.

2.2. Segmentación del Rostro

Como se mencionó previamente, en la primera fase de este proyecto se decidió usar la segmentación mediante el color de la piel como el método más adecuado para la localización del rostro en el ambiente estudiado aquí¹. El espacio de color considerado como el más adecuado es el espacio de color HSL (Hue, Saturation, Luminance), el cual es más cercano a la percepción humana que otros y de este modo colores similares ocupan regiones compactas y distintas del espacio. Por ejemplo en la Figura 2.1, se presenta la distribución del color de la piel en el espacio en cuestión.

El segundo problema, además de la selección del espacio de color correcto, es escoger un rango de valores que representen el color deseado para un amplio rango de diferentes secuencias. es crucial encontrar valores que pueden ser encontrados en diferentes contenidos, dado que en aplicaciones de tiempo real es imposible obtener umbrales individuales sin la interacción del usuario, al menos para la primera imagen.

En el trabajo desarrollado por Eckert [37] se encontraron unos umbrales para los valores de H y S los cuales fueron obtenidos mediante la extracción manual de las regiones del color de la piel y chequeando su distribución en el espacio HSV. Los umbrales obtenidos son presentados en la Ecuación 2.1

$$\begin{aligned} 15^\circ &\leq H \leq 30^\circ \\ 0,25 &\leq S \leq 0,55 \end{aligned} \tag{2.1}$$

después de la segmentación, se produce una imagen como lo que se muestra en las Figuras 2.2 y 2.3.

Resumiendo, el algoritmo completo del proceso de segmentación implementado en este proceso es guiado por los siguientes pasos:

Algoritmo 1 Segmentación del rostro.

Require: Transformar la imagen a HSL o HSV.

//Producir una imagen máscara, donde los píxeles son seleccionados de acuerdo a los umbrales para H y S

for $i = 1$ to $imageHeight$ **do**

for $j = 1$ to $imageWidth$ **do**

if $15^\circ \leq image(i, j, H) \leq 30^\circ$ **and** $0,25 \leq image(i, j, S) \leq 0,55$ **then**

$image(i, j) \leftarrow 1$

else

$image(i, j) \leftarrow 0$

end if

end for

end for

Aplicar un Filtro de mediana.

¹ésta decisión se reevaluó, para cuando se retomó el proyecto a inicio del año 2008

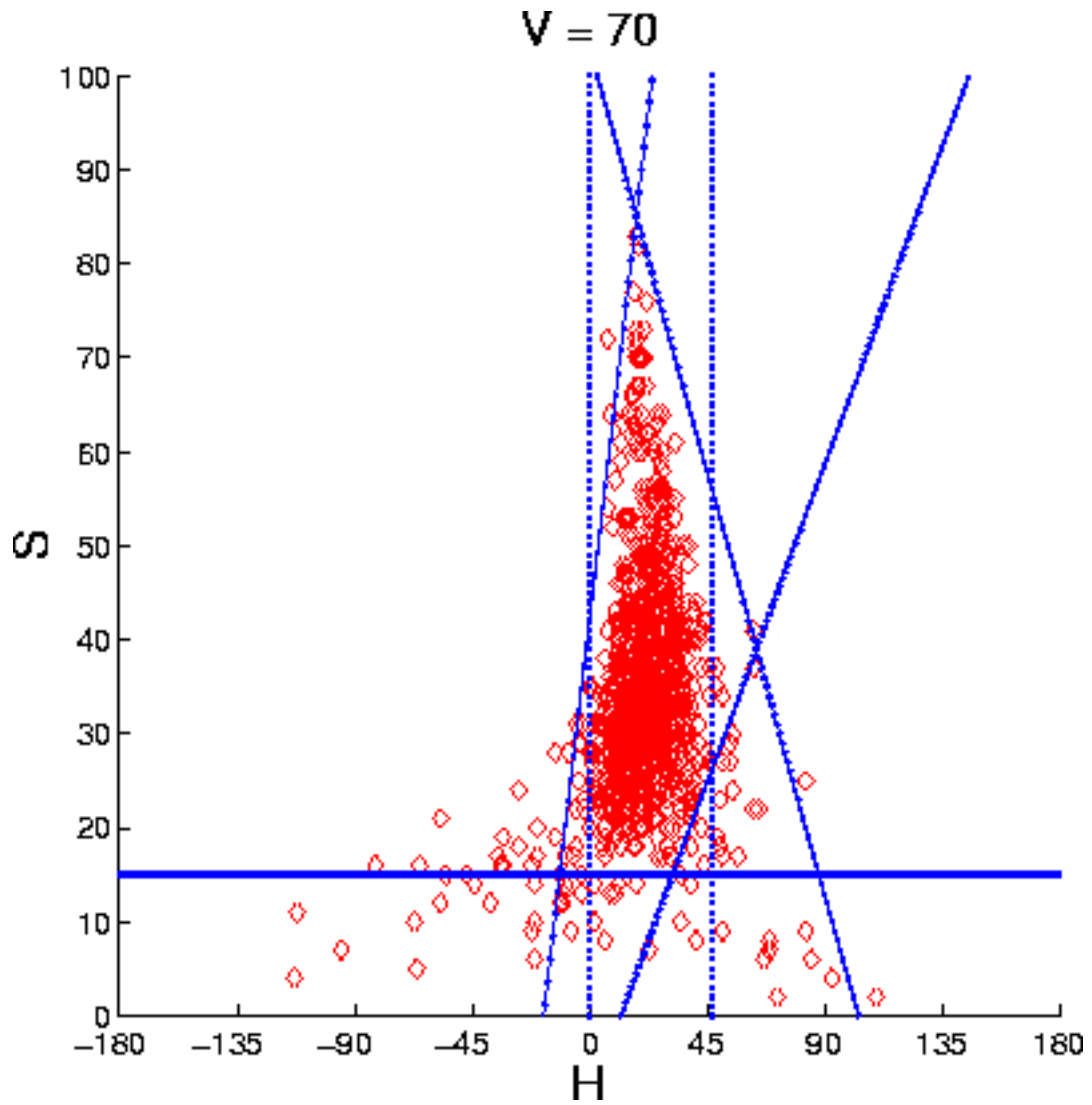


Figura 2.1: Distribución del color de la piel en el espacio H vs. S.



Figura 2.2: Segmentación mediante los umbrales para H y S.



Figura 2.3: Resultado de la detección del rostro.

2.3. Detección del Rostro Mediante el Algoritmo de Viola-Jones

Dentro del marco general de la detección y reconocimiento de objetos, la detección del rostro ha centrado numerosos estudios e investigaciones, motivados tanto por su importancia a nivel neurobiológico, como por el enorme abanico de aplicaciones prácticas en las que sería utilizable (identificación biométrica, monitorización de individuos, interfaces hombre máquina, compresión de vídeo, ...). Fruto de estos trabajos han surgido multitud de algoritmos y métodos para detectar caras en imágenes [38], [4], entre los que, actualmente, se destaca sobremanera el desarrollado por Paul Viola y Michael Jones [1].

Este algoritmo se divide en tres etapas, primero se realiza una transformación de la imagen generando una nueva, llamada imagen integral (introducida por [1]), en la segunda etapa se realiza la extracción de características usando filtros con base Haar, y por último se usa la técnica de boosting (algoritmo de AdaBoost [39]) para la construcción de clasificadores en cascada.

2.3.1. Imagen integral

Esta nueva representación de una imagen fue introducida en [1]. Permite extraer de forma rápida características a diferentes escalas ya que no se trabaja directamente con los valores de intensidad, en lugar de ello, se hace con una imagen acumulativa que se construye a partir de operaciones básicas.



Figura 2.4: Imagen integral.

En la imagen integral (ver Figura 2.4), el punto (x, y) contiene la suma de los píxeles de la parte superior izquierda de la imagen. Esta se calcula a partir de la siguiente ecuación:

$$I_I(x, y) = \sum_{x' \leq x, y' \leq y} I(x', y')$$

Donde $I_I(x, y)$ es la imagen integral y $I(x, y)$ es la imagen original.

2.3.2. Extracción de Características

La extracción de características es una etapa en el reconocimiento de patrones, en la cual las medidas u observaciones son procesadas para encontrar atributos que puedan ser usados para asignar los objetos a determinada clase [40]. En imágenes, las características de cada objeto se extraen al aplicar ciertas funciones que permitan la representación y descripción de los objetos de interés de la imagen (patrones).

La extracción de características para la detección del rostro, es realizada aplicando a la imagen filtros con bases Haar. Estos filtros pueden ser calculados eficientemente sobre la imagen integral, son selectivos en la orientación espacial y frecuencia, y permiten ser modificados en escala y orientación. En la Figura 2.5, se muestran algunos de los filtros usados para la extracción de características.

Los filtros con bases Haar, realizan una codificación de diferencia de intensidades en la imagen, generando características de contornos, puntos y líneas, mediante la captura de contraste entre regiones.

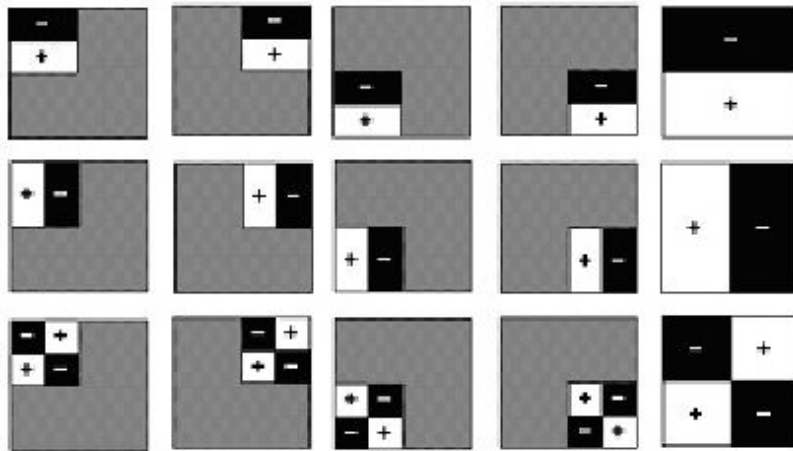


Figura 2.5: Filtros Haar con traslación, rotación y cambios de escala.



Figura 2.6: Característica Haar.

En la Figura 2.6, se muestra la convolución de un filtro con la imagen integral. La característica extraída usando este filtro se puede calcular como la diferencia entre la región oscura y la región clara. La operación se puede realizar en un tiempo constante simplemente adicionando y sustrayendo los valores de los vértices para cada rectángulo. En la Figura 2.6 la suma de los píxeles del rectángulo D se puede calcular como:

$$D = (A + D) - (B + C).$$

2.3.3. Clasificación

Esta etapa dentro del sistema de detección se encarga de realizar la clasificación del vector de características extraído de la imagen de entrada, asignándole la clase con la que se encuentra una mayor similitud, de acuerdo al modelo inducido durante el entrenamiento [41].

En esta etapa se usan clasificadores básicos para construir un clasificador en cascada (*Boosting*), que permite la clasificación rápida de características con porcentajes de detección buenos como se reporta en [1, 42, 43], en la Figura 2.7 se muestra un esquema de un clasificador en cascada.

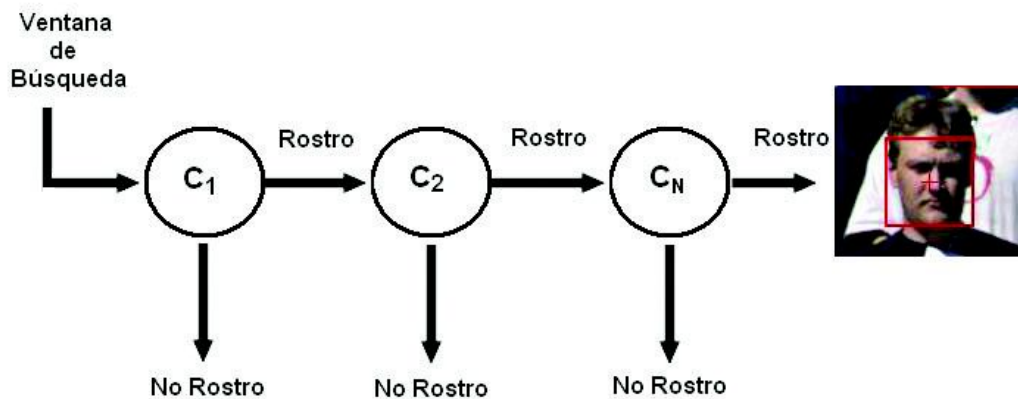


Figura 2.7: Clasificador Boosting.

Boosting fue introducido por [41], este es un método de clasificación que combina varias reglas sencillas para formar una única regla más compleja y precisa. La idea se basa en la afirmación de que varias reglas sencillas, cada una de ellas con una precisión sólo ligeramente superior a una clasificación aleatoria, pueden combinarse para formar una regla de mayor precisión, siempre y cuando se disponga de un número suficiente de muestras de entrenamiento.

Para aplicar la técnica de *Boosting* primero se debe establecer un algoritmo de aprendizaje sencillo (clasificador débil o base), que será llamado repetidas veces para crear diversos clasificadores base. Para el entrenamiento de los distintos clasificadores base se emplea en cada iteración un subconjunto diferente de muestras de entrenamiento y una distribución de pesos diferente sobre las muestras de entrenamiento. Finalmente, estos clasificadores base se combinan en un único clasificador que se espera sea mucho más preciso que cualquiera de los clasificadores base por separado.

En función de los clasificadores base que se utilicen, las distribuciones que se empleen para entrenarlos y el modo de combinarlos, podrán crearse distintas clases del algoritmo genérico de *Boosting*. El algoritmo de *Boosting* empleado por Viola y Jones en su trabajo es conocido como *AdaBoost* [41].

A continuación se muestra el algoritmo propuesto por [41].

- Dadas unas imágenes de ejemplo $(x_1, y_1), \dots, (x_n, y_n)$ donde $y_i = 0, 1$ para los ejemplos positivos y negativos respectivamente.
- Inicialice los pesos $\omega_{1,i} = \frac{1}{2m}, \frac{1}{2l}$ para $y_i = 0, 1$ respectivamente, donde m y l son el número de negativas y positivas respectivamente.
- For $t = 1, \dots, T$:

1. Normalice los pesos $\omega_{t,i} \leftarrow \frac{\omega_{t,i}}{\sum_{j=1}^n \omega_{t,j}}$

2. Seleccione el mejor clasificador débil con respecto al error pesado

$$\epsilon_t = \min_{f,p,\theta} \sum_i \omega_i |h(x_i, f, p, \theta) - y_i|.$$

3. Defina $h_t(x) = h(x, f_t, p_t, \theta_t)$ donde f_t, p_t y θ_t son minimizadores de ϵ_t .

4. actualice los pesos:

$$\omega_{t+1,i} = \omega_{t,i} \beta_t^{1-e_i}$$

donde $e_i = 0$ si el ejemplo x_i es clasificado correctamente, $e_i = 1$ de otra forma, y

$$\beta_t = \frac{\epsilon_t}{1-\epsilon_t}$$

- El clasificador final más fuerte:

$$C(x) = \begin{cases} 1 & \sum_{t=1}^T \alpha_t h_t(x) \geq \frac{1}{2} \sum_{t=1}^T \alpha_t \\ 0 & \text{de otra forma} \end{cases}$$

2.4. Segmentación de los Labios

2.4.1. Bordes Horizontales

Uno de los métodos más comunes para la extracción de las características de la boca es el uso de los valores en escala de grises y la detección de bordes [7, 25, 44]. El paso inicial, en muchas de estas técnicas, consiste en encontrar la posición vertical del centro de la boca dentro de la imagen. Esto generalmente se logra tomando la suma de cada fila y encontrando la fila con la suma mínima (ver Figura 2.8.), note que esto funciona si y sólo si la boca esta cerrada y en una posición frontal, entonces, evaluando los valores de la fila de suma mínima y posiblemente filas cercanas a ésta, puede encontrarse las esquinas de los labios mediante el establecimiento de un umbral o cualquier otra regla que ayude a encontrar las esquinas correspondientes.

Rainer Stiefelhagen en [7] utiliza la proyección horizontal P_h para encontrar la posición vertical de la línea de los labios,

$$P_h(x) = \sum_{y=1}^W I(x, y), \quad 0 \leq x \leq H \quad (2.2)$$

donde $I(x, y)$ es la función de intensidad de la imagen y H y W son el alto y el ancho de la región de búsqueda. Como la línea de los labios es la estructura horizontal más oscura presente en la imagen, entonces la búsqueda de esta se reduce a encontrar el mínimo global dentro de la función P_h . La Figura 2.8 ilustra esto.

Otro método común que usa las imágenes en escala de grises como base para la búsqueda de las esquinas, y que ha sido usado con más éxito, es el uso de bordes horizontales [25]. Esta idea es que el área de la boca tiene un alto contenido de bordes, especialmente en la dirección horizontal, esto nuevamente sólo se cumple para cuando la boca esta cerrada y en posición frontal. Estos bordes, pueden ser extraídos fácilmente mediante cualquier máscara detectora de bordes en la dirección horizontal, como por ejemplo un operador prewitt de 3×3 , y la imagen resultante puede ser umbralizada con un valor de borde apropiado, y se usa un método similar al anterior para la búsqueda de las esquinas.

2.4.2. Tono, saturación y valor (HSV)

El espacio de color de Tono, la Saturación y Valor (conocido también como intensidad) ha sido utilizado también para la extracción de información de los labios en imágenes [37]. La principal razón por la cual el espacio HSV es preferido, es por la desagregación de la iluminación respecto del color, de tal forma que variaciones en la iluminación no deben causar gran variación en el tono. Coaniz *et al.*, en [45] presentan un método basado en el espacio HSV el cual será presentado a continuación:

La probabilidad de un píxel a pertenecer a la región de los labios, está basada en un valor de tono predefinido h_0 que es representativo del tono de los labios y,

$$f(h) = \begin{cases} 1 - \frac{(h-h_0)^2}{w^2} & , |h - h_0| \leq w, \\ 0 & , \text{en otro caso} \end{cases} \quad (2.3)$$

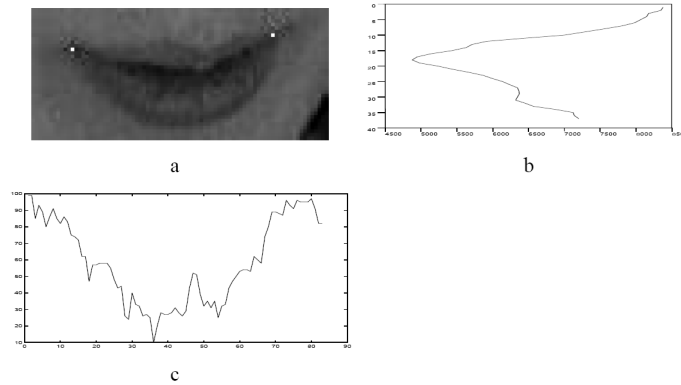


Figura 2.8: Extracción de las esquinas de los labios usando valores en escala de grises. a) esquinas encontradas, b) suma de las escalas de grises, c) valores de grises para la fila de suma mínima.

donde h representa el valor de tono actual y w controla la distancia en la cual los valores de tono de alrededor caen a cero. Este método puede ser usado para identificar varias características de los labios, como ancho y alto. Lo más notable es que la boca es difícilmente visible con respecto al área que la rodea cuando se observa la transformación de tono, que es la aplicación de la ecuación 2.3 a una imagen.

2.4.3. Extracción de Características de los Labios Usando Exclusión de Rojos

La última sección muestra que muchas de las técnicas actuales basadas en píxeles no identifican de manera adecuada las esquinas de los labios, o incluso en algunos casos, la región de los labios. Esto nos lleva a explorar otras técnicas de extracción de los labios. Esta técnica en particular, no se enfoca en buscar en el espectro del color rojo, en lugar de ello se enfoca en los valores de los colores verde y azul. La lógica detrás de esto es que los labios son predominantemente rojos lo que indica que cualquier contraste que pueda presentarse será visto en los rangos de color verde o azul, osea Exclusión de rojos. Así, después de hacer convolución con un filtro Gaussiano para remover cualquier ruido, los colores verde y azul son combinados así,

$$\log \left(\frac{G}{B} \right) \leq \beta. \quad (2.4)$$

Usando la escala logarítmica, se realza el contraste entre áreas distintivas, y variando el valor β el área de la boca y las características de los labios pueden ser plenamente identificados. En [23] este parámetro es calculado de manera manual para cada sujeto en particular. Éste método no fue implementado dado que la imagen se trata en RGB y se está trabajando en espacio HSL.

2.4.4. Extracción automática de características de los labios para verificación “de vida” en autenticación de audio y vídeo

Ésta técnica se utiliza en un sistema de reconocimiento de personas mediante autenticación audiovisual, el sistema trata de evitar los ataques de repetición de la voz mediante grabaciones y de representación del rostro mediante fotografías a través de una verificación “de vida”, osea, verifican si lo que se está diciendo coincide verdaderamente con el movimiento facial y las características acústicas dinámicas de un rostro hablando como son los movimientos articulatorios de los dientes, lengua y labios.

las características dinámicas de los labios son determinadas usando un sistema de seguimiento automático de los labios, el cual localiza los labios en la secuencia de vídeo y entonces extrae los parámetros dinámicos de los labios.

La detección y seguimiento de características faciales como los labios, es complicado en sistemas de visión por computador. Esta área de investigación tiene muchas aplicaciones en sistemas de identificación de rostros, detección de orientación de la cabeza, interacción hombre máquina, teleconferencia, etc. Para detectar los labios en un rostro, se necesitan resolver dos problemas: Segmentación espacial, que es encontrar y seguir los labios, y reconocimiento, que es estimar las características relacionadas con la configuración de los labios para la clasificación. El objetivo principal de muchos de los métodos propuestos para extracción de características de los labios, particularmente en el dominio de reconocimiento audiovisual de voz, es asistir en la tarea de reconocimiento de la voz incluyendo la modalidad visual.

Reconocimiento y detección de los labios

La detección de los labios es llevada a cabo en el primer fotograma de la secuencia de vídeo y el seguimiento de la región de los labios en fotogramas subsecuentes es completada mediante la proyección de las marcas encontradas en el primer fotograma. Esto es seguido por las mediciones sobre los límites de la región de los labios basado en la detección de bordes del pseudo tono y el tracking. La ventaja de la extracción de características faciales basados en el espacio de color HSV es que es más simple y potente en comparación con métodos basados en plantillas deformables, snakes e imágenes en pirámide.

El esquema de detección del rostro consiste de tres etapas. la primera etapa es para clasificar cada píxel en la imagen como de piel o no piel. La segunda etapa es identificar diferentes regiones de piel a través de análisis de conectividad. La ultima etapa es determinar si la región es un rostro o no.

Para la detección de la piel fue generado un modelo estadístico del color de la piel a través de entrenamiento supervisado, usando un conjunto de regiones del color de la piel. Los vectores de color C_r y C_b fueron concatenados como filas de la siguiente manera:

$$\underline{x} = (C_{r11}, \dots, C_{r1m}, C_{r21}, \dots, C_{rnm}, C_{b11}, \dots, C_{b1m}, C_{b21}, \dots, C_{bnm})^T \quad (2.5)$$

El histograma muestra que la distribución de color de la piel para diferentes personas está agrupada en este espacio de color y puede ser representado por un modelo gaussiano $N(\underline{\mu}, C)$ con media $\underline{\mu} = E[\underline{x}]$ y matriz de covarianza $C = E[(\underline{x} - \underline{\mu})(\underline{x} - \underline{\mu})^T]$.

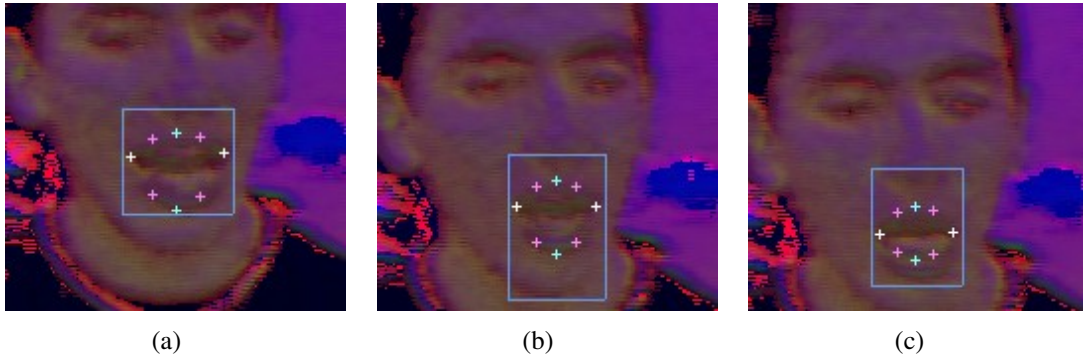


Figura 2.9: imágenes de la boca con puntos faciales extraídos.

La imagen de probabilidad de la piel es entonces obtenida y umbralizada para obtener una imagen binaria. Aplicando una regla de relación de aspecto y ajuste de plantilla con un rostro promedio, se verifica que la región sea verdaderamente un rostro.

Extracción de puntos Claves en la Región de los Labios

Una vez la región de la cara es localizada, la región de los labios es detectada usando umbralización por tono y saturación. Es fácil detectar los labios en el espacio de color de HSX ($X = L, I, V$) porque los valores de tono y saturación para los labios es muy uniforme para un amplio rango de color de los labios e igualmente para condiciones de iluminación variable y diferentes colores de piel [6]. Se obtiene una imagen binaria B a partir de las imágenes de tono H y saturación S usando umbrales $H_0 = 0,4$ y $S_0 = 0,125$ tal que $B_{ij} = 1$ para $H_{ij} > H_0$ y $S_{ij} > S_0$, y $B_{ij} = 0$ en cualquier otro caso.

En la Figura A.1 se presentan los resultados de la extracción de la boca y los puntos claves sobre diferentes posturas labiales.

2.5. Modelos Estadísticos de Apariencia

2.5.1. Introducción

Muchos problemas en la interpretación de imágenes de diferentes tipos (medicas, antropométricas, etc.), involucra la necesidad de un sistema automático que interprete la imagen que se le presenta, o sea, recuperar la estructura de la imagen y saber su significado. Esto involucra necesariamente el uso de modelos que describan y etiqueten la estructura esperada. Las aplicaciones reales se caracterizan porque a menudo hay que trabajar con estructuras complejas, imágenes que están ruidosas y posiblemente con información incompleta, bajo estas características es prácticamente imposible interpretar una imagen dada sin un conocimiento a priori de la anatomía del objeto que se busca.

Los métodos basados en modelos ofrecen potenciales soluciones a todas estas dificultades. El conocimiento a priori del problema puede resolver la confusión potencial causada por la complejidad estructural, generar tolerancia a datos ruidosos o incompletos, y proveer una forma de etiquetar las estructuras recuperadas. Se requiere aplicar el conocimiento de las formas esperadas, sus relaciones espaciales, y su apariencia de píxeles para restringir el sistema automático a interpretaciones plausibles. Un ejemplo puede ser un modelo del rostro, capaz de generar imágenes convincentes de cualquier individuo, cambiando sus expresiones. Usando tal modelo, la interpretación de imágenes, puede ser planteada como un problema de ajuste: dada una imagen a interpretar, las estructuras que se buscan, pueden ser localizadas y etiquetadas ajustando los parámetros del modelo de tal forma que se genera una imagen “producida” la cual es tan similar como es posible a la imagen real.

Cómo las aplicaciones reales implican a menudo tratar con clases de objetos los cuales no son idénticos, se requiere tratar con la variabilidad. Esto nos deja frente a la idea de modelos deformables –modelos los cuales mantienen las características esenciales de los objetos que representan, pero se pueden deformar para ajustarse a un rango de ejemplos–. Hay dos características principales deseables en un modelo deformable. Primero, debe ser general, es decir, debe ser capaz de generar cualquier ejemplo posible de la clase que el representa. Segundo, debe ser específico, es decir, debe generar ejemplos “válidos”, debido a que el objetivo de usar técnicas basadas en modelos es limitar la atención del sistema a interpretaciones posibles. Para obtener modelos específicos de objetos variables, se requiere adquirir el conocimiento de como varían estos.

Una técnica poderosa es aprender la variación de un conjunto de entrenamiento anotado que sea confiable. Más abajo se describe brevemente como se pueden construir modelos estadísticos para representar la forma y la textura (el patrón de intensidades de píxel) de estructuras de interés. Estos modelos se pueden generalizar desde el conjunto de entrenamiento y ser usados para ajustarse a nuevas imágenes, localizando la estructura, para la cual fueron entrenados, en la imagen. Se describen dos técnicas. La primera, Modelos de Forma Activa (ASM por su sigla en inglés), se concentra en ajustar un modelo de forma a una imagen, ajustando el modelo a los bordes de la estructura objetivo. La segunda técnica, Modelos de Apariencia Activa (AAM por su sigla en inglés), intenta sintetizar la apariencia completa de la imagen objetivo, escogiendo parámetros los cuales minimicen la diferencia entre la imagen objetivo y una imagen generada por el modelo. Ambos algoritmos han probado ser rápidos, precisos y confiables.

La variabilidad inter e intra-personal inherente en las estructuras biológicas hace que la inter-

pretación de imágenes faciales una tarea difícil. En los últimos años ha sido considerable el interés en métodos que usen modelos deformables para interpretar imágenes. Una motivación es obtener un desempeño robusto usando el modelo para restringir la soluciones a sólo ejemplos válidos de la estructura modelada. Los algoritmos de ajuste de modelos pueden ser clasificados como “basados en forma” en los cuales un modelo deformable representa, y se ajusta a los bordes u otra característica distribuida, y los “basados en apariencia”, en los cuales el modelo representa la región completa de la imagen cubierta por la estructura del modelo.

2.5.2. Modelos Estadísticos de Apariencia

Un modelo de apariencia puede representar la variabilidad de forma y textura que se presentan en un conjunto de entrenamiento. El conjunto de entrenamiento consiste de imágenes etiquetadas, donde puntos clave (landmark) son marcadas sobre cada objeto de muestra. Por ejemplo, para construir un modelo del rostro en imágenes 2D se necesita cierto número de imágenes marcadas con puntos en posiciones clave para resaltar las características principales. La Figura 2.10 presenta la imagen de un rostro, etiquetada con 68 puntos, los cuales resaltan sus características principales.



Figura 2.10: Etiquetas utilizadas para el algoritmo de AAMs.

Dado el conjunto de entrenamiento se puede generar un modelo estadístico de la variación de la forma aplicando PCA al conjunto de vectores que describen la forma en el conjunto de entrenamiento. Los puntos etiquetados, \mathbf{x} , sobre un objeto describen la forma de ese objeto. Cualquier muestra puede ser aproximada usando:

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}_s \mathbf{b}_s \tag{2.6}$$

donde \bar{x} es el vector media de la forma, \mathbf{P}_s es un conjunto de modos ortogonales de la variación de la forma y \mathbf{b}_s es un vector de parámetros de la forma.

Para construir un modelo estadístico de la apariencia se deforma cada imagen ejemplo tal que sus puntos de control coincidan con la forma media (usando un algoritmo de triangulación). Entonces se muestrea la información de intensidad de la imagen normalizada sobre la forma cubierta por la forma media. Para minimizar el efecto de la variación global de iluminación, se normalizan las muestras resultantes.

Aplicando PCA a los datos normalizados se obtiene un modelo lineal:

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{P}_g \mathbf{b}_g \quad (2.7)$$

donde $\bar{\mathbf{g}}$ es el vector de apariencia medio normalizado, \mathbf{P}_g es un conjunto de modos ortogonales de variación de apariencia, y \mathbf{b}_g es el conjunto de parámetros de la variación de la apariencia. La forma y la apariencia de cualquier imagen de ejemplo puede ser representado por los vectores \mathbf{b}_s y \mathbf{b}_g . Dado que puede haber correlación entre la forma y las variaciones de intensidad, se concatenan los vectores, se aplica PCA nuevamente y se obtiene un modelo de la forma:

$$\begin{pmatrix} \mathbf{W}_s \mathbf{b}_s \\ \mathbf{b}_g \end{pmatrix} = \mathbf{b} = \begin{pmatrix} \mathbf{Q}_s \\ \mathbf{Q}_g \end{pmatrix} \mathbf{c} = \mathbf{Q} \mathbf{c} \quad (2.8)$$

donde \mathbf{W}_s es una matriz diagonal de pesos para cada parámetro de la forma, \mathbf{Q} es un conjunto de modos ortogonales y \mathbf{c} es un vector de parámetros de apariencia que controla la forma y las intensidades del modelo.

Nótese que la naturaleza lineal del modelo permite expresar la forma y la textura directamente como una función de \mathbf{c}

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}_s \mathbf{W}_s^{-1} \mathbf{Q}_s \mathbf{c}, \quad \mathbf{g} = \bar{\mathbf{g}} + \mathbf{P}_g \mathbf{Q}_g \mathbf{c} \quad (2.9)$$

una forma \mathbf{X} , en una imagen, puede ser generada aplicando una transformación apropiada a los puntos $\mathbf{x} : \mathbf{X} = \mathbf{S}_t(\mathbf{x})$. Normalmente S_t será una transformación de similitud descrita por un escalamiento s una rotación, θ , y una traslación (t_x, t_y) . Por linealidad se representa el escalamiento como (s_x, s_y) donde $s_x = (s \cos(\theta) - 1)$, $s_y = (s \sin(\theta))$. El vector de parámetros de pose $t = (s_x, s_y, t_x, t_y)^T$ es entonces cero para la transformación identidad y $S_{t+\delta t}(x) \approx S_t(S_{\delta t}(x))$. La textura en la imagen es generada aplicando un escalamiento y un offset a las intensidades, $\mathbf{g}_{im} = \mathbf{T}_u(\mathbf{g}) = (\mathbf{u}_1 + \mathbf{1}) \mathbf{g}_{im} + \mathbf{u}_2 \mathbf{1}$, donde \mathbf{u} es el vector de transformación de parámetros, definido tal que $\mathbf{u} = \mathbf{0}$ es la transformación identidad y $T_{\mathbf{u}+\delta \mathbf{u}}(\mathbf{g}) \approx T_{\mathbf{u}}(T_{\delta \mathbf{u}}(\mathbf{g}))$. Una reconstrucción total está dada mediante la generación de la textura en la forma media, entrelazando (warping) ésta tal que los puntos del modelo caigan sobre los puntos de la imagen, \mathbf{X} .

2.5.3. Modelos de Forma Activa

Dada una aproximación regular inicial de una forma, una instancia de un modelo puede se ajustada a una imagen. Escogiendo un conjunto \mathbf{b} de parámetros para el modelo, se puede definir la forma

del objeto en un espacio coordenado centrado en el objeto. Se puede crear una instancia X del modelo en el espacio de la imagen mediante la definición de la posición, orientación y escala. Este ajuste puede ser mejorado mediante una aproximación iterativa que ajuste la instancia X a una imagen.

En la práctica se busca a lo largo de la normal a cada punto del modelo (Figura 2.11). Si se asume que el límite del modelo corresponde a un borde, se puede buscar simplemente el borde más fuerte a lo largo de la normal. La posición de éste nos da la nueva localización del punto en el modelo.

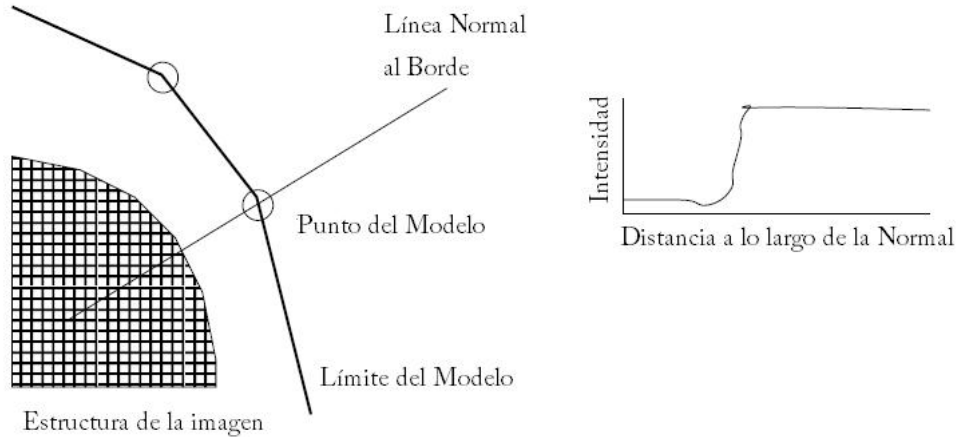


Figura 2.11: Buscar a lo largo de la normal a cada punto del modelo.

Sin embargo, los puntos del modelo no están ubicados siempre en los bordes más fuertes del vecindario de píxeles, estos pueden representar un borde secundario más débil o alguna otra estructura de la imagen. La mejor manera de ajustar es aprender lo que se está buscando en la imagen mediante un conjunto de entrenamiento. Esto se obtiene mediante el muestreo a lo largo de la normal de cada punto del modelo en el conjunto de entrenamiento, y construyendo un modelo estadístico de la estructura de escala de grises.

Modelando la estructura local

Dado un punto en el modelo, se muestrea a lo largo de los píxeles normales al punto, a cada lado de este, en la i -ésima imagen de entrenamiento. Se tienen $2k + 1$ muestras las cuales pueden ser puestas en un vector g_i . Para reducir los efectos de los cambios de intensidad se muestrea la derivada a lo largo de la línea normal al punto, en lugar de muestrear los valores de intensidad directamente. Se normaliza entonces la muestra dividiendo por la suma de los valores absolutos de las $2k + 1$ muestras.

$$g_i \rightarrow \frac{1}{\sum_j |g_{ij}|} g_i \quad (2.10)$$

Se repite para cada imagen de entrenamiento, y se obtiene un conjunto de muestras normalizadas g_i para el modelo de puntos. Se asume que estas muestras están distribuida como una distribución

gaussiana multivariada, y se estima la media $\hat{\mathbf{g}}$ y la covarianza $\mathbf{S}_{\mathbf{g}}$. Esto produce un modelo estadístico para el perfil de escala de grises alrededor del punto. Esto se repite para cada punto del modelos entregando así, un modelo para cada punto.

Modelos de Forma Activa Multiresolución

Para mejorar la eficiencia y robustez del algoritmo, este puede ser implementado en un esquema multiresolución. Esto involucra buscar el objeto en una imagen de baja resolución, y luego se refina la búsqueda en una serie de imágenes de mayor resolución. Esto conlleva a un algoritmo más rápido, y más inmune a caer en una estructura errónea de la imagen. La resolución se escala en potencias de dos (2).

2.5.4. Modelos de Apariencia Activa

Esta sección presenta el algoritmo básico de ajuste de Modelos de Apariencia Activa (AAMs). Una descripción detallada de los algoritmos de ajuste utilizados en esta tesis pueden ser consultados en [31].

Búsqueda de AAMs

Los parámetros del modelo de apariencia, \mathbf{c} , y los parámetros de transformación de la forma, \mathbf{t} , definen la posición de los puntos del modelo en el cuadro de la imagen \mathbf{X} , los cuales definen la forma de la parte de la imagen que será representada por el modelo. Durante el ajuste se muestrean los píxeles en esta región de la imagen, \mathbf{g}_{im} , y se proyectan en el modelo de textura, $\mathbf{g}_s = T^{-1}(\mathbf{g}_{im})$. la textura actual del modelo está dada por $\mathbf{g}_m = \bar{\mathbf{g}} + \mathbf{P}_g \mathbf{Q}_g \mathbf{c}$. La diferencia entre el modelo y la imagen (medido en el espacio normalizado de la textura) está dado por

$$\mathbf{r}(\mathbf{p}) = \mathbf{g}_s - \mathbf{g}_m \quad (2.11)$$

donde \mathbf{p} son los parámetros del modelo, $\mathbf{p}^T = (\mathbf{c}^T | \mathbf{t}^T | \mathbf{u}^T)$.

Una medida escalar de la diferencia es la suma de los cuadrados de los elementos en \mathbf{r} , $E(\mathbf{p}) = \mathbf{r}^T \mathbf{r}$. La expansión de Taylor de primer orden de la ecuación 2.11 da

$$\mathbf{r}(\mathbf{p} + \delta\mathbf{p}) = \mathbf{r}(\mathbf{p}) + \frac{\partial \mathbf{r}}{\partial \mathbf{p}} \delta\mathbf{p} \quad (2.12)$$

Donde el ij -ésimo elemento de la matriz $\frac{\partial \mathbf{r}}{\partial \mathbf{p}}$ es $\frac{dr_i}{dp_j}$.

El error residual durante el ajuste es \mathbf{r} . Se requiere escoger un $\delta\mathbf{p}$ tal que minimize $|\mathbf{r}(\mathbf{p} + \delta\mathbf{p})|^2$. Igualando la Ecuación 2.12 a cero se obtiene la solución RMS,

$$\delta\mathbf{p} = -\mathbf{R}\mathbf{r}(\mathbf{p}) \text{ donde } \mathbf{R} = \left(\frac{\partial \mathbf{r}^T}{\partial \mathbf{p}} \frac{\partial \mathbf{r}}{\partial \mathbf{p}} \right)^{-1} \frac{\partial \mathbf{r}^T}{\partial \mathbf{p}} \quad (2.13)$$

En un esquema de optimización estandar sería necesario recalculer $\frac{\partial r}{\partial \mathbf{p}}$ a cada paso, lo cual es una operación costosa. Sin embargo, se asume que éste está siendo calculado en un espacio de referencia normalizado, entonces ésta puede ser considerada aproximadamente fija. De este modo puede ser estimada una sola vez del conjunto de entrenamiento. Se estima $\frac{\partial r}{\partial \mathbf{p}}$ mediante diferenciación numérica, desplazando sistemáticamente cada parámetro desde el valor óptimo conocido sobre imágenes típicas y calculando un promedio sobre el conjunto de entrenamiento. De esta forma se puede precalcular \mathbf{R} y usar este en todas las búsquedas subsecuentes con el modelo.

Refinamiento Iterativo del Modelo

Usando la Ecuación 2.13 se puede sugerir una corrección en los parámetros del modelo basado en un residual \mathbf{r} . Esto permite construir un algoritmo iterativo para resolver el problema de optimización. Dada una estimación inicial de los parámetros del modelo \mathbf{c} , la pose \mathbf{t} , la transformación de textura \mathbf{u} , y la muestra de la imagen en la estimación actual, \mathbf{g}_{im} , un paso del procedimiento iterativo es como sigue:

1. Proyectar la muestra de textura en el modelo de textura usando $\mathbf{g}_s = T_{\mathbf{u}}^{-1}(\mathbf{g}_{im})$.
2. Evaluar el vector de error, $\mathbf{r} = \mathbf{g}_s - \mathbf{g}_m$, y el error actual, $E = |\mathbf{r}|^2$.
3. Calcular los desplazamientos predichos, $\delta \mathbf{p} = -\mathbf{R}\mathbf{r}(\mathbf{p})$.
4. Actualizar los parámetros del modelo $\mathbf{p} \rightarrow \mathbf{p} + k\delta \mathbf{p}$, donde inicialmente $k = 1$.
5. Calcular los nuevos puntos, \mathbf{X}' y la textura del modelo \mathbf{g}'_m .
6. Muestrear la imagen en los nuevos puntos para obtener \mathbf{g}'_{im} .
7. Calcular el nuevo vector de error, $\mathbf{r}' = T_{\mathbf{u}'}^{-1}(\mathbf{g}'_{im}) - \mathbf{g}'_m$.
8. Si $|\mathbf{r}'|^2 < E$ entonces se acepta la nueva estimación, sino, intentar con $k = 0,5$, $k = 0,25$, etc.

Este procedimiento se repite hasta que se cumpla el criterio de convergencia del error o que éste no mejore a pesar de los cambios en k . Al igual que en los ASMs aquí también se puede utilizar la técnica de búsqueda multiresolución.

La base de desarrollo de este trabajo se fundamenta en el procesamiento de imágenes y de vídeo, apuntando al objetivo último de lograr una interfaz hombre máquina que sea amigable y útil en la vida diaria al usuario final.

3.1. Introducción

El algoritmo desarrollado para este trabajo está dividido en dos partes: la parte de procesamiento de imágenes, en donde se debe hacer la detección, segmentación, seguimiento de las características faciales e interpretación de las mismas; y la parte de la máquina de estados en donde se determina el comando que se debe enviar al robot.

Dentro de estos se destaca por obvias razones el algoritmo de procesamiento de imágenes, ya que este es el problema básico a resolver cuando se desarrolla una interfaz hombre máquina de este tipo. El algoritmo de visión fue dividido, como se mencionó arriba, en cuatro problemas básicos, la detección del rostro, la segmentación de las características faciales, el seguimiento de éstas y la interpretación de las posturas labiales. Para cada uno de estos se implementaron diferentes técnicas encontradas en la literatura, de las cuales las que presenten mejor desempeño serán las que se integrarán en el algoritmo final de control del robot.

3.2. Descripción General

En la Figura 3.1 se muestra una diagrama general de la forma en la que fluye la información en el algoritmo implementado. La etapa inicial es la adquisición, ésta se describe mediante dos técnicas de acuerdo a las necesidades de cada algoritmo. Una de ellas es descrita en el Apéndice A y la segunda se hizo simplemente mediante una cámara WEB e iluminación día. Esta información es llevada al computador mediante una tarjeta de vídeo o a través de una interfaz USB. Luego de adquirir la imagen esta se utiliza como insumo para el algoritmo. Lo primero que se hace es la detección del rostro, para este caso se tiene la restricción de que en la imagen se encuentra sólo un rostro. Aquí, se utilizaron dos técnicas distintas para la detección: la propuesta por [37], la cual está basada en técnicas segmentación del color de la piel en el espacio HSV; y la planteada por [1] la cual utiliza técnicas de Clasificación por Boosting para la detección de rostros dentro de una imagen. Para la segmentación se utilizaron 3 técnicas, una basada en color la cual define umbrales para el color de los labios en el espacio HSX ($X = L, V, I$), otra basada en mediciones y restricciones heurísticas la cual usa como suministro la imagen arrojada por el algoritmo de detección de la piel propuesto en [37], y la última técnica utiliza Modelos de Apariencia Activa para detectar las características faciales. En la etapa de seguimiento se hace uso de dos técnicas

de acuerdo a la técnica usada para la segmentación; para las técnicas de segmentación basadas en color y heurística se optó como técnica de seguimiento la búsqueda exhaustiva, o sea, repetir todo el proceso de segmentación en cada fotograma dentro de la secuencia, ya que esta búsqueda tarda muy poco tiempo y permite procesar la imagen en tiempo real; para la segmentación por Modelos de Apariencia Activa se utilizó como técnica para el seguimiento los mismos modelos de apariencia activa ya que la actualización del modelo es sencilla entre fotogramas consecutivos. En la extracción de características se hizo el cálculo de algunas métricas faciales y se extrajeron los puntos que conforman el contorno de los labios para determinar a partir de estos las órdenes para el robot. Por último, en lo que se refiere a procesamiento de imágenes, la información extraída pasa a la etapa de clasificación en donde se utilizaron dos técnicas, una propuesta por [46], en donde se utilizan métricas del rostro para determinar la expresión facial, y la otra basada en la distancia de la posición actual de la boca respecto de una posición patrón inicial.

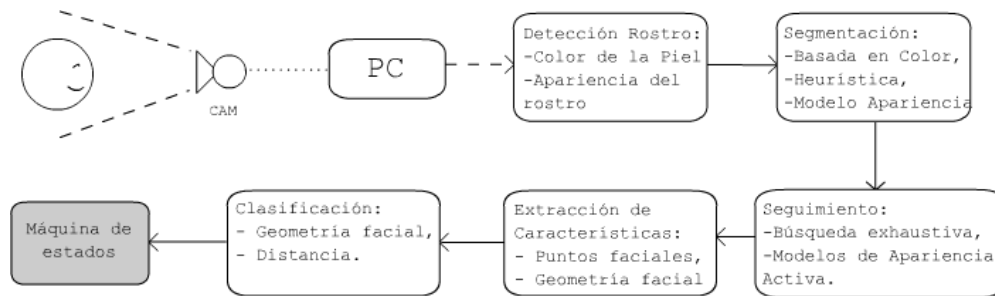


Figura 3.1: Descripción General del Algoritmo.

3.3. Detección de rostros basados en el color de la piel

El color de la piel humana ha sido usado, y ha probado ser una característica muy efectiva en muchas aplicaciones desde detección de rostros hasta seguimiento de las manos. Aunque diferentes personas tienen diferentes colores de piel, muchos estudios han demostrado que la mayor diferencia se presenta en la intensidad y no en su crominancia. Para la detección del rostro dentro de la imagen se utilizó el espacio de color HSL y la técnica presentada por [37], la cual es referenciada en la Sección 2.2 y su algoritmo descrito en el Algoritmo 1.

En las Figuras 3.2, 3.3, 3.4 se presenta el resultado de la aplicación del algoritmo propuesto por [37] para detección de píxeles de piel, algo notorio es que aunque presenta cierta invariabilidad al tono de piel de las personas, es altamente dependiente de la iluminación de la escena. En la Figura 3.2 la escena se encuentra iluminada por la luz de las lámparas de techo que posee la habitación (ver Figura A.1), para este caso el algoritmo funciona aunque de manera pobre, para la Figura 3.3, se agregó una luz en frente del sujeto (ver Figura A.2) y esto mejoró notablemente el desempeño del algoritmo, en la Figura 3.4 no se utilizó ningún tipo de luz artificial, sólo se adquirió la imagen con la luz que ofrecía el día, en este caso se puede notar que el algoritmo de Eckert [37] no logra sortear el problema de iluminación que está presente.

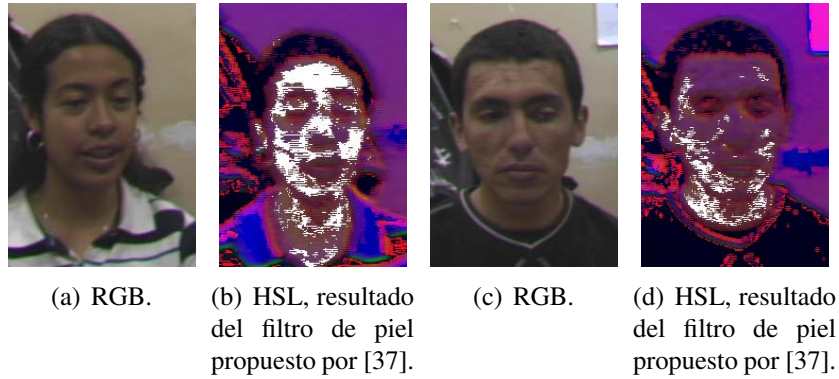


Figura 3.2: imágenes adquiridas bajo condiciones de iluminación no controlada (i.e.: con la iluminación artificial propia de la habitación).



Figura 3.3: imágenes adquiridas bajo condiciones de iluminación semicontroladas (i.e.: iluminación artificial de la habitación + Luz frontal para eliminar las sombras).

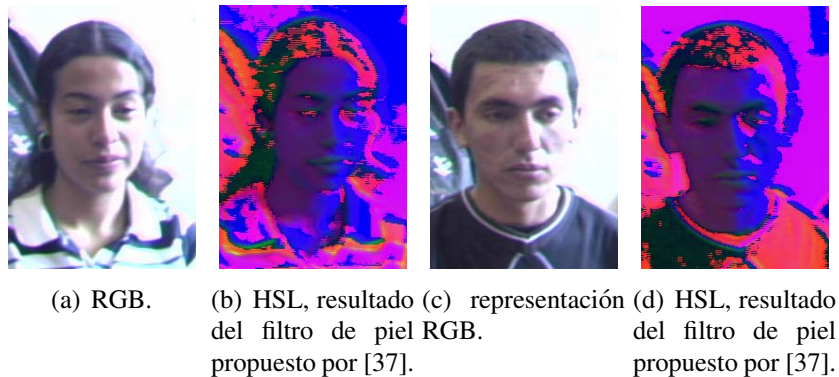


Figura 3.4: imágenes adquiridas bajo condiciones de iluminación no controladas (i.e.: luz del día que ingresa por una ventana).

Esto plantea de entrada un problema con el algoritmo de detección de rostros lo cual nos lleva

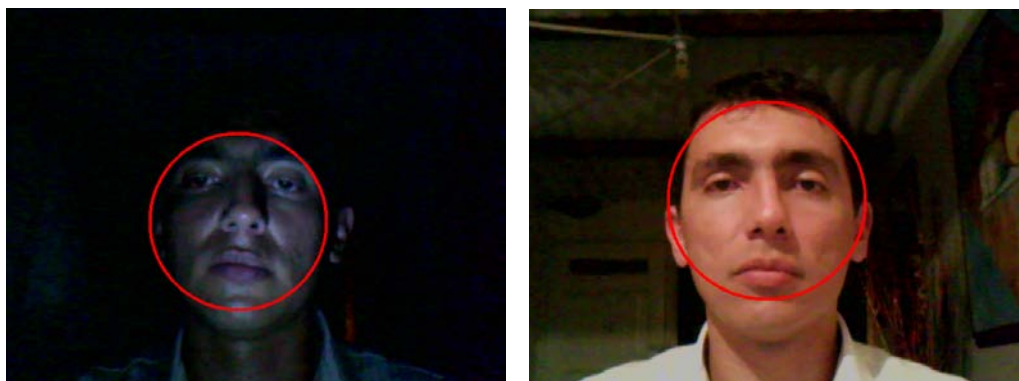
a que se deben mantener muy controladas las condiciones de iluminación si no se quieren tener problemas con la detección.

3.4. Detección de rostros basados en el detector de Viola-Jones

El algoritmo de Viola-Jones es un detector muy potente que se ha mostrado eficaz encontrando rostros, características faciales, vehículos u otros objetos en una imagen. Este algoritmo se divide en tres etapas: primero se realiza una transformación de la imagen generando una nueva llamada imagen integral (introducida por [1]), en la segunda etapa se realiza la extracción de características usando filtros con base Haar, y por último se usa la técnica de boosting (algoritmo de AdaBoost [39]) para la construcción de clasificadores en cascada. Una descripción más detallada puede ser encontrada en la Sección 2.3.



Figura 3.5: Rostros detectados con diferentes expresiones faciales.



(a) imagen adquirida con iluminación de una linterna tipo LED. (b) imagen adquirida con iluminación de techo.

Figura 3.6: Respuesta del detector a imágenes adquiridas bajo condiciones de iluminación diferentes.

Las Figuras 3.5 y 3.6, presentan ejemplos de detecciones de rostros en posición frontal logradas mediante el detector de Viola-Jones [1], nótese que ésta técnica es invariante al espacio de color, incluso puede soportar variaciones drásticas de iluminación como se puede ver en la Figura 3.6. Un análisis detallado sobre las ventajas y desventajas del detector de Viola-Jones puede ser encontrado en [47].

3.5. Segmentación de características faciales con técnicas basadas en color

Para la segmentación de los labios con técnicas basadas en color se utiliza como insumo la imagen arrojada por el algoritmo de detección de rostros basados en el color de la piel (ver Sección 3.3) con algunos procesos para acentuar la región de piel dentro de la imagen.

3.5.1. Algoritmo de piel

A partir de una región definida como el rectángulo que encierra los píxeles etiquetados como piel en el paso anterior (ver Figura 3.7), se utiliza nuevamente el filtro de piel de [37] pero esta vez utilizando sólo la componente de color H y ampliando los límites del filtro para obtener una condición de la siguiente manera:

$$0^\circ \leq H \leq 45^\circ \quad (3.1)$$



Figura 3.7: rectángulo que encierra la región de piel.

Estos límites fueron definidos a partir de análisis de la gráfica de distribución de color de la piel presentada en la Figura 2.2. Aplicando el filtro de la Ecuación 3.1 se obtiene una imagen como la presentada en la Figura 3.8.

La caja limitante del rostro se calcula a partir de un simple conteo de filas y columnas de píxeles blancos sobre la imagen binarizada: si un valor mínimo de píxeles es superado entonces se ha llegado a uno de los límites del rostro. Este simple algoritmo funciona bien para las restricciones de luz y de número de rostros (sólo uno) en la escena. A partir de esta imagen (ver Figuras 3.7 y 3.8) se hace la búsqueda de los labios.

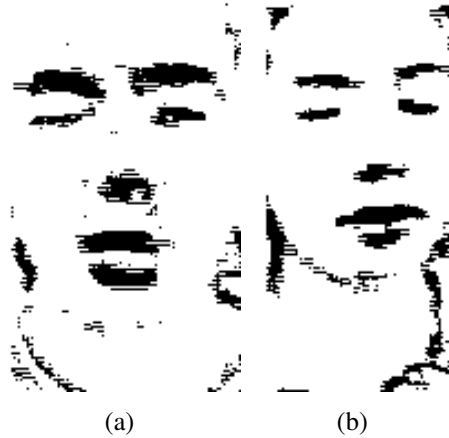


Figura 3.8: resultado de la aplicación del filtro definido por la Ecuación 3.1.

3.5.2. Detección de la región de la boca

La detección de los labios y posterior extracción del contorno se divide en dos etapas, una la detección del área de los labios (es decir, la porción rectangular dentro de la imagen en donde se encuentra la boca), y dos, la detección de las comisuras, borde superior e inferior externos de los labios, los cuales se hicieron mediante una variación de la técnica presentada en [7].

Para la detección del área de la boca se utilizó una técnica heurística a partir del conocimiento a priori de que en la imagen, existe sólo un rostro en vista casi frontal, por lo tanto, si se asume que el rostro en la imagen es un rostro adulto de proporciones antropométricas normales, la boca se encuentra del centro hacia abajo de la imagen como se muestra en la Figura 3.9.

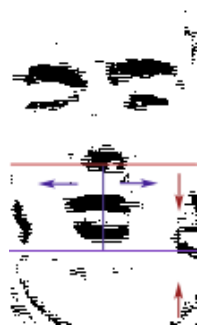


Figura 3.9: Búsqueda del área de la boca.

Para explicar el método utilizado se hará uso de la Figura 3.9. Primero se define la búsqueda para los límites izquierdo y derecho de la boca, el cual se indica por el conjunto de líneas azules, estos límites están definidos por las comisuras de la misma, entonces, lo que se hace es: iniciando desde el centro de la imagen (definido por la línea vertical azul) y hacia afuera en dirección de las flechas, se hace el conteo de píxeles negros en dirección vertical para cada columna (del centro hacia afuera), cuando se llega a la primera columna que no tenga píxeles negros o posea una cantidad muy pequeña de estos se sabrá que se ha llegado al límite (izquierdo o derecho) de la boca. Luego

de esto se procede a la búsqueda de los límites superior e inferior de la boca de acuerdo a como se indica por las flechas y la línea roja, entonces: se comienza desde la línea roja hacia abajo y desde el límite inferior de la imagen hacia arriba recorriendo la imagen por filas; nuevamente se hace el conteo de píxeles negros de cada fila y cuando se ha llegado a un número de filas consecutivas con una cantidad de píxeles negros superior a la mitad de la distancia entre los recién hallados límites izquierdo y derecho de la boca se dice que se ha encontrado el límite superior/inferior de la boca. Algunos de los resultados de este algoritmo se presenta en la Figura 3.10.

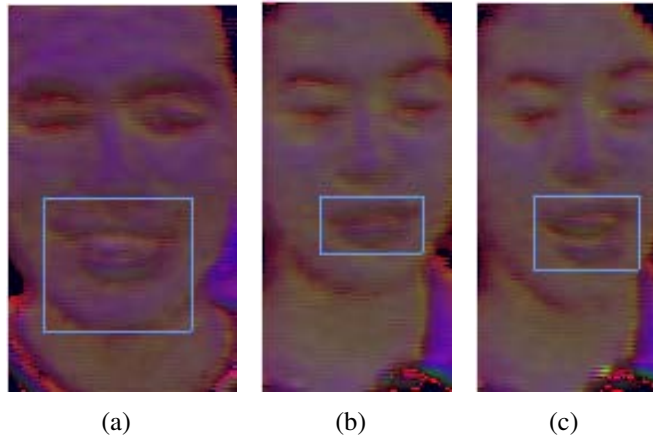


Figura 3.10: Detección de la región de la boca.

3.5.3. Extracción del contorno de los labios

Como se menciona arriba, se hizo una variación a la técnica planteada por [7] el cual utiliza la proyección horizontal de las filas para calcular en donde se encuentran las comisuras de los labios. La técnica desarrollada en este trabajo trabaja bajo el supuesto que la boca no se encuentra siempre centrada horizontalmente, esta por efectos de perspectiva, rotación de la cabeza, etc., se puede encontrar en un lugar diferente a la línea horizontal de píxeles que se define en [7], sin embargo, si conserva una posición que no se aleja demasiado de la horizontal.

Para este caso no se busca una única línea horizontal de los labios sino que se buscan dos líneas horizontales, una para cada comisura, lo que se hace es aplicar el algoritmo descrito por [7] en la primera y tercera porción del área encontrada para los labios (ver Figura 3.11.) para de esta forma hallar la fila con la suma mínima y así encontrar la posición vertical de cada una de las comisuras. Para la posición horizontal se hace uso de la imagen binarizada resultado del filtro de piel (ver Figura 3.8) en donde se busca el punto de cambio entre píxeles negros y blancos sobre la línea hallada en el paso anterior; de esta forma se determina en donde se encuentran las comisuras de los labios.

Para la búsqueda de los límites superior e inferior de los labios, se utiliza un método inspirado en la proyección vertical de píxeles de [7], sólo que aquí se busca a partir de la línea central entre las dos comisuras en un vecindario de 3 píxeles a la derecha y 3 píxeles a la izquierda (líneas azules en la Figura 3.11.)

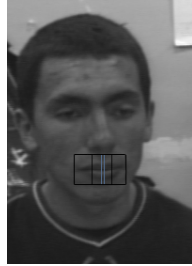


Figura 3.11: Búsqueda de las esquinas de los labios y los límites superior e inferior de estos.

Después de obtener estos límites se procede de igual forma para calcular los límites superior e inferior de los labios en la mitad derecha e izquierda de la boca. El resultado de este procedimiento puede ser visto en la Figura 3.12.



Figura 3.12: Resultado del algoritmo de detección del contorno de la boca.

3.6. Segmentación de características faciales con modelos de apariencia activa

Los modelos Estadísticos de Apariencia los cuales fueron explicados de manera general en la sección 2.5, son utilizados en diferentes aplicaciones en donde se requiere interpretar la imagen que se está presentando. Para esta aplicación se optó por utilizar los modelos de apariencia activa para la segmentación de características faciales debido a que estos han sido utilizados con gran éxito para llevar a cabo esta tarea en diferentes aplicaciones [28, 34, 33, 31].

Para este trabajo se implementó el algoritmo de AAMs presentado en [31] el cual utiliza el modelo inverso composicional para deformación y ajuste (warping) de imágenes propuesto por Simon Baker e Iain Matthews en [48], debido a que ofrece ventajas sobre el algoritmo original propuesto en [28] en la velocidad de procesamiento ya que este traslada gran parte de los cálculos que son computacionalmente intensivos a la etapa de entrenamiento del algoritmo. La descripción completa del algoritmo inverso composicional para AAMs puede ser encontrado en [31].

3.6.1. Entrenamiento del modelo

Dado que los modelos de apariencia activa se clasifican como un modelo estadístico de apariencia, éste requiere una etapa de entrenamiento en donde se le presenten al modelo las diferentes variaciones que debe aprender, estas pueden ser de variación de forma, de variación de apariencia, de variación de iluminación, etc. En la Figura 3.13 se presentan varios ejemplos que deben ser aprendidos por el modelo de apariencia activa. Estas imágenes etiquetadas marcan 68 puntos claves del rostro, de los cuales 19 pertenecen a la boca. Aquí se presentan en escala de grises debido a que la herramienta de etiquetado carga la imagen de esta forma, pero el modelo se entrena con imágenes a color.

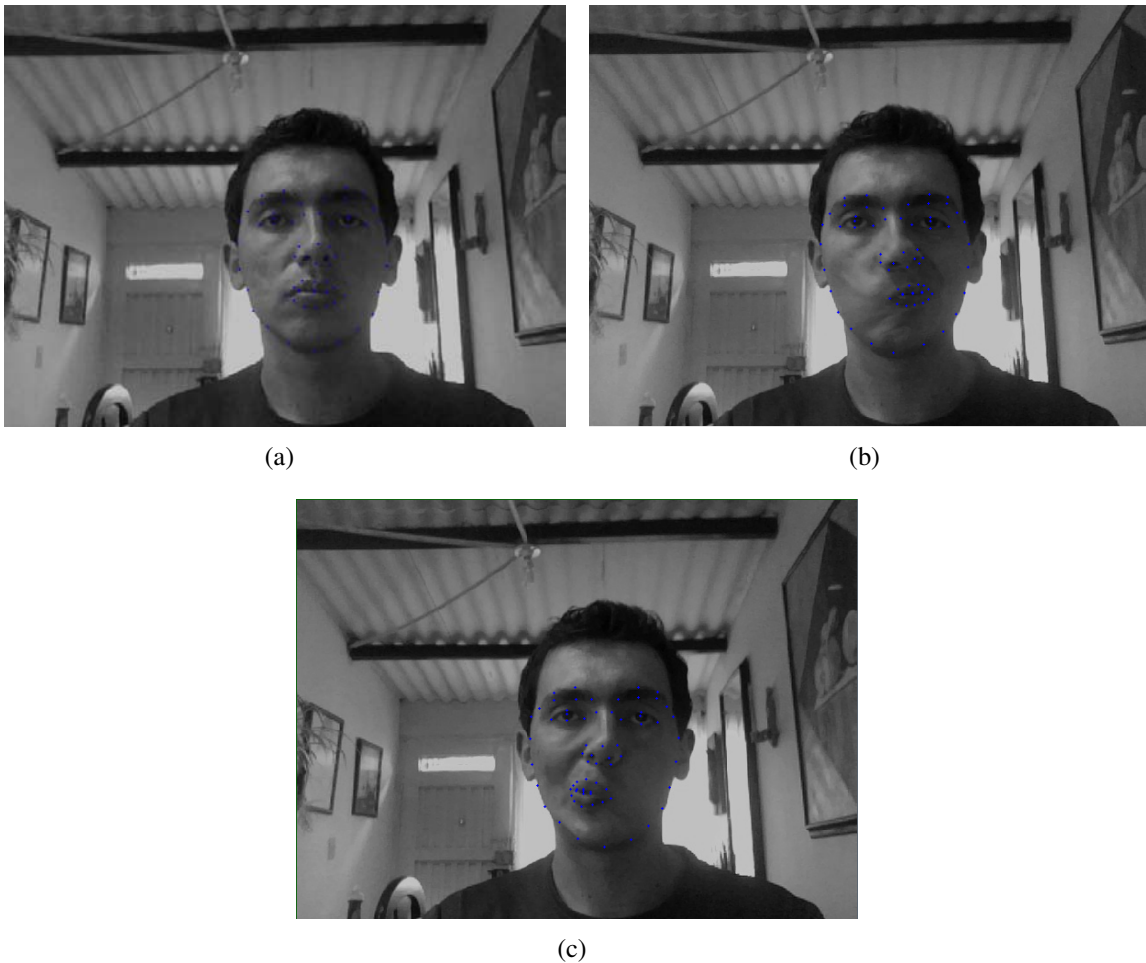


Figura 3.13: Imágenes de ejemplo para el entrenamiento del modelo de apariencia activa.

El modelo requiere que le sean presentadas todas las variaciones que se van a presentar en la etapa de ajuste, o sea, cuando el modelo se ajusta a una imagen de entrada desconocida, ya que si no se cumple esta condición el ajuste será pobre y en ocasiones impredecible.

El modelo se entrena aplicando PCA a las muestras de forma y de apariencia y se obtienen modelos de forma y apariencia como se describe en la ecuación 2.6 para la forma y en la ecuación 2.7 para

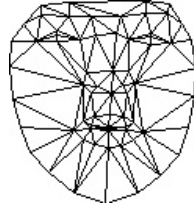


Figura 3.14: Modelo de forma promedio \bar{x} .

la apariencia, resultados de este entrenamiento se presentan en las Figuras 3.14 y 3.15

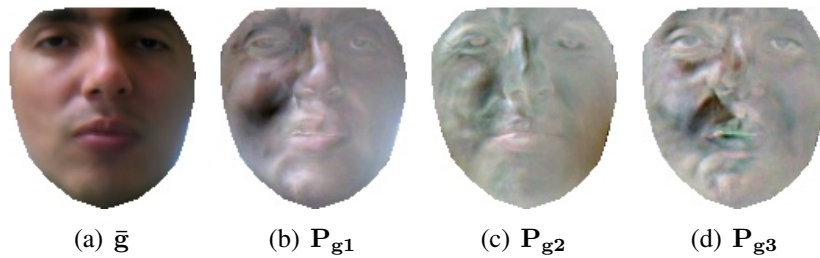


Figura 3.15: Variaciones de apariencia del Modelo de Apariencia Activa.

3.6.2. Extracción de los labios

Debido a que el modelo de apariencia activa AAM del rostro posee en sí mismo los puntos de la boca, para la extracción/segmentación de los labios, sólo se requiere determinar los puntos que describen esta forma lo cual hace que este paso sea algo muy simple. En la Figura 3.16 se presenta el resultado del ajuste del Modelo de Apariencia Activa a una imagen de entrada nueva, en donde se ven claramente los puntos que describen el contorno de los labios. El modelo de apariencia activa utilizado contiene 19 puntos para describir los labios, 12 para el contorno externo, 8 para el contorno interno (6 + 2 de las comisuras de los labios) y uno para el centro de la boca, estos puntos pueden ser vistos en la Figura 3.17.

3.7. Seguimiento de los labios en la secuencia

Basados en el tiempo que utilizan los algoritmos de detección de los labios para completar su tarea, ≈ 50 milisegundos para el algoritmo basado en el color de la piel en una imagen de 200x300 pixeles y ≈ 12 milisegundos para el algoritmo de modelos de apariencia activa en una imagen de 640x480, y a que el objeto de interés en la imagen, para este caso el rostro y los labios, no presenta cambios bruscos de velocidad y/o posición, se decidió utilizar como técnicas de seguimiento la búsqueda exhaustiva y los modelos de apariencia activa respectivamente de acuerdo a la técnica de detección.

Para el algoritmo basado en características de color simplemente se hace la búsqueda completa en cada nuevo fotograma de entrada, en la Figura 3.18 se presenta una secuencia de imágenes en



Figura 3.16: Ajuste del Modelo de Apariencia Activa a una imagen de entrada en donde se ven los puntos que describen el contorno de los labios.

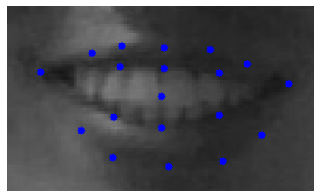


Figura 3.17: Puntos del modelo que describen el contorno de los labios.

donde se realizó el seguimiento de los puntos de la boca mediante búsqueda exhaustiva.

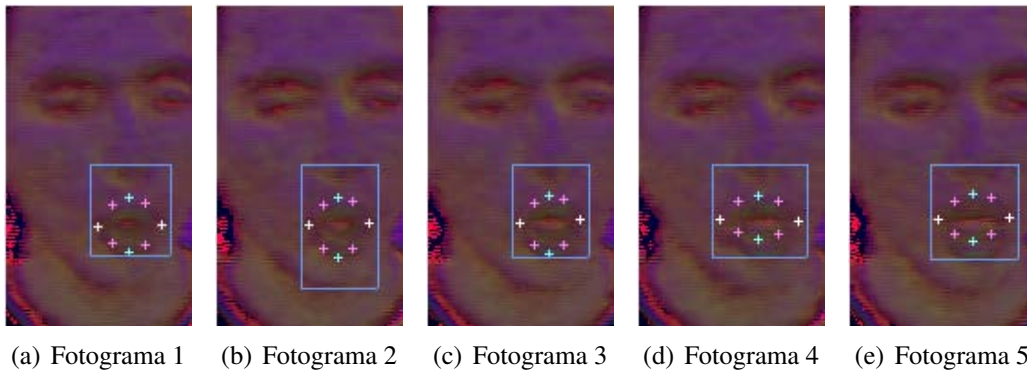


Figura 3.18: Resultado de seguimiento del algoritmo basado en color.

En la Figura 3.18 pueden verse algunas fallas en la detección de los labios debido a que la técnica utilizada allí no contiene ningún tipo de conocimiento incorporado, sobre la posición en donde se encontraban los puntos de la boca en el fotograma inmediatamente anterior.

El modelo de apariencia activa se utilizó como algoritmo de seguimiento ya que a partir de los resultados parciales, el algoritmo posee una alta velocidad de convergencia, a que los parámetros del modelo encontrados para el fotograma actual pueden ser utilizados para la búsqueda en el fotograma siguiente, y además, varios autores [35] [30], [49], han utilizado esta misma técnica para el seguimiento en sus aplicaciones, algunos incluso en 3D. En la Figura 3.19 se presenta una secuencia de imágenes consecutivas en donde se realizó el seguimiento mediante el algoritmo de modelos de apariencia activa. Aquí se puede ver claramente el nivel de precisión que logra alcanzar este algoritmo.

3.8. Clasificación de los gestos para determinar las órdenes

Con base en los resultados parciales de los algoritmos basados en color y apariencia, es notoria la superioridad de los modelos de apariencia activa sobre los algoritmos basados en color, por esto y por la capacidad de extensibilidad de los AAMs, se decidió continuar el desarrollo utilizando como insumo los resultados que arroja esta técnica.

Para el control del brazo robótico se definieron seis gestos faciales básicos para el control de los tres grados de libertad del brazo, cuatro bucales y aprovechando la capacidad de los AAMs para el seguimiento de todas las características faciales se utilizaron dos gestos de los ojos para completar el conjunto de seis órdenes.

Los gestos bucales se utilizaron para controlar las direcciones arriba, abajo, izquierda, derecha y los gestos de los ojos se utilizaron para manejar el zoom in y el zoom out. Con el fin de explicar cada gesto y su intención, estos se presentan en la Figura 3.20.

Para permitir que la interfaz sea útil en la vida real, ésta debe poseer además de los 6 comandos, una secuencia de inicio, de tal forma que el sistema pase de un estado de reposo a un estado de atención, definiendo como estado de reposo, aquél en donde la aplicación funciona haciendo la

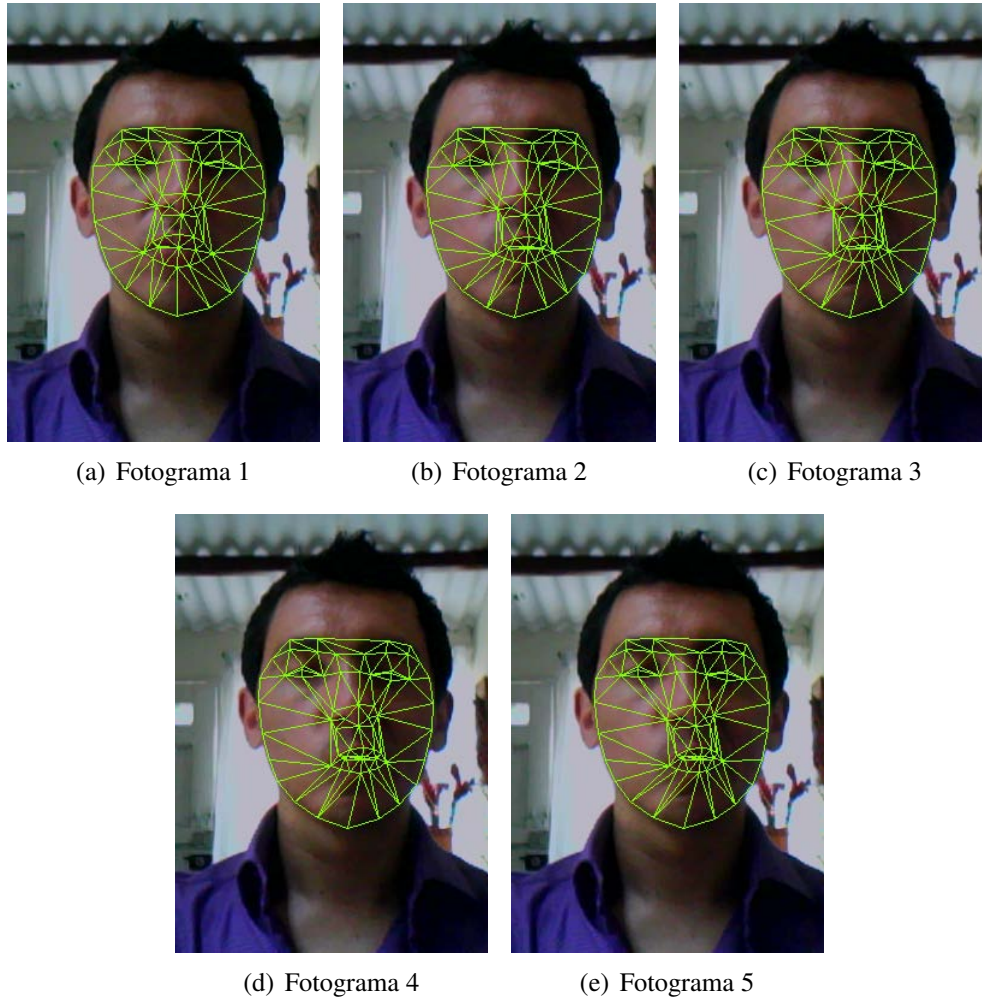


Figura 3.19: Resultado de seguimiento del algoritmo basado en AAMs.

detección y el seguimiento, mas no, la interpretación de los gestos, y como estado de atención aquél en donde cada gesto facial puede ser interpretado para tomar control del brazo robótico. Los gestos aquí utilizados para la secuencia de inicio son los mismos que se utilizan para mover el brazo robótico de izquierda a derecha. El sistema para interpretar los gestos puede ser visto como una máquina de estados finitos, en los cuales se ingresa de acuerdo con el valor de una o varias entradas. Esta máquina se presenta en la Figura 3.21.

Inicialmente se parte del estado de reposo, entonces, si la boca se sostiene con el gesto hacia la izquierda se pasa al estado de pre-atención, luego si la boca se sostiene con el gesto hacia la derecha, se pasa al estado de atención. Aquí, ya se puede controlar el brazo mediante los comandos preestablecidos, para volver al estado de reposo, se debe mantener el rostro en un gesto serio durante 4 segundos. Como se puede ver en la Figura 3.21, después de estar en el estado de atención, existen diversos caminos para pasar de un estado a otro sin regresar a un estado intermedio, esto es, que se puede mover el brazo casi como si se estuviera manipulando con un joystick.



Figura 3.20: Gestos para manejar los 3 grados de libertad del robot.

3.8.1. Clasificación de los gestos

Para la clasificación de los gestos para comandar el robot se hace lo siguiente:
para controlar las direcciones arriba, abajo, izquierda y derecha

- calcular el centro de masa de los puntos que describen la boca en el modelo,
- calcular la posición desde este centro de masa respecto al centro de masa de los mismos puntos en la imagen promedio del modelo.
- de acuerdo con la posición obtenida se determina el comando. Si la dirección es una combinación de dos posibles comandos (ej.: arriba-izquierda) prevalece el de mayor distancia hasta el centro de masa.

para controlar el zoom:

- se determina la distancia entre el parpado superior e inferior de cada ojo.
- se calcula la relación del tamaño de un ojo con respecto al otro.
- si esta relación es superior a un umbral se determina que el ojo está cerrado.
- se ejecuta el comando asociado a cada ojo.

En el capítulo 4, se presentan las medidas de desempeño de los algoritmos que se decidieron utilizar para el prototipo final.

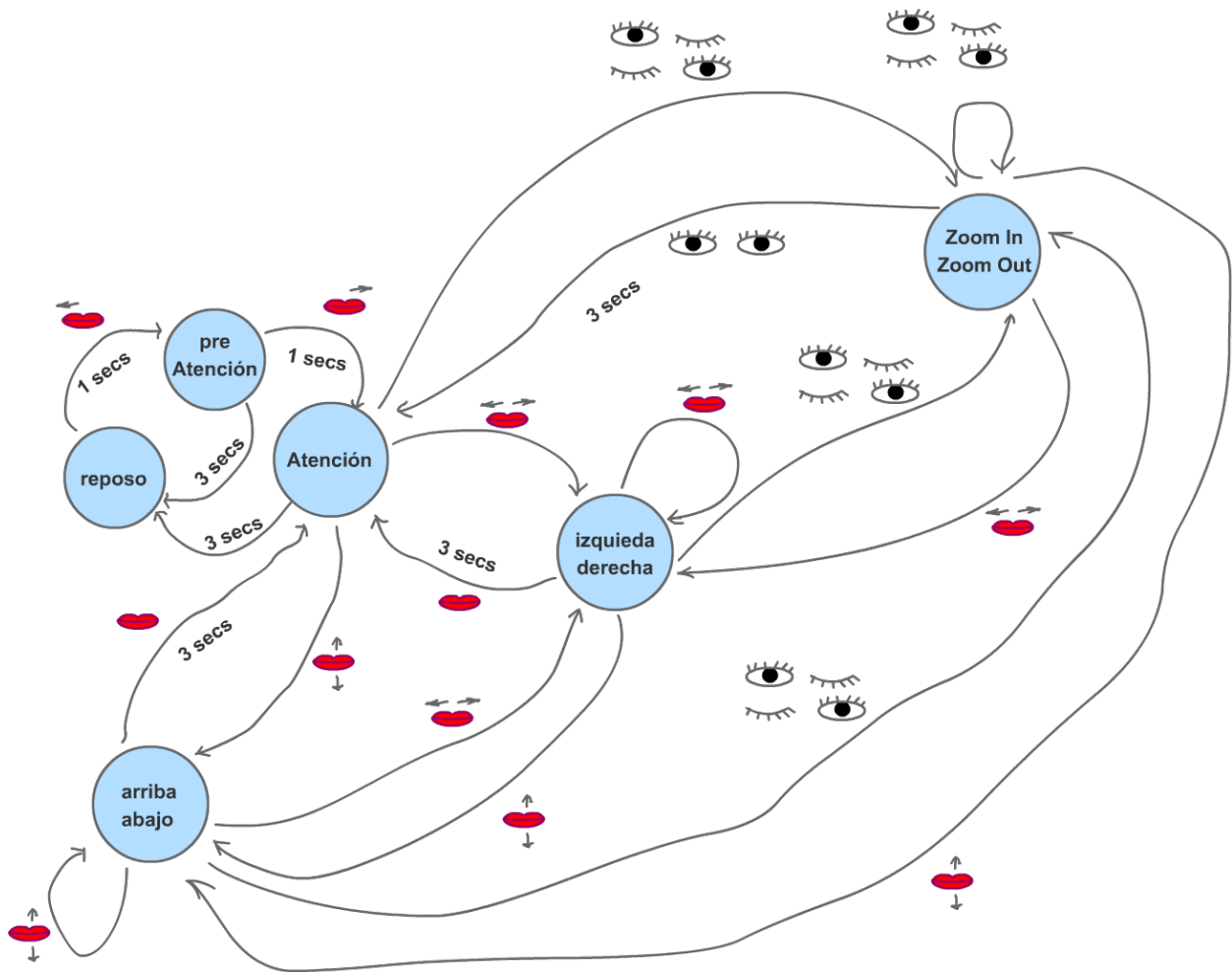


Figura 3.21: Diagrama de estados para controlar el brazo robótico.

Capítulo 4

Resultados

En este capítulo se evalúa el desempeño de los algoritmos basados en apariencia los cuales arrojaron los mejores resultados parciales en el desarrollo de este trabajo.

4.1. Hardware y software utilizado

Para el desarrollo de este trabajo se utilizó un computador portátil convencional marca DELL modelo Vostro 1510 con las siguientes características:

- Procesador Intel[®] Core[®] 2 Duo (1.8 GHz, 2 MB de caché de nivel 2)
- Sistema operativo Windows Vista[®] Home Basic.
- 4 GB de memoria DDR2 SDRAM de doble canal a 667 MHz.
- Cámara integrada de 1,3 megapíxeles(1280x1024 píxeles).

El lenguaje de programación utilizado es C++ y el compilador Microsoft[®] Visual C++ 2005.

4.2. Detección de rostros

El algoritmo para la detección del rostro que se utilizó en esta tesis es el propuesto por [50] que fue mejorado posteriormente por [51] y que está implementado dentro de la librería OpenCV[®] desarrollada por Intel[®].

En la Figura 4.1 se presenta un gráfico de tiempo de detección en una secuencia de vídeo típica, en donde se puede observar un tiempo promedio de $208ms$, un tiempo máximo de $326,651ms$ y un tiempo mínimo de $189,727ms$ en una imagen de 640×480 píxeles, este tiempo puede reducirse considerablemente haciendo un escalamiento en la imagen de tal forma que el área a procesar sea menor reduciendo así el tiempo de procesamiento del algoritmo, en la Tabla 4.1 se presentan algunos tiempos de respuesta para diferentes escalas. Este escalamiento no reduce la precisión del detector, al menos no para la aplicación actual.

Como puede verse en la Tabla 4.1, el tiempo de detección del rostro para el algoritmo puede llevarse a un valor bajo de tal forma que no represente un problema a la hora de llevar la aplicación

Escala	T Máximo	T Mínimo	T Promedio
1	326.651	189.727	207.991
1/2	175.457	41.535	48.567
1/4	57.618	9.383	11.547

Cuadro 4.1: Tiempos de respuesta en ms para el algoritmo de Viola Jones.

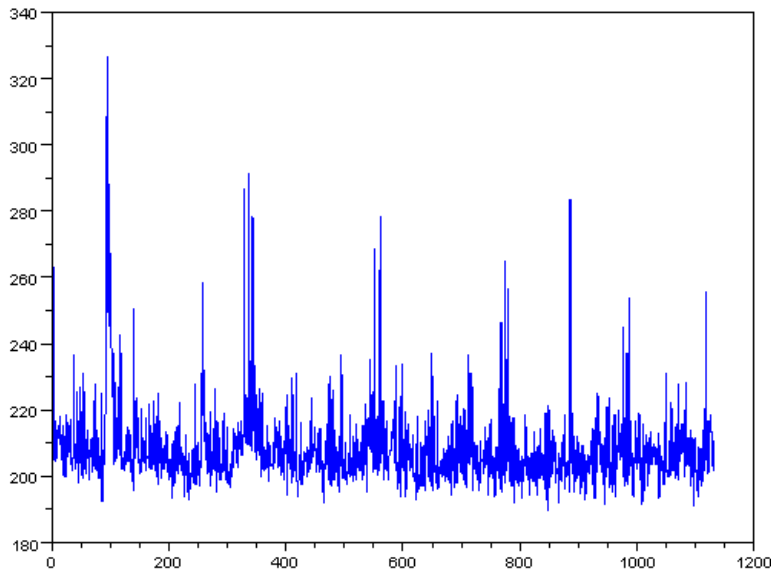


Figura 4.1: Tiempos de respuesta en ms para el algoritmo de Viola - Jones en una imagen de 640x480 píxeles.

a un ambiente de tiempo real. Sin embargo, para esta aplicación no es necesario hacer la detección del rostro en cada fotograma, basta con hacerlo en el fotograma inicial, el resto lo hace el algoritmo de modelos de apariencia activa que se utiliza como algoritmo de seguimiento. Este algoritmo sólo es utilizado de nuevo, cuando el algoritmo de seguimiento pierde las características faciales.

El algoritmo de Viola-Jones ha sido ampliamente utilizado y probado en aplicaciones de detección del rostro dentro de una imagen y en secuencias de vídeo, así que no se considera relevante hacer pruebas exhaustivas sobre el rendimiento de éste, para mayor información sobre el desempeño del algoritmo aquí utilizado, se recomienda revisar [52], [53], [1].

4.3. Detección y seguimiento de características faciales

Esta tarea conjunta se llevó a cabo mediante un modelo estadístico de apariencia, concretamente un Modelo de Apariencia Activa. El algoritmo utilizado es la implementación de la aproximación propuesta por [31], la cual adopta el modelo composicional inverso de ajuste de imágenes prop-

uesto por ellos mismos en [48], como una variación del algoritmo de alineación de imágenes propuesto por Lucas-Kanade en [32]. Por lo tanto no se evaluará cuantitativamente este algoritmo con respecto a otro algoritmo de detección o seguimiento de características. Para ver un análisis detallado sobre el desempeño del algoritmo como Modelo de Apariencia Activa véase [31]. Dado que esta tesis pretende realizar la implementación de una interfaz hombre máquina de tiempo real que pueda interpretar gestos, entonces desde esa óptica serán presentados los resultados.

4.3.1. Entrenamiento

Para el entrenamiento se utilizó un conjunto de 53 imágenes Ground-truth (etiquetadas manualmente), cada una con 68 puntos distribuidos así:

- 15 para el contorno del rostro,
- 22 para los ojos y las cejas,
- 12 para la nariz,
- 19 para los labios.

Estas imágenes se adquirieron sin ningún tipo de control sobre la iluminación de la escena, sólo cuidando que las características faciales fueran visibles al momento de la adquisición. A continuación se presentan las características de las imágenes utilizadas para el entrenamiento.

- formato de imagen: JPEG,
- alto de la imagen: 480,
- ancho de la imagen: 640,
- profundidad de bits: 24.

Además de las imágenes para el entrenamiento, se adquirieron y grabaron también secuencias de vídeo para el proceso de pruebas de los algoritmos, estos fueron capturados en formato AVI sin compresión haciendo uso de la utilidad de manipulación de vídeo de la librería OpenCV[®]. A continuación se presentan las características de las secuencias de vídeo utilizadas para el desarrollo del presente trabajo.

- formato de video: AVI,
- compresion: 0 (ninguna),
- fotogramas por segundo: 30,
- alto de la imagen: 480,
- ancho de la imagen: 640,

- profundidad de bits: 24,
- tamaño de imagen: 921600(640x480x3).

La calidad de las imágenes adquiridas tanto en el entrenamiento como en las pruebas, es lo suficientemente buena para lograr el alcance que se pretende con este trabajo. Por lo tanto no fue necesario el desarrollo de algoritmos adicionales para el almacenamiento de las secuencias de prueba y entrenamiento.

Como se definió en la sección 2.5.2, ecuaciones 2.6 y 2.7, se aplica PCA a los datos de entrenamiento, y para este conjunto, el algoritmo arroja 7 vectores propios que describirán las variaciones de forma y 23 que se encargarán de las variaciones de apariencia. En las Figuras 3.14 y 3.15, se presentan la forma y la apariencia promedio junto con los primeros tres modos de variación para este modelo.

4.3.2. Ajuste en el primer fotograma

en la Figura 4.2 se presenta la evolución del modelo de apariencia activa mientras se ajusta al rostro en el primer fotograma.

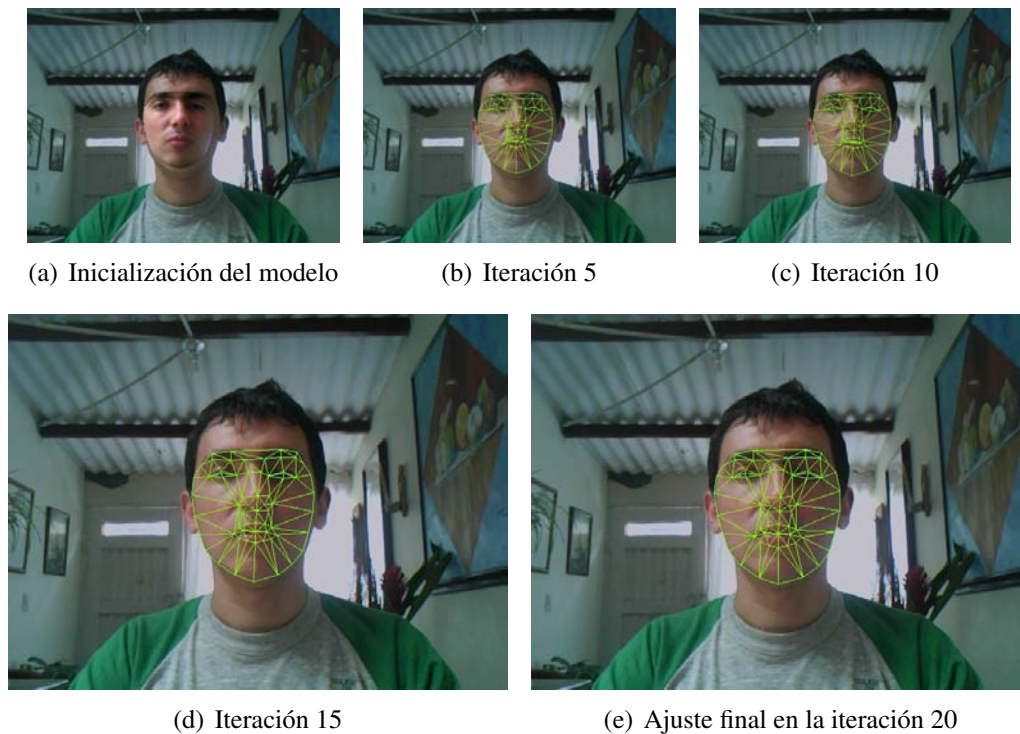


Figura 4.2: Proceso de iteración del modelo de apariencia activa, para ajustarse a la primera imagen luego de la detección del rostro por parte del algoritmo de Viola-Jones.

El error de ajuste en la región de los labios es de 2.3 píxeles como máximo, y 0.41 como mínimo en cada uno de los 19 puntos que describen ésta. El tiempo que se toma para ajustarse a esta primera

imagen es $\approx 60ms$. Este tiempo es alto pensando en el sistema de tiempo real, pero más adelante se mostrará que este tiempo de ajuste se ve reducido altamente gracias al algoritmo de seguimiento.

4.4. Modelo de apariencia activa como algoritmo de seguimiento

Como se expuso en el capítulo 3, el modelo de apariencia activa utilizado para la detección de las características faciales, fue utilizado también como algoritmo de seguimiento con resultados bastante satisfactorios en cuanto a velocidad de procesamiento y capacidad de seguimiento de las características faciales. En la Figura 4.3 se presentan algunos ejemplos de ajuste y seguimiento del algoritmo y en la Tabla 4.4 se presentan los tiempos, el número de iteraciones y el error en píxeles para el ajuste en la etapa de seguimiento.

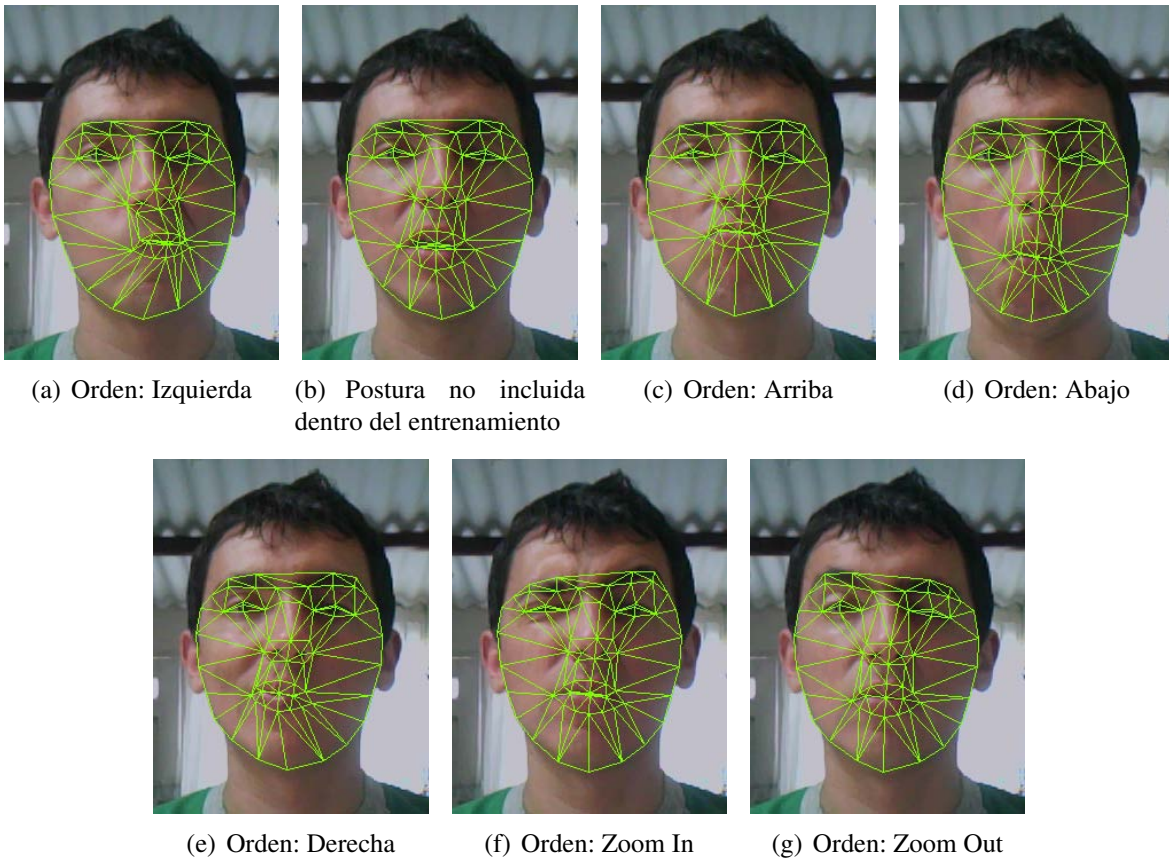


Figura 4.3: Ejemplos de ajuste y seguimiento del Modelo de Apariencia Activa.

El número máximo de iteraciones presentado en la Tabla 4.4, fue de 12 en esta secuencia de prueba, este número junto con el error máximo de píxeles, corresponde al ajuste presentado en la Figura 4.3(b), el cual se debe a que la postura labial a la que se ajustó el modelo, no se encontraba presente dentro de las imágenes de entrenamiento del algoritmo. Sin embargo, el Modelo de apariencia acti-

	Mínimo	Máximo	Promedio
Iteraciones	1	12	2
Tiempo(ms)	3,148	42,446	6,529
Error en Píxeles	0	5	1,5

va, sorteó medianamente bien este problema (desde el punto de vista de que no se alejó demasiado del contorno real de los labios).

4.4.1. Errores de Ajuste: No todo es “color” de rosa

Ya se han presentado las diferentes ventajas de los modelos de apariencia activa aquí utilizados, pero como en la mayoría de las técnicas, existen casos para los cuales el algoritmo falla. Debido a que los modelos de apariencia activa se basan en modelos aprendidos mediante imágenes de entrenamiento previas, estos fallarán cuando se presentan casos para los cuales no fueron entrenados, por ejemplo posturas o poses no contempladas en el entrenamiento, condiciones de iluminación diferentes a las con que se entrenó el algoritmo, etc. En la Figura 4.4 se presenta un ejemplo de error en el ajuste para un cambio en la iluminación de la escena.

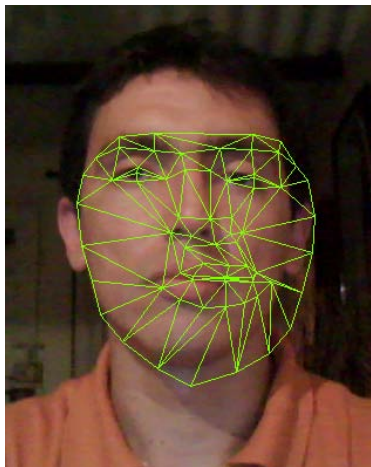


Figura 4.4: Error en el ajuste del modelo de apariencia activa

Este error puede ser corregido, presentándole al algoritmo cuando menos una (1) imagen de entrenamiento en donde se presente este tipo de iluminación. En la Figura 4.5 se muestra el ajuste después de haber entrenado el modelo con una imagen que incluía esta iluminación.

4.5. Clasificación de las órdenes para el control del brazo

Para la clasificación de los gestos faciales se utilizó el método planteado en el capítulo 3, adicionalmente se debieron agregar restricciones de distancia (umbrales) a partir de las cuales una postura se consideraba válida, en la Tabla 4.2 se presentan las distancias umbral elegidas para cada postura.

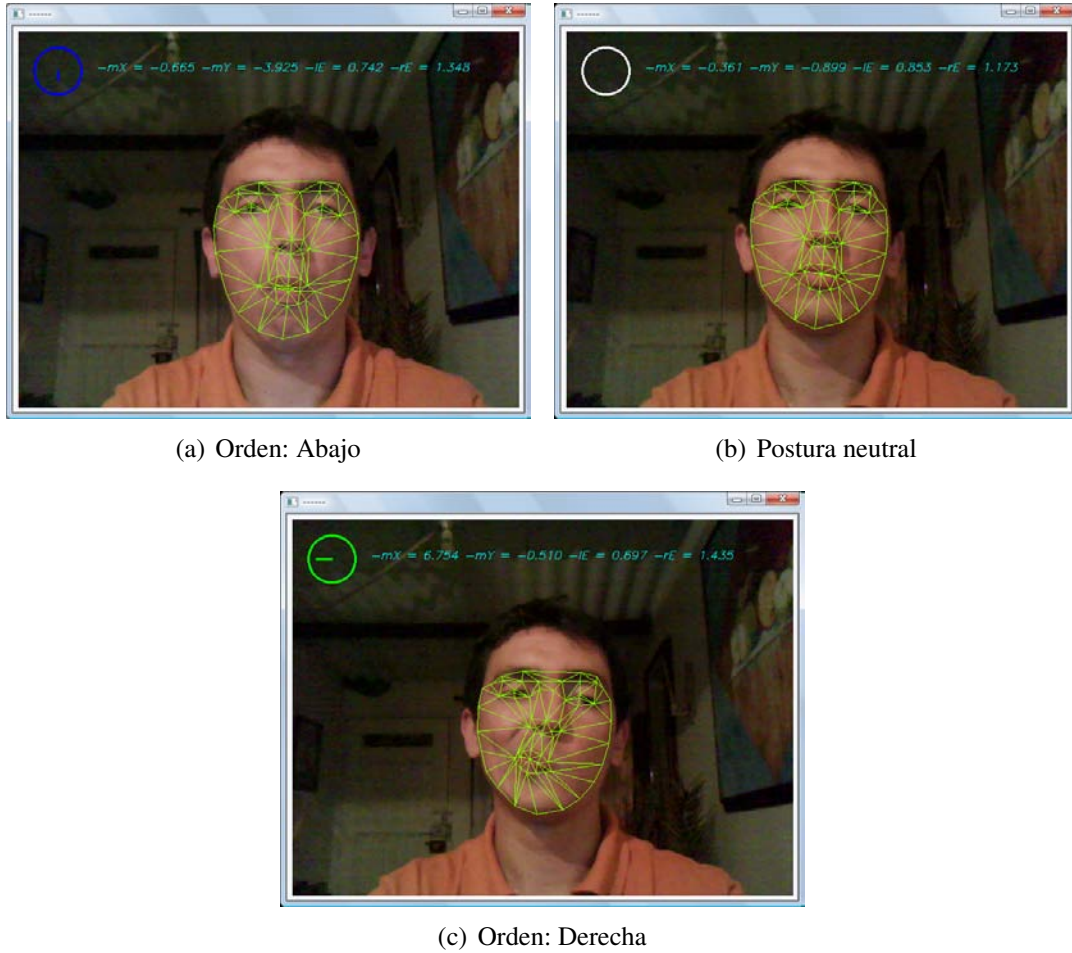


Figura 4.5: Ejemplos de ajuste y seguimiento del Modelo de Apariencia Activa bajo condiciones de iluminación diferentes.

Estas distancias y relaciones de aspecto se calcularon a partir de pruebas empíricas sacadas de las secuencias de vídeo de entrenamiento. La definición de estos umbrales, permite generar una zona muerta en donde el algoritmo no interpretará los gestos, evitando de esta forma ambigüedad en la órdenes generadas.

En la Figura 4.6 se muestran algunas imágenes en donde se presenta la correcta clasificación de los comandos, nótese que aunque en algunos casos el ajuste del modelo, especialmente en el área de la boca no es muy preciso, los gestos faciales siguen siendo correctamente clasificados. Lo cual es el objetivo final de esta aplicación. No se calcularon tiempos de desempeño de este algoritmo ya que su complejidad computacional no es significativa, puesto que se basa sólo en unos cuantos condicionales(sentencias *if*).

Orden	Umbral(píxeles)
Arriba	2,5
Abajo	2,5
Derecha	3
Izquierda	3
Zoom In	1,5
Zoom Out	1,5

Cuadro 4.2: Umbrales de distancia a partir de los cuales una orden comienza a ser válida. El umbral para los ojos es una relación de aspecto de uno contra el otro.

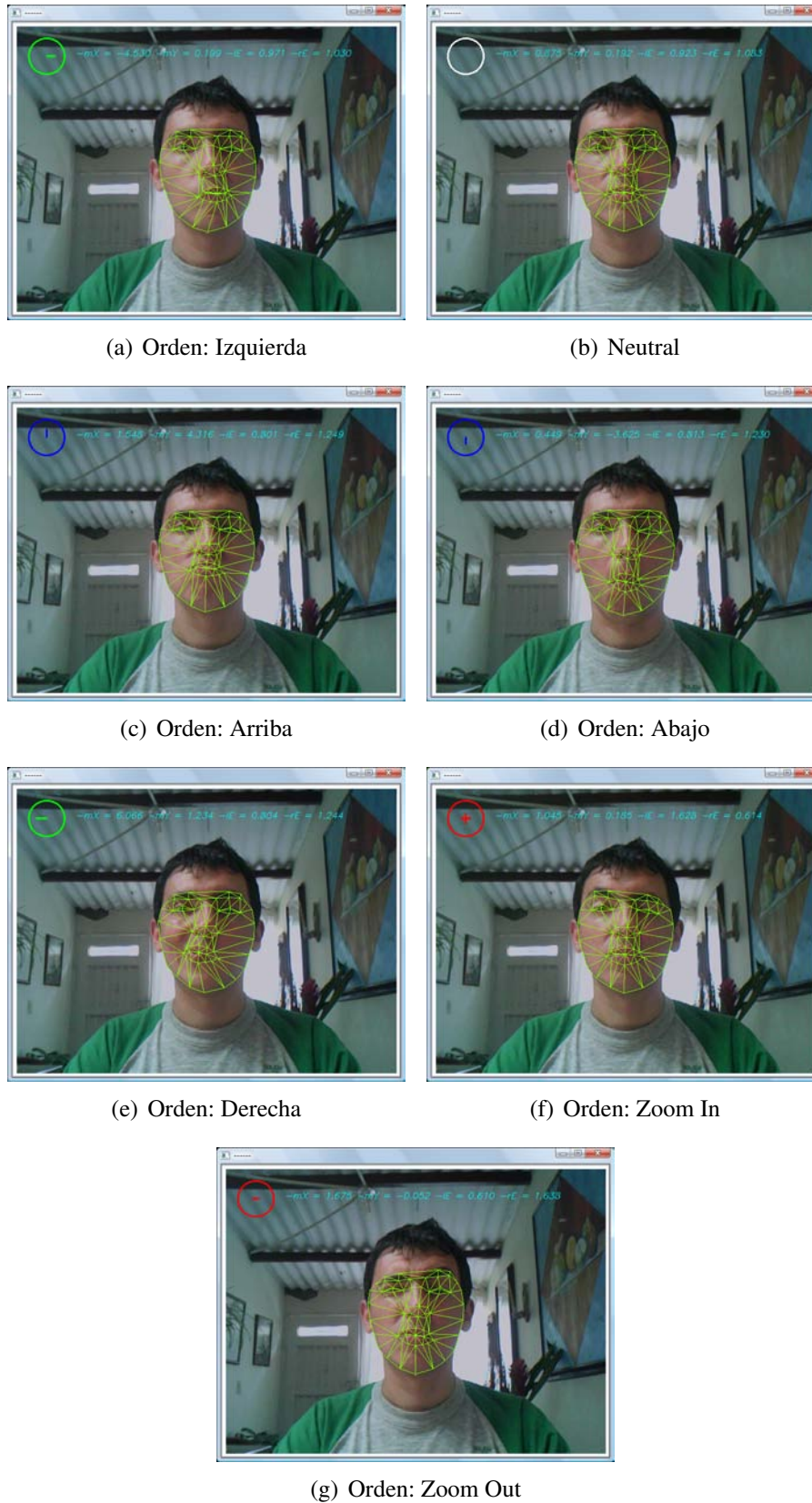


Figura 4.6: Ejemplos de clasificación de las órdenes para el control del brazo.

Capítulo 5

Conclusiones

Se desarrolló un sistema de tiempo real, basado en visión artificial que puede seguir los gestos faciales de una persona y no requiere ningún tipo de dispositivo de contacto, tales como sensores colocados en el rostro o la cabeza. El cirujano puede de manera fácil y precisa, controlar el brazo de robot, haciendo simplemente los gestos faciales adecuados, sin tener que usar interruptores o comandos de voz para iniciar la secuencia de control.

El algoritmo de seguimiento basado en modelos de apariencia activa, permite además de segmentar, seguir, y extraer características faciales de una manera robusta, un procesamiento de la imagen de hasta 286 fotogramas por segundo máximo y 23 como mínimo y junto con la técnica de clasificación, permite determinar además de la dirección, una magnitud asociada a esta, con lo cual se puede en un momento dado controlar la velocidad del movimiento, dándole así mayor control y precisión al movimiento.

El sistema desarrollado permite hacer de manera efectiva y en tiempo real, la detección del rostro, la segmentación, el seguimiento, y la clasificación de gestos faciales para el control del brazo robótico, pero se debe tener en cuenta que todo este proceso se desarrolla sobre un sistema de coordenadas de dos dimensiones (2D), lo cual obliga a estar en frente de la cámara y en posición frontal si se quiere controlar el brazo.

En un futuro cercano, se piensa trabajar extendiendo la aplicación a un sistema de tres dimensiones (3D) en donde se manejarán más de una cámara, permitiendo así que el control del brazo no esté sujeto a estar de frente a la cámara. También se quiere trabajar en un sistema automático o semiautomático de etiquetado de imágenes para el entrenamiento, debido a que actualmente esta actividad se lleva a cabo de manera totalmente manual.

Apéndice A

Adquisición

En la gran mayoría de los sistemas de procesamiento digital de señales, una de las etapas más cruciales es la adquisición de la señal; de ésta depende que tanto preproceso y adecuación debe ser aplicado a la señal que se está tratando. en el caso del procesamiento digital de imágenes, ésta es quizás la más importante dado que de la calidad de ella dependen gran parte de los resultados. Para el procesamiento de vídeo es igualmente crucial, más, cuando se requiera de una aplicación que funcione en tiempo real; de ésta dependen que tan pesados o livianos –desde el punto de vista computacional– deben ser las etapas de filtrado segmentación. Una forma de aligerar el preproceso y la segmentación de la imagen, es controlar los más posible las condiciones de adquisición, restringiendo la escena de tal forma que sea especialmente diseñada para la aplicación específica. Para el caso particular de nuestro sistema de adquisición, ésta se llevó a cabo dentro de un ambiente poco controlado en su etapa inicial, en donde, aunque se utilizó una fuente de luz artificial para iluminar la escena, no se tenía control sobre ésta. Posterior a ello y dados los diferentes problemas que daba la iluminación que se estaba utilizando(poca uniformidad en las áreas de la piel debido a sombras), se opto por usar una fuente de luz adicional, la cual nos permitió mejorar la calidad de la segmentación y la baja en el costo computacional de la misma.

A continuación se presentan los detalles sobre la etapa de adquisición, que involucra la disposición espacial de los dispositivos, más los espacios de color usados.

A.1. Condiciones de Adquisición

En la etapa inicial se utilizó un ambiente cerrado en donde la única fuente de luz era, las lamparas de luz blanca propias del recinto en donde se ejecutaba la prueba, las cuales estaban ubicadas en el techo de éste, y sobre las cuales no se tenía control alguno.

A.1.1. Arquitectura de la Escena

Para la adquisición de la secuencia de imágenes se llevó a cabo una configuración como la que se muestra en la Figura A.1, en donde se tuvieron en cuenta factores como: posición de la cámara respecto del objetivo, uniformidad de la luz, distancia del objetivo respecto del fondo, uniformidad del fondo, etc. Dado que la secuencia de imágenes que se quiere adquirir requiere de unas condiciones controladas, se tuvieron en cuenta consideraciones similares a las utilizadas para la toma de fotografías, las cuales son presentadas de manera amplia en el trabajo desarrollado por [44]

A.1.2. Características de los equipos de adquisición

Los recursos utilizados para la adquisición de las imágenes son los que la universidad tiene actualmente a disposición. Entre otros están: una cámara de vídeo JVC, una tarjeta de adquisición National Instruments y un equipo PC.

A.1.3. Modificaciones de la Arquitectura de la Escena

Los resultados parciales de los algoritmos de detección del rostro y la boca presentaban problemas debido a la ausencia de control sobre la fuente de iluminación lo cual generó poca uniformidad de luz en la región de la cara, la presencia de sombras en algunas de las regiones de piel dentro del rostro. Antes de pensar en cambiar los algoritmos de detección por unos más robustos se decidió modificar la arquitectura de adquisición de la escena, buscando mejorar estos aspectos que hacían que la adquisición no fuera lo suficientemente buena. Para ello se decidió introducir una fuente de luz adicional que homogeneizara la cantidad de luz en el rostro, y sobre la cual se tuviera control.

La adición de ésta nueva fuente de luz, mejoró notablemente la calidad de la imagen adquirida y permitió conservar los algoritmos de detección del rostro y de los labios que se habían planteado inicialmente. La configuración de la escena presentada en la Figura A.1 y en la Figura A.2.

A.1.4. Espacio de Color Usado en la Adquisición

El espacio de color utilizado para la adquisición de las secuencias de vídeo fue el HSV. Dado que es uno de los más utilizados para la detección de la piel, y a que la tarjeta de adquisición de vídeo usada nos entrega la información de color en este espacio directamente sin la necesidad de hacer una transformación adicional. El espacio de color HSV además de ser el espacio más parecido a la percepción humana del color, también agrupa los píxeles del color de la piel en un cluster bien definido, lo que nos permite una segmentación simple de estos, ie. mediante umbralización.



Figura A.1: esquemas de adquisición inicial.



Figura A.2: esquemas de adquisición mejorado.

Espacios de color para la segmentación del color de la piel

Para la detección de colores específicos es necesario tener un espacio de color que represente el color precisamente e independientemente de la intensidad de la iluminación. los espacios de color más importantes pueden ser divididos en dos grupos principales:

Espacios de color Perceptualmente uniformes ■ L*a*b* y Luv : Representaciones del sistema Munsell [37].

espacios de color perceptualmente no uniformes ■ Diagrama de crominancia CIE (XYZ): sistema de primarias hipotéticas que representan todos los colores posibles en un área hiperbólica.

- RGB y CMY espacios de color tridimensional usado para video, monitores o imágenes digitalizadas.
- HSV, HSL, HSI, y HSC: espacios que están orientados por las necesidades del usuario tales como que sus componentes son más intuitivos.
- YUV, YCbCr y YIQ: utilizados para aplicaciones de video o tareas de transmisión de television.

la característica de uniformidad perceptual puede ser muy útil para tareas de segmentación, porque tales modelos intentan ser compatibles con la percepción del color del ojo humano. Como las transiciones entre colores son lineales en el espacio, los colores forman regiones compactas, lo cual simplifica la segmentación. La definición de L*a*b* está basada en un sistema de color intermedio conocido como espacio CIE XYZ:

$$\begin{aligned} L^* &= 116f\left(\frac{Y}{Y_0}\right) - 16 \\ a^* &= 500 \left[f\left(\frac{X}{X_0}\right) - f\left(\frac{Y}{Y_0}\right) \right] \\ b^* &= 200 \left[f\left(\frac{Y}{Y_0}\right) - f\left(\frac{Z}{Z_0}\right) \right] \end{aligned} \quad (\text{B.1})$$

donde

$$f(q) = \begin{cases} q^{\frac{1}{3}}, & \text{si } q > 0,008856 \\ 7,787q + \frac{16}{116}, & \text{en otro caso} \end{cases} \quad (\text{B.2})$$

X_0, Y_0, Z_0 representan el blanco de referencia [44]

de acuerdo con el trabajo desarrollado por [44] el espacio de color L^*, a^*, b^* produce resultados muy precisos, pero dado el elevado costo computacional que exige la transformación del espacio de color se decidió no usarlo para el desarrollo de éste trabajo.

Entre los espacios de color no uniformes, el uso directo del espacio de color YCbCr puede producir los resultados más rápidos si la imagen original esta almacenada en este formato por lo cual no sería necesario hacer alguna transformación. Pero [37] no considera que éste sea apropiado por que no permite la determinación de umbrales válidos generales que resulten en segmentaciones con bajo ruido. Los autores en [13] encontraron la transformación a otros espacios de color demasiado costosa y proponen un rango de valores para Cb y Cr que en sus propios resultados experimentales son muy amplios, i.e. la segmentación contiene demasiadas regiones ruidosas debido a colores similares en el fondo. para lidiar con estas regiones, los autores proponen cuatro procesos adicionales para clasificar las regiones, entre otras costosos procesos estadísticos. En vez de esto, los espacios de color del grupo 2 c) parecen ser más útiles por que sus componentes intuitivos son más similares a la percepción humana y representan regiones de color compactas. Aquí, el costo de calculo para transformaciones son mucho menores que para $L^*a^*b^*$, ellos admiten una separación total de luminancia y crominancia y el área cubierta por los tonos de piel es mas compacta. concretamente el espacio de Color HSV, HSL, etc. ha sido probado en comparación con el $L^*a^*b^*$ y el RGB en tareas de segmentación del color de la piel y ha demostrado ser mejor.

A continuación se presentan algunas de las transformaciones de espacios de color más utilizadas en la detección de la piel.

Transformación de YCbCr a RGB

Las señales Y, Cb,Cr son construidas teóricamente de las primitivas RGB de acuerdo con la siguiente expresión:

$$\begin{bmatrix} E'_Y \\ E'_R - E'_Y \\ E'_B - E'_Y \end{bmatrix} = \begin{bmatrix} 0,299 & 0,587 & 0,114 \\ 0,701 & -0,587 & -0,114 \\ -0,299 & -0,587 & 0,886 \end{bmatrix} \cdot \begin{bmatrix} E'_R \\ E'_G \\ E'_B \end{bmatrix} \quad (\text{B.3})$$

donde E'_R, E'_G, E'_B son señales primarias. Después de la normalización, corrección de gama y cuantificación de señales de 8 bits (E'_R, E'_G, E'_B a $E'_{RD}, E'_{GD}, E'_{BD}$) se obtenida la ecuación B.4

$$\begin{bmatrix} Y \\ C_R \\ C_B \end{bmatrix} = \frac{1}{256} \begin{bmatrix} 77 & 150 & 29 \\ 131 & -110 & -21 \\ -44 & -87 & 131 \end{bmatrix} \cdot \begin{bmatrix} E'_{RD} \\ E'_{GD} \\ E'_{BD} \end{bmatrix} \quad (\text{B.4})$$

B.1. Transformación de RGB a HSV

la transformación de RGB a HSV puede ser expresado en una manera sencilla en forma de un pseudoalgoritmo encontrado en [37]

Algoritmo 2 Transformación de espacio RGB a HSV

```
max  $\leftarrow$  Max(R, G, B)
min  $\leftarrow$  Min(R, G, B)
v  $\leftarrow$  max
if max  $\neq$  0 then
    s  $\leftarrow$  (max - min)/max
else
    s  $\leftarrow$  0
end if
if s = 0 then
    h  $\leftarrow$  INDEFINIDO
else
    delta  $\leftarrow$  max - min
    if r = max then
        h = (G - B)/delta
    else if G = max then
        h = 2 + (B - R)/delta
    else if B = max then
        h = 4 + (R - G)/delta
    end if
    h  $\leftarrow$  h * 60
    if h < 0 then
        h  $\leftarrow$  h + 360
    end if
end if
```

Apéndice C

Herramientas de Software

C.1. OpenCV

OpenCV (Open source Computer Vision library) es una librería de código abierto desarrollada por Intel. Esta librería proporciona funciones de alto nivel para procesamiento de imágenes. Permite a los programadores crear aplicaciones poderosas en el dominio de la visión artificial. OpenCV ofrece muchos tipos de datos de alto-nivel como árboles, gráficos, matrices, etc.

C.1.1. Rasgos de OpenCV

OpenCV implementa una gran variedad de herramientas para la interpretación de imágenes. Es compatible con Intel Image Processing Library (IPL) que implementa algunas operaciones en imágenes digitales. Incluye primitivas como binarización, filtrado, estadísticas de imágenes, pirámides, OpenCV es principalmente una librería que implementa algoritmos para técnicas de calibración (Calibración de Cámaras), detección de rasgos, para rastrear (Flujo Óptico), análisis de forma (Geometría, Contorno), análisis de movimiento (Plantillas de Movimiento, Estimadores), reconstrucción 3D (Transformación de vistas), segmentación de objetos y reconocimiento (Histograma, etc.).

El rasgo esencial de la librería junto con la funcionalidad y la calidad es su desempeño. Los algoritmos están basados en estructuras de datos muy flexibles, acoplados con estructuras IPL; más de la mitad de las funciones ha sido optimizada aprovechándose de la Arquitectura de Intel.

OpenCV usa la estructura Iplimage para crear y manejar imágenes. Esta estructura tiene gran cantidad de campos, algunos de ellos son más importantes que otros. Por ejemplo el width es la anchura del Iplimage, height es la altura, depth es la profundidad en bits y nChannels el número de canales (uno por cada nivel de gris de las imágenes y tres para las imágenes a color).

OpenCV en cuanto a análisis de movimiento y seguimiento de objetos, ofrece una funcionalidad interesante. Incorpora funciones básicas para modelar el fondo para su posterior sustracción, generar imágenes de movimiento MHI (Motion History Images) para determinar dónde hubo movimiento y en qué dirección, algoritmos de flujo óptico, etc.

OpenCV viene con una interfaz gráfica llamada highGUI. Esta interfaz gráfica es muy importante porque se necesita bajo OpenCV para visualizar imágenes.

C.1.2. Inconvenientes de OpenCV

Dadas las grandes posibilidades que ofrece OpenCV para el tratamiento de imágenes, calibración de cámaras, y otras muchas aplicaciones más como por ejemplo, para simular una prótesis ocular basada en un implante cortical y estudiar el funcionamiento de las retinas artificiales, etc.

Quizá de los pocos inconvenientes que se pueden encontrar en ella sea en el caso del seguimiento de objetos, en el cual, el principal inconveniente que es que no ofrece un producto completo, tan sólo algunas piezas que sirven como base para montar sobre ellas un producto final.

Sin embargo, la presencia de funciones muy interesantes, y las posibilidades ya comentadas que ofrece la librería hacen que estos inconvenientes no sean realmente significantes.

C.2. Herramienta de Modelado de Software - UML

UML (Unified Modeling Language) es un lenguaje que permite modelar, construir y documentar los elementos que forman un sistema software orientado a objetos. Se ha convertido en el estándar de facto de la industria, debido a que ha sido impulsado por los autores de los tres métodos más usados de orientación a objetos: Grady Booch, Ivar Jacobson y Jim Rumbaugh. Estos autores fueron contratados por la empresa Rational Software Co. para crear una notación unificada en la que basar la construcción de sus herramientas CASE. En el proceso de creación de UML han participado, no obstante, otras empresas de gran peso en la industria como Microsoft, Hewlett-Packard, Oracle o IBM, así como grupos de analistas y desarrolladores.

Bibliografía

- [1] Paul VIOLA and Michael JONES, “Robust real-time face detection,” *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004. VII, 3, 10, 12, 24, 27, 28, 40
- [2] T.; Koara K. Nishikawa, A.; Hosoi, “*FAce MOUSE*: A novel human-machine interface for controlling the position of a laparoscope,” *IEEE Transactions on Robotics and Automation*, vol. 19, no. 5, pp. 825 – 841, Oct. 2003. IX, X, XI, 1
- [3] J. Fernández-Lozano I. García-Morales R. Molina-Mesa C. Pérezdel-Pulgar J. Serón-Barba M. Azouaghe Víctor F. Muñoz, J. Gómez de Gabriel, “Design and control of a robotic assistant for laparoscopic surgery,” Tech. Rep. 4, Instituto de Automática y Robótica Avanzada de Andalucía. Universidad de Málaga Severo Ochoa, Parque Tecnológico de Andalucía. Málaga, 2000. 1
- [4] and Narendra Ahuja Ming-Hsuan Yang, David J. Kriegman, “Detecting faces in images: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, JANUARY 2002. 1, 2, 3, 10
- [5] Pentland Alex Basu Sumit, Oliver Nuria, “3d modeling and tracking of human lip motion,” Tech. Rep. 442, MIT Media laboratory Perceptual Computing Section, 1998. 1, 4
- [6] Timothy F. Cootes Iain Matthews, “Extraction of visual features for lipreading,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, FEBRUARY 2002. 1, 4, 17
- [7] Jie Yang Rainer Stiefelhagen, Uwe Meier, “Real time lip tracking for lipreading,” Tech. Rep., University of Karlsruhe, Interactive Systems Laboratories, Germany, 1998. 1, 4, 14, 29, 30
- [8] Neil A. Thacker Juergen Luetttin, “Locating and tracking facial speech features,” *Proceedings of the international Conference on Pattern Recognition*, vol. 10, 1996. 1
- [9] N. Oliver, A. Pentland, and F. Berard, “Lafter: a real-time face and lips tracker with facial expression recognition,” 2000. 1
- [10] M. Kass, A. Witkin, and D. Terzopoulos, “Snakes: Active contour models,” *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, JANUARY 1988. 1

-
- [11] Hao Jiang and Mark S. Drew, “A predictive contour inertia snake model for general video tracking,” Tech. Rep., School of Computing Science, Simon Fraser University, Vancouver, B.C., Canada V5A 1S6, 1999. 1
- [12] Jain Anil K. Hsu Rein-Lien, Abdel-Mottaleb Mohamed, “Face detection in color images,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, MAY 2002. 1
- [13] Mayank Vatsa Richa Singh Sanjay Kr. Singh, D. S. Chauhan, “A robust skin color based face detection algorithm,” *Tamkang Journal of Science and Engineering*, vol. 6, no. 4, pp. 227–234, 2003. 1, 2, 3, 54
- [14] Kaucic Robert and Blake Andrew, “Accurate real time, unadorned lip tracking,” Tech. Rep., Dept of Engineering Science, University of Oxford, 1998. 1, 4
- [15] Beet Steve W. Luetttin Juergen, Thacker Neil A., “Active shape models for visual speech feature extraction,” Tech. Rep. 95/44, University of Sheffield, UK. Electronic Systems Group, 1996. 1
- [16] Alla Andreeva Vladimir Vezhnevets, Vassili Sazonov, “A survey on pixel-based skin color detection techniques,” Tech. Rep., Graphics and Media Laboratory, Faculty of Computational Mathematics and Cybernetics, Moscow State University, Moscow, Russia. 1, 2, 3
- [17] Rothkrantz Leon J.M. Pantic Maja, “Automatic analysis of facial expressions: the state of the art,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 696–710, DECEMBER 2000. 3
- [18] B. Soin P.A. Finlay A. Gordan S. Aiono, J.M. Gilbert, “Controlled trial of the introduction of a robotic camera assistant (endoassistTM) for laparoscopic cholecystectomy,” *Surgical Endoscopy*, vol. 16, no. 9, pp. 1267–1270, SEPTEMBER 2002. 3
- [19] Cohn Jeffrey F. Tian Ying-li, Kanade Takeo, “Robust lip tracking by combining shape, color and motion,” Tech. Rep., Robotics Institute, Carnegie Mellon University, Department of Psychology, University of Pittsburgh, 2000. 3, 4
- [20] Wang Shi Lin Leung Shu-Hung and Lau Wing-Hong, “Lip image segmentation using fuzzy clustering incorporating an elliptic shape function,” *IEEE Transactions on Image Processing*, vol. 13, no. 1, JANUARY 2004. 3, 4
- [21] Lau Wing Hong Liew Alan Wee-Chung, Leung Shu Hung, “Segmentation of color lip images by spatial fuzzy clustering,” *IEEE transactions on Fuzzy Systems*, vol. 11, no. 4, 2003. 3, 4
- [22] Rambaruth Ratna Porter Robert Mark, “Face detection,” UK PATENT APPLICATION GB 2 395 779 A, JUNE 2004. 3
- [23] Trent W Lewis and David M W Powers, “Lip feature extraction using red exclusion,” in *Selected papers from Pan-Sydney Workshop on Visual Information Processing*, Peter Eades and Jesse Jin, Eds., Sydney, Australia, 2001, ACS. 4, 15

-
- [24] P. Perez H. Li R. Forchheimer C. Kervrann, F. Davoine and C. Labit, “Generalized likelihood ratio-based face detection and extraction of lip features,” in *AVBPA+97m*, Crans Montana, Ed., march 1997. 4
- [25] G. Chetty and Wagner, “Automated lip feature extraction for liveness verification in audio-video authentication,” New Zealand, 2004, *Proc. Image and Vision Computing*, pp 17-22. 4, 14
- [26] Josef Kittler M. Ulises Ramos Sanchez, jiri Matas, “Statistical chromaticity-based lip tracking with b-splines,” Tech. Rep., Department of Electronic and Electrical engineering, University of Surrey, 1997. 4
- [27] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, “Active shape models-their training and application,” *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, January 1995. 4
- [28] T. F. Cootes, G. J. Edwards, and C. J. Taylor, *Active Appearance Models*, 1998. 4, 5, 6, 31
- [29] M. B. Stegmann, R. Fisker, B. K. Ersbøll, H. H. Thodberg, and L. Hyldstrup, “Active appearance models: Theory and cases,” in *in Proc. 9th Danish Conf. Pattern Recognition and Image Analysis*, 2000, vol. 2000, pp. 49–57. 4, 5
- [30] Jing Xiao, Simon Baker, Iain Matthews, and Takeo Kanade, “Real-time combined 2d+3d active appearance models,” in *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2004, pp. 535–542. 4, 5, 35
- [31] Iain Matthews and Simon Baker, “Active appearance models revisited,” *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 135–164, November 2004. 5, 6, 22, 31, 40, 41
- [32] B. D. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision,” in *IJCAI81*, 1981, pp. 674–679. 5, 41
- [33] G.J. Edwards, C.J. Taylor, T.F. Cootes, and Manchester M Pt, “Interpreting face images using active appearance models,” 1998. 5, 31
- [34] T.F. Cootes, K. Walker, and C.J. Taylor, “View-based active appearance models,” *Automatic Face and Gesture Recognition, IEEE International Conference on*, vol. 0, pp. 227, 2000. 5, 31
- [35] Fadi Dornaika and Jörgen Ahlberg, “Fast and reliable active appearance model search for 3d face tracking,” *IEEE Transactions on Systems, Man and Cybernetics, Part B*, vol. 34, pp. 1838–1853, 2004. 5, 35
- [36] James L. Crowley and Joelle Coutaz, “Vision for man machine interaction,” in *EHCI*, 1995, pp. 28–45. 6

-
- [37] Martina Eckert, *ADVANCED MOTION COMPENSATION FOR VIDEO CODING*, Ph.D. thesis, ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE TELECOMUNICACIÓN, 2002. 6, 7, 14, 24, 25, 26, 28, 53, 54
- [38] Erik Hjelmås and Boon Kee Low, “Face detection: A survey,” *Computer Vision and Image Understanding*, vol. 83, no. 3, pp. 236–274, Sept. 2001. 10
- [39] Yoav Freund and Robert E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” in *European Conference on Computational Learning Theory*, 1995, pp. 23–37. 10, 27
- [40] “Ieee standar glossary for image processing and pattern recognition terminology,” New York, USA, 1990, Published by the Institute of Electrical and Electronics Engineers. 11
- [41] K; Prasad G; Subbanna B Prem and Sumam D, “Human face detection and tracking using color modeling and connected component operators,” 2000, Karnataka Regional Engineering College. 12, 13
- [42] GOSHTASBY Ardeshir y GARCÍA Oscar LI, Yadong, “Detecting and tracking human faces in videos.,” in *ICPR*, 2000, pp. 1807–1810. 12
- [43] Henry A. Rowley, Shumeet Baluja, and Takeo Kanade, “Neural network-based face detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 23–38, 1998. 12
- [44] Ivonne Mejia Gomez, “Extracción automática de características faciales para el estudio antropométrico en niños entre 5 y 10 años de la ciudad de manizales,” 2004. 14, 49, 54
- [45] Tarcisio Coianiz, Lorenzo Torresani, and Bruno Caprile, “2d deformable models for visual speech analysis,” in *In NATO Advanced Study Institute: Speechreading by Man and Machine*. 2002, pp. 391–398, Springer Verlag. 14
- [46] Eva Cerezo, Isabelle Hupont, Cristina Manresa-Yee, Javier Varona, Sandra Baldassarri, Francisco J. Perales, and Francisco J. Seron, “Real-time facial expression recognition for natural interaction,” in *IbPRIA '07: Proceedings of the 3rd Iberian conference on Pattern Recognition and Image Analysis, Part II*, Berlin, Heidelberg, 2007, pp. 40–47, Springer-Verlag. 25
- [47] R. Lienhart, A. Kuranov, and V. Pisarevsky, “2003, empirical analysis of detection cascades of boosted classifiers for rapid object detection,” *DAGM 25th Pattern Recognition Symposium*, 2003. 28
- [48] Simon Baker and Iain Matthews, “Lucas-kanade 20 years on: A unifying framework,” *International Journal of Computer Vision*, vol. 56, pp. 221–255, 2004. 31, 41
- [49] Hesam Najafi, Yakup Genc, and Nassir Navab, “Fusion of 3d and appearance models for fast object detection and pose estimation,” in *In ACCV*, 2006, pp. 415–426. 35

-
- [50] Paul Viola and Michael Jones, “Robust real-time object detection,” in *International Journal of Computer Vision*, 2001. 39
- [51] R. Lienhart and J. Maydt, “An extended set of haar-like features for rapid object detection,” 2002, vol. 1, pp. I-900–I-903 vol.1. 39
- [52] Modesto Castrillón, “Gias - grupo de inteligencia artificial y sistemas,” <http://gias720.dis.ulpgc.es/Gias/modesto.html>. 40
- [53] Marco Castrillón, Oscar Déniz, Cayetano Guerra, and Mario Hernández, “Encara2: Real-time detection of multiple faces at different resolutions in video streams.,” *J. Visual Communication and Image Representation*, vol. 18, no. 2, pp. 130–140, 2007. 40
- [54] E. Begin R. Hurteau, S. DeSantis and M. Gagner, “Laparoscopic surgery assisted by a robotic cameraman: Concept and experimental results,” in *IEEE Int. Conf. Robotics and Automation*, San Diego, CA, MAY 1994, pp. 2286–2289.
- [55] J. G. DeGabriel J. F. Lozano-E. Sanchez-Badajoz-A. Garcia-Cerezo-R. Toscano V. F. Muñoz, C. Vara-Thorbeck and A. Jimenez-Garrido, “A medical robotic assistant for minimally invasive surgery,” in *IEEE Int. Conf. Robotics and Automation*, San Francisco, CA, APRIL 2000, pp. 2901–2906.
- [56] R. Toscano J. Gomez J. Fernandez-M. Felices C. Vara-Thorbeck, V. F. Muñoz and A. Garcia-Cerezo, “A new robotic endoscope manipulator a preliminary trial to evaluate the performance of a voice-operated industrial robot and a human assistant in several simulated and real endoscopic operations,” *Surgical Endoscopy*, vol. 15, pp. 924–927, SEPTEMBER 2001.
- [57] L. Jacobs A. Halverson D. Uecker Y. Wang J.M. Sackier, C. Wooters, “Voice activation of a surgical robotic assistant,” *The American Journal of Surgery*, vol. 174, no. 4, pp. 406–409, OCTOBER 1997.
- [58] P. G. Schulam J. A. Cadeddu-B. R. Lee-R.G. Moore M. E. Allaf, S. V. Jackman and L. R. Kavoussi, “Laparoscopic visual field. voice vs. foot pedal interfaces for control of the aesop robot,” *Surgical Endoscopy*, vol. 12, no. 12, pp. 1415–1418, DECEMBER 1998.
- [59] Christophe Doignon Michel F. de Mathelin Guillaume Morel-Joël Leroy Luc Soler lexandre Krupa, Jacques Gangloff and Jacques Marescaux, “Autonomous 3-d positioning of surgical instruments in robotized laparoscopic surgery using visual servoing,” *IEEE TRANSACTIONS ON ROBOTICS AND AUTOMATION*, vol. 19, no. 5, OCTOBER 2003.
- [60] D. Prats A. Casals, J. Amat and E. Laporte, “Vision guided robotic system for laparoscopic surgery,” in *IFAC Int. Congr. Advanced Robotics*, Barcelona - Spain, 1995, pp. 33–36.
- [61] K. Arbter G.-Q. Wei and G. Hirzinger, “Real-time visual servoing for laparoscopic surgery,” *IEEE Eng. Med. Biol. Mag.*, vol. 16, pp. 40–45, JANUARY 1997.
-

-
- [62] D. R. Uecker Y. F. Wang and Y. Wang, “A new framework for vision enabled and robotically assisted minimally invasive surgery,” *Comput. Med. Imaging and Graphics*, vol. 22, pp. 429–437, 1998.
- [63] B. Moghaddam and A. Pentland, “Probabilistic visual learning for object recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 696–710, JULY 1997.
- [64] sony Ericson Mobile communications, “Noise reduction and audio-visual speech activity detection,” PATENT APPLICATION EP 1 443 498 A1, August 2004.
- [65] María José LADO TOURIÑO and Arturo José MÉNDEZ PENÍN, “Identificación y catalogación de imágenes de interfaz,” *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.
- [66] Carlos Alejo RAMÍREZ and Manuel David PÉREZ, “Detección de caras y análisis de expresiones faciales,” .
- [67] Federico LECUMBERRY, “Cálculo de disparidad y segmentación de objetos en secuencias de video,” Tesis de maestría en ingeniería eléctrica, Universidad de la república Montevideo, Uruguay, 2005.
- [68] Rogério Schmidt Feris, Teófilo Emídio de Campos, and Roberto Marcondes Cesar Junior, “Detection and tracking of facial features in video sequences,” in *MICAI*, Osvaldo Cairó, Luis Enrique Sucar, and Francisco J. Cantu, Eds. 2000, vol. 1793 of *Lecture Notes in Computer Science*, pp. 127–135, Springer.
- [69] Nariman Habili, “Automatic segmentation of the face and hands in sign language video sequences,” 2001.
- [70] M. Bartlett, G. Littlewort, I. Fasel, and J. Movellan, “Real time face detection and facial expression recognition: Development and application to human-computer interaction,” 2003.
- [71] P. Silapachote, D. R. Karuppiah, and A. Hanson, “Feature selection using adaboost for face expression recognition,” in *Proceedings of the Fourth IASTED International Conference on Visualization, Imaging, and Image Processing*, Marbella, Spain, September 2004, pp. 84–89.
- [72] Rachid BELAROUSSI and Maurice MILGRAM, “Face detecting and skin color based tracking: a comparative study,” 2007.
- [73] Thilak R. Kumar, S. Kumar Raja, and A. G. Ramakrishnan, “Eye detection using color cues and projection functions,” in *ICIP (3)*, 2002, pp. 337–340.
- [74] J. Huang and Harry Wechsler, “Eye detection using optimal wavelet packets and radial basis functions (RBFs),” *IJPRAI*, vol. 13, no. 7, pp. 1009–1026, 1999.
- [75] Saad A. Sirohey and Azriel Rosenfeld, “Eye detection in a face image using linear and nonlinear filters,” *Pattern Recognition*, vol. 34, no. 7, pp. 1367–1391, 2001.

-
- [76] Guo-Can Feng and Pong Chi Yuen, “Multi-cues eye detection on gray intensity image,” *Pattern Recognition*, vol. 34, no. 5, pp. 1033–1046, 2001.
- [77] I. Pitas K. Sobottka, “A novel method for automatic face segmentation facial feature extraction and tracking,” *Signal Processing: Image Communication*, vol. 12, no. 3, pp. 263–281, 1998.
- [78] Hui Wu, Wei-Ngan Chin, and Joxan Jaffar, “An efficient distributed deadlock avoidance algorithm for the AND model,” *Software Engineering*, vol. 28, no. 1, pp. 18–29, 2002.
- [79] Jean-Christophe Terrillon, Hideo Fukamachi, Shigeru Akamatsu, and Mahdad N. Shirazi, “Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images,” in *FG '00: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000*, Washington, DC, USA, 2000, p. 54, IEEE Computer Society.
- [80] Gonzálo Pajares Martínsanz; Jesús M. de la Cruz Garcá, *Visión por computador: Imágenes digitales y aplicaciones*, Editorial Ra-ma, 2001.
- [81] BP Chen, J. Tiddeman, “Robust facial feature tracking under various illuminations,” in *13th IEEE International Conference on Image Processing (ICIP)*. 2006, IEEE.
- [82] S. Kimura and M. Yachida, “Facial expression recognition and its degree estimation,” in *CVPR97*, 1997, pp. 295–300.
- [83] J.M. Wertz M.A. Perrott M.A. Sayette, Jeffrey Cohn and D.J. Parrott, ,” *A psychometric evaluation of the Facial Action Coding System for assessing spontaneous expression*, vol. 25, pp. 167 – 186, 2001.
- [84] Irfan A. Essa, “Coding, analysis, interpretation, and recognition of facial expressions,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 757–763, 1998.
- [85] Sheryl M. Ehrlich, Diane J. Schiano, and Kyle Sheridan, “Communicating facial affect: it’s not the realism, it’s the motion,” in *CHI '00: CHI '00 extended abstracts on Human factors in computing systems*, New York, NY, USA, 2000, pp. 251–252, ACM Press.
- [86] M. Pantic and I. Patras, “Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences,” *Systems, Man, and Cybernetics, Part B, IEEE Transactions on*, vol. 36, no. 2, pp. 433–449, 2006.
- [87] Nicu Sebe, Michael S. Lew, Ira Cohen, Ashutosh Garg, and Thomas S. Huang, “Emotion recognition using a cauchy naive bayes classifier,” in *in Proc. ICPR*, 2002, pp. 17–20.
- [88] Ying-Li Tian Takeo, Ying li Tian, Takeo Kanade, and Jeffrey F. Cohn, “Eye-state action unit detection by gabor wavelets,” in *In Proceedings of International Conference on Multi-modal Interfaces (ICMI 2000*, 2000, pp. 143–150.