Reconocimiento de rostros basado en características invariantes

por

Claudia Cruz Pérez

Tesis sometida como requisito parcial para obtener el grado de

MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

en el

Instituto Nacional de Astrofísica, Óptica y Electrónica
Agosto 2008
Tonantzintla, Puebla.

Supervisada por:

Dr. Luis Enrique Sucar Succar
Dr. Eduardo Morales Manzanares

Resumen

La habilidad de reconocer personas es un punto fundamental para mejorar la interacción humano-robot en robots de servicio. Se han propuesto múltiples enfoques para el reconocimiento de rostros; sin embargo, estos asumen condiciones poco realistas para un robot de servicio, como tener una imagen con un rostro centrado bajo condiciones controladas de iluminación.

En esta tesis se propone un sistema de reconocimiento de rostros en ambientes con condiciones interiores realistas y un tiempo de respuesta adecuado para un robot móvil. El sistema es capaz de aprender en línea un nuevo rostro basado en una sola imagen, la cual es usada posteriormente para reconocer a la persona bajo diferentes condiciones en el ambiente. Una etapa de preprocesamiento es usada para reducir el efecto de las diferentes condiciones de iluminación, y entonces identificar tres regiones en el rostro: ojo izquierdo, ojo derecho y nariz-boca. Características SIFT son extraidas de cada región y son almacenadas en un vector de características, el cual es usado para su reconocimiento. La estrategia de correspondencia es capaz de descartar rostros desconocidos y un enfoque Bayesiano mejora la precisión sobre múltiples imágenes.

Tres conjuntos de experimentos son propuestos: reconocimiento en video, reconocimiento al incrementar el número de sujetos en la base de rostros y el número de cuadros de cada sujeto, y finalmente un experimento para evaluar el desempeño en el robot de servicio Markovito. Resultados en experimentos con diez personas muestran

que este método es capaz de aprender diferentes rostros y reconocerlos en un promedio de tres segundos con un 96.65 % de precisión y 57.32 % de recuerdo, lo cual significa un resultado competitivo de acuerdo al estado del arte.

Abstract

The ability to recognize people is a key element for improving human-robot interaction in service robots. There are many approaches for face recognition; however, these assume unrealistic conditions for a service robot, like having an image with a centered face under controlled illumination.

In this thesis a novel face recognition system is developed assuming realistic indoor environments and an adequate response time for a mobile robot. The system is able to learn on-line a new face based on a single frame, which is later used to recognize the person even under different environmental conditions. A preprocessing step is used to reduce the effect of different illumination conditions, and then identify three regions in the face: left eye, right eye and nose–mouth. SIFT features are extracted from each region and stored in a feature vector, which is used for recognition. The matching strategy is able to discard unknown faces and the recognition process uses a Bayesian approach over several frames to improve accuracy.

Three sets of experiments were performed: recognition in a video sequence, recognition increasing the number of individuals in a data base of faces, and, finally an experiment to evaluate the performance on the service robot Markovito. Results in experiments with ten people show that this method is able to learn and recognize different faces on average in three seconds with 96.65 % of precision and 57.32 % of recall, which is a significant result according to the state of the art.

Agradecimientos

A Dios por esta maravillosa vida que me ha concedio.

A mis asesores Dr. Luis Enrique Sucar y Dr. Eduardo Morales por la guía para realizar este trabajo.

A mis sinodales Dra. Angélica Muñoz, Dr. Francisco Martínez y Dr. Carlos Alberto Reyes por los útiles comentarios para mejorar esta tesis.

A mi familia por su cariño invaluable, son paz cuando no tengo calma.

A Fer por su compañía y paciencia en esas largas jornadas de trabajo, te amo cielo.

A mis amigos y mis compañeros del grupo de robótica, gracias por escucharme.

Para mis padres, Ely y Fer,

gracias por formar parte de mi vida.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

Una de las principales tareas a resolver dentro del área de visión por computadora en las últimas décadas ha sido el reconocimiento de rostros. La diversidad de sus aplicaciones comerciales, la constante creciente en los recursos computacionales y los retos en el campo de investigación han atraído a numerosos grupos de neurofisiólogos, psicólogos y científicos de la informática a una suma de esfuerzos en vía de resolver tan desafiante problema. En particular, en la robótica de servicio la habilidad de reconocer individuos es un proceso necesario para la interacción autónoma entre usuario-robot en favor de desarrollar sistemas robóticos capaces de realizar tareas como guías de museo (Burgard et al. (1998); Kim et al. (2004)), mensajería en oficinas (Avilés et al. (2007)), y cuidado de adultos mayores como el trabajo presentado por Pineau et al. (2002), entre otras.

1.1. Reconocimiento de rostros

El problema de reconocimiento de rostros a través del procesamiento de imágenes puede plantearse de la siguiente manera: dadas una imagen o video de una escena, el objetivo del reconocimiento de rostros es la identificación o verificación, de una o más personas en la escena, utilizando una base de datos de individuos conocidos. Algunos

de los retos a los cuales se enfrenta el reconocimiento de rostros son:

- **Pose.** La imagen de un rostro varía dependiendo de la posición del rostro respecto a la cámara. Un rostro puede aparecer en la imagen de frente, de perfil, rotado sobre el plano de la imagen o con alguna inclinación. Por consiguiente la apariencia de un mismo rostro puede cambiar significativamente al variar la pose.

- **Presencia o ausencia de componentes estructurales.** La apariencia de un sujeto puede verse dramáticamente afectada al ser modificada con presencia o ausencia de anteojos, barba, bigotes, gorras, maquillaje, entre otros.

- **Expresión facial.** Dada la naturaleza no rígida del rostro, su apariencia puede ser modificada con las gesticulaciones faciales.

- **Oclusiones.** Partes del rostro pueden ser ocluidas por otros objetos o personas ocultando elementos clave para su reconocimiento.

- **Condiciones de la imagen.** En el proceso de adquisición de las imágenes podemos encontrar problemas como las condiciones de iluminación y ruido en las imágenes, que es un efecto indeseable que consiste en la aparicion aleatoria de señales ajenas a la imagen original, esto puede ser generado por errores en el sensor de adquisición o al medio de transmisión de la señal.

Un sistema de reconocimiento de rostros se puede dividir conceptual y funcionalmente en tres bloques. En la primera etapa se realiza la detección del rostro en la imagen, que consiste en determinar la presencia de un rostro en la imagen, en caso de existir uno o más rostros se obtienen sus posiciones en la imagen. En la segunda etapa se realiza la extracción de características. Finalmente, en la tercera se aplica algún algoritmo de reconocimiento de rostros. En la figura 1.1 se muestra este esquema.

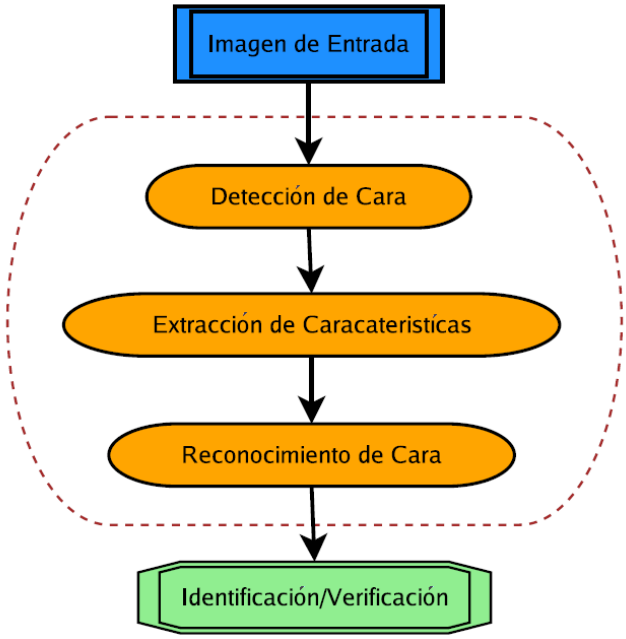


Figura 1.1: Diagrama de bloques representativo del problema de reconocimiento automático de rostros.

1.2. Antecedentes

La creciente necesidad de sistemas de interacción humano-robot más amigables ha impulsado el desarrollo de sistemas de reconocimiento adecuados a plataformas de robótica móvil. Si bien es cierto que los sistemas de reconocimiento actuales han alcanzado un nivel de madurez aceptable, la mayoría de ellos asume condiciones poco factibles en el marco de trabajo de robótica móvil; un ejemplo es el contar con una imagen con iluminación uniforme y el rostro centrado. Más aún la capacidad de aprender nuevos individuos en tiempo de ejecución en muchos de estos sistemas se vuelve inviable por la cantidad de ejemplos de entrenamiento requeridos y el tiempo de generación de los modelos. Es por ello que el problema de reconocimiento de rostros continua siendo vigente a más de cuatro décadas de sus primeros intentos de solución.

Entre los primeros trabajos que abordan una solución semi-automática se destaca el presentado por Goldstein et al. (1971) donde se sistematiza el marcado de 21 medidas

subjetivas, entre ellas el color del pelo y el ancho de los labios. Sin duda, una de las mayores limitantes de estos sistemas reside en la dificultad de automatizar el proceso de extracción y confrontación de características. Más adelante Kohonen (1989) presenta un sistema de reconocimiento de caras que demuestra el uso de redes neuronales para el reconocimento con base en imágenes de rostros normalizados. Incluyó el análisis de componentes principales (PCA) mediante el uso de los vectores propios de la matriz de autocorrelación de la imagen de cara normalizada. En 1989, Kirby y Sirovich introducen un método matemático que simplifica el cálculo de las *eigenfaces*. Muestran la eficiencia de la representación de las imágenes de caras alineadas y normalizadas mediante bases de dimensión menor a 100. En 1991, Turk y Pentland utilizan el hecho de que la representación mediante *eigenfaces* minimiza el error cuadrático medio para detectar y localizar caras en imágenes naturales y a distintas escalas. Con estas ideas los autores logran una aplicación confiable en tiempo de ejecución.

A mediados de la década del noventa, aparece una gran cantidad de métodos que intentan ampliar sus condiciones de funcionamiento de iluminación, pose, expresión, entre otras. Estos avances fueron impulsados por un crecimiento del área de reconocimiento de patrones y un marcado avance tecnológico. Algunos de estos métodos abordan generalizciones del trabajo de Turk y Pentland como el Análisis Discriminante Lineal (LDA) y el Análisis de Componentes Independiente (ICA). Otros optan por enfoques como las redes neuronales, enfoques evolutivos, modelos ocultos de Markov, entre muchos otros.

Recientemente el problema del reconocimiento de rostros ha sido abordado en video. Algunos de estos enfoques consideran aspectos en el tiempo de respuesta, así como variaciones en la iluminación y ángulo de rotación del rostro. En estos trabajos se dividie el problema en tres etapas principales: detección del rostro, seguimiento y reconocimiento. Sin duda, uno de los algoritmos más populares para abordar el problema de la detección de rostros es el propuesto por Viola y Jones (2001a) que mejora

el rendimiento de clasificadores muy simples al ser combinados en uno fuerte, además de utilizar una representación de la imagen que reduce el cálculo de las características. Debido a dicha capacidad de calcular un gran número de características en un corto tiempo de ejecución, la representación de la imagen integral, resulta ser utilizada para resolver el problema de detección, seguimiento y reconocimiento propuesto por Grabner et al. (2007) .

Recientemente se abordó el problema de reconocimiento de rostros por Bicego et al. (2006) y por Luo et al. (2007) haciendo uso de descriptores característicos SIFT propuestos por Lowe (2004) para la representación de los rostros. El método SIFT extrae características invariantes de una imagen, estas características son robustas a variaciones en su rotación, escala, iluminación, cambios de perspectiva 3D y ciertas oclusiones. Sin embargo; estos trabajos únicamente contemplan el uso de imágenes del rostro en donde la localización de los ojos es la misma para todas las imágenes, con condiciones de iluminación uniformes y rotaciones máximas de $\pm 30^{\circ}$ del rostro.

1.3. Retos computacionales

Dentro de los principales retos computacionales a los cuales se enfrenta el reconocimiento de rostros en video podemos destacar:

- La detección del rostro en una secuencia de imágenes correspondiente a un video, así como un enfoque adecuado para poder aprovechar la información de las imágenes.
- El método de reconocimiento de rostros debe ser robusto a cambios en la iluminación, escala, expresiones faciales, rotaciones y presencia de accesorios (como gafas, gorras, bigotes, barba, entre otros).
- En el caso de implementación en plataformas de robótica móvil, el tiempo de respuesta resulta un factor crítico, así como el aprendizaje de nuevos rostros en

tiempo de ejecución.

Teniendo presentes estas consideraciones, en esta tesis se abordan estos retos. En esta tesis se aborda el problema de la representación de los rostros mediante el uso de puntos característicos SIFT, ya que han mostrado un buen desempeño en tiempo de ejecución en problemas de robótica móvil. Además del uso de un enfoque Bayesiano para la incorporción de evidencia de múltiples imágenes para reforzar la precisión en el reconocimiento. Más aún, el enfoque propuesto en esta tesis requiere únicamente una imagen del rostro del individuo que se desea registrar, lo cual resulta de suma utilidad para su aplicación en robótica.

1.4. Objetivos

1.4.1. Objetivo general

El objetivo general de la tesis es desarrollar un sistema de reconocimiento de rostros para robots móviles de servicio, con base en la extracción de características invariantes mediante la metodología SIFT.

1.4.2. Objetivos particulares

- Selección e implementación de un algoritmo para la detección de rostros y seguimiento en una secuencia de imágenes.
- Definición de una estrategia para correspondencia de los puntos característicos SIFT aplicado al problema de reconocimiento de rostros.
- Desarrollo de un sistema de reconocimiento de rostros en video robusto a ligeros cambios en la expresión facial, iluminación, oclusiones parciales y rotaciones de perfil de $\pm 15^{\circ}$.

1.5. Alcances y limitaciones

Se plantea desarrollar un sistema de reconocimiento de rostros que sea capaz de:

- Reconocer rostros frontales y con rotaciones en la pose de hasta $\pm 15^{\circ}$
- Reconocer rostros con presencia y ausencia de anteojos no obscuros.
- Reconocer rostros con variaciones en la iluminación de ambientes interiores.
- Aprender nuevos individuos en tiempo de ejecución.
- Realizar el reconocimiento de rostros en un tiempo adecuado para una plataforma de robótica móvil.

Dentro de las restricciones que se contemplaron para el desarrollo de la tesis, la principal es que el proceso de detección, seguimiento y reconocimiento se realizaran para una persona por imagen. A diferencia de los trabajos de Bicego y Luo et al; en este trabajo se utiliza una secuencia de imágenes correspondientes a un video, por tal motivo no se cuenta con una imagen del rostro bien centrada, además de tener condiciones de iluminación uniforme.

1.6. Descripción del sistema

El sistema de reconocimiento de rostros propuesto consta de tres módulo principales. El primero de ellos realiza la detección de rostros utilizando el algoritmo de detección rápida de objetos propuesto por Viola y Jones (2001a). La segunda etapa consiste de un seguidor de rostros, que busca reducir el espacio de búsqueda con base a la localización de rostros en imágenes previas. La tercera etapa consta del método de reconocimiento de rostros la cual utiliza características invariantes SIFT para la representación del rostro. Para determinar la identidad de un individuo se plantea el uso de un enfoque Bayesiano para integrar la evidencia generada en imágenes previas y la

evidencia de la imagen actual.

Finalmente, se diseñaron tres conjuntos de pruebas para evaluar el desempeño del sistema. En el primer conjunto se evaluá la precisión y recuerdo del reconocedor en video, se realizaron pruebas con diez personas en un ambiente interior sin controlar la iluminación. En estas pruebas se analiza el comportamiento del sistema cuando se intenta reconocer a un sujeto desconocido, alcanzando niveles de precisión entre el 92.77 y el 99.4 %. En las pruebas realizadas para evaluar el reconocimiento de personas conocidas, el sistema logra precisiones entre el 96.65 y el 100 % en el conjunto de pruebas. El segundo conjunto de pruebas analiza el comportamiento del sistema al incrementar el número de individuos en la base de rostros, para estas pruebas se utilizó la base de rostros de Yale Extendida que cuenta con 28 personas obteniendo una precisión entre 100 y 85.4 %. En el tercer conjunto de pruebas se evaluá el comportamiento del sistema en el robot de servicio Markovito alcanzando una precisión del 87.76 % y 40.17 % de recuerdo.

1.7. Organización del documento

En el Capítulo 2 se presenta una revisión de los métodos más relevantes para el reconocimiento de rostros. En el Capítulo 3 se realiza una revisión de algunos de los métodos de detección de rostros en el área. En el Capítulo 4 se detalla el método de detección a utilizar, así como la descripción del método de seguimiento propuesto. En el Capítulo 5 se describe el proceso de reconocimiento a detalle, esto incluye el preprocesamiento que reciben las imágenes, el proceso de extracción de características invariantes SIFT, la estrategía de correspondencia entre puntos característicos de la imagen a evaluar con las registradas en la base de datos y, finalmente, la incorporación de la regla de Bayes para robustecer el reconocimiento de rostros con base en la evidencia de la imagen actual y la generada en imágenes previas. En el Capítulo 6 se detallan los experimentos para prueba del método en imágenes estáticas y en video. Por último, en

el Capítulo 7 se presentan las conclusiones y trabajo futuro.

Capítulo 2

Detección de rostros

Uno de los principales procesos que existe en el reconocimiento de rostros en secuencias de video es la detección de rostros para su posterior reconocimiento y seguimiento. En este capítulo se describen y analizan algunos métodos de detección de rostros. Sin embargo, cabe señalar que la mayor aportación de esta tesis radica en el método de reconocimiento. Es importante destacar que gran parte del correcto funcionamiento del reconocimiento recae en el detector de rostros, ya que funge como entrada para la etapa de seguimiento y posteriormente para la del reconocimiento.

2.1. Introducción

El problema de detección de rostros puede ser planteado como: dada una imágen, se desea saber si existen rostros en dicha imagen, además de la localización de las coordenadas de dichos rostros. Entre los principales problemas a los cuales se enfrenta la detección de rostros se pueden mencionar:

- Pose y orientación de la cara.
- Tamaño de la cara.

- Presencia de componentes estructurales (lentes, barba, gorro, etc).
- Expresión de la cara.
- Oclusiones parciales.
- Problemas de iluminación.
- Ruido en el proceso de adquisición de las imágenes y pérdida de fidelidad por los métodos de compresión de video.
- Cantidad desconocida de caras en la imagen.

Por otro lado, el seguimiento es el proceso de localización de un objeto en movimiento en tiempo de ejecución usando un dispositivo de captura de video. El objetivo del algoritmo de seguimiento, es analizar los cuadros de video e indicar la localización posible del objeto en futuros cuadros.

El estado del arte en detección de rostros es sumamente amplio, muchos enfoques han sido propuestos en la búsqueda de mejorar los resultados obtenidos. En el trabajo de Yang et al. (2002), se presentan una revisión general de algunos de los métodos más relevantes, así como una clasificación de ellos:

1. **Métodos basados en conocimiento.** Estos se basan en el modelado del conocimiento humano de las características que conforman el rostro. Este conocimiento se usa para formar reglas que permitan distinguir entre rostros y no-rostros.
2. **Métodos basados en características invariantes.** Buscan las características del rostro que son persistentes a diferentes condiciones de iluminación y pose como lo es el color de la piel y la textura.
3. **Métodos de correspondencia de plantillas.** Utilizan varios patrones del rostro que describen completa o parcialmente sus características. Para detectar un rostro

se calcula la correlación entre una imagen de entrada y los patrones con los que se cuenta.

4. **Métodos basados en la apariencia.** Se crean patrones o modelos a partir de un conjunto de imágenes de entrenamiento tomando los valores de sus pixeles.

2.2. Detección de rostros

2.2.1. Redes Neuronales

Uno de los trabajos más significativos utilizando redes neuronales es el propuesto por Rowley et al. (1998). Este enfoque se divide en dos etapas, en la primera de ellas se efectuá una mejora a la imagen para después usar varias redes neuronales especializadas en diferentes regiones del rostro. En la segunda etapa se combinan las detecciones de la primera etapa usando heurísticas y otras redes neuronales para reducir el números de falsos positivos. En este trabajo, los autores reportan hasta un 90.5 % de detecciones correctas en el conjunto de prueba CMU[1] con 130 imágenes con 507 rostros.

2.2.2. Clasificador Bayesiano

Schneiderman y Kanade (1998), utilizan un clasificador Bayesiano simple y análisis de componentes principales (PCA) para la detección del rostro. En este sistema se calcula la probabilidad de la apariencia y posición de los patrones del rostro, así como una normalización de la intensidad de cada región, alcanzando una precisión del 92.5 % en el conjunto CMU+MIT. Posteriormente, Schneiderman y Kanade (2000) mejoran el desempeño de su sistema incluyendo características *wavelet* y la posición de dichas características. Utilizan clasificadores especializados en las distintas poses u orientaciones de los objetos a detectar, reportando un 90.2 % de detecciones frontales correctas y un 92.8 % en detecciones de rostros de perfil.

[1]Los conjuntos de prueba MIT+CMU se encuentran disponible en http://vasc.ri.cmu.edu/idb/html/face/frontalimages/

2.2.3. *EigenRostros*

El más representivo de los trabajos bajo este enfoque es presentado por Turk y Pentland (1991), en el cual se usa análisis de componentes principales (PCA) sobre un conjunto de imágenes de rostros. Cuando se proyecta la imagen de un rostro sobre el subespacio creado con PCA, la imagen no cambia mucho; pero cuando se proyectan imágenes de no-rostros, estás varian considerablemente. Entonces, para realizar las detecciones de rostros se calcula la distancia entre la región a evaluar y el subespacio de rostros. La distancia será menor para regiones con rostros, que aquellas en las que no se encuentre rostro. Para los experimentos se utilizaron imágenes de 16 individuos con diferentes configuraciones en la pose, iluminación, escala y resolución (2592 imágenes en total), alcanzando precisiones del 96 % en sus variaciones de iluminación, 85 % en cambios de la orientación y un 64 % en la variación del tamaño.

2.2.4. Máquinas de Soporte de Vectores

Cortes y Vapnik (1995) presentan la idea del modelo para una máquina de soporte vectorial (SVM) aplicado a la detección de rostros es la elección de un margen de separación que mejor distinga el conjunto de rostros del conjunto de no-rostros. Para la evaluación del sistema se utilizaron dos conjuntos de datos, el primero formado por 313 imágenes de alta resolución con sólo un rostro por imagen, mientras el segundo conjunto se formó por imágenes de diferentes resoluciones con un total de 155 rostros. La precisión alcanzada en el primer conjunto fue de un 97.1 %, sin embargo en el segundo conjunto este porcentaje se vió reducido a un 74.2 %.

2.2.5. Algoritmos de Boosting

Estos algoritmos buscan mejorar el desempeño de clasificadores muy simples al ser combinados en uno fuerte. En el trabajo de Viola y Jones (2001a), se utiliza el algoritmo AdaBoost propuesto por Freund y Schapire (1995) para seleccionar características

Haar para detectar rostros de frente. Además del uso de de una representación para las imágenes conocida como imagen integral, la cual contiene en cada elemento la suma de todos los píxeles arriba y a la izquierda del pixel correspondiente de la imagen original, con lo cual se reduce el tiempo de cálculo de las características. Los autores también introdujeron el uso de clasificadores en cascada, donde el primer clasificador es muy simple pero rápido y el último es muy complejo pero muy exacto (contando con un total de 32 clasificadores en cascada). El primer nivel de la cascada consiste de 2 características Haar y los 31 restantes usan un total de 4295 características. El conjunto de entrenamiento consistió de 4916 imágenes de rostros y 10000 de no-rostros. En esté trabajo Viola y Jones reportan un 93.9 % de detecciones correctas con el conjunto CMU+MIT con un desempeño de 15 fps (fotogramas por segundo).

2.3. Análisis

En el trabajo de Yang et al. (2002), se presenta un conjunto de pruebas para algunos métodos de detección de rostros. Se realizaron pruebas con dos conjuntos de prueba tomados de las bases de datos CMU+MIT, el primer conjunto de pruebas, *test1*, con 125 imágenes con 483 rostros y el segundo *test2* con 23 imágenes con 136 rostros. En la tabla 2.1 se muestran los resultados obtenidos para dichos métodos.

Los métodos mostrados reflejan porcentajes de detección de rostros superiores al 90 %; sin embargo, el guiarse únicamente por los porcentajes de precisión no reflejan la eficiencia global del método, ya que puede contar con un gran número de rostros sin reconocerlos, como el caso de las redes neuronales de Rowley et al. (1998) con un 92.5 % pero con 862 de falsas detecciones. En el caso de usar el detector de Viola y Jones (2001a) con las características Haar extendidas de Lienhart y Maydt (2002), se obtiene una precisión de 95.55 % con un 0.05 % de falsas detecciones con una velocidad de ejecución de 15 fps (fotogramas por segundo). Algunos factores a considerar para la selección del algoritmo de detección de rostros son:

Métodos	Test1 % Precisión	Test1 FD	Test2 % Precisión	Test2 FD
Basados en distribución, Sung y Poggio (1998)	N/A	N/A	81.9	13
Redes Neuronales, Rowley et al. (1998)	92.5	862	90.3	42
Clasificador Naive Bayes, Schneiderman y Kanade (1998)	93	88	91.2	12
SVM, Osuna et al. (1997)	N/A	N/A	74.2	20
Discriminante lineal Fisher, Yang et al. (2000a)	93.6	74	91.5	1
SNoW con características multi-escala, Yang et al. (2000b)	94.8	78	94.1	3
Aprendizaje Inductivo, Duta y Jain (1998)	90.0	N/A	N/A	N/A

Tabla 2.1: Resultados de precisión y falsas detecciones (FD) para algunos métodos de detección de rostros presentados en Yang et al. (2002).

- Muchos de los resultados reportados se basan en diferentes conjuntos de entrenamiento y distintos parámetros de ajuste. El número y variedad de los ejemplos de entrenamiento afectan directamente el desempeño.

- El tiempo de entrenamiento y ejecución, los cuales son de vital importancia en aplicaciones de tiempo real.

- El tamaño de rostros a detectar en las imágenes varía en los algoritmos. Por lo tanto, en los resultados lo que para algunos autores se considera un rechazo correcto, para otros implica una penalización en el recuerdo del sistema.

En particular, la aplicación de robot mensajero para la cual se requiere la etapa de detección de rostros nos obliga a adoptar un método con buen desempeño en tiempo de ejecución. Debido a que en la etapa de reconocimiento se verifica la presencia de los ojos en la imagen, se puede aceptar un pequeño margen de error en los falsos positivos, ya que estos serán descartados en la etapa de reconocimiento. Por lo tanto, en esta tesis se empleará el detector de objetos rápidos propuesto por Viola y Jones (2001a) , utilizando el conjunto de características Haar extendidas de Lienhart y Maydt (2002).

Capítulo 3

Reconocimiento de rostros

3.1. Introducción

El reconocimiento de rostros ha sido un problema recurrentemente abordado en las últimas decadas por números grupos de investigación alrededor del mundo. El problema de reconocimiento de rostros a través del procesamiento de imágenes puede plantearse de la siguiente manera: dadas una imagen o video de una escena, el objetivo del reconocimiento de rostros es la identificación o verificación, de una o más personas en la escena, utilizando una base de datos de individuos conocidos. Por identificación se entiende el proceso de asociar a un individuo una única etiqueta correspondiente a su rostro. Varios factores deben ser considerados al introducirse en este amplio campo de investigación, uno de los principales es el considerar si se trabajará con imágenes estáticas, o bien con video. Es por ello que en este capítulo se ofrece una revisión a estas dos formas de abordar el problema de reconocimiento de rostros.

De acuerdo a Zhao et al. (2003), los métodos de reconocimiento de rostros en imágenes estáticas se dividen en las siguientes categorías:

- **Métodos holísticos.** Se utiliza toda la imagen del rostro como unidad básica de procesamiento, que sirve finalmente como entrada al sistema de reconocimiento.

- **Métodos basados en características locales.** Se extraen características locales, como pueden ser ojos, nariz, boca, etc. Sus posiciones y estadísticas locales constituyen la entrada al sistema de reconocimiento.

- **Métodos híbridos.** Realizan el reconocimiento combinando las características particulares de un rostro y la imagen del rostro global, como se sugiere por Thompson (1980); Bartlett y Searcy (Cognitive Psychology); Yin (1969) y Wechsler et al. (1998), lo realizan los seres humanos.

Por otro lado, estudios como los de O'Toole et al. (2002); Wechsler et al. (1998) y Knight y Johnston (1997) muestran que el movimiento ayuda al reconocimiento de rostros familiares; además de demostrar que los humanos podemos reconocer rostros animados mejor que un conjunto de imágenes aleatorias de la misma persona. Si bien el reconocimiento de rostros aplicado a secuencias de video surge originalmente como una extensión al reconocimiento en imágenes estáticas, en el estudio publicado por Zhao et al. (2003) se identifican tres etapas principales a considerar al trabajar con video:

- Detección de rostro y estimación de la pose.

- Seguimiento de rostro.

- Extracción de características y modelado del rostro.

Se describen en particular los trabajos que explotan la metodología SIFT aplicada al problema de reconocimiento de rostros ya que este trabajo de tesis está basado en estas características. Antes de presentar dichos trabajos se describen las principales bases de datos usadas en el área.

3.2. Principales bases de datos

Para comparar métodos de reconocimiento en imágenes estáticas se han construido bases de datos de imágenes bajo diferentes condiciones y características. Antes de men-

cionar los principales métodos de reconocimiento en imágenes estáticas, se desbriben algunos de los conjuntos de prueba más comunes.

3.2.1. Base de Datos Yale

La base de rostros Yale, construida por Belhumeur et al. (1997) está formada por 165 imágenes en escala de gris de 15 individuos. Cada sesión de 11 imágenes por individuo contempla variaciones en la expresión facial, fuente de iluminación y uso de anteojos en los individuos. En la figura 3.1 se muestra una sesión completa para un sujeto de la base de datos.

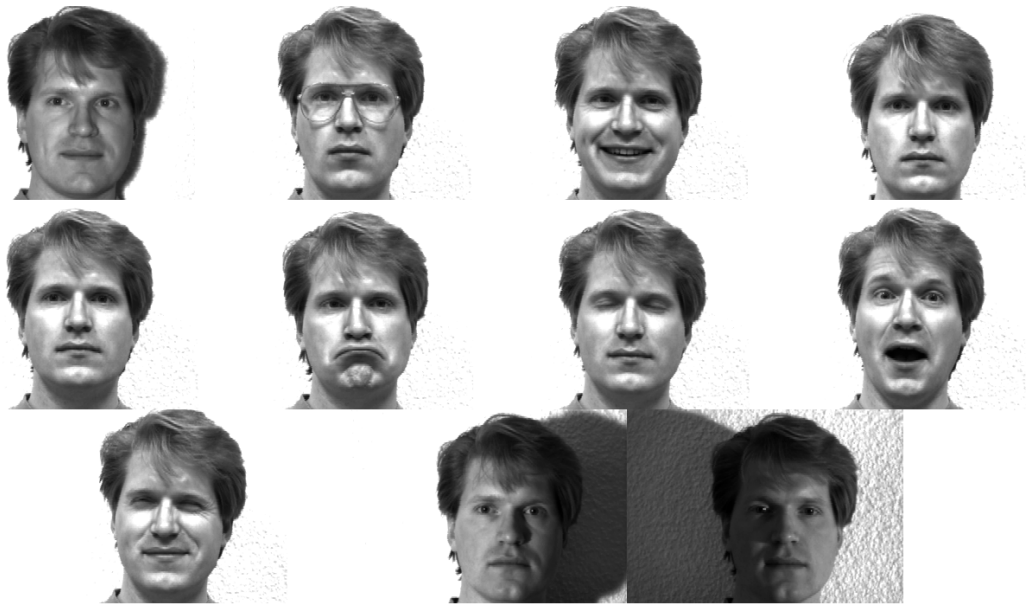


Figura 3.1: Ejemplo de sesión completa para un individuo de la base de datos de Yale de Belhumeur et al. (1997).

3.2.2. Base de Datos Yale Extendida B

La base de datos de rostros de Yale Extendida B construida por Georghiades et al. (2001); Lee et al. (2005) contiene 16128 imágenes de 640×480 en escala de gris de

28 individuos bajo 9 poses y 64 condiciones de iluminación diferentes. Se incluye también un conjunto de imágenes recortadas con únicamente el rostro de los individuos. En la figura 3.2 se muestran algunas imágenes recortadas de la base de rostros de Yale Extendida B.

Figura 3.2: Imágenes de la base de rostros de Yale Extendida B de Georghiades et al. (2001); Lee et al. (2005).

3.2.3. Base de Datos CAS-PEAL

La base de datos de rostros CAS-PEAL construida por Cao y Shan (2004) y Gao et al. (2008) contiene 99,594 imágenes de 1040 individuos (595 hombres y 445 mujeres) con variaciones en la pose, expresión, accesorios e iluminación. Nueve cámaras fueron colocadas en un semicírculo para capturar simultáneamente en pose a cada individuo. También se capturaron 18 imágenes variando la pose del rostro hacia arriba y hacia abajo. Se consideraron 5 clases de expresiones faciales, 6 clases de accesorios (3

lentes y 3 gorras) y 15 direcciones de iluminación. En la figura 3.5 se muestran ejemplos de imágenes de esta base de rostros.

Figura 3.3: Ejemplo de imágenes de CAS-PEAL de Cao y Shan (2004); Gao et al. (2008).

3.2.4. Base de Datos FERET

La base de datos FERET contiene una galeria de imágenes que incluyen 1196 imágenes de 1196 personas y cuatro clases de conjuntos de prueba: fafb (1195 imágenes con variación en expresión), fafc (194 imágenes con variación en iluminación), dupI (722 imágenes tomadas en menos de 18 meses) y dupII (234 imágenes tomadas en los 18 meses siguientes). En la figura 3.4 se muestran algunos ejemplos de imágenes

de dicha base de rostros.

Figura 3.4: Ejemplo de imágenes de FERET de Phillips et al. (1997).

3.2.5. Base de Datos ORL

Formada por el grupo de trabajo de los laboratorios AT&T, esta base de rostros está formada por 10 imágenes diferentes de 40 individuos. Las imágenes fueron tomadas en diferentes periodos de tiempo, variando la iluminación, la expresión facial y accesorios. El tamaño de cada imágen es de 92×112 pixeles en escala de grises. Un ejemplo de imágenes tomadas para un individuo de esta base de rostros se muestra en la figura 3.5.

Figura 3.5: Ejemplo de imágenes de ORL de Samaria y Harter (1994).

3.3. Reconocimiento en imágenes estáticas

3.3.1. Métodos holísticos

Múltiples enfoques han sido utilizados considerando una imagen como la unidad de entrada al sistema; entre estos se destacan los basados en el análisis de componentes. En estos trabajos se considera la distribución de imágenes que contienen caras y se captura la variabilidad de estas imágenes.

En el trabajo de Turk y Pentland (1991) se presenta un método para reconocimiento de rostros con base en el análisis de componentes principales, PCA. En dicho método se busca extraer de un conjunto de imágenes de entrenamiento, un subespacio cuya base maximice la varianza del espacio original. A estos vectores se les denomina *Eigenfaces* dado que son los vectores propios correspondientes a los valores propios más grandes de la matriz de covarianza de las imágenes. Dado que se trabaja con las proyecciones de los rostros originales en el subespacio generado, se logra reducir considerablemente la dimensión del problema. Posteriormente se establece una métrica para medir la similitud entre dos vectores dados. En los experimentos realizados por Turk y Pentland, se contemplan variaciones en la iluminación, escala y orientación con una base de datos de 2,500 imágenes de 16 individuos, reportando resultados de hasta un 96 % de precisión con variaciones de iluminación, 85 % para las orientaciones y sólo un 64 % de precisión para el escalamiento de las imágenes.

De igual forma que el PCA, el Análisis Discriminante Lineal (LDA) intenta llevar el espacio de caras a un subespacio de baja dimensionalidad que aumente la separabilidad de las clases presentes. Presentado por Belhumeur et al. (1997), la idea del algoritmo es encontrar la base de vectores en un subespacio que mejor discrimine entre las diferentes clases. Particularmente se utilizan todas las muestras de todas las clases y se calcula la matriz de dispersión entre clases distintas y la matriz de dispersión en la misma clase.

Se busca maximizar la relación entre el determinante de la matriz inter-clase y el determinante de la matriz intra-clase. Los elementos de la base que maximiza la relación se denominan *Fisherfaces*. En los experimentos realizados por Belhumeur et al. (1997) se evaluó el desempeño del algoritmo en comparación con el presentado por los *Eigenfaces*, formando el conjunto de prueba de Yale. Resultados en este conjunto de prueba muestran que la variación del desempeño del método de *Eigenfaces* así como en el propuesto, depende del número de componentes considerados, mostrando tasas de error promedio para el método de *Eigenfaces* de un 25 %, con tan sólo un 7.3 % del método que utiliza *Fisherfaces* con variaciones en el número de componentes principales del 50, 100 y 150.

Por otro lado, Liu y Wechsler (2000) proponen abordar el problema de reconocimiento de rostros con técnicas evolutivas. Este método, al igual que PCA y LDA se basa en el análisis de componentes. *Evolutionary Pursuit*, plantea una manera novedosa de obtener una base de vectores eficiente para la representación de las imágenes de caras. Para encontrar la base, se realiza una búsqueda a manera de maximizar una función de aptitud que mide al mismo tiempo la precisión de la clasificación y la habilidad de generalización del sistema. Como el problema de buscar la base óptima es de alta dimensionalidad, se utiliza un algoritmo genético. En dicho trabajo se muestra una comparativa entre el Análisis de Componentes Principales (PCA), el Análisis Discriminante Lineal (LDA) y la Búsqueda Evolutiva (EP), en el cual se muestra una significativa reducción de la dimensionalidad de las bases utilizadas, reportando una precisión del 92.14 % en un subconjunto de 1107 imágenes de la base de rostros FERET.

3.3.2. Métodos basados en características locales

En dichos métodos se extraen características independientes como lo son los ojos, nariz y boca para su posterior reconocimiento. De igual forma que en los holísticos, la

familia de los métodos basados en características locales cuentan con un amplio rango de métodos.

En Edwards et al. (1998) se utilizan *Active Appearance Model* (AAM) que es un modelo estadístico de la forma y la apariencia en niveles de gris del objeto de interés que se puede generalizar a casi cualquier ejemplo válido de dicho objeto. Un ejemplo válido en el problema de reconocimiento de rostros es una imagen de un rostro conocido. Ajustar dicho modelo a una imagen implica encontrar los parámetros del modelo que minimize la diferencia entre la imagen y una síntesis del modelo, proyectado en la imagen. El modelo se genera combinando modelos de variación de la forma y la apariencia de la cara. Éstos modelos se construyen a partir de imágenes de prueba donde se marcan puntos de interés (puede ser en forma manual o automática). Los modelos de forma y apariencia se generan considerando distintos puntos marcados en la imagen como vectores y aplicando PCA a dicha información. Cada ejemplo de forma y apariencia puede entonces ser representado por los coeficientes correspondientes. Dado que existe correlación entre la variación de la forma y la apariencia, se aplica PCA a la concatenación de los vectores de los modelos anteriores y con ello se construye el modelo final. Los resultados obtenidos por este enfoque reflejan un 88 % de precisión en la tarea de reconocimiento de rostros con una base de datos propia formada por 400 imágenes, en la figura 3.6 se muestran algunos ejemplos de imágenes que forman la base de rostros usada por Gareth *et al.*

Figura 3.6: Imágenes de la base de rostros usada por Edwards et al. (1998)

Otro enfoque utilizado es el *Elastic Bunch Graph Matching* (EBGM) el cual con-

siste en extraer una representación de la cara en forma de grafo y el reconocimiento se realiza comparando los grafos correspondientes a las distintas imágenes. Se define un conjunto de puntos principales, como por ejemplo las pupilas, esquinas de la boca, etc. Un grafo etiquetado representando una cara consiste en lo siguiente: *N* nodos ubicados en los puntos principales y las aristas que se forman entre parejas de nodos. Cada nodo es etiquetado con los denominados *jets* y cada arista es etiquetada con la distancia entre los nodos correspondientes. Los *jets* se basan en una transformada *wavelet* definida como la convolución de la imagen con una familia de núcleos de Gabor con distintas frecuencias y orientaciones. En Wiskott et al. (1997), este concepto es usado para el reconocimietno de rostros efectuando los experimentos en la base de datos FERET Grother et al. (2003), alcanzando porcentajes de precisión de un 98 % en la mejor de sus configuraciones; sin embargo, al involucrarse situaciones de prueba con conjuntos de rostros de perfil y medio perfil, su porcentaje se ve drásticamente reducido a un 12 %.

3.3.3. Análisis

Después de más de 40 años de investigación en el reconocimiento de rostros han surgido un gran número de métodos y sistemas. Cabe destacar el hecho de que cada uno de estos métodos cuenta con sus ventajas y desventajas, y por lo tanto la selección del método a utilizar debe considerar la aplicación en la cual se implanta. Por ejemplo, en aplicaciones que tengan como entrada al reconocimiento de rostros en imágenes muy pequeñas, los métodos basados en características locales resultan una mala selección. Otra consideración en el proceso de selección del algoritmo es la cantidad de ejemplos de entrenamiento que este necesita. Finalmente podemos mencionar que la tendencia es desarrollar métodos híbridos que mezclen las ventajas de los holísticos y los basados en características locales, ya que esta es la forma sugerida por Wechsler et al. (1998); O'Toole et al. (2002); Knight y Johnston (1997) en la cual los seres humanos realizamos el proceso de reconocimiento.

En general los métodos de reconocimiento en imágenes estáticas han mostrado buenos resultados, en bases de rostros con diferentes condiciones en pose, iluminación, expresión, algunos incluyen accesorios como gorras o anteojos; sin embargo, en su gran mayoría únicamente manejan imágenes del rostro, con el mismo tamaño y escala. Muchos métodos requieren una gran cantidad de imágenes de entrenamiento. Estos factores limitan el uso de estos métodos para aplicaciones de robótica donde el tiempo de respuesta y las condiciones no controladas son aspectos importantes a considerar.

3.4. Reconocimiento en secuencias de video

3.4.1. Retos del reconocimiento en secuencia de video

Entre los principales retos a los cuales se enfrenta el reconocimiento de rostro aplicado a secuencias de video podemos considerar:

1. La calidad de una imagen obtenida por el dispositivo de captura no siempre es la más adecuada. Dicha calidad se ve afectada por factores como la resolución que se define como el número de píxeles que componen la imagen, y la profundidad de color la cual hace referencia a la cantidad de bits usados para representar el color de un pixel. Esto implica imágenes con poca calidad (por ejemplo, 160 $\times$ 120 pixeles), cuyos objetos no se encuentran bien definidos o pierden detalles importantes para el proceso de reconocimiento.

2. El tamaño de la imagen del rostro que el detector encuentra puede ser muy pequeño, lo cual difiere de la mayoría de los métodos de reconocimiento de rostros en imágenes estáticas, en los cuales el tamaño de la imagen se encuentra predefinido.

3. Las características extraídas en el proceso de detección de rostros, en muchos casos no pueden ser utilizadas para la etapa del reconocimiento por la diferencia entre los procesos de detección y reconocimiento.

3.4.2. Sistemas para reconocimiento en video

Una propuesta reciente para reconocimiento de personas que considera restricciones en el tiempo de respuesta y el ambiente en el cual se desenvuelve, es el presentado por Apostoloff y Zisserman (2007). Para dicho trabajo se considera el problema de reconocimiento de rostros en una secuencia de video en tres etapas: detección, seguimiento e identificación. En la primera etapa se utilizan una cascada de detectores de rostros Viola y Jones (2001a). Si un nuevo rostro es detectado, entonces es comparado con aquellos que están siendo seguidos en ese momento, y si este no se traslapa por más de un cuarto del área del rostro, se inicializa un nuevo proceso de seguimiento para dicho rostro. Para el seguimiento, un conjunto de imágenes de la nueva persona son recolectados a lo largo de las siguientes imágenes. Cuando un número determinado de muestras ha sido recolectado (12 aproximadamente), entonces un conjunto de entrenamiento es generado con diferentes traslaciones y escalas sobre una rejilla uniforme. Posteriormente, un regresor basado en *kernel* es entrenado para predecir la localización en *(x,y)* del individuo. Finalmente, en la identificación se usa un modelo de estructura pictórica para extraer 13 características faciales del rostro. La región del rostro es entonces normalizada con respecto a un rostro canónico de 80×80 pixeles. Entonces, una transformación afín es computada entre las características faciales del rostro a comparar y las características del rostro canónico. Para cada característica facial, un parche de 15 pixeles de diámetro es extraído, así la representación del rostro es la concatenación de dichos parches. Para clasificar los rostros los autores utilizan un clasificador *random-ferns*[1] con 40 *ferns* de 17 niveles, donde la evaluación de un nodo es una simple comparación entre dos elementos del descriptor. Cuando suficientes muestras del seguidor son colectadas (típicamente 10), la identificación se realiza marginalizando[2] sobre el resultado del clasificador en cada imagen del seguidor.

[1] Un *random-fern* Ozuysal et al. (2007) es una modificación al clasificador *random-forest* Kam (1995), el cual se construye con un conjunto de clasificadores de árbol. La clasificación de un nuevo objeto se coloca como entrada en todos los árboles del bosque. Cada árbol proporciona una clasificación, y entonces votan por cierta clase. El bosque escoje la clasificación con base en la mayoría de los votos.

[2] Restar importancia a descriptores faciales cuyas probabilidades *posteriori* resulten poco relevantes para la clasificación del rostro.

Para la evaluación se emplearon secuencias de video de programas de televisión dividiendo las pruebas en dos grupos, los intra-episodios (imágenes de un episodio se utilizan para entrenamiento, el resto del episodio para prueba) y los inter-episodios (entrenamiento con imágenes de episodios diferentes a los de prueba). Los mejores resultados reportados por el sistema arrojan un $97 \pm 2\,\%$ de precisión con un 20 % de recuerdo en los experimentos intra-episodios; sin embargo, este desempeño se ve reducido al efecutarse las pruebas inter-episodio, obteniendo una precisión de aproximadamente el 90 % conservando un recuerdo del 20 %. En los experimentos, más de cinco personas pueden ser seguidas a 15 fps (fotogramas por segundo) en una máquina multi-núcleo a 1.86 GHZ. Se realizaron las pruebas con una base de datos de 10 rostros. En la figura 3.7 se muestran algunos resultados del sistema.

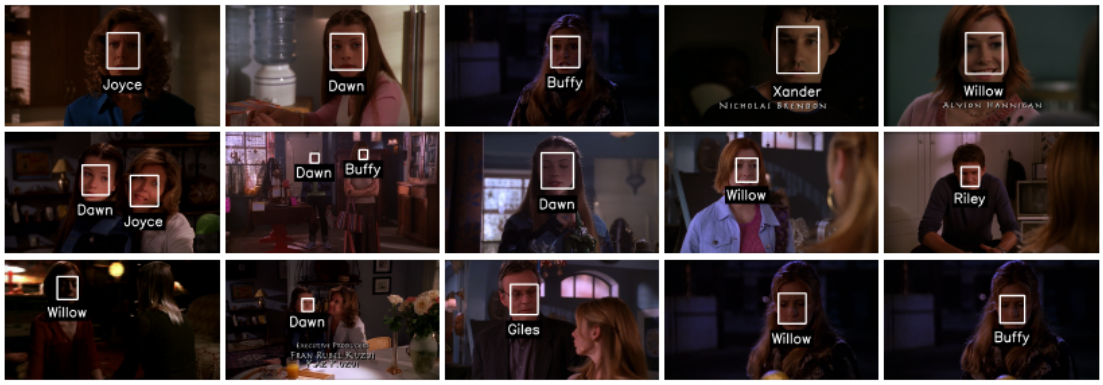


Figura 3.7: Ejemplos de clasificación del sistema de Apostoloff y Zisserman. Las primeras dos filas muestran ejemplos de clasificación correctos, mientras la última presenta casos de reconocimiento incorrectos de Apostoloff y Zisserman (2007).

En Grabner et al. (2007) se busca abordar el problema de reconocimiento de rostros aplicado a una plataforma robótica. Los autores consideran el problema de detección, seguimiento y reconocimiento como un problema de clasificación binaria. Para la representación del rostro en las tres etapas (detección, seguimiento y reconocimiento) Grabner *et al.* utilizan características Haar (ver sección 4.1), calculadas con base en la

imagen integral propuesta en Viola y Jones (2001a), con lo cual se busca la reducción en el costo computacional. En la fase de detección, el problema se plantea como la clasificación entre rostros y fondo. En el seguimiento, el problema se reduce a la clasificación entre la imagen de un rostro correctamente centrado o una imagen con solo una pequeña porción del rostro. Para mejorar la precisión del seguimiento se actualiza el clasificador de seguimiento con la región de la imagen actual, con lo cual se consideran pequeñas variaciones en iluminación y pose entre imágenes.

Posteriormente, la identificación se plantea como un problema de clasificación multiclase entre los rostros conocidos y un conjunto de rostros de entrenamiento no conocidos. Los experimentos fueron realizados con el robot Flea, el cual cuenta con una cámara binocular, procesando las imágenes a 640×480 pixeles efectuando el procesamiento en un equipo multi-núcleo a 2 GHz. Los resultados reportados únicamente indican que todo el proceso se ejecuta a 12 fps; sin embargo, no se reporta la precisión alcanzada. La situación del robot con respecto de los individuos tampoco es clara, pero se infiere que el robot permanece fijo y tres individuos caminan alrededor del robot. En la figura 3.8 se muestran algunos resultados de dicho sistema.

3.4.3. Análisis

El desempeño alcanzado por los sistemas presentados en esta sección, muestran la viabilidad del uso de un esquema de tres módulos: **detección, seguimiento y reconocimiento.** Por lo tanto, para este trabajo de tesis se abordará el mismo esquema. A diferencia de los métodos planteados para este trabajo se considera la opción de aprender individuos en tiempo de ejecución con tan sólo un ejemplo de entrenamiento por sujeto, lo cual contrasta con el trabajo presentado por Apostoloff y Zisserman (2007), quien utiliza la mitad de las imágenes de un capítulo de televisión para entrenar su sistema. Otro factor a considerar es el ambiente de ejecución, ya que en Apostoloff y Zisserman (2007) se utilizan imágenes de capítulos de televisión los cuales cuentan

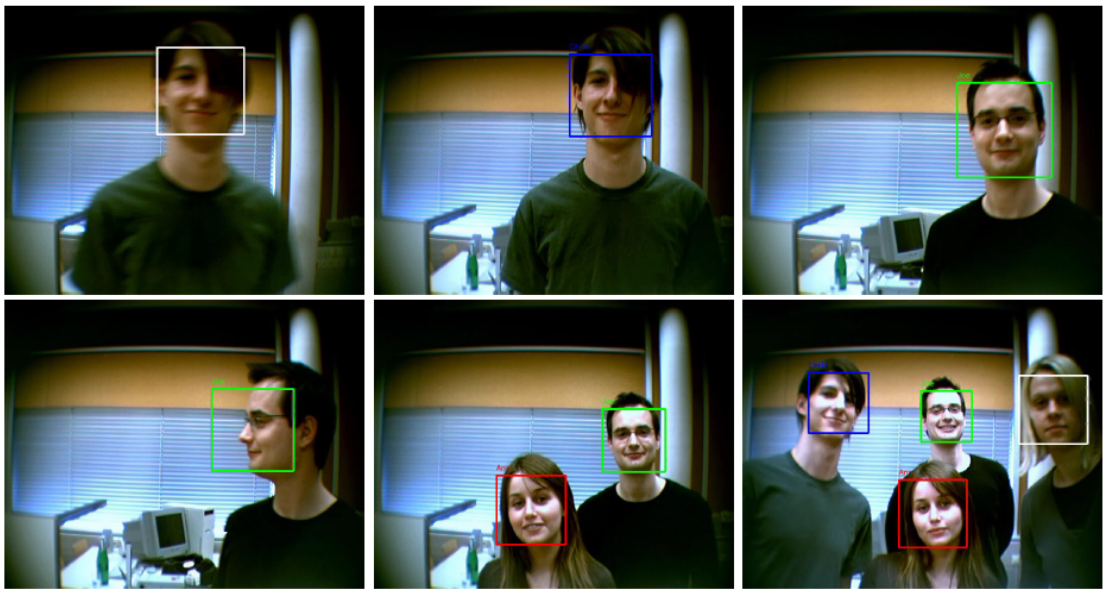


Figura 3.8: Ejemplos de detección de rostros de Flea (Tomado de Grabner et al. (2007)).

muchas veces con fuentes controladas de iluminación; mientras que en esta tesis se consideran ambientes interiores sin control de iluminación adicional. En contraste con el trabajo de Grabner et al. (2007), se realizarán pruebas en tiempo de ejecución con una base de rostros conocidos más extensa que la utilizada en dicho trabajo, la cual contó únicamente con tres individuos, además de que las situaciones de evaluación del trabajo mencionado no resultan del todo claras.

3.5. Reconocimiento basado en SIFT

3.5.1. Introducción

Introducido por primera vez por Lowe (1999) en 1999, SIFT (*Scale Invariant Feature Transform*) es un algoritmo para la extracción de características distintivas e invariantes en imágenes. Estas características son invariantes a rotación y escalamiento, así como robustas a un rango de proyecciones 3D, cierta presencia de ruido en la imagen, y algunos cambios en la iluminación.

Cada punto característico SIFT está compuesto por cuatro partes; su localización (las coordenadas (x, y) en la imagen en donde fue encontrado el punto SIFT), escala, orientación dominante y finalmente un descriptor característico (vector de 128 enteros). Entre las ventajas que el autor remarca podemos distinguir las siguientes:

- La localidad de las características, ya que son independientes entre sí, se brinda robustez a oclusiones parciales y entornos conglomerados sin previa segmentación.
- La particularidad de las características ya que cuentan con un descriptor conformado por 128 elementos, con lo cual se pueden comparar con otros puntos SIFT de otros objetos en grandes bases de datos.
- La cantidad de características generadas incluso en imágenes pequeñas.
- Su eficiencia en tiempo de cómputo

En las siguientes secciones se presentan dos trabajos que utilizan la métodología SIFT para resolver el problema de reconocimiento de rostros. Posteriormente en el Capítulo 5 se describe a detalle el métodos SIFT.

3.5.2. Correspondencia con rejillas

Un primer intento de resolver el problema de reconocimiento de rostros empleando SIFT, fue propuesto por Bicego et al. (2006). En ese trabajo, la extracción de puntos invariantes SIFT se realiza de la misma forma que la descrita por Lowe (2004), pero con la diferencia que Bicego *et al.* se enfocan en una estrategia de correspondencia de los puntos adecuada al problema de reconocimiento de rostros. Se estudia el desempeño de tres estrategias de emparejamiento diferentes:

- **Correspondencia con base en distancia mínima entre puntos.**
  En esta estrategia se calculan todas las distancias euclidianas entre los descripto-

res característicos de dos rostros, tomando preferencia por el que cuente con la menor distancia.

- **Correspondencia de ojos y boca.**

  Considerando que la mayor cantidad de información sobre la identidad de una persona está ubicada en la zona de los ojos y la boca, se busca mapear información en áreas coincidentes en dos rostros. Para esta estrategia las coordenadas de los ojos coinciden en los mismos puntos para todas las imágenes y con base en estas coordenadas se subdividen dos regiones, una para los ojos y otra exclusivamente para la boca. Entonces, cuando se desea comparar dos rostros, se calcula la distancia mínima entre todos los puntos de la región de los ojos del sujeto desconocido, contra todos los puntos de la región de los ojos de la imagen en la base de rostros a comparar. De forma similar se calcula la distancia mínima entre los puntos para la zona de la boca en ambas imágenes. Posteriormente estas distancias son promediadas obteniendo la evaluación de similitud entre dichos rostros.

- **Correspondencia con rejillas.**

  En esta estrategia se subdivide la imagen del rostro en ocho subimágenes de 1/4 del ancho de la imagen $\times$ 1/2 de su alto. Al querer comparar dos imágenes de rostros, entonces se considera la distancia mínima entre pares de subimágenes correspondientes; estás distancias mínimas son promediadas finalmente para dar la evaluación de similitud entre los rostros.

Para la evaluación se utilizaron imágenes de la base de datos BANCA construida por Bailly-Baillire et al. (2003) con 52 sujetos diferentes (26 hombres y 26 mujeres). Se consideran 12 sesiones de imágenes por cada persona, cuatro de ellas en situaciones adversas, cuatro en situaciones controladas y cuatro en situaciones degradadas. Se extrajeron cinco imágenes de cada sesión las cuales fueron utilizadas como conjunto de entrenamiento y pruebas. Todas las imágenes fueron escaladas a 210×200 pixeles,

además de ajustar la posición de los ojos para hacerlas coincidir a ciertas coordenadas, finalmente se aplicó ecualización de histograma para mejorar el contraste de la imagen. Se dividieron las imágenes en 2 grupos G1 y G2, y se calculó la tasa de error para ambos grupos. En los resultados reportados se puede apreciar un 94 % de precisión en las pruebas con el grupo G1, mientras G2 reporta un porcentaje de precisión de aproximadamente 97 % para el método de correspondencia con rejillas, el cual mostro el mejor comportamiento.

3.5.3. Correspondencia con agrupamiento

Partiendo de los resultados obtenidos por Bicego et al. (2006), Luo et al. (2007) proponen un método de correspondencia de características basadose en el agrupamiento automático de características por el algoritmo K-medias. Se inicializan k centroides con coordenadas (x, y) de manera aleatoria dentro de la imagen, posteriormente considera la posición en (x, y) de cada característica SIFT y se busca cuál es el centroide más cercano a dicha característica, se recalculan los valores para los centroides y se reconstruyen las subregiones. En el caso en que los centroides de las k-subregiones no se modifiquen, se concluye el algoritmo de agrupamiento; en caso contrario, se siguen ajustando los centroides y las subregiones. En su trabajo Luo *et al.* consideran la construcción de cinco subregiones (una por cada ojo, una para la nariz y dos para las comisuras de la boca).

Una vez que se cuenta con las subregiones generadas por el algoritmo K-medias, se obtiene un atributo de similaridad S entre dos rostros con base en una combinación de similaridades locales y globales, el esquema general del procedimiento se muestra en la figura 3.9.

En este esquema, una imagen I esta formada por $(f_1^1, ..., f_1^{m_1}, f_2^1, ..., f_2^{m_2}, ..., f_5^{m_5})$ vectores descriptores de las características SIFT (f_i^j refleja el *j-ésimo* descriptor SIFT

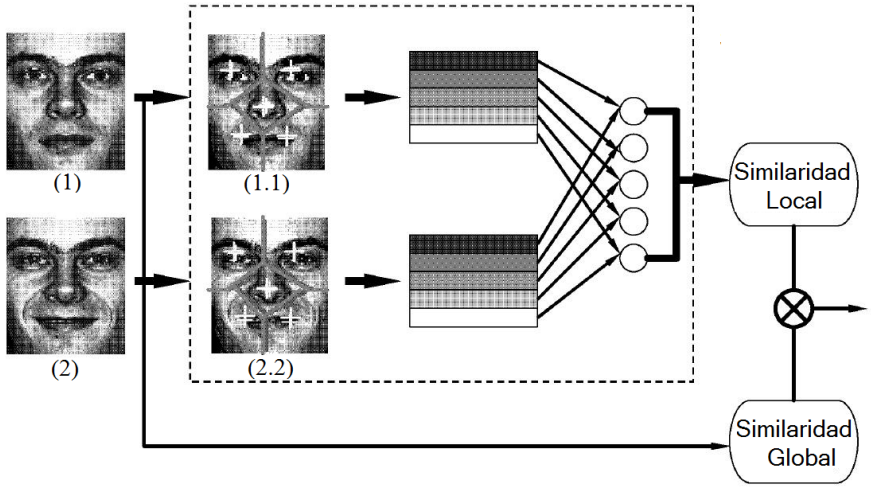


Figura 3.9: Esquema general del método de correspondencia por agrupamiento. En (1) y (2) se muestran la imagen en la base de rostros y la imagen a comparar respectivamente. En (1.1) se generan de forma automática las cinco subregiones con base en el algoritmo k-medias, de forma similar para (2.2), dichas subregiones son representadas por los recuadros sombreados en la imágen. Entonces se obtiene la similaridad entre regiones correspondientes usando la distancia euclidiana entre los puntos SIFT de cada subregión, representadas por los círculos blancos. Estas similaridades son combinadas para obtener una similaridad local. Por otro lado, se calcula la similaridad global entre las imágenes (1) y (2) sin las restricciones espaciales generadas por las subregiones. Finalmente, las similaridades local y global son combinadas para un atributo de similitud, que se obtiene mediante el operador $\otimes$ que es una multiplicación entre ambas similaridades. Mientras mayor sea dicha similitud, se sugiere que la identidad del individuo corresponde al de la base de rostros (Tomado de Luo et al. (2007)).

en la *i-ésima* subregión). Si se cuenta con dos imágenes I_t, I_r, la similaridad local es calculada en términos de la ecuación 2.1.

$$S_L(I_t, I_r) = \frac{1}{k}\sum_{i=1}^{k} max(d(f_{ti}^x, f_{ri}^y) \times w_i) \tag{3.1}$$

donde $x \in [1, ..., m_{ti}]$, $y \in [1, ..., m_{ri}]$, k es el número de subregiones, w_i es un peso de ajuste asignado a la *i-ésima* subregión y $d(f_{ti}^x, f_{ri}^y)$ denota la similaridad de dos vectores SIFT. Posteriormente, la similaridad global es obtenida con la ecuación 2.2.

$$S_G(I_t, I_r) = \frac{match(I_t, I_r)}{\sum_{i=1}^{k} m_{ri}} \tag{3.2}$$

donde $match(I_t, I_r)$ calcula el número de características cuyo cociente entre el mayor y el segundo mayor, es menor a un umbral. Finalmente la similaridad entre los rostros es dada en términos de $S = S_L \times S_G$. En los experimentos Luo *et al.* utilizan 1196 rostros de la base de datos FERET, Phillips et al. (1997), y 2538 de CAS-PEAL, Cao y Shan (2004), las imágenes fueron normalizadas a 150×130 y 75×65 pixeles respectivamente, además de ser enmascaradas para conseguir únicamente el rostro en ellas. Para las imágenes de FERET se manejaron cuatro grupos de prueba en los cuales se obtuvieron resultados del 97 % de precisión en imágenes con variaciones en expresión, 47 % en variaciones de iluminación, 61 % en imágenes tomadas en un periodo menor a los 18 meses y un 53 % en imágenes con más de 18 meses de las mismas personas. Las pruebas efectuadas en la base de datos CAS-PEAL incluyen el uso de accesorios (anteojos), cambios en la expresión, así como variaciones de 15º y 30º de la pose del rostro. Los resultados reportados arrojan 93.1 %, 94.7 %, 99.9 % y 70.6 %, respectivamente. Ejemplos de las imágenes de la base de datos CAS-PEAL son mostradas en la figura 3.10.

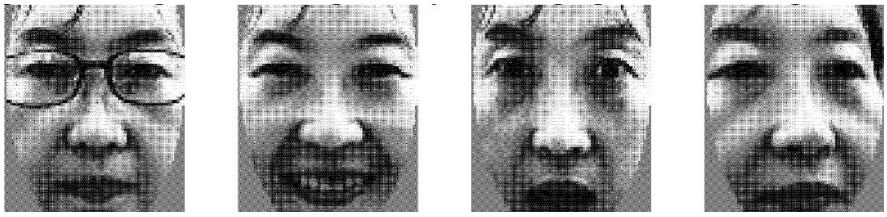


Figura 3.10: Ejemplos de imágenes normalizadas y enmascaradas de la base CAS-PEAL usadas por Luo et al. (2007).

3.5.4. Análisis

El uso de descriptores característicos SIFT ha mostrado un buen desempeño aplicado al problema de reconocimiento de rostros, debido a su robustez ante variaciones de iluminación, rotaciones, transformaciones afines y oclusiones parciales de los objetos. La mayor aportación de estos trabajos radica en la estrategia de emparejamiento de los descriptores, ya que la similitud entre puntos semejantes (como lo son los encontrados en los ojos) puede llevar a malas clasificaciones. Sin embargo, los métodos descritos anteriormente asumen bases de rostros con condiciones de iluminación controladas, la localización de los ojos en las mismas coordenadas para todas las imágenes e imágenes enmascaradas (ajustadas al rostro) con las mismas resoluciones. Dentro de las principales aportaciones de esta tesis recae en la forma que estos puntos sean agrupados de forma automática, así como la fusión de la evidencia provista en varias imágenes de una secuencia.

3.6. Conclusiones

En este capítulo se describieron algunos de los métodos de reconocimiento de rostros más significativos en el área. Se agruparon de acuerdo con la entrada de procesamiento en: reconocimiento de rostros en imágenes estáticas y reconocimiento de rostros en video.

Los métodos de reconocimiento de rostros que trabajan con imágenes estáticas han

mostrado un excelente desempeño en bases de rostros estándar como las presentadas en la sección 3.2; sin embargo, su éxito se ve ligado al tipo de imágenes que utiliza, ya que son de un tamaño y escala fijos, con condiciones de iluminación controlada así como de pose, lo cual resulta poco probable para su uso en robótica móvil. Por otro lado los métodos de reconocimiento de rostros que utilizan video han contemplado muchas de dichas condiciones adversas; sin embargo, mejoras en el tiempo de respuesta y recuerdo de dichos sistemas son requeridos para la aplicación de un robót de servicio, además de no consideran el aprendizaje en tiempo de ejecución de nuevos individuos.

En particular los métodos de reconocimiento de rostros que utilizan descriptores característicos SIFT presentan resultados alentadores en precisión; sin embargo, consideran situaciones poco comunes para una aplicación de robótica móvil ya que no se cuenta con imágenes del rostro centradas, ni con el mismo tamaño para todas las imágenes.

Sin duda existen numerosas técnicas para abordar el problema de reconocimiento de rostros; sin embargo, muchas de estas se ven limitadas a funcionar en ambientes con variación en la iluminación o pose del rostro. La variación en las expresiones faciales y ausencia/presencia de accesorios en los individuos suman complejidad a la tarea de reconocimiento de rostros. En particular al abordar el reconocimiento de rostros en video, añade la posibilidad de reforzar la precisión a través de la evidencia dada en las imágenes; sin embargo, nuevos problemas como la detección y seguimiento de los individuos deben ser considerados también. Dichos procesos aumentan sin duda el costo computacional del sistema, lo cual es un punto crítico para la mayor parte de aplicaciones orientadas a la robótica de servicio. Por lo tanto en esta tesis se parte del buen resultado obtenido por los métodos de reconocimiento basados en características SIFT y el refuerzo en la evidencia que aporta al reconocimiento del rostro con el video para desarrollar un método de reconocimiento de rostros para aplicaciones de robótica móvil.

Capítulo 4

Detección y seguimiento de rostros

En este capítulo se abordan las técnicas usadas para la implementación de las fases de detección y seguimiento. Con base en el análisis realizado en el capítulo 2, se decidió utilizar una implementación del método propuesto por Viola y Jones (2001a) para ambas fases.

4.1. Características Haar

Las características Haar son características digitales usadas en reconocimiento de objetos. Adquieren su nombre de las *wavelets* Haar wavelet (2008), las cuales son funciones que consisten en un breve impulso positivo seguido de un breve impulso negativo. En imágenes digitales, las características Haar son definidas por la diferencia entre la suma de todos los pixeles de dos regiones. En la figura 4.1 se muestran las características Haar básicas. En Papageorgiou et al. (1998), se presentan un marco general de entrenamiento para la detección de objetos basándose en una representación *wavelet* de un objeto, derivado de un análisis de las instancias de la clase. Esto es aprendiendo una clase de objeto en términos de un subconjunto de características Haar básicas. Para sus pruebas, los autores utilizan 3 características Haar básicas de 2×2 y de 4×4 pixeles en imágenes de rostros de 19×19, lo que obtiene un total de 1734 características

Haar en todas las posibles posiciones del rostro por imagen. En Viola y Jones (2001b,a) se proponen el uso de un algoritmo AdaBoost para la selección de las características Haar para la representación de los objetos, así como una representación de la imágen que acelera el proceso de cómputo de las características Haar. Posteriormente, Lienhart y Maydt (2002) extendieron las características Haar, agregando características rotadas en 45°. Para la implementación del algoritmo de seguimiento de esta tesis se utiliza el conjunto de características extendidas de Lienhart y Maydt mostradas en la figura 4.2.

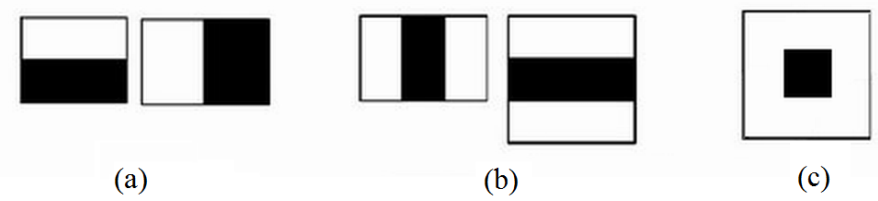


Figura 4.1: Características Haar básicas, en (a) se calcula la diferencia entre las 2 regiones, en (b) se calcula la diferencia entre la region central y los extremos, en (c) se calcula la diferencia entre la región central y la región que la rodea.

4.2. Imagen Integral

Una de las mayores aportaciones del trabajo de Viola y Jones es la representación de las imágenes para acelerar el cálculo de las características Haar, denominada **Imagen Integral**. Esta imagen integral contiene en cada píxel la suma de todos los pixeles arriba y a la izquierda de la imagen original. Con este método se logran obtener los valores requeridos para el cálculo de la sumatoria de una región con tan sólo tres o cuatro operaciones básicas. Por ejemplo, para obtener el valor de la región *D* de la figura 4.3, se suma el valor del píxel 4 con el valor del píxel 1, y se resta la suma del píxel 2 con el 3. Por lo tanto, sólo se requieren 6 referencias a la imagen integral para calcular el valor de una característica Haar con dos regiones adyacentes.

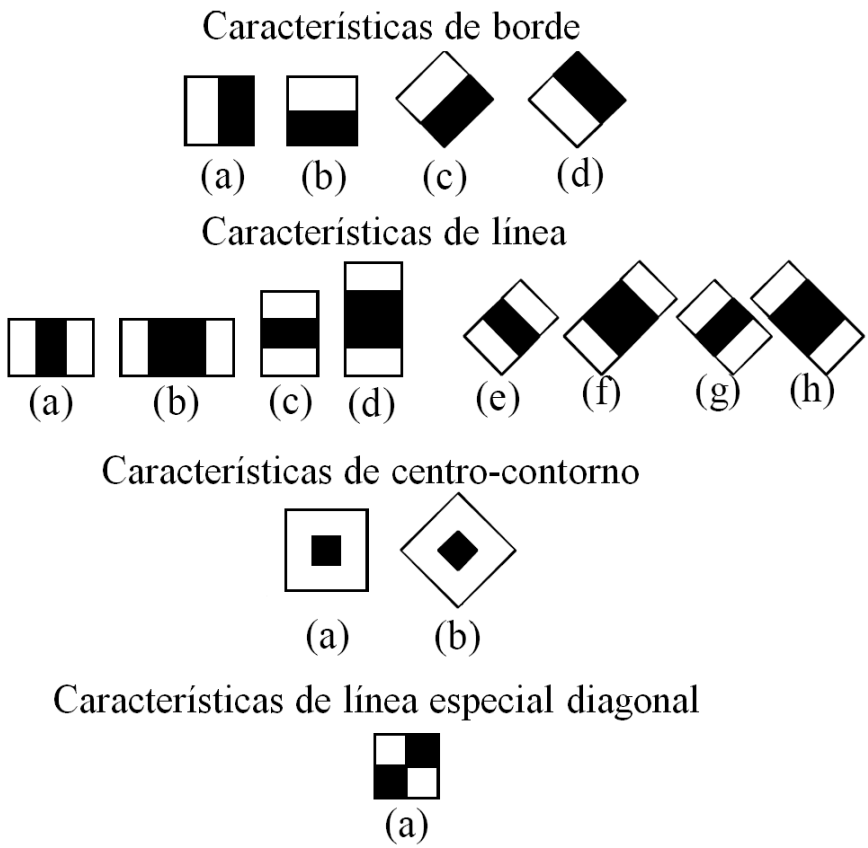


Figura 4.2: Características Haar extendidas de Lienhart y Maydt (2002).

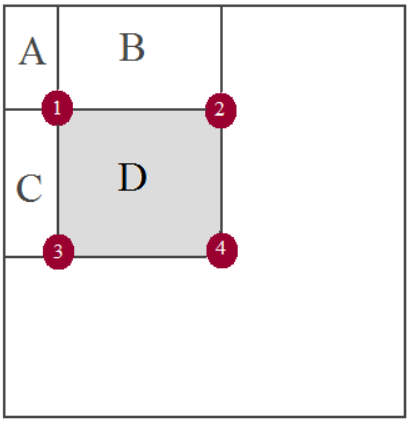


Figura 4.3: Cálculo de la región D con valores de los pixeles (4+1) - (2+3) de la imagen integral.

4.3. AdaBoost aplicado a detección de rostros

Adaptive Boosting, introducido Freund y Schapire (1995), es un algoritmo de *Boosting* que busca mejorar los resultados de clasificación mediante la combinación secuencial de clasificadores débiles. Los clasificadores débiles son llamados así porque no se espera que tengan un error de clasificación muy bajo. Inicialmente, a todos los ejemplos se les asigna un peso igual; sin embargo, cada vez que se genera un clasificador se cambian los pesos favoreciendo a los ejemplos mal clasificados. Esto brinda la capacidad para utilizar el nivel de errror de cada algoritmo y poner mayor atención a dichos ejemplos.

Para la selección de características Haar, Viola y Jones (2001a) utilizan el algoritmo de AdaBoost descrito en el algoritmo 1. El proceso general de reconocimiento es un árbol de desición conocido como cascada, donde el primer clasificador es muy simple pero rápido y el último es más complejo pero exacto, ver figura 4.4.

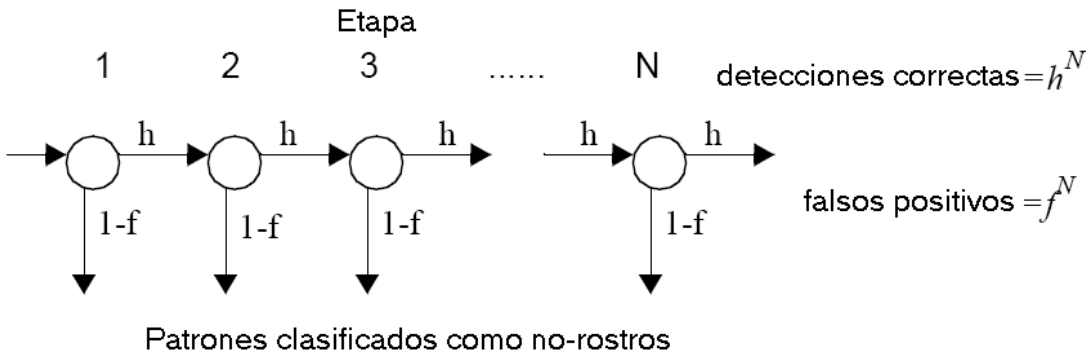


Figura 4.4: N niveles del detector de cascada especializado. Se toma un conjunto de N clasificadores débiles basados en una sóla característica Haar de manera que se construye un clasificador fuerte como una combinación lineal de los N débiles.

En la implementación del clasificador se usó una cascada de 20 clasificadores de árboles de decisión binarios de un sólo nivel. El conjunto de entrenamiento consistió de 4916 imágenes de rostros y 10000 imágenes de no-rostros.

Algoritmo 1 Algoritmo AdaBoost

Entrada: Conjunto de imágenes $(x_1, y_1), ..., (x_n, y_n)$, donde x_i representa la imagen y y_i la etiqueta para dicha imagen 0 para ejemplos negativos y 1 para positivos.

Salida: Clasificador fuerte $h(x)$

Variables locales: T clasificadores débiles

Inicializar los pesos $w_{1,i} = \frac{1}{2m}, \frac{1}{2l}$ para $y_i = \{0, 1\}$ donde *m* y *l* son el número de ejemplos negativos y positivos respectivamente.

para $t = 1, ..., T$ **hacer**

Normalizar los pesos

$$w_{t,i} \leftarrow \frac{w_{t,i}}{\sum_{j=1}^{n} w_{t,j}}$$

Para cada característica Haar, *j*, entrenar un clasificador h_j. Calcular el error con respecto a w_t con

$$\epsilon_j = \sum_i |h_j(x_i) - y_i|$$

Elegir al clasificador, h_t, con el error ϵ_j más bajo.

Actualizar los pesos

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

donde $e_i = 0$ si el ejemplo x_i es clasificado correctamente, $e_i = 1$ en cualquier otro caso, y $\beta_t = \frac{\epsilon_t}{1-\epsilon_t}$

fin para

El clasificador fuerte es:

$$h(x) = \begin{cases} 1 & si \sum_{t=1}^{T} \alpha_t h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} \alpha_t \\ 0 & \text{en caso contrario} \end{cases}$$

donde $\alpha_t = log\frac{1}{\beta_t}$

devolver $h(x)$

4.4. Seguimiento

La principal dificultad en el seguimiento en video es la asociación de la localización del objetivo en cuadros consecutivos, especialmente cuando dichos objetos se mueven relativamente rápido (superior a 5 Km/h que es la velocidad al caminar de un humano promedio).

Para este trabajo partiremos de la información proporcionada por el detector de rostros para definir un algoritmo de seguimiento de rostros. Después de detectado un rostro, su tamaño (definido por una región rectangular) es utilizado para definir una ventana de búsqueda en el siguiente cuadro, con lo cual se busca reducir el espacio de búsqueda. Esta ventana de búsqueda es definida incrementando cada lado rectangular por 2/3 de su longitud previa. Si en un cuadro el detector de rostros no retorna algún rostro, entonces se realiza la búsqueda para el siguiente cuadro en la imagen completa. Pese a su sencillez, esté método es sumamente robusto a diferentes condiciones de iluminación e incluso en algunas oclusiones parciales. En la figura 4.5 se muestran algunos resultados del algoritmo de seguimiento.

4.5. Conclusiones

La selección y análisis de los algoritmos de detección de rostros juega un papel sumamente importante en el éxito del reconocimiento de rostros, ya que sirve de entrada a dicho proceso. La selección del algoritmo debe contemplar factores propios de la aplicación en la cual será implementado. En el caso de esta tesis se consideran importantes factores como el tiempo de respuesta, la capacidad de funcionar en condiciones de iluminación variables y oclusiones parciales del rostro. Pese a que se desea una precisión aceptable en el algoritmo, esta no juega un papel crítico en su selección, ya que en etapas posteriores se considera una estrategia para el descarte de falsos positivos. Uno de

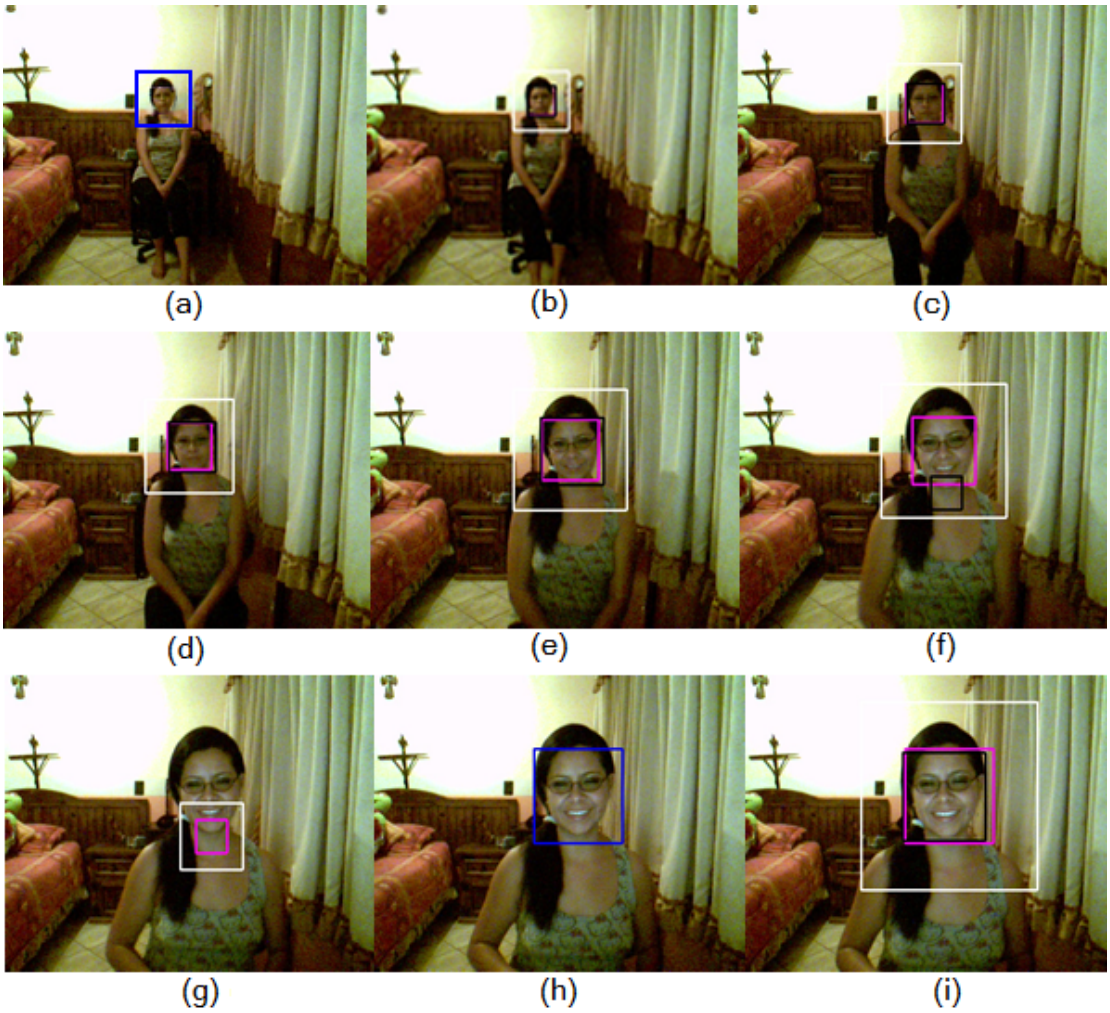


Figura 4.5: El proceso de seguimiento inicia buscando el rostro en la imagen completa (a); si un rostro es detectado, la región de interés es remarcada por un cuadro azul. Para el cuadro (b), la búsqueda del rostro se realiza sobre una ventana de búsqueda (cuadro blanco) incrementada en un factor de 2/3 del cuadro del rostro detectado en la imagen anterior (cuadro rosa), si el rostro es localizado en esta ventana de búsqueda se remarca con un cuadro negro. En (f) se muestra un cuadro rosa que refleja la localización del rostro en (e), el cuadro blanco muestra la ventana de búsqueda para (f) y el cuadro negro la región con rostro retornada por el detector. Entonces para (g) se reajusta la ventana de búsqueda; sin embargo, no se presenta el recuadro negro ya que el detector no retorna una región con rostro. En (h) se vuelve a realizar la búsqueda sobre la imagen completa mostrando un recuadro azul para la región con el rostro detectado.

los algoritmos que mejor cubre estas características es el propuesto por Viola y Jones. Debido al buen desempeño mostrado por la detección con Adaboost, se plantea el uso de un seguidor de personas con base en él. Una reducción en el espacio de búsqueda se consigue con la definición de una ventana de búsqueda con la información del cuadro anterior.

Capítulo 5

Sistema de reconocimiento de rostros

En este capítulo se describe el método desarrollado para reconocimiento de rostros. En la figura 5.1 se observa un esquema general del sistema de reconocimiento. Como entrada del sistema se cuenta con el video obtenido de la cámara del robot, a estas imágenes se les aplica el proceso de detección y seguimiento descritos en el capítulo 4, cuya salida genera una imagen del rostro que sirve de entrada para el proceso de reconocimiento.

El proceso de reconocimiento se divide en cuatro etapas principales. En la primera de ellas se realiza un **preprocesamiento de la imagen**, con el cual se busca aumentar la cantidad de puntos característicos SIFT en la imagen. En el siguiente paso se realiza una eliminación de falsos positivos arrojados por la fase de detección y seguimiento, buscando en la imagen la posición de uno o ambos ojos, con esta información a su vez se realiza el proceso de **selección de los puntos característicos SIFT** que servirán como descriptores del rostro. Ya con los puntos SIFT del rostro, se convierte esta información a un **vector de características similares** que refleja la similitud de la imagen a comparar con todas las imágenes de los individuos registrados en el sistema. Considerando que en circunstancias comunes tanto individuos conocidos como desconocidos se encontrarán en el ambiente, se establece un criterio de descarte para posibles imáge-

nes de personas desconocidas. Para lograr una mayor precisión en el reconocimiento, **se utiliza un enfoque probabilístico para incluir información de imágenes previas con la imagen actual** y entonces determinar la identidad del sujeto. A continuación se describe la metodología para la solución del problema de reconocimiento de rostros.

5.1. Método SIFT

La detección de puntos característicos invariantes resulta una tarea crucial en el área de visión por computadora, ya que estos pueden ayudar a resolver problemas de correspondencia entre diferentes puntos de vista de un objeto o una escena. Por lo tanto, su campo de aplicación es muy amplio siendo usados en procesos de alineación de imágenes, reconstrucción en 3D, seguimiento de objetos, navegación de robots, entre otras.

En particular, *Scale Invariant Feature Transform*, SIFT propuesto por Lowe (2004) busca la extracción de características que resulten invariantes a rotaciones y escalas en la imagen, robustas ante ciertos cambios en las condiciones de iluminación y oclusiones, así como algunos cambios en la perspectiva de los objetos o presencia de ruido en las imágenes. En la figura 5.2 se muestran resultados de correspondencia entre los puntos característicos de las imágenes de entrenamiento, y escenas reales con situaciones de cambios en su escala, rotación, iluminación, oclusión y ruido añadido por el proceso de adquisición de las imágenes; las líneas verdes unen pares de descriptores con la mayor similitud.

El proceso para extraer los descriptores característicos SIFT consta de cuatro etapas: **(i) Detección del espacio-escala**, que corresponde a la construcción de una pirámide de imágenes escaladas y suavizadas con un filtro Gaussiano para buscar puntos máximos y mínimos en las diferentes escalas de la imagen, dichos puntos son candidatos a puntos invariantes en la imagen que serán analizados en las siguientes etapas. **(ii) Localiza-**

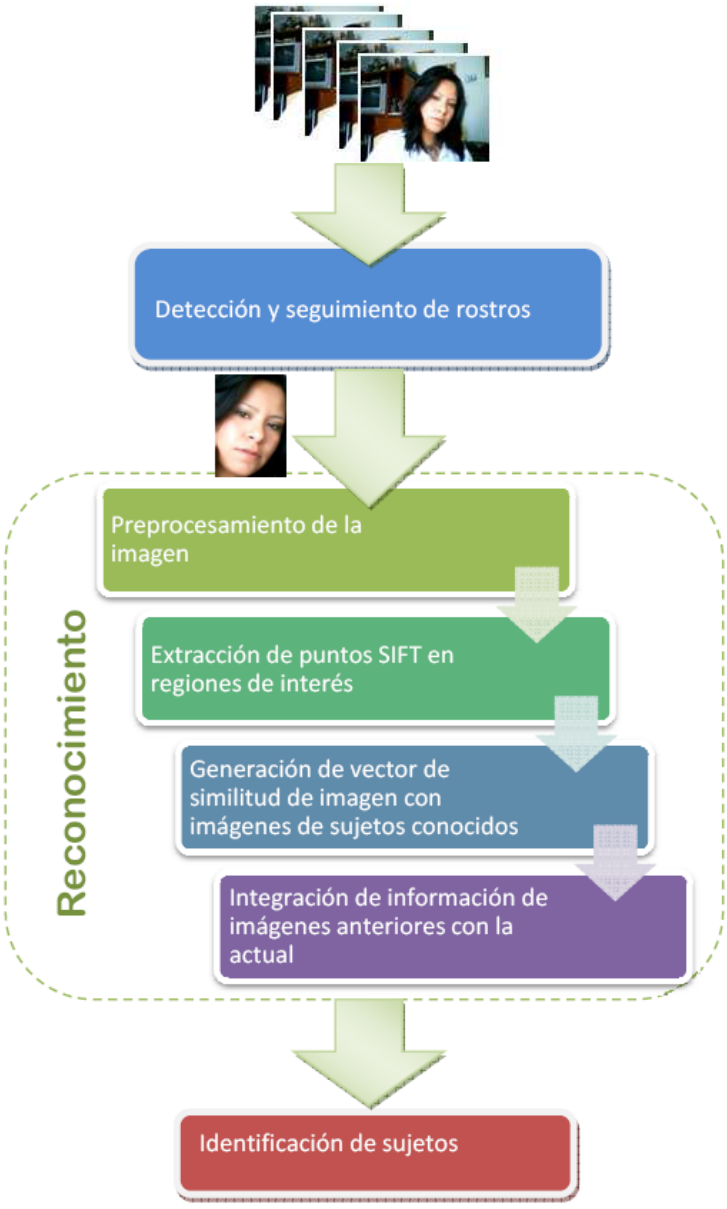
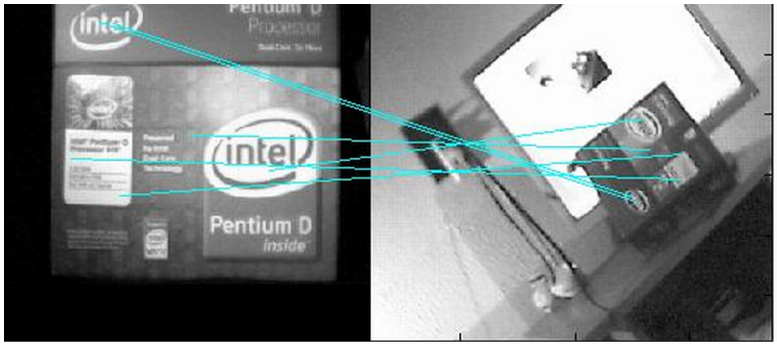


Figura 5.1: Diagrama de bloques del proceso de reconocimiento de rostros. Como entrada al sistema se tiene una secuencia de imágenes correspondientes a un video. Entonces se realiza la detección y el seguimiento del rostro. Para el reconocimiento se cuenta con una imagen únicamente del rostro, la cual recibirá un preprocesamiento para su posterior extracción de puntos característicos. Entonces se formarán las estructuras necesarias para su procesamiento con las imágenes de la base de rostros. Para mejorar la precisión del reconocedor se integra la información de varios cuadros mediante un enfoque Bayesiano.



(a)

(b)

Figura 5.2: Resultados de correspondencia con descriptores SIFT. En la (a) se muestra a la izquierda el objeto de entrenamiento, y a la derecha una escena en donde se situa el objeto con rotación, escalamiento y cambio de iluminación. En (b) se presenta una oclusión parcial.

ción de puntos característicos, los puntos generados en el paso anterior son evaluados para medir su estabilidad de acuerdo a su escala, localización y radio, con ello se busca eliminar puntos característicos con bajo contraste (sensibles al ruido en la imagen o que se encuentren mal ubicados en los bordes de los objetos). **(iii) Asignación de la orientación del puntos característicos**, para todos los puntos no rechazados en (ii) se calcula su orientación y magnitud con lo que se busca hacerlos invariantes a rotaciones. **(iv) Descripción de las características**, un descriptor de 128 elementos es generado para cada punto tomando en cuenta la información de la región local de la imagen (sus pixeles vecinos), con lo que se obtiene invarianza a ciertos cambios en la iluminación y cambios en la perspectiva 3D. Finalmente un punto característico SIFT estará formado por su localización (x, y) en la imágen original, la magnitud y la escala calculadas en (iii) y el descriptor de 128 elementos calculado en (iv). A continuación se describen a más detalle estas cuatro etapas.

1. **Detección del espacio-escala**

   La primera etapa de la detección de puntos es identificar localizaciones y escalas que puedan asignarse repetidamente bajo diferentes vistas del mismo objeto. Para ello, los puntos de interés son identificados tomando los extremos en el espacio-escala[1]. La representación del espacio-escala es una familia de imágenes derivadas, definidas por la función:

   $$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y)$$

   dode $*$ es la función de convolución de la imagen original, $I(x, y)$, con un *kernel* Gaussiano $G(x, y, \sigma)$ con diferentes varianzas σ. Está función representa una es-

[1] El espacio-escala es una marco de trabajo para la representación de una señal multi-escala desarrollada para la visión por computadora, en dicha teoría el manejo de estructuras de imágenes en diferentes escalas, representando una imagen como una familia de imágenes suavizadas.

cala en dicho espacio. Las escalas son substraidas para formar una pirámide de Gaussianas como:

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y)$$
$$= L(x, y, k\sigma) - L(x, y, \sigma)$$

donde k es un factor constante entre dos escalas. Este proceso se ilustra en la figura 5.3. Un conjunto de escalas forma un octavo, después del cual la imagen es submuestreada por un factor de 2 para servir de entrada al siguiente octavo, esto permite encontrar puntos candidatos a puntos SIFT en diferentes escalas. Finalmente, máximos y mínimos de la DoG (*Difference of Gaussians*) son detectados comparando cada pixel con sus 26 vecinos en una región de 3×3 pixeles en la escala actual y las dos escalas adyacentes (ver figura 5.4). Estos puntos máximos y mínimos representan puntos candidatos a ser puntos SIFT, los cuales serán calculados o descartados en etapas posteriores.

2. **Localización de puntos característicos**

Para cada posición candidata de puntos de interés se genera un modelo detallado para determinar su estabilidad de acuerdo a su escala y localización. Estos puntos son seleccionados con base en sus medidas de estabilidad definida por:

$$D(\hat{x}) = D + \frac{1}{2}\frac{\delta D^T}{\delta x}\hat{x}$$

donde D es la representa el nivel de la pirámide de Gaussianas en la que el punto fue localizado, y $\hat{x} = (x, y, \sigma)$ representa al punto. Si los valores de $|D(\hat{x})|$ para un punto candidato rebazan un umbral de 0.03 entonces es considerado para las siguientes etapas. Con esto se logra rechazar puntos sensibles al ruido o mal

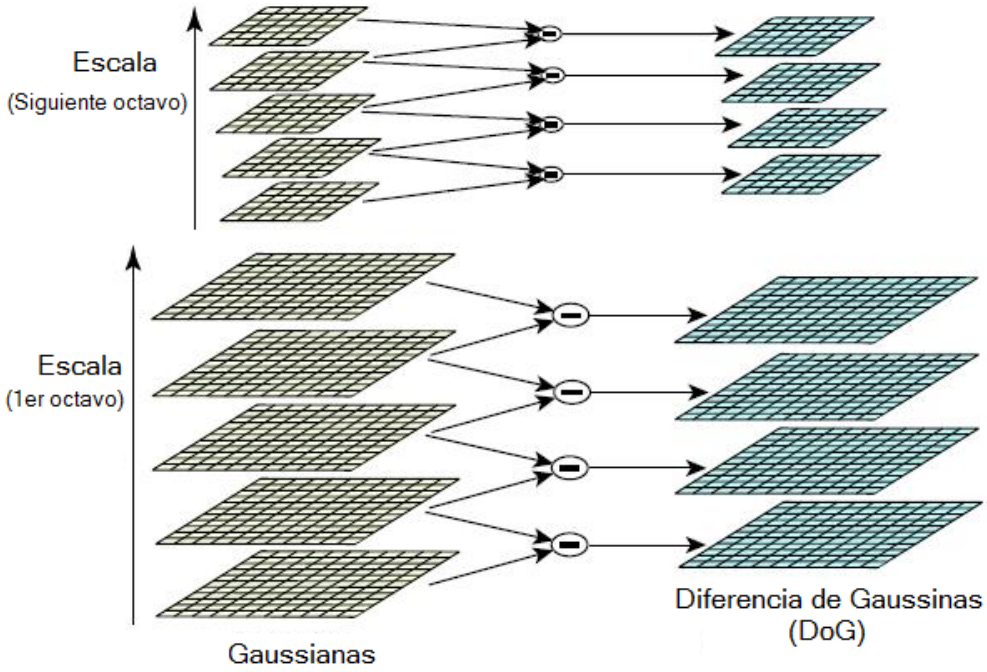


Figura 5.3: Para cada octavo (conjunto de imágenes del espacio-escala), la imagen inicial se convoluciona repetidamente con un *kernel* Gaussiano para producir el conjunto de imágenes espacio-escala (mostrado en la izquierda). Gaussianas adyacentes son sustraidas para generar la pirámide de diferencias de Gaussianas, DoG (derecha). Después de cada octavo, la imagen Gaussiana es submuestreada por un factor de 2, y el proceso es repetido (Tomado de Lowe (2004)).

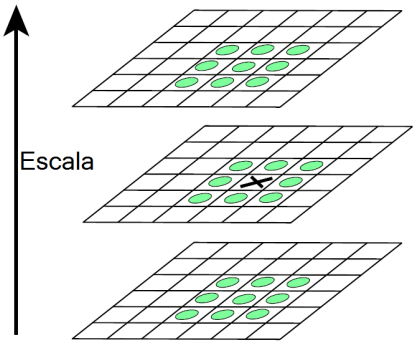


Figura 5.4: Los puntos de interés son detectados con los máximos y mínimos en la DoG. Este proceso se realiza comparando los valores del pixel con sus 9 vecinos en la escala superior e inferior, así como los 8 vecinos de su escala, tomada de Lowe (2004).

localizados a lo largo de los bordes.

3. **Asignación de la orientación**

Para cada punto característico una orientación es asignada, basada en las direcciones de los gradientes de la imagen local. Estás son calculadas con las ecuaciones 5.1 y 5.2.

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2} \tag{5.1}$$

$$\theta(x,y) = tan^{-1}\left(\frac{L(x,y+1) - L(x,y-1)}{L(x+1,y) - L(x-1,y)}\right) \tag{5.2}$$

donde $m(x,y)$ es el gradiente de magnitud, y $\theta(x,y)$ su orientación. $L(x,y)$, representa el valor del pixel (x,y) para cada escala.

4. **Descripción de las características**

Los gradientes m y θ, calculados en el paso anterior son medidos en la escala seleccionada en una región alrededor de cada punto de interés. Estos gradientes son transformados en una representación que permite niveles de significancia para la distorsión de forma y cambios de iluminación. Esto se realiza definiendo una vecindad de 16×16 gradientes agrupados en regiones de 4×4 gradientes. Para cada región, un histograma de 8 orientaciones posibles es definido, entonces cada gradiente de la región es sumado al elemento del histograma de mayor similitud en ángulo, ver figura 5.5.

Entonces las $8 \times 4 \times 4$ orientaciones de las 16 regiones son concatenadas para for-

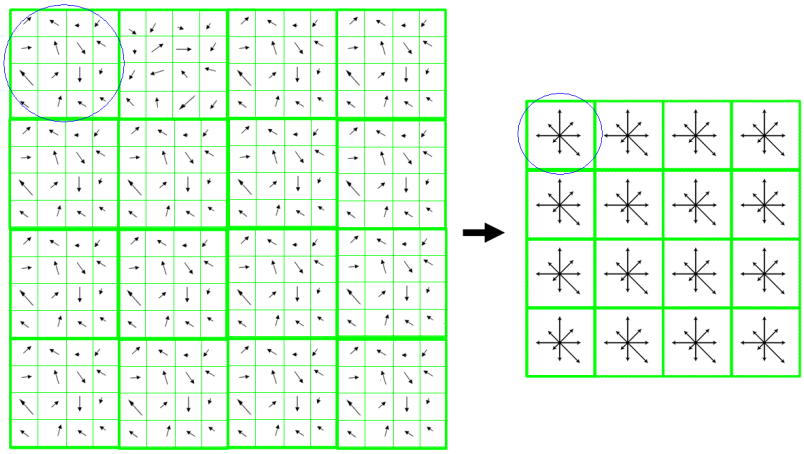


Figura 5.5: Se genera una cuadrícula de 16×16 gradientes de magnitud, agrupados en regiones de 4×4. Cada región es convertida a un histograma de ocho direcciones, entonces cada gradiente se mapea a alguna de estas direcciones. Finalmente, el descriptor SIFT es formado por la unión de las 8 direcciones de cada uno de las 16 regiones dando como resultado un vector de 128 elementos.

mar el descriptor característico de 128 elementos de los puntos SIFT. Finalmente se incluye al punto SIFT la posición en (x, y) donde fue localizado el punto, la escala en la que encontró y la orientación asignada en el paso previo.

5.2. Preprocesamiento de imagen

Una vez que un rostro es localizado, este recibe un preprocesamiento antes de ingresar a la etapa de reconocimiento con el objetivo de reducir el efecto de diferentes condiciones de iluminación. Como primer paso la imagen es mejorada mediante la ecualización de su histograma. Posteriormente, un algoritmo de compensación es aplicado.

5.2.1. Equalización del histograma

La ecualización del histograma es un método que usualmente incrementa el contraste en imágenes cuya distribución del histograma no es uniforme Awcock y Thomas. A través de este ajuste, las intensidades pueden ser mejor distribuidas sobre el histograma. La ecualización del histograma permite un mapeo de niveles de intensidad de una escala de niveles de gris *p*, en una escala de niveles de gris *q* con una distribución uniforme. Esto permite a las imágenes mejorar su contraste[2] sin distorcionar la imagen, ver figura 5.5, donde (a) es una imagen con bajo contraste y (b) su histograma no distribuido en toda la escala de grises, en (c) se muestra la imagen ecualizada con su correspondiente histograma en (d) ya extendido en todo el rango de valores. Esté método es útil en imágenes con fondos o frentes que son muy brillantes o muy obscuros.

Si consideramos una imagen en escala de gris, y sea n_i el número de ocurrencias del nivel de gris i. La probabilidad de ocurrencia de un pixel de escala i en la imagen es:

$$p(x_i) = \frac{n_i}{N}, i \in 0, ..., L-1 \tag{5.3}$$

donde N, es el número total de pixeles en la imagen y L el total de escalas de gris presentes en la imagen. Entonces, se calula la *función de distribución acumulativa*, c, con respecto de p, de la siguiente manera.

$$c(i) = \sum_{j=0}^{i} p(x_j), \tag{5.4}$$

Lo que se busca, es crear una transformación de la forma $y = T(x)$ tal que produzca una escala y, para cada escala de gris x en la imagen original, tal que la función de probabilidad acumulativa de y, sea distribuida a través de un rango de valores. Esta transformación es definida por:

[2]El contraste se define como la tasa de cambio de la luminancia relativa de los elementos de la imagen.

$$y_i = T(x_i) = c(i) \tag{5.5}$$

Pero considerando que T, mapea en un rango de valores entre 0 y 1, una última transformación es requerida para mapear dichos valores al conjunto de valores de la escala de gris *L* original. Esta es definida por:

$$y_i{}' = y_i \cdot (max - min) + min \tag{5.6}$$

con max y min los valores máximo y mínimo de la escala L. En la figura 5.6 se muestra un ejemplo con la imagen de entrada al proceso de ecualización de histograma, así como su respectivo histograma. Posteriormente, se muestra el resultado en la imagen y el histograma al mejorar el contraste con la ecualización.

5.2.2. Compensación de iluminación

El método de compensación utilizado por Ramírez-García (2006) es descrito a continuación. La imagen ecualizada, *ImgFace*, es dividida en 16 regiones regulares. Se obtienen los valores promedio para cada una de las regiones, almacenando el mayor de sus valores, *hv*, ecuación 5.7. La imagen de ajuste, *ImgAdj*, es entonces escalada utilizando el método de interpolación bilinear hasta obtener una imagen de tamaño igual al de la imagen original. La diferencia entre el mayor de los valores promedios, *hv*, y el valor de cada pixel de la imagen de ajuste, $ImgAdj_{k\times k}$, son usados para redefinir a la imagen de ajuste definida por la ecuación 5.8.

$$\mathsf{a}hv = max(ImgAdj_{4,4}) \tag{5.7}$$

$$ImgAdj_{x,y} = hv - ImgAdj_{x,y} \tag{5.8}$$

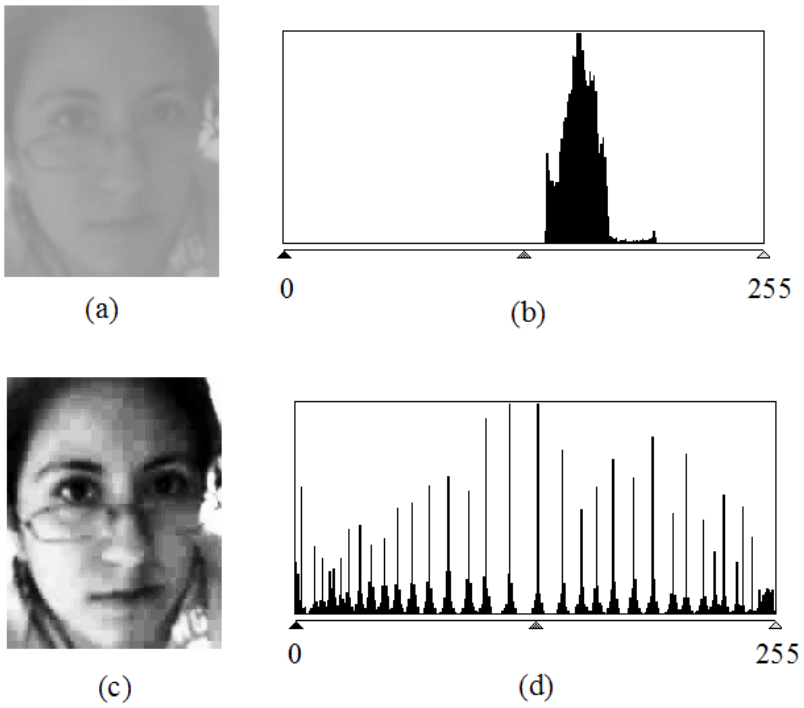


Figura 5.6: En (a) la imagen original con su histograma de valores (b), en (c) la imagen ecualizada con una notoria mejora en el contraste y en (d) el histograma de valores de la imagen ecualizada.

Finalmente el complemento de la imagen interpolada es sumada al rostro original para obtener así una imagen con una iluminación uniforme, *ImgFin*, usando la ecuación 5.9.

$$ImgFin_{x,y} = ImgFace_{x,y} + ImgAdj_{x,y} \tag{5.9}$$

Este proceso se ilustra en la figura 5.7, que muestra la imagen original, la imagen obtenida por los 16 valores promedio de las regiones, la imagen de ajuste y el resultado de la compensación de iluminación.

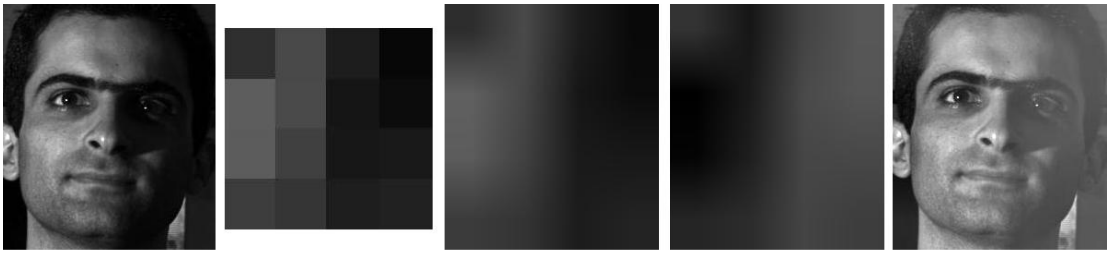


Figura 5.7: Compensación de iluminación: (a) imagen original, (b) imagen de ajuste de 4×4 pixeles, (c) imagen de ajuste interpolada, (d) imagen complemento y (e) resultado de la compensación.

Como resultado final de aplicar la ecualización del histograma y la compensación de la iluminación, en la figura 5.8 se puede apreciar el incremento en los puntos característicos entre las imágenes originales y las mejoradas.

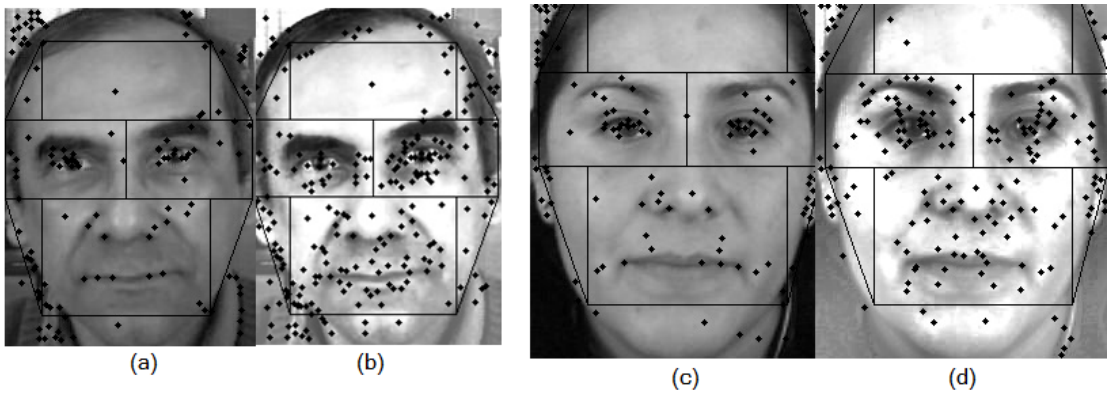


Figura 5.8: En (b) el resultado de aplicar ecualización de histograma seguido de la compensación de iluminación en (a), de forma similar para la imagen (c) el resultado del procesamiento se muestra en (d). Los puntos negros representan los puntos SIFT localizados en cada imagen.

5.3. Detección de características

Tomando la imagen del rostro ya procesada por la ecualización del histograma y seguida por la compensación de la iluminación, se busca eliminar los falsos positivos retornados por el detector de rostos. Para ello se utiliza el método de detección propuesto por Viola y Jones (2001a), pero en esta ocasión entrenado para buscar patrones de ojos en la imagen. Está tarea se realiza entrenando el clasificador Adaboost con un conjunto de ejemplos positivos de imágenes de ojos, y un conjunto negativo de imágenes sin ojos. De tal forma que si no se detecta al menos un ojo, entonces se descarta la imagen del rostro.

Una vez que el clasificador detecta la presencia de uno ó ambos ojos en el rostro, retorna sus coordenadas relativas a la imagen. Esta información se despliega con dos parches rectangulares en cada ojo como se muestra en la figura 5.9, dicha información corresponde a la posición (x,y), ancho y alto del área de cada ojo. Con base en esos datos se busca la generación automática de tres regiones de interés en el rostro: una correspondiente al ojo izquierdo, otra para el ojo derecho y una última para la nariz-boca. Las coordenadas, el ancho y el alto de cada ojo sirven como entrada al algoritmo

2, con el cual se genera un vector que contiene las coordenas necesarias para definir las tres regiones de interés . En la figura 5.10 (a), se muestra un ejemplo de los puntos retornados por el algoritmo 2, mientras que en (b) se muestran las combinaciones de los puntos retornados por el vector para la generación de las regiones de interés.

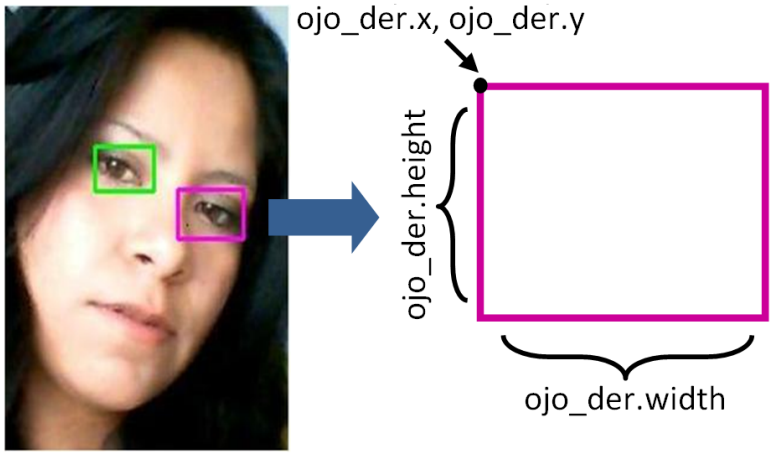


Figura 5.9: Información retornada por el detector de ojos. Coordenadas (x,y), ancho y alto del área que contiene al ojo.

Posteriormente, se extraen los puntos característicos SIFT de dicha imagen y de estos únicamente son considerados para el reconocimiento aquellos puntos que coincidan en alguna de las tres regiones. En la figura 5.11 se muestra este proceso en la fase de detección de ojos, extracción de puntos característicos SIFT y la agrupación en las áreas de interés para el reconocimiento.

5.4. Representación de la base de datos de individuos

Una de las principales ventajas del método de reconocimieto de rostros propuestos es la flexibilidad que ofrece para almacenar nuevos individuos en tiempo de ejecución.

Cuando un nuevo sujeto desea ser incluido en la base de datos de individuos co-

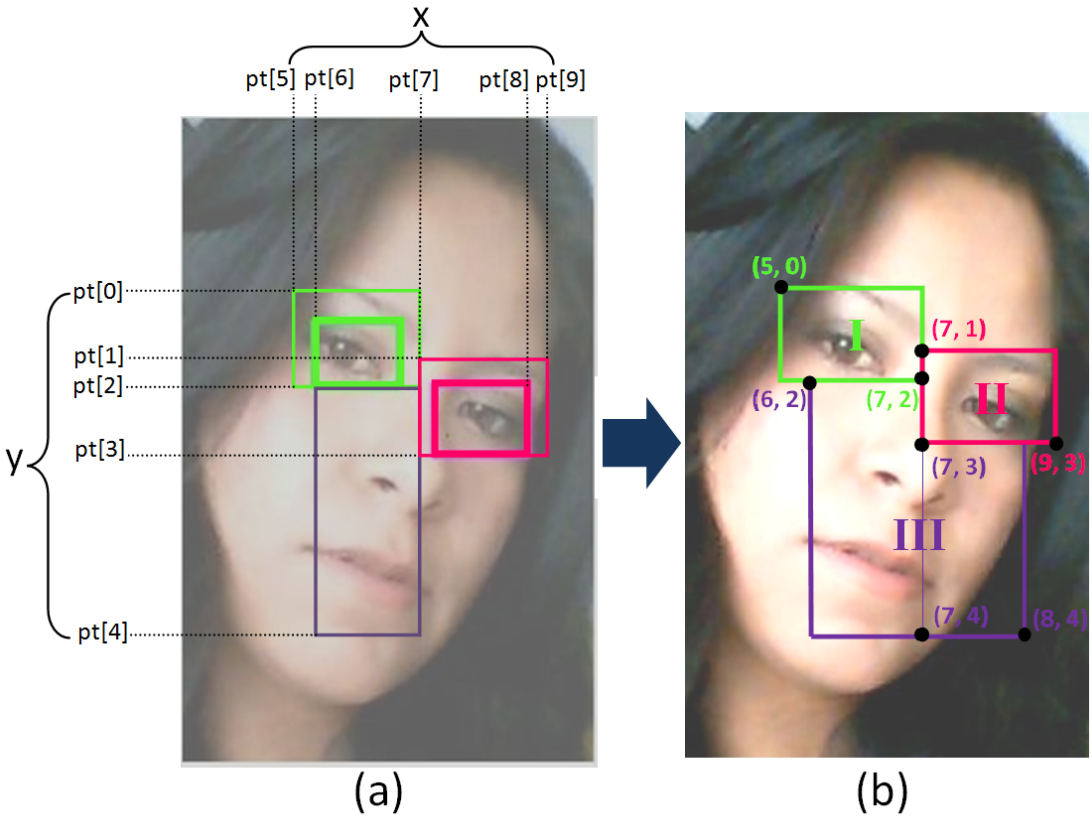


Figura 5.10: En (a) se muestran los puntos generados por el algoritmo 2, los elementos pt[0] al pt[4] son coordenadas para y, mientras que los elementos pt[5] al pt[9] refieren a las coordenadas en x que son usadas en (b) para definir las tres regiones de interés: I para el ojo izquierdo, II para el ojo derecho y III para la región nariz-boca.

Algoritmo 2 Algoritmo para la generación de regiones de interés

Entrada: Información del detector ojo_izq.x, ojo_izq.y, ojo_izq.height, ojo_izq.width, ojo_der.x, ojo_der.y, ojo_der.height, ojo_der.width
Salida: Vector pt[]
Variables locales: T clasificadores débiles
width1 = ojo izq.x + ojo izq.width
bridge = ojo der.x - width1
average = (ojo der.height + ojo izq.height)/5
width2 = ojo der.x + ojo der.width
si ojo izq.height $<$ ojo der.height **entonces**
  third = ojo izq.height/3
si no
  third = ojo der.height/3
fin si
si ojo der.y $<$ ojo izq.y **entonces**
  mayor = ojo der.y
si no
  mayor = ojo izq.y
fin si
si bridge $\geq$ 0 **entonces**
  half = bridge/2
  bridge = width1 + half
si no
  half = (bridge $\times$ -1)/2
  bridge = ojo der.x + half
fin si
pt[0] = ojo izq.y - average
pt[1] = ojo der.y - average
pt[2] = ojo izq.y + ojo izq.height
pt[3] = ojo der.y + ojo der.height
pt[4] = mayor + width2 - ojo izq.x
pt[5] = ojo izq.x - average
pt[6] = ojo izq.x
pt[7] = bridge
pt[8] = width2
pt[9] = width2 + average
devolver pt[]

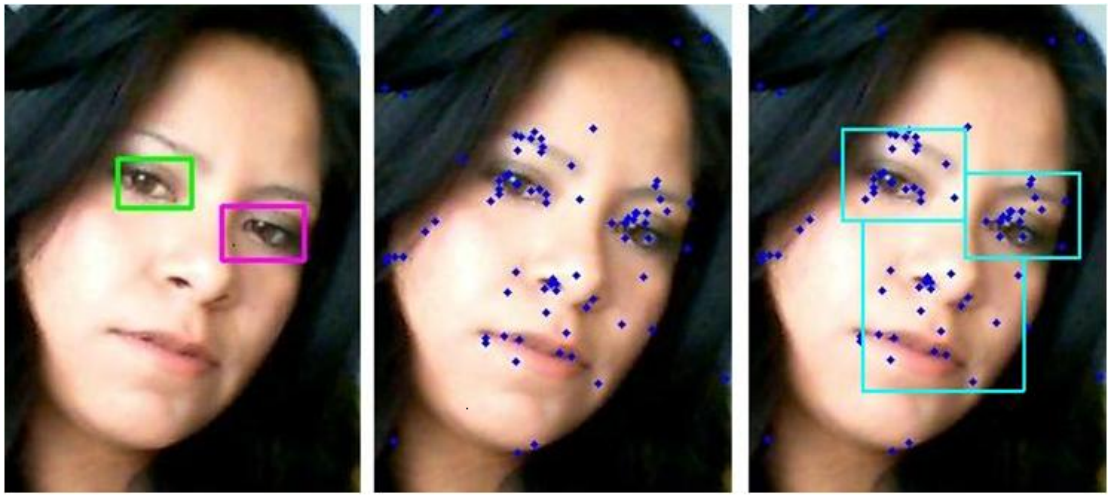


Figura 5.11: Extracción de características SIFT: (a) se muestra la fase de detección de ojos utilizando el algoritmo de Viola y Jones, (b) la localización de los puntos SIFT en (x,y), (c) la generación de las tres áreas de interés para el reconocimiento.

nocidos, se debe esperar hasta que aparezca en alguna imagen del video su rostro y sean detectados los ojos. Se definen las tres regiones de interés y se extraen los puntos característicos SIFT para cada individuo en la base de datos. Entonces se genera un vector de características que almacena para cada sujeto la sumatoria de los puntos SIFT de sus tres regiones de interés definido por: $\vec{TP} = \{tp_1, tp_2, ..., tp_n\}$ donde tp_i es el número de puntos característicos del sujeto i. Además, se define un vector de umbrales, $\vec{T} = \{t_1, t_2, ..., t_n\}$, donde cada entrada t_i es un porcentaje α de los puntos SIFT totales de la imagen del rostro i. Por ejemplo, si el primer individuo tiene 180 puntos característicos SIFT, y se define $\alpha = 10$, entonces tp_1 y t_1 tendrán valores de 180 y 18 respectivamente.

Con estas estructuras se logra incluir un nuevo sujeto a la base de rostros, como un nuevo índice en cada uno de los vectores $\vec{TP}$ y $\vec{T}$. Por tal motivo, se puede decir que el proceso de añadir nuevos individuos a la base de datos sólo requiere una imagen del sujeto, lo que resulta ideal para aplicaciones de robótica móvil.

5.5. Estrategia de correspondencia

La estrategia de correspondencia entre dos puntos SIFT se basa en el criterio usado por Lowe (2004). Un punto SIFT está formado por sus coordenadas *(x,y)* en la imagen, la escala en la que fue encontrado, su orientación y un descriptor con 128 elementos . En Lowe (2004) se considera que dos puntos característicos son similares si la distancia euclidiana entre pares de descriptores es menor a cierto umbral.

Entonces, la cantidad de puntos similares entre dos imágenes está definido por la ecuación 5.10:

$$simililud(I_{BD}, I_{new}) = \sum_{\forall\, reg \in\ Cara} match(I_{BD}^{reg}, I_{new}^{reg}) \tag{5.10}$$

donde I_{BD} representa la imagen almacenada en la base de rostros, I_{new} es la imagen obtenida en el cuadro de video a comparar, reg recorre las regiones de interés que conforman a $Cara = \{ojo\_derecho, ojo\_izquierdo, nariz\_boca\}$, y $match(I_x, I_y)$ es la función que retorna la cantidad de puntos similares entre dos regiones de acuerdo a la distancias euclidiana entre todos los puntos de dichas regiones. Con esto se logra que puntos característicos que definen a zonas parecidas como lo son las regiones de los ojos, no se confundan en el proceso de correspondencia.

Una vez que se tiene el número total de puntos similares entre un rostro conocido y el rostro a identificar, se genera un vector de similaridades, $\vec{S} = \{s_1, s_2, ..., s_n\}$ para los n rostros almacenados en la base de datos, donde s_i se obtiene como $similitud(I_i, I_{new})$ y representa la cantidad de puntos similares entre el rostro nuevo y el rostro del sujeto i.

Dado que se contempla que en el ambiente del robot se encuentren individuos conocidos y desconocidos, se observa un comportamiento particular para los vectores de similitud cuando se presentan dichas situaciones, ver . En donde los primeros 10 cua-

dros se presenta la cantidad de puntos característicos entre cada uno de los individuos de la base de rostros y el individuo a identificar (no registrado en la base de rostros), en estos vectores se puede observar que se conservan intervalos de valores limitados (entre 1 y 15 puntos similares). En contraste en los cuadros 11 al 23 se presenta el comportamiento de los vectores de similitud cuando se tiene registrado al individuo a reconocer, variando sus intervalos entre 1 y 111 puntos similares.

En las figuras 5.12 y 5.13 se grafican algunos de los vectores correspondientes a situaciones en donde no se tiene registrado al individuo en la base de rostros, y situaciones en donde si se tiene registrado al individuo, respectivamente. Como puede verse en la gráfica 5.12 se sugiere que un vector con valores en un pequeño rango, pueden corresponder a un individuo no registrado en la base de rostros; mientras que los comportamientos de pico, sugieren la presencia de un sujeto conocido en la base de rostros. Por tal motivo se plantean tres criterios para descartar vectores de similaridad que puedan contener sujetos desconocidos:

1. **Criterio 1**

   Al menos un elemento del vector de similitud es mayor que su correspondiente en la base de rostros.

   $$\text{Si } s_i > t_i \text{ para cualquier } i = 1, ..., n$$

   donde $s_i \in \vec{S}$ y $t_i \in \vec{T}$.

2. **Criterio 2**

   Si se cumple el critero 1, y además se cumple que la diferencia entre el máximo de los valores de similitud de $\vec{S}$ y el $2do$ más grande es supera por el doble al segundo.

   $$max(\vec{S}) - 2nd(\vec{S}) \geq 2 \times 2nd(\vec{S})$$

	Cuadro	VECTOR DE SIMILITUD						MAX	2nd	AVR	C1	C2	C3
DESCONOCIDO	1	1	0.1	1	2	2		2	2	1.03	NO	NO	NO
	2	2	1	2	4	3		4	3	2.00	NO	NO	NO
	3	1	1	3	4	6		6	4	2.25	NO	NO	NO
	4	3	2	6	3	7		7	6	3.50	NO	NO	NO
	5	5	3	10	8	4		10	8	5.00	SI	NO	NO
	6	1	3	10	4	5		10	5	3.25	SI	NO	SI
	7	5	4	11	6	6		11	6	5.25	SI	NO	NO
	8	3	2	12	4	8		12	8	4.25	SI	NO	NO
	9	9	10	15	10	8		15	4	9.25	SI	SI	NO
	10	2	6	10	2	7		10	7	4.25	SI	NO	NO
CONOCIDO	11	3	2	12	7	7	63	63	12	6.20	SI	SI	SI
	12	2	3	9	6	6	65	65	9	5.20	SI	SI	SI
	13	2	3	9	6	5	111	111	9	5.00	SI	SI	SI
	14	3	2	6	4	5	29	29	6	4.00	SI	SI	SI
	15	2	2	4	2	3	13	13	4	2.60	SI	SI	SI
	16	4	7	1	4	1	4	7	4	2.80	NO	NO	NO
	17	3	2	3	2	5	12	12	5	3.00	SI	NO	SI
	18	0.1	3	7	9	3	37	37	9	4.42	SI	SI	SI
	19	6	3	8	2	4	37	37	8	4.60	SI	SI	SI
	20	7	3	8	4	2	53	53	8	4.80	SI	SI	SI
	21	5	4	10	7	4	52	52	10	6.00	SI	SI	SI
	22	4	9	10	14	2	47	47	14	7.80	SI	SI	SI
	23	6	11	7	5	8	54	54	11	7.40	SI	SI	SI

Tabla 5.1: Comportamiento en los vectores de similitud para individuos no registrados en la base de rostros y individuos registrados. Los diez primeros cuadros representan a individuos no registrados, como puede verse los valores para dichos vectores se mantienen en un rango pequeño de valores; mientras que para los últimos trece se muestra un pico ya que se tiene al individuo registrado en la base de rostros. Con los tres criterios se propone una estrategia para seleccionar cuadros con una alta probabilidad de tener un sujeto conocido. El C1considera que al menos un valor del vector rebase el $n\,\%$ de los puntos en el vector de umbrales. C2 busca que la diferencia entre el mayor y el segundo más grande sea del doble del segundo. Finalmente C3 considera que el valor promedio de los puntos, exceptuando al máximo sea del doble de dicho valor promedio. Con estos criterios se puede ver que criterio C1 es más relajado y permitiría el paso de vectores con sujetos desconocidos, mientras C2 y C3 resultan más estrictos.

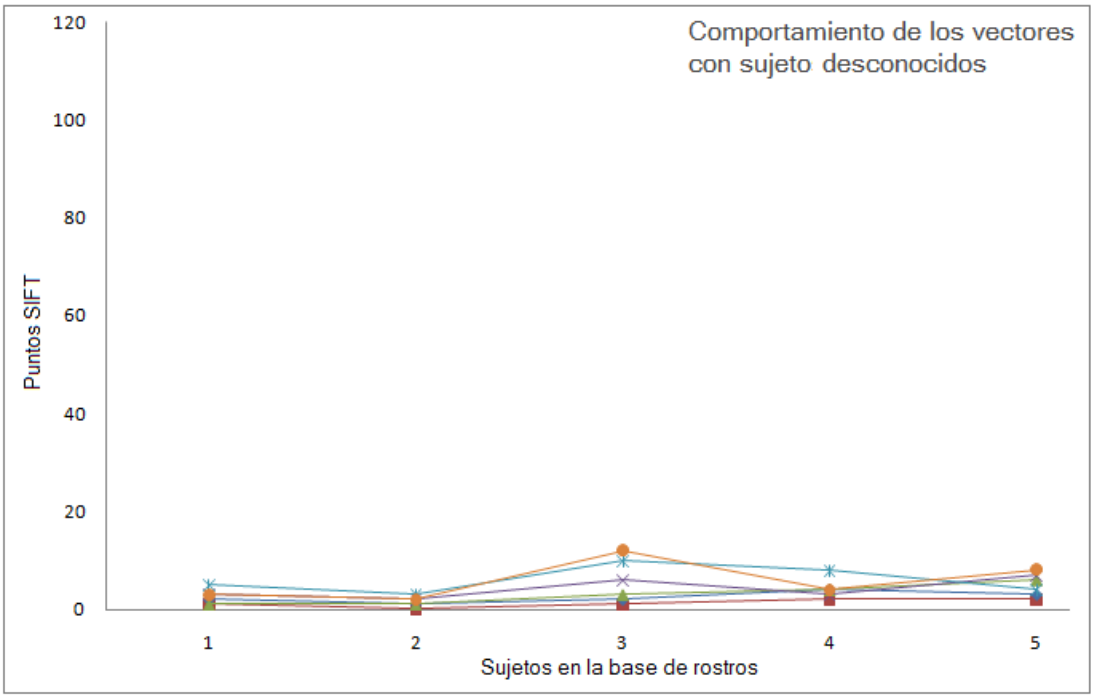


Figura 5.12: Comportamiento de vectores de similitud para individuos desconocidos a la base de rostros con valores entre [0.1,15] puntos SIFT.

3. **Criterio 3**

   Si se cumple el critero 1, y además se cumple que la diferencia entre el máximo de los valores de similitud de $\vec{S}$ y el promedio del resto es más grande por el doble de dicho promedio.

$$max(\vec{S}) - avrg(\vec{S}) \geq 2 \times avrg(\vec{S})$$

   donde $avrg(\vec{S})$ es el promedio de todos los $s_i \in \vec{S}$, excepto el máximo.

En la tabla 5.1 se presentan el comportamiento de aplicar los criterios 1, 2 y 3 en las columnas C1, C2 y C3. Como puede observarse el criterio 1 es más relajado y permite el paso de algunos vectores correspondientes a individuos desconocidos a la etapa de reconocimiento descrita en la siguiente sección; mientras que los criterios 2 y 3 resultan más estrictos y eliminan la mayor parte de los vectores de los sujetos no registrados. Sin embargo; para fines de la aplicación del robot mensajero, se determinó utilizar

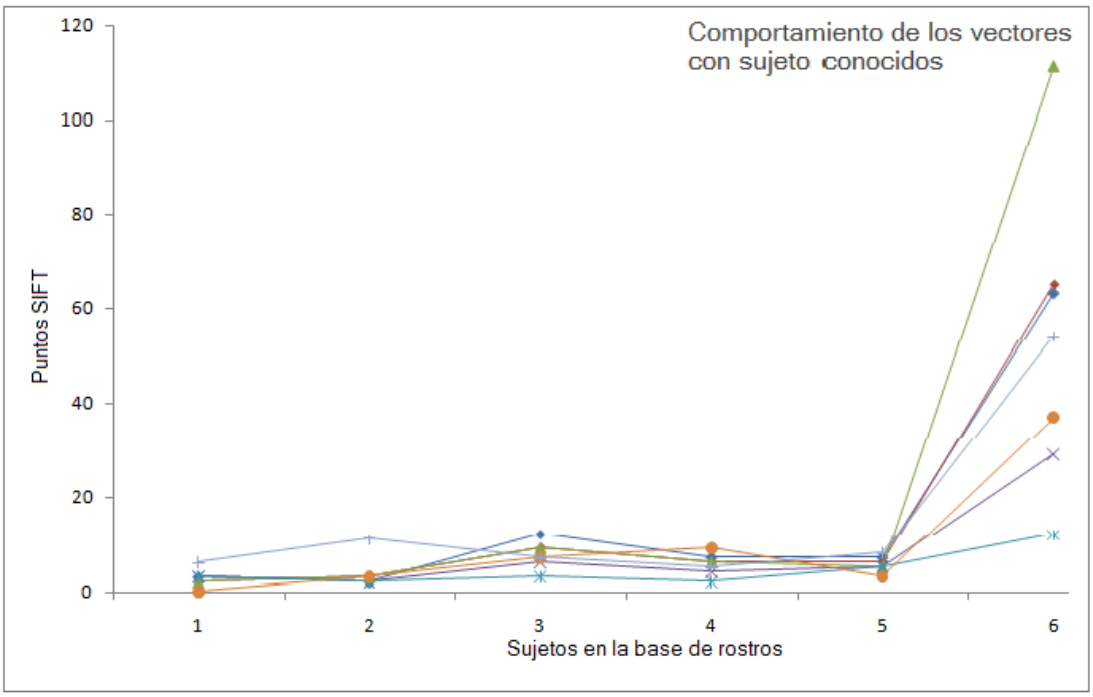


Figura 5.13: Comportamiento de vectores de similitud para individuos registrados en la base de rostros con valores entre [0.1,111] puntos SIFT.

únicamente el criterio 1, ya que las imágenes obtenidas por el video suelen generar pocos puntos característicos SIFT, y los criterios 2 y 3 descartarían una gran cantidad de vectores de similitud con lo que se disminuirian los porcentajes de recuerdo del sistema; mientras que seleccionar los criterios 2 o 3 aumentarían los porcentajes de precisión. Estos experimentos son detallados en el capítulo 6.

5.6. Reconocimiento en video

5.6.1. Teorema de Bayes

El teorema de Bayes ofrece un método estadístico para calcular una probabilidad condicional. Este teorema es de gran utilidad para evaluar una probabilidad *a posteriori* partiendo de probabilidades simples, y así poder revisar la estimación de la probabilidad *a priori* de un evento.

El teorema de Bayes parte de una situación en la que es posible conocer las probabilidades de que ocurran una serie de sucesos A_i. A esta probabilidad se añade un suceso B cuya ocurrencia proporciona cierta información, porque las probabilidades de ocurrencia de B son distintas según el suceso A_i que haya ocurrido. Conociendo que ha ocurrido el suceso B, la fórmula del teorema de Bayes nos indica cómo modifica esta información las probabilidades de los sucesos A_i, reflejada en la ecuación 5.11.

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{P(B)} = \frac{P(A_i)P(B|A_i)}{\sum_{j=1}^{n} P(A_j)P(B|Aj)} \tag{5.11}$$

5.6.2. Reconocimiento

Debido a que este sistema trabaja con video, se busca utilizar la información provista por varias imágenes en favor de lograr una mejor precisión en el proceso de reconocimiento. Esto se realiza usando un enfoque Bayesiano para realizar el reconocimiento del rostro. La idea es evaluar la probabilidad de tener rostro f_i dada la información de las características de la imagen I, denotada por $P(f_i|I)$, y entonces actualizar las probabilidades usando información de varios cuadros con base en la fórmula de Bayes.

$$P(f_i|I) = \frac{P(f_i)P(I|f_i)}{P(I)} = \frac{P(f_i)P(I|f_i)}{\sum_{k=1}^{n} P(I|f_k)P(f_k)} \tag{5.12}$$

donde $P(f_i)$ es la probabilidad *a priori* que se inicializa como $\frac{1}{n}$ para los n rostros almacenados. Para incorporar la información de los cuadros previos para el ajuste de las probabilidades en el cuadro actual, la $P(f_i)^t$ (la probabilidad *a priori* de un rostro i en el cuadro t) se toma como $P(f_i|I)^{t-1}$ (la probabilidad *posteriori* de un rostro i dada la evidencia s en el cuadro $t-1$).

$P(I|f_i)$ es estimado del vector de similaridades. Se aplicaron dos enfoques para realizar esta conversión, uno absoluto y uno relativo. La probabilidad absoluta, P_{abs},

toma el porcentaje de similaridades para cada rostro con respecto a las similaridades de todos los rostros:

$$P_{abs}(I|f_i) = \frac{s_i}{\sum_{k=1}^{n} s_k} \tag{5.13}$$

donde s_i es en número de puntos similares entre la imagen actual y el $i-ésimo$ rostro de la base de datos. El esquema relativo, P_{rel}, toma el porcentaje de similaridades relativas con respecto al número total de puntos característicos reconocidos en cada rostro, con respecto a las similaridades relativas para todos los rostros:

$$P_{rel}(I|f_i) = \frac{\frac{s_i}{tp_i}}{\sum_{k=1}^{n} \frac{s_k}{tp_k}} \tag{5.14}$$

donde $tp_i \in \vec{TP}$ es el número total de puntos SIFT del $i - ésimo$ rostro en la base de datos. Para los experimentos se probaron los dos esquemas de probabilidad; sin embargo, para la implementación en el robot de servicio Markovito se optó por un esquema de probabilidada absoluta ya que en combinación con un umbral de 5 % en $\vec{T}$, reportan una precisión del 96.65 % con un recuerdo del 57.32 %.

Para el reconocimiento, todas las probabilidades $P(f_i|I)$ serán almacenadas en un vector global de probabilidades denotado como:

$$\vec{P} = \{P(f_1|I), P(f_2|I), ..., P(f_n|I)\}$$

Entonces, una persona será reconocida cuando su probabilidad sea claramente superior a la del resto de los rostros. Esta criterio se considerá cubierto con la siguiente condición:

$$max(\vec{P}) - 2nd(\vec{P}) \geq 2 \times 2nd(\vec{P})$$

donde la función max calcula el máximo valor de probabilidad para el vector $\vec{P}$, y $2nd$ el segundo elemento con mayor probabilidad. Si esta condición se cumple, entonces se obtiene el índice del elemento de $\vec{P}$ con la mayor probabilidad como $persona = ind\_max(\vec{P})$, y se dice que se ha reconocido al sujeto, $persona$, en la imagen.

En cuanto un rostro es reconocido, las probabilidades son reiniciadas a $\frac{1}{n}$ para cada elemento $P(f_i|I)$. De igual forma se reinician las probabilidades si han pasado 10 cuadros en los que ningún rostro haya sido localizado, o bien los ojos no sean encontrados. El reinicializar las probabilidades contribuye a la posibilidad de detectar un rostro diferente si el usuario cambia, ya que los valores de los vectores de probabilidad se encuentran muy elevados para el sujeto inicial, mientras que para el resto de ellos es muy bajo y no sería posible elevar dicho valor sin reinicialización.

5.7. Conclusiones

En este capítulo se expuso un método para reconocimiento de rostros que contempla varias etapas. Primero la mejora de imágenes que incrementa la cantidad de puntos SIFT en las imágenes, con lo cual se aumenta el número de puntos característicos similares entre la imagen de entrada y las imágenes de rostros almacenadas en la base de rostros conocidos. Este preprocesamiento de las imágenes robustece la extracción de las características, lo cual resulta una aportación significativa.

Con el uso de un detector de ojos no sólo se rechazan falsos positivos del detector de rostros, sino que también se definen tres regiones de correspondencia para los puntos SIFT: región ojo derecho, región ojo izquierdo y región nariz/boca . Estas regiones eliminan los empates entre puntos muy similares como lo son los generados en los ojos. Si bien es cierto que la restricción de localización en coordenadas *(x,y)* de la imagen ya se ha utilizado con anteriordad Bicego et al. (2006); Luo et al. (2007), la definición

automática y generación de regiones de interés mediante la localización de los ojos en la imagen y biometría representa una aportación de esta tesis.

Se presentó un esquema de representación que refleja las similitudes entre la imagen a evaluar y las existentes en la base de rostros conocidos, mediante un vector de similaridades. Este esquema mostró no sólo ser flexible para la inclusión de nuevos individuos en tiempo de ejecución, sino que permitió descartar imágenes con alta probabilidad de tener sujetos desconocidos en la base de rostros. El uso de un enfoque Bayesiano permitió el refuerzo de hipótesis sobre la identidad del individuo, mezclando información de imágenes posteriores con la evidencia mostrada en la imagen actual.

En general, el uso de características SIFT, la detección de ojos con Adaboost y el enfoque Bayesiano para incorporación de evidencia, brindan a este esquema la robustez necesaria para su uso en plataformas de robótica móvil, lo cual se demostrará en el siguiente capítulo.

Capítulo 6

Experimentos y resultados

Se diseñaron diversos experimentos para la evaluación del sistema de reconocimiento desarrollado. Dado que el sistema de reconocimiento busca ser aplicado en una plataforma robótica, el primer conjunto de experimentos evalúa su desempeño al utilizar la cámara monocular del robot PeopleBot. En el segundo grupo de experimentos se busca analizar el comportamiento del sistema de reconocimiento de rostros al ser evaluado con una base de datos estándar con variaciones en iluminación, así como el desempeño con el incremento de sujetos en la base de datos. Finalmente se propone un experimento de reconocimiento de rostros con el robot de servicio Markovito y personas en movimiento.

6.1. Experimentos con video

La incorporación del sistema de reconocimiento de rostros orientado a interacción humano-robot resulta un punto crítico de esta tesis, por lo tanto se contempla un conjunto de experimentos para su prueba. El dispositivo de adquisición de imágenes con el que cuenta el robot PeopleBot es una cámara Canon VC-C4 CCD con un zoom de $16\times$, 200 grados de movilidad sobre el eje x, y 120 grados de inclinación, manejando una resolución de 640×480.

Para los experimentos se grabó un video, en donde cada persona se presenta caminando de frente al robot inmóvil desde una distancia de 5 metros hasta llegar a un metro de distancia. Aproximadamente se capturan 100 cuadros por cada persona. En estos experimentos se consideran a 10 personas. Este experimento fue desarrollado en el laboratorio de robótica sin un control especial de la iluminación para la captura del video.

Debido a que se considera que en el ambiente del robot se encontrarán tanto personas conocidas como desconocidas, se plantea estudiar el comportamiento de los tres criterios de descarte presentados en la sección 5.5. Durante los primeros 50 cuadros de cada individuo, su identidad resultará desconocida para el robot, pasadas estas imágenes se incorpora su imagen a la base de rostros, entonces se trata de reconocer su identidad en los 50 cuadros restantes, ver figura 6.1. Para el primer experimento se evalua a los desconocidos se usan los primeros 50 cuadros de cada individuo (lado izquierdo de la figura 6.1); mientras que para evaluar el comportamiento del reconocimiento se usarán los cuadros del lado derecho de la figura 6.1.

En la figura 6.2 se muestran las imágenes que conforman la base de datos para estos experimentos. Es importante destacar que una vez registradas las imágenes en la base de datos, no recibieron tratamiento adicional para posteriores ejecuciones. Además, ya que se desea analizar el comportamiento del reconocedor con presencia o ausencia de elementos estructurales (anteojos), para todos los individuos se cuenta con algunos cuadros donde aparecen con anteojos y otros donde no.

Tanto para los experimentos con video como para los posteriores, se busca evaluar únicamente el desempeño del reconocimiento (no considerando las etapas de detección y seguimiento), por tal motivo sólo se considerarán para las evaluaciones de desempeño aquellos cuadros en los que se logre una detección de rostro y ojos exitosa, aquellos

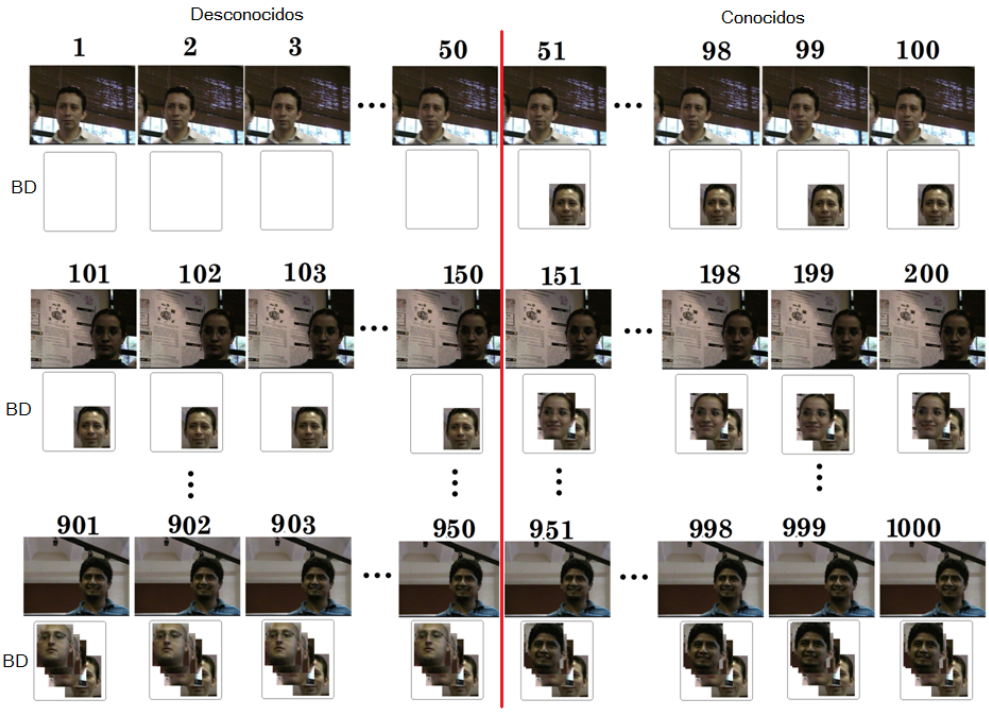


Figura 6.1: Esquema general del video usado en los experimentos 1 y 2. Para cada sujeto se tienen 100 cuadros consecutivos de aparición. Durante los primeros 50 cuadros la base de rostros no tendrá registrado al sujeto y se intentará reconocerlo, posteriormente se registrará y se continuará con el proceso de reconocimiento. Para la evaluación del experimento de desconocidos se utilizarán los cuadros del lado izquierdo de la línea roja (1-50, 101-150, ..., 901-950), mientras que para el experimento de personas conocidas se utilizarán los cuadros del lado derecho (51-100, 151-200 ,..., 951-1000).

cuadros que no cubran con estas características serán despreciados y no se contabilizan en el recuerdo del sistema. Para las pruebas de este sistema se considera la precisión y recuerdo como:

$$recuerdo = \frac{a}{a+b} \tag{6.1}$$

$$precisión = \frac{a}{a+c} \tag{6.2}$$

donde a representa el número de instancias correctamente reconocidas, b el número de instancias incorrectamente rechazadas y c es el número de incorrectamente reconocidos.

Figura 6.2: Imágenes almacenadas en la base de datos del robot.

6.1.1. Experimento 1: Desconocidos

En el primer experimento se analizó el comportamiento del módulo de reconocimiento en circunstancias en donde no se conocía al sujeto que se colocaba frente a la cámara (los 50 primeros cuadros de la secuencia de video). Dado que los criterios descritos en la sección 5.5 sirven como filtro para distinguir si el vector de similitudes refleja un sujeto conocido o desconocido, se realizaron pruebas con cada uno de ellos. De igual forma se probaron diferentes umbrales para la generación del vector de umbrales, que indica el porcentaje mínimo de puntos similares entre las imágenes de la base de datos y la imagen del sujeto a identificar. En la tabla 6.1 se muestran los resultados

obtenidos para diferentes umbrales y criterios, así como los esquemas de probabilidad absoluta y relativa.

Umbral para $\vec{T}$	C1-A	C1-R	C2-A	C2-R	C3-A	C3-R
5 %	0.9277	0.8675	0.9819	0.9880	0.9699	0.9578
7 %	0.9337	0.8976	0.9819	0.9880	0.9759	0.9759
9 %	0.9759	0.9277	0.9880	**0.9940**	0.9880	0.9819
11 %	0.9880	0.9398	0.9880	**0.9940**	0.9880	0.9819
13 %	0.9819	0.9578	0.9880	**0.9940**	0.9880	0.9880
15 %	0.9880	0.9699	**0.9940**	**0.9940**	**0.9940**	**0.9940**
20 %	**0.9940**	**0.9940**	**0.9940**	**0.9940**	**0.9940**	**0.9940**

Tabla 6.1: Resultados de precisión para personas desconocidas, con diferentes umbrales donde *C1, C2* and *C3* corresponden a los tres diferentes criterios de descarte de vectores de similaridad, A y R son los esquemas para obtener la probabilidad *absoluta* y *relativa*.

En los resultados obtenidos en la tabla 6.1 se puede apreciar que considerando que en varios casos se logra obtener porcentajes de precisión de hasta 99.40 %, esto significa que de 100 imágenes de sujetos desconocidos al sistema, aproximadamente a 1 de ellos lo reconocerá incorrectamente como algún sujeto de la base de datos. Los resultados para este experimento varían entre un 86.75 % y un 99.4 % de precisión para el mejor de los casos.

6.1.2. Experimento 2: Conocidos

En este experimento se evaluó el desempeño del módulo de reconocimiento para sujetos conocidos. Una vez transcurridos los 50 cuadros donde el sistema no conocía al sujeto, se registró a esa persona en el sistema para su reconocimiento. De manera similar se realizaron pruebas con diferentes umbrales, criterios y esquemas de probabilidades. Los resultados de precisión se muestran en la tabla 6.2, mientras la tabla 6.3 muestra los resultados de recuerdo.

Umbral	C1-A	C1-R	C2-A	C2-R	C3-A	C3-R
5 %	0.9665	0.9825	0.9741	0.9940	0.9776	0.9854
7 %	0.9690	0.9858	0.9783	0.9937	0.9813	0.9896
9 %	0.9802	0.9948	0.9817	**1.0000**	0.9844	0.9944
11 %	0.9890	1.0000	0.9934	**1.0000**	0.9943	0.9939
13 %	0.9882	0.9939	0.9929	**1.0000**	0.9939	0.9935
15 %	0.9876	0.9935	0.9925	**1.0000**	0.9936	0.9932
20 %	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**

Tabla 6.2: Precisión para personas conocidas, con diferentes umbrales donde *C1, C2* y *C3* corresponden a los tres diferentes criterios aplicables, y A y R son los esquemas para obtener la probabilidad *absoluta* y *relativa*.

Umbral	C1-A	C1-R	C2-A	C2-R	C3-A	C3-R
5 %	**0.5732**	0.5504	0.4631	0.4024	0.5369	0.4975
7 %	0.5421	0.5098	0.4423	0.3854	0.5160	0.4670
9 %	0.4865	0.4634	0.3946	0.3528	0.4632	0.4317
11 %	0.4401	0.4185	0.3659	0.3260	0.4268	0.4000
13 %	0.4083	0.3951	0.3415	0.3041	0.3976	0.3732
15 %	0.3888	0.3756	0.3244	0.2920	0.3780	0.3585
20 %	0.3358	0.3236	0.2847	0.2603	0.3309	0.3187

Tabla 6.3: Recuerdo para personas conocidas, con diferentes umbrales donde *C1, C2* y *C3* corresponden a los tres diferentes criterios aplicables, y A y R son los esquemas para obtener la probabilidad *absoluta* y *relativa*.

En los resultados se puede observar que al ser aplicado el criterio 2 con un umbral de 11 % de puntos SIFT similares y un esquema de probabilidad relativa se alcanza una precisión del 100 % con un recuerdo del 41.85 %, que mejora los resultados presentados por Apostoloff y Zisserman (2007), donde se obtiene en la mejor configuración del sistema un $97 \pm 2\,\%$ de precisión con un 20 % de recuerdo; sin embargo no es posible hacer una comparación directa ya que las condiciones e imágenes no son las mismas.

Umbral	C1-A	C1-R	C2-A	C2-R	C3-A	C3-R
5 %	**0.8499**	0.8491	0.7980	0.7682	0.8398	0.8239
7 %	0.8372	0.8307	0.7874	0.7553	0.8314	0.8086
9 %	0.8148	0.8092	0.7566	0.7316	0.8036	0.7888
11 %	0.7916	0.7825	0.7396	0.7075	0.7855	0.7664
13 %	0.7696	0.7627	0.7187	0.6861	0.7645	0.7456
15 %	0.7550	0.7476	0.7030	0.6734	0.7495	0.7335
20 %	0.7165	0.7052	0.6655	0.6377	0.7120	0.7005

Tabla 6.4: F-measure para personas conocidas, con diferentes umbrales donde *C1, C2* y *C3* corresponden a los tres diferentes criterios aplicables, y A y R son los esquemas para obtener la probabilidad *absoluta* y *relativa*.

Otra forma de presentar los resultados es mostrada en la tabla 6.4 donde se calcula el *F-measure* que es una medida que busca un representar la relación entre la precisión y el recuerdo de un sistema. Como puede verse en dicha tabla, el mejor compromiso entre precisión y recuerdo se obtiene al combinar un umbral de 5 % para el vector de similitud, un esquema de probabilidad absoluta y utilizando únicamente el criterio 1 de descarte para vectores de similitud con un porcentaje de 84.99 %.

6.1.3. Análisis

Como era de esperarse, al incrementar el umbral para la generación del vector de umbrales se incrementa la precisión pero se disminuye el recuerdo. Sin embargo, los valores de precisión tanto para personas conocidas como desconocidas muestran ser competitivos con los presentados en el estado del arte con nuestro sistema. Para personas desconocidas los rangos de precisión varian entre 86.75 % y el 99.4 %; mientras que para personas conocidas se consigue una precisión 100 % con un recuerdo del 41.85 % hasta 96.65 % con recuerdo del 57.32 %, dependiendo de la selección de criterio, umbral y esquema de probabilidad.

Considerando que la principal aplicación del reconocimiento es la interacción humano-robot, no es crítico que el sistema reconozca al sujeto en cada cuadro. El tiempo invertido en la detección, seguimiento y reconocimiento por cuadro consume en promedio 1.2 segundos, y en promedio se requieren 2.5 cuadros para tomar una decisión sobre la identidad de la persona, lo que significa que se puede realizar el reconocimiento de una persona en aproximadamente 3 segundos. Tanto los experimentos para sujetos conocidos como desconocidos fueron ejecutados en un equipo Pentium D a 2.8 Ghz con 1 Gb de memoria RAM. Algunos resultados de la ejecución del sistema se muestran en la figura 6.3.

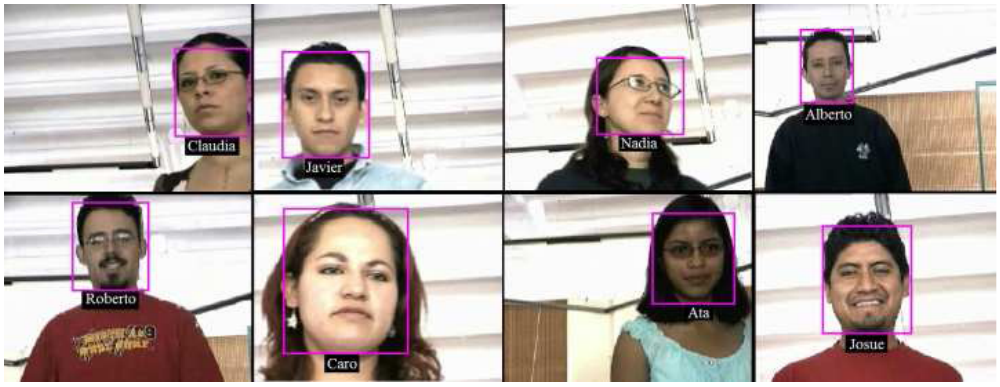


Figura 6.3: Algunos resultados del reconocimiento en video tomado de la cámara del robot. Como puede observarse en las imágenes se tienen diferentes condiciones de iluminación, escala, expresión, pose. Para todos los sujetos se probó con imágenes con y sin anteojos; mientras que en la base de rostros se cuenta con únicamente tres sujetos con anteojos.

6.2. Experimentos de escalabilidad

En estos experimentos se busca analizar el comportamiento del módulo de reconocimiento de rostros en imágenes estáticas al aumentar el número de personas en la base de datos de rostros.

En Grabner et al. (2007) utilizan la base de datos de AT&T, que está formada por 40 sujetos con 10 imágenes de 92×112 de cada individuo con pequeñas variaciones en el ángulo de rotación del rostro, así como en las expresiones. Para sus experimentos se dividió la base de datos tomando un 70 % para entrenamiento y 30 % para pruebas. Para probar el desempeño del sistema al aumentar el número de sujetos en la base de datos, se seleccionaron de forma aleatoria 20 imágenes del 30 % de pruebas, para cada vez que se incrementó la base de rostros (de 1 a 40). Los resultados obtenidos para dichos experimentos muestran tasas de reconocimiento de entre el 98.1 % y el 88.6 % aproximadamente, al incrementar de 1 a 40 personas su base de datos.

De manera similar, en esta tesis se busca evaluar la escalabilidad del sistema al aumentar el número de sujetos conocidos en la base de datos. Sin embargo, se optó por utilizar una base de datos distinta que brinde escenarios más adversos, ya que cuentan con diferentes iluminaciones y no se tiene únicamente la imagen del rostro como las usadas por Grabner et al. (2007). Por tal motivo se seleccionó la base de datos de Yale Extendida que cuenta con 16128 imágenes de 28 sujetos bajo 9 diferentes poses y 64 condiciones de iluminación diferentes. Para los experimentos se seleccionó un conjunto de 21 imágenes por cada individuo, 20 imágenes para pruebas y 1 imagen para la base de rostros. En la figura 6.4 se muestra un ejemplo de algunas de las imágenes utilizadas en la prueba, en el centro de la figura se muestra la imagen a utilizar para el entrenamiento de la base de datos.

6.2.1. Experimento 3: Incrementar individuos en la base de rostros

Como un primer experimento se plantea analizar el comportamiento del reconocedor al aumentar el número de personas conocidas en la base de datos. Debido a que el sistema de reconocimiento de rostros está planteado para trabajar con video, se usó como entrada el conjunto de 20 imágenes por persona de manera secuencial.

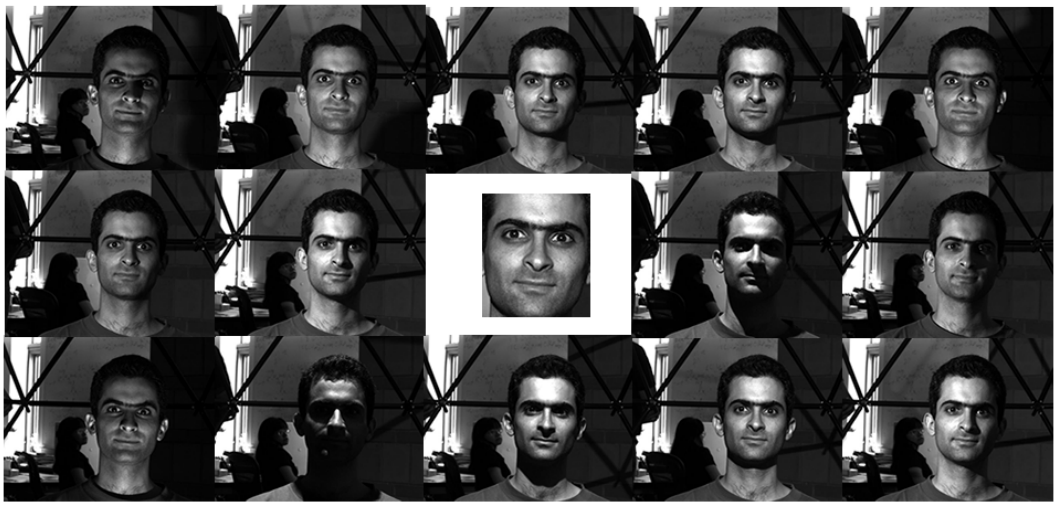


Figura 6.4: Ejemplo de imágenes de la base de datos de Yale Extendida B utilizadas para pruebas. En el centro de la imagen se muestra la imagen para entrenamiento en los experimentos Georghiades et al. (2001).

Inicialmente, el sistema comenzará con el sujeto 1 en la base de datos de sujetos conocidos, y se evaluará su desempeño al compararse con las 20 imágenes de prueba del sujeto 1. Posteriormente, se añadirá el sujeto 2 a la base de datos y se realizará la prueba con las 20 imágenes de cada uno de los sujetos conocidos por el sistema, y así sucesivamente hasta alcanzar a los 28 sujetos conocidos. Este esquema puede ser representado por la figura 6.5.

Es importante mencionar que como se está evaluando el desempeño del reconocimiento de rostros, únicamente serán consideradas aquellas imágenes en donde se logró una detección de rostro y ojos exitosa. Como puede observarse en las gráficas de la figura 6.6 y 6.7, la precisión al aumentarse la cantidad de sujetos en la base de datos se mantiene estable con reconocimientos correctos superiores al 94.5 %; sin embargo, el recuerdo se ve afectado ya que consume una mayor cantidad de cuadros el deliberar una identidad para el sujeto.

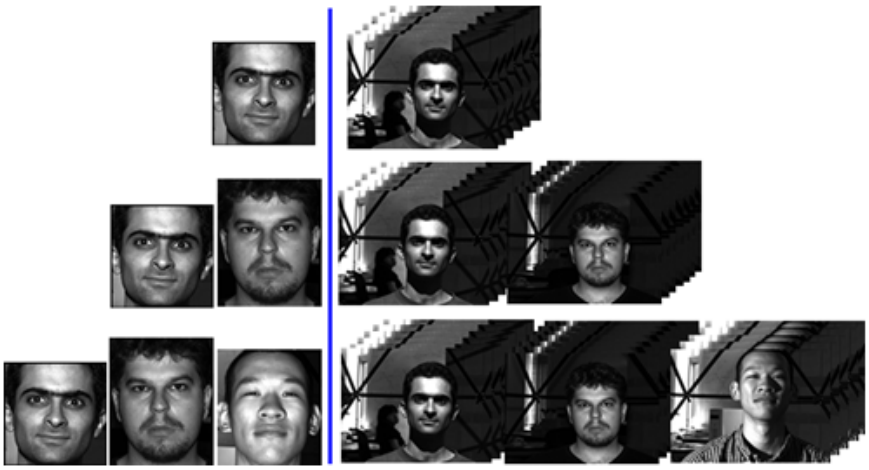


Figura 6.5: Representación del experimento 3. Inicialmente la base de rostros conocidos incluye únicamente un individuo, se toma una muestra de 20 imágenes de la base de rostros Yale B y se realiza el reconocimiento. Posteriormente, se añade a la base de rostros una imagen del sujeto 2, entonces se realizan pruebas de reconocimiento con las 20 imágenes del sujeto 1 y 20 imágenes del sujeto 2. Este proceso se repite para los 28 individuos que forman la base de rostros de Yale B.

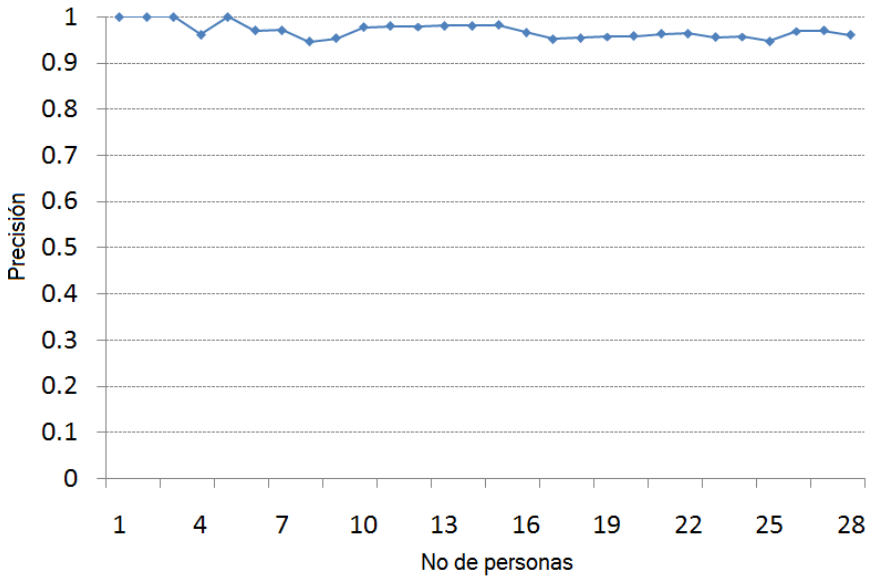


Figura 6.6: Resultados de precisión del reconocimiento para la base de rostros Yale Extendida al incorporar sujetos en la base de rostros y evaluarlo únicamente con sujetos conocidos.

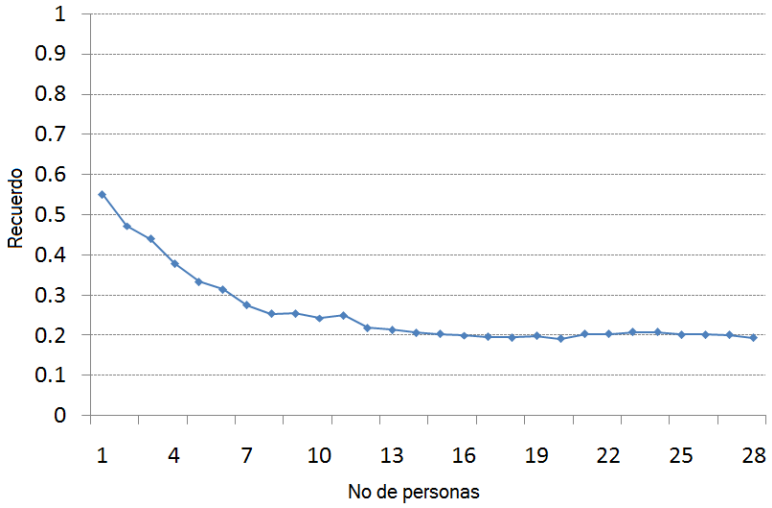


Figura 6.7: Resultados de recuerdo del reconocimiento para la base de rostros Yale Extendida al incorporar sujetos en la base de rostros y evaluarlo únicamente con sujetos conocidos.

6.2.2. Experimento 4: Comportamiento al incrementar el número de cuadros

En este experimento se busca evaluar el desempeño del sistema de reconocimiento al aumentar la evidencia de imágenes por persona. De igual forma se analiza el comportamiento del reconocedor cuando se presentan sujetos desconocidos al sistema. Para todas las pruebas se tienen 10 sujetos conocidos en la base de datos (etiquetados del sujeto 1 al 10). Para pruebas se tienen 20 imágenes por cada uno de los 28 individuos de la base de rostros. Se presentarán *n* imágenes aleatorias de cada una de las 28 personas simulando un video de $n \times 28$ cuadros para cada prueba. Se realizaron experimentos con valores de $n = 1, 2, 10, 15, 20$. El objetivo de esta prueba es analizar si a mayor número de apariciones de la persona en el video, la precisión del sistema aumenta. El experimento fue repetido 10 veces para cada valor de n para obtener resultados estadísticamente significativos. En las figuras 6.8 y 6.9 se muestran la precisión y el recuerdo promedio de las 10 ejecuciones para cada uno de los subconjuntos de n imágenes. En todos los casos los 10 primeros individuos son los registrados en la base de datos del sistema, mientras los siguientes 18 son desconocidos.

En las gráficas 6.8 y 6.9, los 10 primeros valores corresponden a los sujetos conocidos por el sistema, mientras las personas 11 a 28 son desconocidas en todo momento al sistema. Como puede observase, al incrementar la cantidad de imágenes por individuo la precisión y recuerdo del sistema aumentaron, hasta un máximo de 20 imágenes, donde se mostró un comportamiento similar al de 15 imágenes por individuo. Los peores resultados se obtienen al efectuar el reconocimiento con una sola imagen por cada sujeto, con rangos entre un 20 % de precisión con 20 % de recuerdo, hasta una precisión del 60 % con 9.5 % de recuerdo. Mientras que al aumentar el número de imágenes en 15 y 20, los mejores valores de precisión alcanzan hasta un 100 % con recuerdo de 45.5 % y 47.3 %.

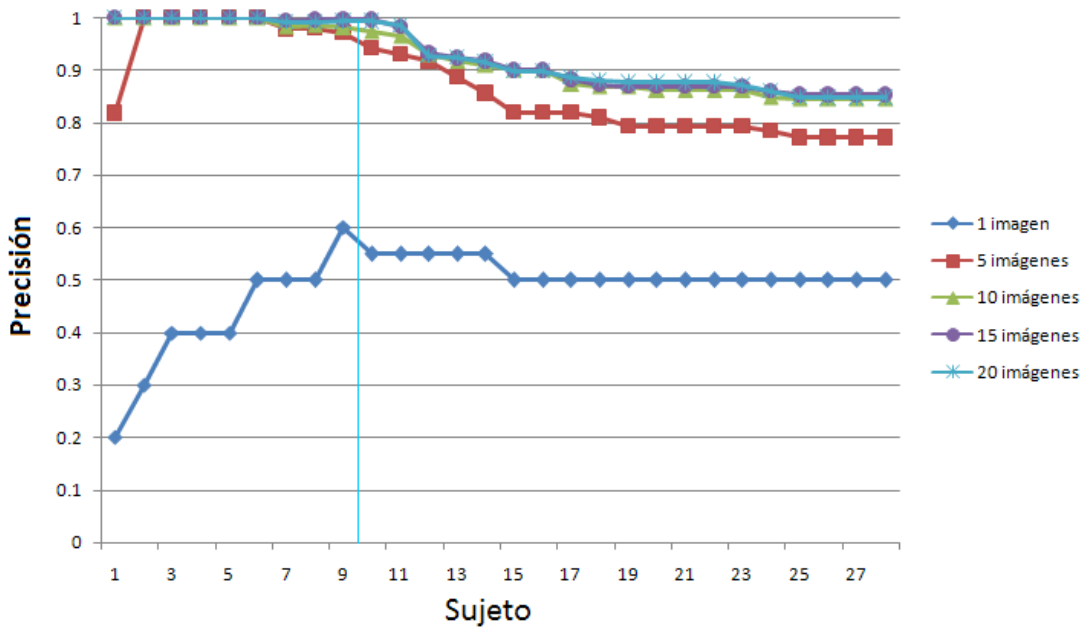


Figura 6.8: Resultados de precisión del reconocimiento de nuestro sistema al realizar el experimento con diferente número de imágenes con sujetos conocidos y desconocidos. Las 10 primeras personas son las conocidas, mientras los sujetos 11 al 28 son desconocidos.

En la figura 6.10 se presenta una gráfica de los resultados obtenidos por Grabner

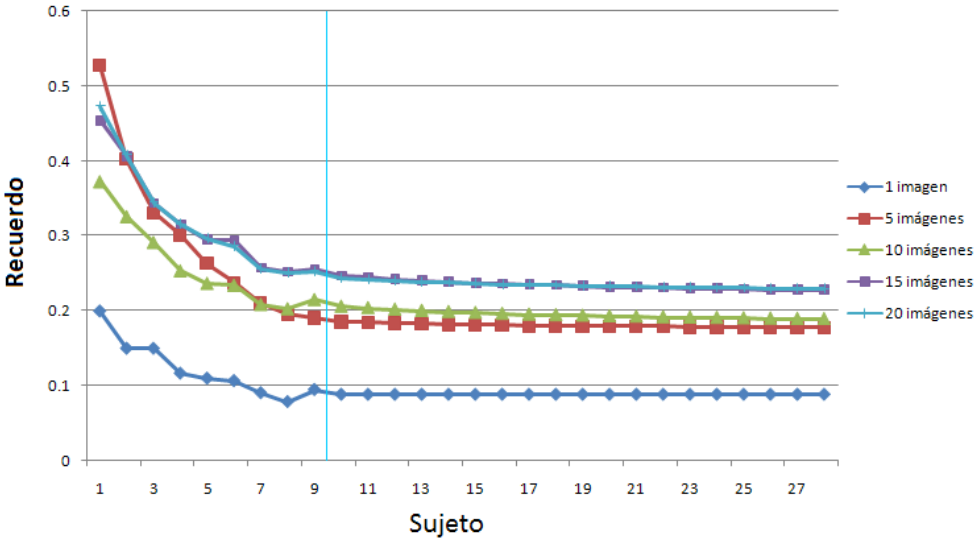


Figura 6.9: Resultados de recuerdo del reconocimiento de nuestro sistema al realizar el experimento con diferente número de imágenes con sujetos conocidos y desconocidos. Las 10 primeras personas son las conocidas, mientras los sujetos 11 al 28 son desconocidos.

et al. (2007), como puede observarse en dicho trabajo se alcanzan precisiones entre un 98 % y 88 % para 40 individuos el comportamiento incremental de 1 a 40 individuos en la base de rostros; mientras que en este trabajo se alcanzan porcentajes similares oscilando entre 100 y 85.4 % de precisión para 28 individuos. Hay que recordar que las condiciones de prueba de las imágenes usadas en la evaluación de esta tesis representan un mayor reto que las utilizada en Grabner et al. (2007), ya que manejan una amplia variedad de iluminaciones y no se cuenta con sólo la imagen del rostro. Los mejores resultados en nuestro sistema se obtienen al evaluarse 10 sujetos en la base de rostros e intentar reconocer a los 28 individuos (10 conocidos y 18 desconocidos); sin embargo los valores de precisión varian entre un 85.4 y 100 % de precisión para los 28 individuos (Ver gráfica 6.8)

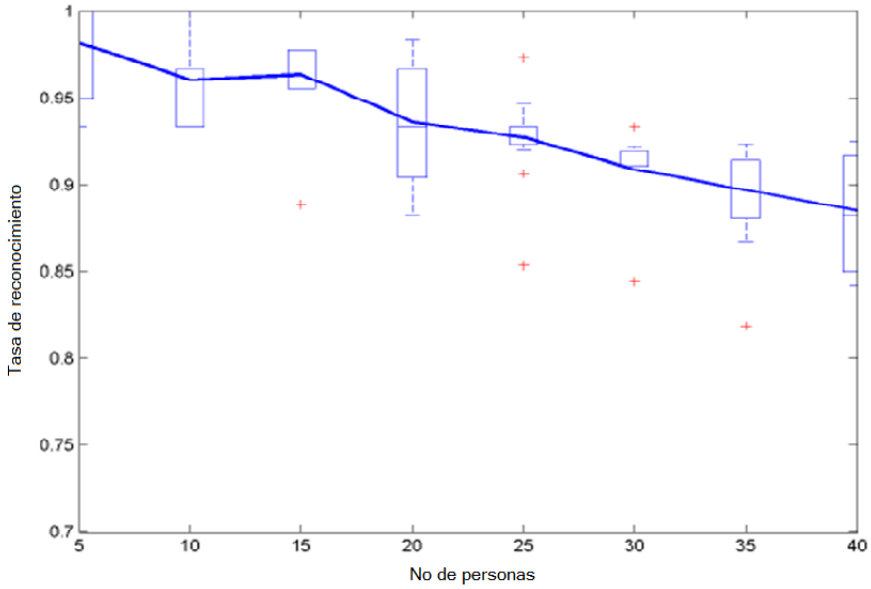


Figura 6.10: Resultados de precisión al aumentar el número de sujetos (Tomado de Grabner et al. (2007)).

6.3. Experimentos con el robot móvil Markovito

Markovito es un robot de servicio basado en una platafora PeopleBot, está dotado de 2 anillos de sonares, una cámara Pan-Tilt Canon VCC5, 2 motores con encoders, un láser, 2 sensores infrarojos, un micrófono bidireccional, así como una computadora integrada en la a,rquitectura del robot. La parte del procesamiento de las imágenes es realizada en una computadora portátil core duo a 2.0 GHz con 2 Gb de memoria RAM.

Para evaluar el desempeño del sistema de reconocimiento en móvimiento se plantea una prueba en donde tanto Markovito como cinco personas (todas ellas previamente registradas en la base de rostros) caminarán en el ambiente. El escenario de pruebas fue el laboratorio de robótica, y la ruta seguida por Markovito y los participantes se encuentran marcadas en la figura 6.11. Para evaluarlo en una situación de iluminación adversa, se seleccionó realizar la grabación a las 6:30 pm cuando el sol se encuentra en su ocaso y las lámparas del laboratorio fueron encendidas (iluminación no estructurada), contando así con varias niveles de iluminación. Las imágenes fueron tomadas con

una resolución de 320×240 pixeles y el video fue editado para considerar en su mayoría imágenes con presencia de personas. Los individuos en esta prueba se encontraron fijos al momento de la grabación, aunque en algunos cuadros se les solicitó caminar lentamente hacia el robot (aún en movimiento) cuya velocidad aproximada era 0.8 metros por segundo, a una distancia entre 1 y 5 metros. En la figura 6.12 se muestra la base de rostros usada para la prueba, como puede observarse el sistema no esta limitado a su funcionamiento con un sólo tamaño de imágenes.

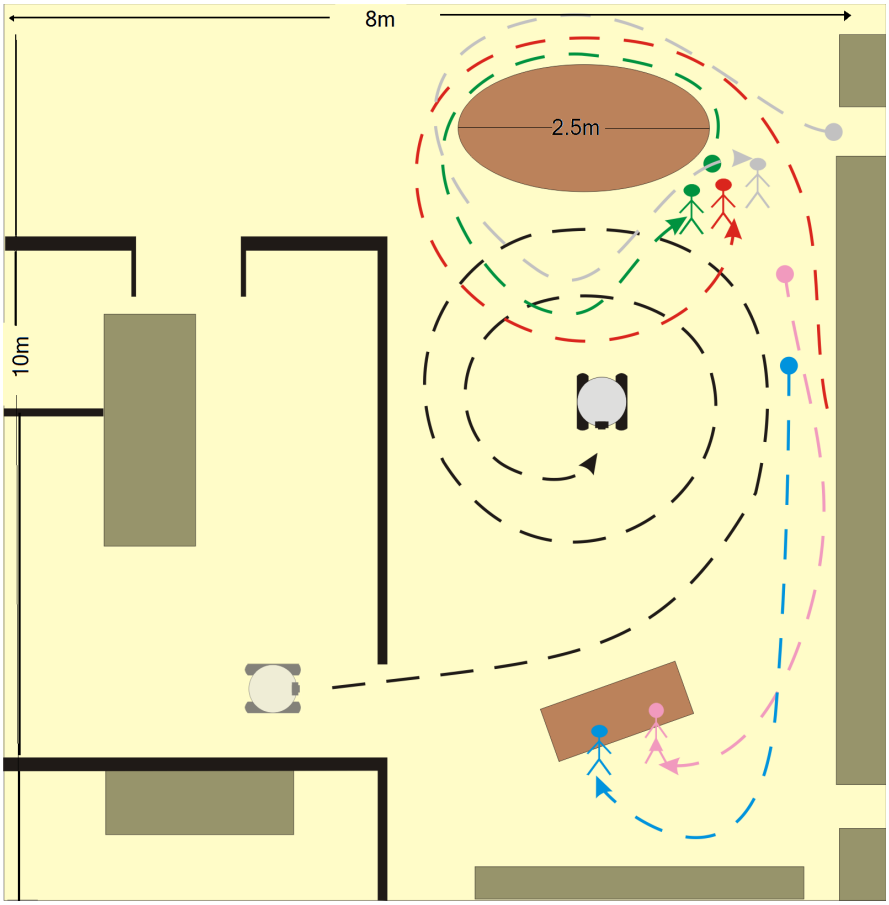


Figura 6.11: Diagrama general del recorrido realizado por Markovito y los cinco participantes en el experimento. La línea punteada negra indica la trayectoria seguida por Markovito (la imágen degradada señala la posición inicial del robot). Cada una de las líneas punteadas de colores muestran las trayectorias seguidas por los participates, los círculos indican las posiciones iniciales.

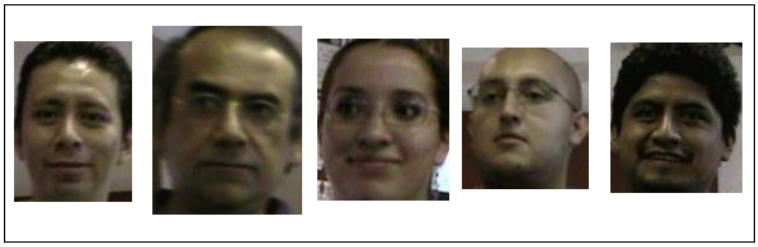


Figura 6.12: Base de rostros utilizada por Markovito. El tamaño de las imágenes es de 100×120 pixeles.

Posteriormente se utilizó el video como entrada al sistema considerando para la evaluación de precisión y recuerdo, únicamente cuadros en los cuales el rostro y los ojos fueran localizados. En la figura 6.13 se muestran ejemplos de las imágenes capturadas por Markovito para la prueba. En la prueba se utilizó un video de 250 cuadros con 50 imágenes consecutivas correspondientes a cada sujeto. Los resultados obtenidos por el sistema se muestran en la tabla 6.5. Dichos resultados presentan un avance importante de acuerdo a los trabajos presentados en el estado del arte, ya que en contraste con los experimentados realizados por Grabner et al. (2007) en el robot móvil Flea, ellos no reportan resultados con su robot en movimiento, son únicamente los participantes los que se mueven alrededor del robot; sin embargo, no se reporta claramente la precisión alcanzada en dicho experimento.

	No cuadros
Cuadros sin detecciones de rostros u ojos	25
Cuadros con reconocimientos correctos	86
Cuadros con reconocimientos incorrectos	11
Cuadros sin suficiente evidencia p/reconocimiento	128
Precisión	**87.76 %**
Recuerdo	**40.17 %**

Tabla 6.5: Resultados de reconocimiento de rostros en Markovito.

Figura 6.13: Ejemplo de imágenes de prueba tomadas por Markovito en movimiento, bajo diferentes condiciones de pose e iluminación.

6.4. Conclusiones

Se plantearon tres conjuntos de experimentos para evaluar el comportamiento del módulo de reconocimiento de rostros: reconocimiento en video para personas conocidas y desconocidas, reconocimiento al incrementar el número de individuos y cuadros, y reconocimiento en el robot móvil Markovito.

El primer conjunto de experimentos consistió en evaluar el desempeño del sistema cuando se presentan sujetos no registrados en la base de rostros y cuando si se encontraban registrados. Para estas pruebas se utilizó la cámara del robot Markovito, en el laboratorio (un ambiente interior sin control de la iluminación adicional), con un conjunto de 10 individuos. Se consideraron variar los porcentajes del vector de umbrales con valores de 5,7,9,11,13,15 y 20 %, los esquemas de mapeo de probabilidad absoluta y probabilidad relativa para la generación del vector de probabilidades, así como los tres criterios de descarte para los vectores de similitud. Para estas configuraciones en los experimentos, se alcanzó hasta un 99.4 % de precisión para el correcto etiquetamiento de DESCONOCIDO para sujetos no registrados en la base de rostros; mientras que para sujetos registrados en la base de rostros se obtiene hasta un 100 % de reconocimientos correctos. Es importante considerar para la elección de los parámetros como criterio, umbral y esquema de probabilidad, un compromiso entre la precisión y el recuerdo del sistema. El tiempo promedio por imagen que consume todo el proceso es de 1.2 segundo, considerando condiciones de iluminación estándar en el laboratorio, generada por fuentes no estructuradas como lo son luz entrante por las ventanas y la luz producida por algunas lámparas de diferentes luminicencia; el reconocimiento requiere en promedio 2.5 imágenes, con lo que el reconocimiento de un sujeto llevaría aproximadamente 3 segundos.

El segundo conjunto de experimentos evalúa el desempeño del sistema al variar el número de cuadros en el video en el que se presenta a un individuo, así como su

comportamiento al aumentar el número de sujetos en la base de rostros. Para ello se utilizó la base de datos de Yale Extendida que representa un desafio por considerar situaciones de iluminación adversas. Se consideró el número de sujetos de la base de datos para realizar pruebas sobre el comportamiento del sistema al incrementar la cantidad de personas en la base de datos. De los experimentos realizados en esta etapa se puede concluir que incrementar el número de imágenes de prueba mejora el desempeño del reconocedor; sin embargo, este comportamiento tiene un umbral donde los resultados permanecen similares. El mejor escenario para el sistema en estas pruebas fue al evaluar el reconocedor únicamente con sujetos previamente registrados en la base de datos, cuyo comportamiento en cuanto a precisión se mantiene aproximadamente constante pese a incrementar la cantidad de sujetos hasta 28 con un porcentaje promedio de precisión de 96.96 %.

El tercer experimento consiste en evaluar el desempeño del sistema con el robot Markovito, en un ambiente donde tanto Markovito como los cinco participantes se encontraban en movimiento. Pruebas realizadas con el robot en condiciones adversas de iluminación, con variaciones en la pose y las expresiones faciales, mostraron un desempeño del 87.76 % de precisión con 40.17 % de recuerdo lo cual resulta adecuado para la aplicación del robot mensajero, ya que presenta resultados de recuerdo superiores los mostrados por el estado del arte y con ello se puede mejorar los tiempos de respuesta.

Capítulo 7

Conclusiones y Trabajo Futuro

La necesidad de una interacción humano-robot más natural motivo en esta tesis a desarrollar un método de reconocimiento de rostros capaz de trabajar con video en condiciones de iluminación no uniformes, con rostros en diversas escalas y rotaciones de perfil de $\pm 30°$. En comparación con otros trabajos similares, los cuales trabajan con episodios de televisión Apostoloff y Zisserman (2007), se logran niveles de precisión significativos con un recuerdo superior en situaciones de evaluación similares. De igual forma se presentan resultados competitivos al comparar el desempeño de nuestro sistema con el utilizado en el robot móvil Flea, de Grabner et al. (2007), al medirse la precisión al aumentar el número de sujetos en la base de rostros.

7.1. Conclusiones

En esta tesis se presentó un sistema para el reconocimiento de rostros para determinar la identidad de un sujeto de entre una base de rostros conocidos, el sistema de reconocimiento formado por tres módulos: detección, seguimiento y reconocimiento. El sistema mostró ser robusto a trabajar con imágenes de rostros sin control de iluminación, lo cual es una situación común en robótica móvil. Además de ser capaz de aprender nuevos individuos en tiempo de ejecución con tan sólo una imagen por per-

sona. Se mostró que el sistema es capaz de realizar el reconocimiento con ausencia o presencia de elementos estructurales como lo son anteojos, ya que en las pruebas realizadas se contemplaba que no todos los sujetos contaran con anteojos en la base de rostros y en pruebas se presentaron situaciones con y sin anteojos con porcentajes de precisión superiores al 96.65 % (ver algunos ejemplos de reconocimiento con anteojos y sin anteojos en la figura 6.3). Para la etapa de detección y seguimiento se desarrolló un algoritmo de seguimiento de rostros basando en el detector de objetos de Viola y Jones (2001a) el cual cuenta con una complejidad del orden $O(n^2)$.

El esquema de reconocimiento de rostros constó de cuatro etapas principales: (i) preprocesamiento de la imagen, (ii) generación de regiones de interés y extracción de puntos característicos SIFT, (iii) generación de vector de similitud y aplicación de criterios de descarte en imágenes con sujetos desconocidos, y (iv) integración de información de cuadros anteriores con la información de la imagen del cuadro actual.

En la etapa de preprocesamiento se implementó un método de mejora para las imágenes que consiste en una ecualización del histograma seguida de una compensación de iluminación, con lo cual se aumentar el número de características invariantes del rostro. La generación de regiones de interés se basó en la localización de los ojos en la imagen y el tamaño de dichas regiones retornadas por el detector, con lo que se generaron las regiones: ojo izquierdo, ojo derecho y nariz-boca. Además con el proceso de detección de ojos se logró descartar imágenes mal detectadas como rostros por el seguidor. En la tercera etapa se genera el vector de similitud, que representa la cantidad de puntos similares entre la imagen nueva y cada una de las almacenadas en la base de rostros. En esta etapa el uso de los criterios de discriminación de vectores de similitud logró descartar del reconocimiento a la mayor parte de cuadros en los cuales se encontraba un sujeto no registrado en la base de rostros; esto permite que sólo las imágenes con suficiente evidencia sean evaluadas para su reconocimiento. Finalmente, en la cuarta etapa se utiliza la regla de Bayes para reforzar las probabilidades de estar

observando a una persona con base a la evidencia de la imagen actual y la información generada en imágenes anteriores, este enfoque mostró buenos resultados tanto en tiempo de ejecución con aproximadamente tres segundos para el reconocimiento de un rostros, como en precisión alcanzando un 96.65 % de precisión. Debido al diseño del sistema de reconocimiento, se mostró que sólo es requerida una imagen por persona para su registro en la base de rostros conocidos. Con lo cual se brinda al sistema la flexibilidad para poder incluir nuevas personas en tiempo de ejecución sin un costoso entrenamiento.

Se realizaron tres conjuntos de pruebas: (i) reconocimiento de rostros en video (para sujeto desconocidos y sujetos conocidos), (ii) reconocimiento al incrementar el número de personas en la base de rostros, así como al variar el número de cuadros de cada individuo en el video, y finalmente (iii) reconocimiento en el robot de servicio Markovito.

Para el primer conjunto de experimentos buscó evaluar dos objetivos: (i) la precisión del sistema cuando se presentaba una persona no registrada en la base de rostros y se asignaba la etiqueta DESCONOCIDO, y (ii) la precisión y recuerdo del sistema cuando se presentaba una persona conocida para su reconocimiento. Se seleccionaron diferentes configuraciones de los parámetros como el umbral de puntos característicos, el criterio de discriminación de vectores de similitud y esquemas de probabilidad; mostrando que se puede alcanzar una precisión del 99.4 % para los desconocidos y un 96.65 % de precisión con 57.32 % de recuerdo para el reconocimiento de personas registradas en la base de rostros. Sin embargo, como era de esperarse colocar parámetros más estrictos generó mejores resultados en la precisión en aproximadamente 3.35 %, pero afectarón los porcentajes de recuerdo en hasta un 31.29 % menos. El tiempo promedio requerido para el reconocimiento es de tres segundos. Los resultados alcanzados muestran ser competitivos con los presentados por Apostoloff y Zisserman (2007) que funciona con episodios de televisión y alcanzan porcentajes de precisión y recuerdo del 97 ± 2 % y 20 %, respectivamente.

El segundo conjunto de experimentos evaluó: (i) el desempeño del sistema al incrementar el número de sujetos en la base de datos y (ii) como afecta la precisión el aumentar o disminuir el número de cuadros consecutivos en los cuales aparece una persona. Para estas pruebas se utilizó la base de rostros de Yale Extendida que presenta condiciones de iluminación variables, reportando un comportamiento estable con precisiones superiores al 94.5 % para cuando se aumentaba el número de sujetos en la base de rostros. Sin embargo, este porcentaje se reduce al realizar las comparaciones con sujetos conocidos y desconocidos hasta un 85.4 %. Por otro lado, al evaluar la precisión al aumentar el número de cuadros por sujeto se puede observar que a mayor número de cuadros mejora la precisión; sin embargo, este comportamiento tiende a no aumentar después de 15 imágenes por sujeto.

Finalmente, se evaluó el desempeño del sistema en el robot de servicio Markovito, planteando una prueba en la cual tanto Markovito como cinco participantes se encontraban en movimiento en el ambiente. Los resultados reportados son de un 87.76 % de precisión con un recuerdo del 40.17 % de recuerdo.

7.2. Aportaciones

Las principales aportaciones de esta tesis son:

- Desarrollo de un algoritmo sencillo de seguimiento de rostros con base en AdaBoost.
- Un método de correspondencia entre los puntos SIFT de las imágenes registradas en la base de rostros y la imagen a reconocer, basado en la localización automática de las regiones de los ojos, nariz y boca. Esto incluye el preprocesamiento de las imágenes con la ecualización del histograma y la compensación de la iluminación, lo que favorece la extracción de puntos SIFT.

- Un criterio efectivo para descartar imágenes de sujetos no registrados en la base de rostros basado en umbrales, con lo cual se busca reducir el tiempo de ejecución.

- La aplicación de un método Bayesiano para el reconocimiento de rostros en un video, el cual hace uso de la información generada en imágenes anteriores y de la evidencia generada en la imagen a reconocer.

En su conjunto estas aportaciones permiten desarrollar un sistema de reconocimiento de rostros robusto y eficiente para su aplicación en robots de servicio.

7.3. Trabajo futuro

Algunas extensiones que se pueden realizar a este trabajo se enlistan a continuación:

- Extender el proceso de seguimiento para más personas en la imagen, de esta forma poder realizar múltiples reconocimientos al mismo tiempo.

- Utilizar de características invariantes SIFT en el proceso de seguimiento para hacerlo más robusto, y en un momento dado reemplazar el actual. Con esto podría abarcarse la detección de rostros con mayor ángulo de rotación.

- Realizar más pruebas con el robot Markovito como por ejemplo:

  - Probar el algoritmo en en ambientes con obstáculos para analizar su comportamiento en conjunto con otros módulos del robot como el de planificación y evación de obstáculos
  - Aumentar el tiempo de la prueba para analizar el comportamiento del sistema durante largos periodos de uso
  - Analizar el aprendizaje de nuevos individuos en tiempo de ejecución durante su recorrido.

- Desarrollar un control automático para la cámara de Markovito

  para seguir y enfocar al rostro que se quiere reconocer.

Referencias

N. Apostoloff y A. Zisserman. Who are you? - real time person identification. En *Proceedings of the British Machine Vision Conference*, páginas 509–518. BMVA, 2007.

H. Avilés, E. Corona, A. Ramírez, B. Vargas, J. Sánchez, L.E. Sucar, y E. Morales. A service robot named markovito. *IEEE 4th Latin American Robotic Symposium*, 2007.

G. W. Awcock y R. Thomas. *Applied Image Processing*. 1996.

E. Bailly-Baillire, S. Bengio, G. Bimbot, M. Hamouz, J. Kittler, J. Marithoz, J. Matas, K. Messer, V. Popovici, F. Pore, B. Ruiz, y J.P. Thiran. The banca database and evalluation protocol. En *International Conference on Audio and Video Based Biometric Person Authentication*, 2003.

J. Bartlett y J. Searcy. Inversion and configuration of faces. 25:281–316, Cognitive Psychology.

P.Ñ. Belhumeur, J. Hespanha, y D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:711–720, 1997.

M. Bicego, A. Lagorio, E. Grosso, y M. Tistarelli. On the use of sift features for face authentication. En *Conference on Computer Vision and Pattern Recognition Workshop*, página 35, 2006.

W. Burgard, A.B. Cremers, Dieter Fox, D. Hähnel, G. Lakemeyer, D. Schulz, W. Steiner, y Sebastian Thrun. The interactive museum tour-guide robot. En *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, 1998.

B. Cao y S. Shan. Baseline evaluations on the cas-peal-r1 face database. *Advances in Biometric Person Authentication*, páginas 370–378, 2004.

C. Cortes y V. Vapnik. Support-vector networks. *Machine Learning*, 3:273–297, 1995.

N. Duta y A.K. Jain. Learning the human face concept from black and white pictures. *Proceedings International Conference Pattern Recognition*, páginas 1365–1367, 1998.

G. J. Edwards, T. F. Cootes, y C. J. Taylor. Face recognition using active appearance models. *Proceedings of the 5th European Conference on Computer Vision*, II:581–595, 1998.

Y. Freund y R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *European Conference on Computational Learning Theory*, páginas 23–37, 1995.

W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, y D. Zhao. The cas-peal large-scale chinese face database and baseline evaluations. *IEEE Trans. on System Man, and Cybernetics (Part A)*, 38:149–161, 2008.

A.S. Georghiades, P.N. Belhumeur, y D.J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 23:643–660, 2001.

A. J. Goldstein, L. D. Harmon, y A. B. Lesk. Identification of human faces. *Proceedings of the IEEE*, 59(5):748760, 1971.

M. Grabner, H. Grabner, J. Pehserl, y P. Korica-Pehserl. Flea, do you remember me? *Lecture Notes in Computer Science*, 4843:657–666, 2007.

P.J. Grother, R.J. Micheals, y P.J. Phillips. Face recognition vendor test 2002 performance metrics. En *Audio- and Video-Based Biometric Person Authentication*, 2003.

H. T. Kam. Random decision forest. *3rd International Conference on Document Analysis and Recognition*, páginas 278–282, 1995.

G. Kim, W. Chung, K.-R. Kim, M. Kim, S. Han, y R.H. Shinn. The autonomous tour-guide robot jinny. *IEEE International Conference on Intelligent Robots and Systems*, 4:3450–3455, 2004.

B. Knight y A. Johnston. The role of movement in face recognition. *Visual Cognition*, 4:265 – 273, 1997.

T. Kohonen. Self-organization and associative memory. *Springer-Verlag New York*, 1989.

K.C. Lee, J. Ho, y D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):684–698, 2005.

R. Lienhart y J. Maydt. An extended set of haar-like features for rapid object detection. En *Proceedings IEEE International Conference on Image Processing, volume 1*, páginas 900–903, 2002.

C.J. Liu y H. Wechsler. Evolutionary pursuit and its application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(6):570–582, 2000.

D. Lowe. Object recognition from local scale-invariant. En *International Conference on Computer Vision*, páginas (2):1150–1157, 1999.

D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 20:91–110, 2004.

J. Luo, Y. Ma, E. Takikawa, S. Lao, M. Kawade, y B. L. Lu. Person-specific sift features for face recognition. En *International Conference on Acoustic, Speech and Signal Processing (2)*, páginas 593–596, 2007.

E. Osuna, R. Freund, y F. Girosi. Training support vector machines: An application to face detection. *IEEE Conference on Computer Vision and Pattern Recognition*, páginas 130–136, 1997.

A. O'Toole, D. Roark, y H. Abdi. Recognitizing moving faces. a psychological and neural synthesis. *Trends in Cognitive Science*, 6:261–266, 2002.

M. Ozuysal, P. Fua, y V. Lepetit. Fast keypoint recognition in ten lines of code. *Computer Vision and Pattern Recognition*, 2007.

P. Papageorgiou, M. Oren, y T. Poggio. A general framework for object detection. En *Proceedings of the Sixth International Conference on Computer Vision*, página 555, 1998.

P. J Phillips, H. Moon, P. Rauss, y S. A. Rizvi. The feret evaluation methodology for face recognition algorithms. *IEEE International Conference on Computer Vision and Pattern Recognition*, páginas 137–143, 1997.

J. Pineau, M. Montemerlo, M. Pollack, N. Roy, y S. Thrun. Towards robotic assistants in nursing homes: challenges and results. En *Workshop notes (WS8: Workshop on Robot as Partner: An Exploration of Social Robots), IEEE International Conference on Robots and Systems*, 2002.

G.A. Ramŕez-García. *Detección de Rostos con Aprendizaje Automático*. Tesis de maestría, INAOE, Puebla, Mexico, 2006.

H. Rowley, S. Baluja, y T. Kanade. Neural networkbased face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:23–38, 1998.

F. Samaria y A. Harter. Parameterisation of a stochastic model for human face identification. *2nd IEEE Workshop on Applications of Computer Vision*, 1994.

H. Schneiderman y T. Kanade. Probabilistic modeling of local appearance and spatial relationship for object recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, páginas 45–51, 1998.

H. Schneiderman y T. Kanade. A statistical approach to 3d object detection applied to faces and cars. *IEEE Conference on Computer Vision and Pattern Recognition*, 1: 746–751, 2000.

K. Sung y T. Poggio. Example-based learning for view-based human face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):39–51, 1998.

P. Thompson. Margaret thatcher - a new ilusion. *Perception*, páginas 483–484, 1980.

M. Turk y A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

P. Viola y M. Jones. Rapid object detection using boosted cascade simple features. En *Proceedings of the Computer Vision and Pattern Recognition*, páginas 511–518, 2001a.

P. Viola y M. Jones. Robust real-time object detection. En *International Workshop on Statistical and Computational Theories of Vision*, 2001b.

Haar wavelet. Wikipedia, http://en.wikipedia.org/wiki/HaarWavelet, 2008. Fecha de última consulta, abril 2008.

H. Wechsler, P.J. Phillips, V. Bruce, F. Fogelman Soulie, y T.S. Huang. Face recognition: From theory to applications. En *Proceeding of NATO-ASI*, 1998.

L. Wiskott, J.M. Fellous, N. Kruger, y C. Von der Malsburg. Face recognition by elastic bunch graph matching. *Pattern Analysis and Mahine Intelligence*, 19:(7):775–779, 1997.

M.H. Yang, N. Ahuja, y D. Kriegman. Mixtures of linear subspaces for face detection. *Proceedings in Fourth International Conference on Automatic Face and Gesture Recognition*, páginas 70–76, 2000a.

M.H. Yang, D.J. Kriegman, y N. Ahuja. Detecting faces in images: A survey. *Pattern Analysis and Mahine Intelligence*, 24:34–58, 2002.

M.H. Yang, D. Roth, y N. Ahuja. A snow-based face detector. *Advances in Neural Information Processing Systems 12, MIT Press*, páginas 855–861, 2000b.

R.K. Yin. Looking at upside-down faces. APA PsychNET, http://www.psycnet.org/index.cfm?fa=buy.optionToBuyid=1969-12269-001, 1969. Fecha de última consulta, Diciembre 2007.

W.Y. Zhao, R. Chellappa, P.J. Philips, y A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Survey*, páginas 399–458, 2003.