DETECCIÓN Y RECONOCIMIENTO DE ROSTROS

TESIS

Que para obtener el grado de
MAESTRÍA EN CIENCIAS EN INGENIERÍA ELÉCTRICA

presenta
Sergio Rogelio Tinoco Martínez

Félix Calderón Solorio
Director de Tesis

Universidad Michoacana de San Nicolás de Hidalgo
División de Estudios de Posgrado de la Facultad de Ingeniería Eléctrica

Agosto 2008

A mi padre Sergio Rogelio Tinoco López

Te amo papá

Resumen

En el presente trabajo se revisa el algoritmo propuesto por Moghaddam et al. en [Moghaddam95a] para la detección automática de rostros y para el proceso complementario de reconocimiento facial en [Moghaddam96]. Aunque no se proporciona una aportación nueva se pudo comprobar en una implementación propia, aplicando varias simplificaciones al desarrollo original de Moghaddam, el desempeño reportado en cuanto al reconocimiento facial se trata. Sin embargo, en cuanto a la combinación de detección y reconocimiento automáticos, el desempeño encontrado fue menor a lo indicado en las referencias citadas, debido principalmente a la dependencia tan alta que el algoritmo de detección presenta con relación al tamaño relativo entre los rostros de las imágenes de prueba utilizadas. Este bajo desempeño se traslada hacia el proceso de reconocimiento, en detrimento del desempeño global.

Las técnicas automáticas de aprendizaje visual referidas se basan en la estimación de densidades de probabilidad en espacios de alta dimensión, utilizando una descomposición de espacios característicos (*eigenspaces*). La densidad de probabilidad de los datos de entrenamiento se modela con una distribución normal/Gaussiana multivariada, la cual se emplea posteriormente para formular un estimador de máxima verosimilitud para la detección de rostros y su reconocimiento automatizado.

Con relación al proceso de reconocimiento facial, el estimador de máxima verosimilitud se emplea para calcular una medida de similitud basada en un análisis Bayesiano de diferencias de imágenes. Se modelan dos clases mutuamente exclusivas de variación entre dos imágenes de rostros: *intrapersonales* (variaciones en la apariencia de un individuo, debidas a cambios en la expresión o en la iluminación) e *interpersonales* (variaciones en la apariencia debidas a diferencias en la identidad). Las funciones de densidad de probabilidad de alta dimensión para cada una de estas clases se obtienen de los datos de entrenamiento empleando la descomposición en espacios característicos antes señalada y, con ambas, se calcula la medida de similitud Bayesiana de la probabilidad *a posteriori* de pertenencia a la clase de diferencias *intrapersonales*, lo cual permite empatar rostros de prueba con aquellos existentes en una base de datos previamente recolectada.

Abstract

In this thesis we make a review of the algorithm proposed by Moghaddam et al. for unsupervised face detection [Moghaddam95a] and for the complementary process of face recognition in [Moghaddam96]. Even if we do not make a contribution, we were able to confirm the reported performance on face recognition in an implementation of our own, applying some simplifications to Moghaddam's development. However, as for face detection and recognition combined, performance found was not even close to what was indicated in these papers. We think this behavior is due to high dependency of the detection algorithm on relative size between faces in general images we used for testing. This low performance is carried over the face recognition process, working against global performance.

The referred unsupervised visual learning techniques are based on density estimation in high-dimensional spaces using an *eigenspace* decomposition. A multivariate Gaussian probability density is used for modeling training data and, later, is used to formulate a maximum-likelihood estimation framework for visual search and target detection for automatic object recognition.

Specifically to face recognition, the maximum-likelihood estimator is used to calculate a similarity measure based on a Bayesian analysis of image differences. We model two mutually exclusive classes of variations between two facial images: *intrapersonal* (variations with respect to different expressions or lighting) and *interpersonal* (variations with respect to a difference in identity).The high-dimensional probability density functions for each respective class are then obtained from training data using the eigenspace decomposition referred above and, with both of them, the Bayesian similarity measure based on the *a posteriori* probability of membership in the *intrapersonal* class is computed. Finally, this measure of similarity is used to rank matches in a previously collected face database.

Contenido

Lista de Figuras

Lista de Tablas

Lista de Símbolos

Símbolos	Descripción
I^t	Imagen de entrenamiento de $m \times n$ pixeles
$\mathbf{x}_i$	Imagen de entrenamiento como vector en $\mathbb{R}^N$
k	Número de imágenes en el conjunto de entrenamiento
$\mathbf{X}$	Imágenes de entrenamiento como matriz en $\mathbb{R}^{N \times k}$
$\bar{\mathbf{x}}$	Imagen promedio del conjunto de entrenamiento
$\tilde{\mathbf{x}}$	Imagen de entrenamiento normalizada respecto a $\bar{\mathbf{x}}$
A	Conjunto de entrenamiento
AA^T	Matriz de covarianzas del conjunto de entrenamiento
$\mathbf{\Sigma}$	Matriz de covarianzas del conjunto de entrenamiento
$\mathbf{\Phi}$	Matriz de vectores característicos de Σ
$\mathbf{v}$	Vectores característicos de Σ
$\mathbf{\Lambda}$	Matriz diagonal de valores característicos de Σ
λ	Valores característicos de Σ
M	Número tomado de vectores característicos principales
$\mathbf{\Phi}_M$	Submatriz de $\mathbf{\Phi}$ con los M vectores característicos principales
F	Subespacio principal o característico
$\bar{F}$	Subespacio complementario ortogonal
$\mathbf{y}_i$	Imagen de entrenamiento $\mathbf{x}_i$ proyectada en F
$\epsilon^2(\mathbf{x})$	Error residual de reconstrucción o $DFFS$
Ω	Tipo de clase de una imagen de entrenamiento
$P(\mathbf{x}\|\Omega)$	Función de verosimilitud de que $\mathbf{x} \in \Omega$
$\hat{P}(\mathbf{x}\|\Omega)$	Estimador de $P(x\|\Omega)$
$P_F(\mathbf{x}\|\Omega)$	Densidad marginal real en F
$\hat{P}_{\bar{F}}(\mathbf{x}\|\Omega)$	Estimador de la densidad marginal real en $\bar{F}$
ρ	Promedio de los valores característicos pertenecientes a $\bar{F}$
$J(\rho)$	Función de costo correspondiente al parámetro ρ
$(i,j)^{ML}$	Estimador de máxima verosimilitud de la posición del objetivo Ω
$(i,j,s)^{ML}$	Estimador de ML de la posición y escala del objetivo Ω
σ	Valores singulares de AA^T (también de A^TA)
r	Número de valores singulares distintos de cero
U	Matriz cuyas columnas son los componentes principales de A
V	Vectores singulares por la derecha de A (las columnas de V)

Símbolos	**Descripción**
$\mathbf{\Sigma}$	Matriz diagonal de valores singulares de AA^T y A^TA
$\mathcal{G}$	Galería de imágenes de rostros conocidos
$\mathcal{P}_\mathcal{G}$	Conjunto de prueba para la verificación
$\mathcal{P}_\mathcal{N}$	Conjunto de prueba para la identificación
E_I	Colección de imágenes de m pixeles de alto por n pixeles de ancho
V_R	Variedad de rostros
V_N	Variedad de no rostro

Capítulo 1

Introducción

1.1. Antecedentes[1]

El ambiente globalizado existente en la actualidad impulsado por el desarrollo tecnológico en la electrónica y en los sistemas de cómputo, principalmente, ha hecho muy tangible la necesidad de formas más seguras y fáciles de utilizar a fin de mantener a salvo nuestra información personal así como pertenencias físicas. En la actualidad todo el mundo necesita un número de identificación para acceder a un cajero automático, una clave de acceso para utilizar una computadora, otra docena para accesar internet y así sucesivamente. Aunque existen métodos confiables de identificación personal biométrica [Dunn07] (análisis de huellas dactilares, rastreadores de retina o iris, etc.), éstos se basan en la cooperación activa de los involucrados, mientras que un sistema de identificación personal basado en el análisis de imágenes frontales o de perfil del rostro muchas veces es efectivo aún sin la cooperación o el conocimiento del participante.

En cuanto a seguridad se trata, las ventajas de utilizar medidas biométricas para verificar la identidad son muchas. Éstas eliminan el uso indebido de tarjetas perdidas o robadas, reemplazan engorrosos números y contraseñas siempre difíciles de recordar y hasta pueden aplicarse como operadores automatizados de control de acceso en edificios o instalaciones de confianza. Más aún, en los sistemas biométricos basados en reconocimiento de rostros no existe la necesidad de tocar un objeto con los dedos o las palmas de las manos, ni la necesidad de presentar los ojos frente a un aparato detector.

[1] Esta sección está fundamentada en los capítulos 1, 2 y 16 de la referencia [Jain05], así como las revisiones del estado del arte en [Yang02] y [Zhao03].

Hoy más que nunca la seguridad es la principal preocupación en aeropuertos y centros de transporte de pasajeros. Aunque es posible controlar las condiciones de iluminación y la orientación de los rostros en estos escenarios (por ejemplo utilizando iluminación controlada en una fila única de pasajeros que intentan abordar), el mayor reto para el reconocimiento de rostros en lugares públicos consiste en el gran número de personas (rostros diferentes) que se deben examinar. Lo que tiene como resultado un porcentaje muy alto de falsas alarmas, generando incomodidad en las personas que son detenidas e interrogadas por el personal de seguridad que utiliza dichos sistemas.

Otras áreas en que se utilizan los sistemas de reconocimiento de rostros han sido las labores de agencias nacionales e internacionales que aplican la ley. Resulta invaluable la posibilidad de buscar e identificar rápidamente al sospecho de un crimen aún con información incompleta de su identidad e, inclusive, contando someramente con un retrato hablado del mismo, elaborado a partir de la descripción verbal de un testigo. Este mismo tipo de búsqueda especializada se ha aplicado para recuperar información de bases de datos, tomando como identificador o clave principal la imagen del rostro. Desde colecciones de fotografías personales; programas de noticias, deportivos o películas; hasta completas bibliotecas digitales de video multilenguaje se han beneficiado de este tipo de tecnología biométrica.

La interacción humana con la computadora en una forma más natural es la manera más conocida en que la ciencia del reconocimiento de rostros se abre paso al público en general. Esto incluye sistemas de cómputo que monitorean continuamente a la persona que está trabajando frente a ella y, si ésta se aparta, bloquean el acceso automáticamente hasta su regreso (o el acceso de una persona autorizada), desbloqueándose al momento y permitiendo su utilización nuevamente. El mismo principio se aplica al acceder a bases de datos personales, sistemas de archivos cifrados, correo electrónico, intranets corporativas, registros médicos, transacciones bancarias en línea, el propio automóvil particular, etc..

Sin embargo el reconocimiento de rostros con el proceso de detección como su primera fase no es simple. Se han planteado muy diversas aproximaciones para su solución pero, en general, la clasificación más aceptada corresponde a la expuesta por Yang et al. en [Yang02], según la cual se clasifican en las siguientes cuatro categorías (considerando que algunos de los métodos se pueden asociar a más de una):

Métodos basados en conocimiento. Estos métodos están basados en reglas que aplican

el conocimiento humano de lo que constituye un rostro. Usualmente las reglas capturan las relaciones entre los rasgos faciales de una persona.

Enfoques de características invariantes. Estos algoritmos tratan de encontrar los elementos estructurales que existen aún y cuando la posición, el punto de vista o la iluminación varían y, entonces, se utilizan para ubicar los rostros dentro de la escena completa.

Métodos de empate de plantillas (*templates*). En ellos se almacena una gran cantidad de patrones estándares de rostros para describir a una cara como un todo o para describir sus rasgos faciales separadamente. Para realizar el proceso de detección y reconocimiento se calcula la correlación entre una imagen de entrada y los patrones almacenados.

Métodos basados en apariencia. En contraste con los métodos de empate de plantillas, en los cuales a éstas las define un experto, las *plantillas* en los métodos basados en la apariencia se aprenden de una gran cantidad de imágenes de ejemplo. Por lo general este tipo de métodos se basan en técnicas de análisis estadístico y aprendizaje automático para encontrar las características relevantes de las imágenes que presentan rostros humanos. Las características aprendidas se presentan en forma de modelos de distribución estadística o funciones discriminantes que se utilizan para el proceso de detección.

En la sección 2.1 que corresponde a la introducción del capítulo 2, se detalla esta clasificación con mayor profundidad y se proporcionan referencias a los trabajos más representativos de cada categoría.

1.2. Planteamiento del Problema

La formulación general del problema de detección y reconocimiento de rostros se puede plantear de la siguiente manera: dada una imagen estática o varias imágenes de video de una escena en particular, se debe determinar la posición (en la imagen) de un solo rostro (*localización*) o de todos los rostros existentes en ella (*detección*); indicando posteriormente la identidad respectiva de los mismos, utilizando como medio de comparación una base de datos de rostros almacenada previamente para tal fin. Se puede utilizar información

Figura 1.1: Variabilidad en la apariencia de un rostro humano debida a la iluminación.

adicional disponible como raza, edad, género, expresión facial, grabaciones de voz, etc.; para restringir la búsqueda a fin de mejorar el proceso de reconocimiento [Zhao03]. En cuanto a la parte de detección también se suele clasificar en dos problemas bien diferenciados: identificación y verificación. En los problemas de *identificación* la entrada al sistema es la imagen de un rostro desconocido, debiéndose reportar la identidad del mismo determinada a partir de una base de datos de individuos conocidos; mientras que en los problemas de *verificación* el sistema necesita confirmar o rechazar la identidad pretendida de la imagen del rostro de entrada al proceso [Yang02].

Aunque se han propuesto muchas técnicas y enfoques que han demostrado eficacia prometedora, las tareas de detección y reconocimiento de rostros todavía son difíciles de completar satisfactoriamente. La variación en la iluminación y posición de la persona con respecto a la cámara han resultado ser los principales problemas en estas tareas. Desafortunadamente son inevitables cuando las imágenes se adquieren en condiciones no controladas como, por ejemplo, en lugares abiertos y al aire libre.

La Figura 1.1 ilustra el problema de la iluminación, cuando el mismo rostro parece diferente debido al cambio en la intensidad o posición de la fuente de luz. Los cambios introducidos por este fenómeno frecuentemente llegan a ser mayores a las diferencias entre individuos, provocando que los sistemas basados en la comparación de imágenes clasifiquen de forma errónea los rostros de entrada [Zhao03].

El desempeño de los sistemas de reconocimiento facial también disminuye significativamente cuando se presentan variaciones en la posición de la persona con respecto al eje óptico de la cámara, en las imágenes de entrada. La Figura 1.2 representa este tipo de variación. Cuando además se combina con alteraciones en la iluminación, la tarea de reconocimiento presenta aún mayor dificultad [Zhao03].

Figura 1.2: Variabilidad en la apariencia de un rostro humano debida a la pose.

1.3. Objetivo y Alcances de la Tesis

El objetivo general de la presente tesis es el desarrollo de un sistema que permita realizar los procesos de detección y reconocimiento de rostros humanos en imágenes estáticas.

1.3.1. Alcances

Las restricciones y consideraciones en el proceso de detección y reconocimiento de rostros respecto del ambiente y los sujetos involucrados en el presente desarrollo son:

Posición. Se asume que las imágenes se toman de una persona en posición vertical y de frente a la cámara, aunque el método implementado resulta eficaz contra pequeñas rotaciones ($< 10\,^{\circ}$) en el plano y hacia dentro y fuera del mismo.

Presencia o ausencia de componentes estructurales. Se asume que las imágenes tomadas al momento del entrenamiento o registro de los sujetos de prueba no variarán con respecto a aquellas tomadas al momento de realizar el proceso de reconocimiento. Esto es, si en el momento del entrenamiento el sujeto presentaba lentes, barba, bigote, etc., al momento del reconocimiento presentará la misma característica.

Expresión Facial. Para el presente trabajo se supone una expresión neutra de las personas inmiscuidas en los experimentos (rostro serio, boca cerrada, ojos abiertos, sin presentar muecas o gestos muy marcados). A pesar de lo anterior, la metodología utilizada resulta robusta con relación a expresiones faciales no muy exageradas (abriendo un poco la boca, por ejemplo).

Oclusión. La implementación realizada no permite la oclusión del rostro en forma alguna.

Condiciones de la imagen. Para el presente trabajo restringimos las condiciones extremas de luz, por lo que sólo se permite la luz natural o la iluminación artificial que la imite.

Localización de rostros. En el presente trabajo de tesis se encara solamente el primer problema, esto es, se da por sentada la existencia de una sola persona en la imagen de entrada al sistema.

Identificación. Nuestra implementación intenta resolver el caso general del reconocimiento de rostros, esto es, intenta resolver el problema de la *identificación* de individuos conocidos o determinar aquéllos que no lo son; en contraposición a la alternativa que consiste en solamente *verificar* la identidad pretendida por un sujeto de prueba.

1.4. Descripción del Sistema

Tal y como proponen Moghaddam et al. en [Moghaddam95a, Moghaddam96], el sistema desarrollado consiste en la adquisición de una imagen del rostro de una persona (fotografía), en una posición vertical de frente sin ningún tipo específico de iluminación (luz natural), en escala de 256 niveles diferentes de gris (8 bits)[2]. Posteriormente la imagen adquirida se somete a una búsqueda multiescala para la localización del rostro y el resultado se lleva a unas dimensiones de ancho y alto predeterminadas. A continuación se realiza otro proceso de búsqueda para ubicar el centro de ambos ojos[3]. Determinados estos dos puntos se utiliza una transformación geométrica de similaridad (traslación, rotación y escalamiento) para alinear el rostro con respecto a un formato estándar (una posición y escala definidos *a priori*). Finalmente se aplican una máscara que elimine la parte externa del rostro con el posible fondo que aún permanezca en la imagen así como una normalización respecto al contraste de la imagen resultante (ecualización por histograma). A partir de aquí se realiza el proceso de reconocimiento (utilizando análisis de componentes principales – *PCA*

[2]El color en una imagen tiene una dependencia muy importante en la iluminación. Por tal motivo se utilizan imágenes exclusivamente en escala de gris, la cual se ve menos afectada por la variación luminosa. Esto también presenta las ventajas de realizar menos cálculos y de manejar menos espacio de almacenamiento por imagen, a diferencia del color, que triplica su uso con respecto a aquélla (en el *modelo de color RGB*, que es el más utilizado).

[3] En la implementación original del Dr. Moghaddam se ubican adicionalmente la punta de la nariz y el centro de la boca, pero en las pruebas realizadas durante la etapa de experimentación, esto generaba una mayor cantidad de errores de detección. Por este motivo se decidió simplificar el proceso localizando solamente el centro de ambos ojos lo cual, para los alcances de la presente investigación, resulta suficiente tal y como se detalla en la sección 2.7.

- *Principal Component Analysis* – [Joliffe86]), comparando la imagen procesada contra una base de datos de rostros que previamente se ha dispuesto para tal fin, con lo cual se sabrá la identidad correspondiente de la persona en análisis (si existe en la base). Todos los rostros que conforman la base de datos referida también se han preparado con el procedimiento descrito, con la finalidad de tener la mayor exactitud posible en la determinación de las identidades de los sujetos de prueba. Gráficamente se representa el proceso anterior en la Figura 1.3.

1.5. Trabajo Previo

Se desea mencionar los siguientes desarrollos previos sobre la detección y el reconocimiento de rostros, por la relevancia que tienen sobre este ámbito de investigación:

***The Open Computer Vision Library* (*OpenCV*)** Esta librería, disponible en internet en [Intel Corp.00], suministra un marco de trabajo de alto nivel para el desarrollo de aplicaciones de visión por computadora en tiempo real: estructuras de datos, procesamiento y análisis de imágenes, análisis estructural, etc. El marco de trabajo facilita en gran medida el aprendizaje e implementación de distintas técnicas de visión, tanto a nivel docente como a nivel de investigador, aislando al desarrollador de las peculiaridades de los distintos sistemas de visión. Concretamente, el conjunto de funciones suministradas por la librería *OpenCV*, escritas en el lenguaje de programación C/C++, se agrupan en las categorías de estructuración y operaciones básicas, procesamiento y análisis de imágenes, análisis estructural, análisis del movimiento y seguimiento de objetos, reconocimiento de objetos, calibración de cámaras de video, reconstrucción tridimensional e interfaces gráficas de usuario así como de adquisición de video. Indudablemente se ha convertido en un estándar para la comunidad científica en el área de visión por computadora y una revisión obligada para cualquier investigador o iniciado en la misma. En virtud de lo anterior, el presente trabajo de tesis inició utilizando la librería *OpenCV* para los procesos de detección y reconocimiento de rostros, aprovechando su implementación del método de las *cascadas de Haar* propuesto por Viola y Jones en [Viola01, Viola04]. Desafortunadamente esta línea de desarrollo se declinó debido a la necesidad de entrenamiento adicional de las citadas cascadas que dicha implementación, en esa época, presentaba (si se pretendía obtener

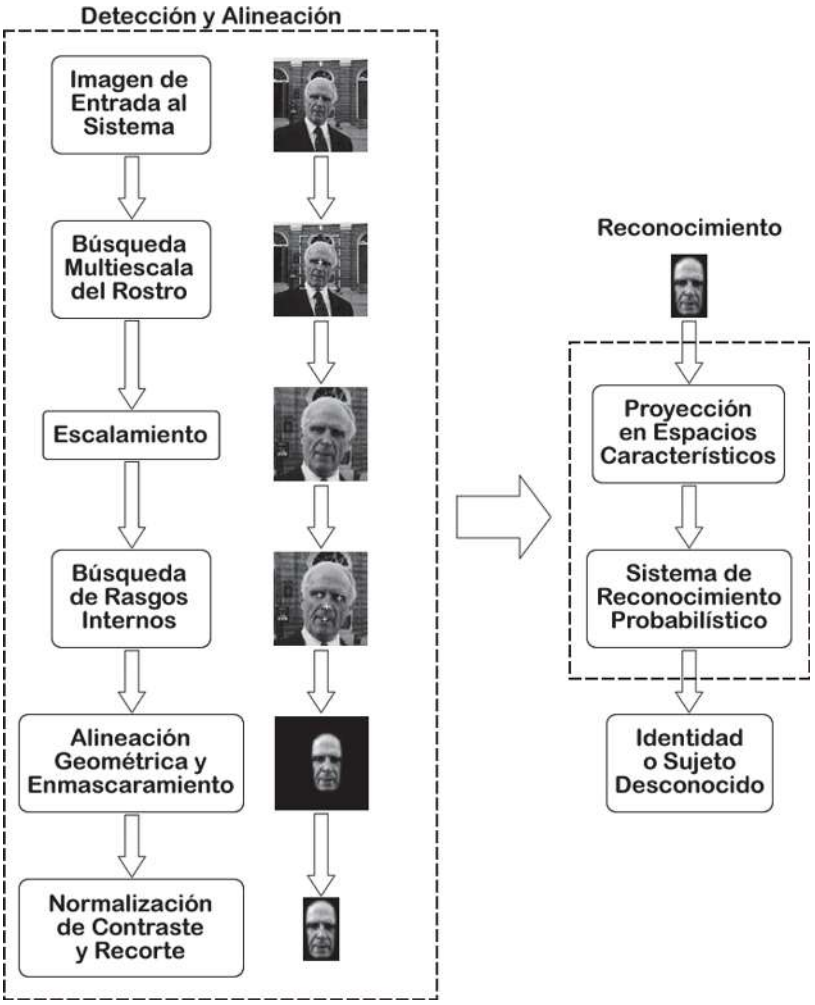


Figura 1.3: El sistema de procesamiento de rostros. (Imágenes tomadas de Moghaddam et al. [Moghaddam96].)

un rendimiento del sistema aceptable).

***The Face Recognition Technology* (*FERET*)** Este programa se realizó entre los años de 1993 y 1997. Fue auspiciado por el Programa de Desarrollo de Tecnología Antinarcóticos del Departamento de Defensa de los Estados Unidos. Sus principales aportes al ámbito de la investigación fueron una metodología estándar para realizar evaluaciones de sistemas de detección y reconocimiento facial (el *protocolo FERET* descrito en [Phillips00]), así como la recolección de una base de datos de imágenes faciales[4] (de acceso público gratuito y descrita en [Phillips98]) que permitiera llevar a cabo tales pruebas. Estas dos acciones permitieron que a partir de ese momento los investigadores alrededor del mundo pudieran establecer líneas de comparación verdaderamente imparciales en cuanto al desarrollo y desempeño de algoritmos y soluciones propuestas a los problemas de detección y reconocimiento facial. La mencionada base de datos cuenta con 14,051 imágenes de 1,199 individuos, con una resolución de 256 por 384 pixeles en escala de grises y divididas en categorías de acuerdo a la expresión facial, las condiciones de iluminación y la pose de los sujetos participantes (disponible en [NIST93]). Un detalle que hasta el momento no se había probado en la mayoría de los trabajos previos, fue la adición de una categoría que presentara imágenes tomadas del mismo individuo con una separación de entre 1 día y 3 años, lo cual resultó un verdadero desafío para los algoritmos y empresas participantes. A la fecha la base de datos *FERET* original en escala de grises es un material extra que se entrega como parte de la nueva base de datos *FERET* de imágenes en color de rostros humanos, también de acceso gratuito y disponible en [NIST03].

Portal web sobre reconocimiento de rostros Desde el año 2005 los doctores Mislav Grgic y Kresimir Delac de la Universidad de Zagreb, Croacia, mantienen un portal web sobre reconocimiento de rostros accesible desde [Grgic05]. En él se concentra información relevante sobre el área del reconocimiento facial y funciona como un repositorio de información para toda la comunidad científica. Es un punto de entrada para todo aquél recién llegado al tema y un recurso centralizado de información.

[4] Lamentablemente, pese a los esfuerzos realizados, no fue posible tener acceso con el tiempo suficiente a esta base de datos. El Instituto Nacional de Estándares y Tecnología de los Estados Unidos (*NIST - National Institute of Standards and Technology*), responsable de la misma, nos proporcionó la autorización de acceso a la base casi simultáneamente con la finalización de los experimentos realizados durante la presente investigación. Por esta razón, la colección de imágenes *FERET* no se utilizó en el desarrollo del presente trabajo.

Se proporciona acceso directo (manteniendo los derechos de autor, por supuesto) a los trabajos más interesantes y relevantes en el área, a algoritmos, bases de datos de imágenes faciales y códigos fuente de programas de computadora. Se presentan enlaces a los desarrollos más recientes, a la calendarización de conferencias, a revistas y libros, empresas que comercializan sistemas de reconocimiento biométrico (incluido facial), así como a sitios en internet de grupos de investigación y noticias en el área, en biometría y en otras cuestiones relacionadas con el mismo ámbito. Este sitio es la fuente principal de recursos que se utilizaron en el desarrollo del presente trabajo.

1.6. Contribuciones

El principal aporte del presente trabajo de tesis consiste en el desarrollo de un sistema automático de detección y reconocimiento facial propio, mismo que respalde los experimentos y pruebas necesarios para determinar que el desempeño reportado del método citado es real. Adicionalmente se tienen contribuciones menores en cuanto a simplificaciones de la implementación original de los algoritmos, las cuales se detallan en el capítulo 2.

1.7. Conclusiones

La necesidad de una mayor seguridad en los sistemas electrónicos y de cómputo actuales ha llevado a la investigación en el área al empleo de las características biométricas de una persona como medio de identificación. La huella dactilar, la palma de la mano, el rostro, la voz, el iris o la pupila de los ojos; características biométricas todas, aventajan a los medios tradicionales de identificación personal al eliminar la necesidad de recordar una gran cantidad de palabras o de portar continuamente un objeto o llave electrónica.

De las características biométricas mencionadas, el rostro conlleva la ventaja adicional de no requerir la colocación de la mano, los dedos o los ojos frente a un aparato rastreador. Incluso este tipo de reconocimiento facial puede realizarse sin el conocimiento o la cooperación de la persona involucrada.

Para *localizar* un rostro en una imagen dada o para *detectar* todos aquellos existentes en una escena, los problemas principales que se deben superar son la iluminación y la posición de las personas con respecto al eje óptico de la cámara. Las variaciones en la apariencia de una persona, introducidas por estos cambios de luz y de pose, pueden llegar

a ser más grandes que, incluso, la variación en las imágenes de diferentes individuos.

Los métodos empleados en la detección y el reconocimiento facial plantean soluciones basadas en reglas establecidas por un experto de lo que constituye la cara de una persona, buscan encontrar características del propio rostro que no varíen bajo cambios de luz o de pose, aplican *plantillas* o patrones de descripción del rostro humano a las imágenes de prueba o, también, obtienen dichos patrones por medios estadísticos o matemáticos para la *identificación* de un sujeto, posiblemente desconocido, o para la *verificación* de la identidad que tal sujeto dice poseer.

El avance tan relevante que en los últimos tiempos han tenido las áreas de la detección e identificación facial, sin lugar a dudas, se fundamenta en desarrollos colectivos a nivel mundial. Ejemplos de esto son la librería abierta de visión por computadora (*OpenCV*) que sirve como plataforma de soporte para desarrollos nuevos y el proyecto *FERET*, el cual implementa el protocolo de evaluación de tales desarrollos proporcionando, además, una base de datos de rostros sobre la cual realizar dicha medición. Aunado a lo anterior, existen diferentes portales en internet que proporcionan el conocimiento condensado de los avances en los últimos años y que sirven para introducirse rápidamente en el estudio de estas áreas.

1.8. Organización del Documento

El esquema de lo que resta de este trabajo de tesis es como sigue: el capítulo 2 describe la metodología seguida para realizar el proceso de detección de rostros humanos en escenas complejas, el capítulo 3 trata la parte correspondiente al proceso de reconocimiento, el capítulo 4 presenta la evaluación del sistema desarrollado con fundamento en los experimentos y pruebas realizados y, finalmente, el capítulo 5 presenta las conclusiones y trabajos futuros.

El capítulo 2 aborda el problema de la detección de rostros humanos en imágenes estáticas empleando estadística y probabilidad. En la introducción del capítulo se da un repaso al estado del arte en el área, después se define el problema, se establece la nomenclatura a utilizar durante el resto de los capítulos, se describe la metodología en que se fundamenta el algoritmo de detección facial y se plantean algunas sencillas modificaciones a la implementación original.

En el capítulo 3 se trata la parte de reconocimiento utilizando *rostros característicos duales* como base para el cálculo de un estimador bayesiano suponiendo una distribución

gaussiana de las imágenes de rostros de entrada al sistema. Se aborda el estado del arte sobre el particular, se plantea el algoritmo a seguir para entrenar al sistema de reconocimiento y para, finalmente, ponerlo a prueba.

La metodología de evaluación del desempeño del sistema desarrollado y los resultados obtenidos de los experimentos en los que ésta se aplica, se muestran en el capítulo 4. Se presentan tres tipos de experimentos: experimentos de reconocimiento con alineación manual del rostro, experimentos de detección automatizada y los experimentos de detección y reconocimiento automáticos combinados.

Las conclusiones y sugerencias para trabajos futuros se presentan en el capítulo 5. Después se tienen los apéndices, las referencias y, finalmente, el glosario de términos. En los apéndices se describe el análisis de componentes principales (*PCA – Principal Component Analysis*); la determinación del valor óptimo del parámetro ρ utilizado en el cálculo del estimador bayesiano, base de los algoritmos empleados para la detección y el reconocimiento facial; y, al final, se proporciona una descripción de las bases de datos de rostros manejadas en los experimentos del capítulo 4.

Capítulo 2

Detección de Rostros

La detección de rostros es el primer paso en cualquier sistema de reconocimiento facial automatizado y su confiabilidad tiene una influencia preponderante en la utilidad del sistema completo. Dada una imagen o un video, el detector de rostros ideal debe ser capaz de identificar y localizar todas las caras presentes en ellos sin importar su posición, escala, orientación, edad y expresión. Mas aún, la detección no debe depender de la iluminación existente ni del contenido de la imagen o el video [Jain05].

La detección facial se puede realizar basándose en muchas de las características de la cara humana: el color de la piel (en imágenes y video a color), el movimiento (en video), la forma de la cabeza o del rostro, la apariencia facial o una combinación de estos parámetros. El proceso consiste en analizar por medio de una ventana una imagen de entrada, en todas sus posibles ubicaciones y en todas sus posibles escalas (dentro de la escala original de la imagen, por supuesto). La detección se resume, pues, en clasificar el patrón en la ventana como un rostro o como algo que no es un rostro (*problema de dos clases: rostro y no rostro*). Este clasificador se aprende de un cierto número de ejemplos, tanto faciales como no faciales, utilizando métodos de aprendizaje estadístico [Jain05].

2.1. Introducción

Como se menciona en el capítulo anterior, una de las clasificaciones de los métodos de detección y reconocimiento de rostros más aceptadas en la actualidad es la propuesta por Yang et al. en [Yang02]. Esta clasificación divide los métodos en cuatro categorías: métodos basados en conocimiento, enfoques de características invariantes, métodos de empate

de plantillas y métodos basados en apariencia; las cuales se describen a continuación en profundidad así como sus trabajos más representativos.

2.1.1. Métodos Basados en Conocimiento

En este enfoque los métodos de detección se desarrollan con base en las reglas establecidas por el conocimiento del investigador de los rostros humanos. Por ejemplo, en una imagen cualquiera una cara presenta frecuentemente dos ojos simétricos entre sí, una nariz y una boca. Las relaciones entre estas características se pueden representar por sus distancias y posiciones relativas. Al inicio se extraen las características faciales de una imagen de entrada y después se identifican algunas regiones como posibles candidatas a contener un rostro basándose en las reglas establecidas. Posteriormente se realiza un proceso de verificación para reducir detecciones erróneas [Yang02].

Ejemplo de este tipo de métodos es el empleado por Yang y Huang en [Yang94], quiénes utilizaron un método jerárquico de tres niveles basado en conocimiento para la detección facial. Las reglas en el nivel más alto son descripciones generales de cómo luciría un rostro humano, mientras que en el nivel inferior el enfoque es hacia los detalles de las características del mismo. Se crea una jerarquía multiresolución de imágenes a través de promedios y submuestreo, como se ejemplifica en la Figura 2.1, luego se aplican las reglas para localizar las regiones candidatas. Por ejemplo: "la parte central del rostro tiene cuatro celdas con una intensidad básicamente uniforme" (la zona marcada con una retícula blanca en la Figura 2.1(d)); "la parte superior que circunda un rostro tiene básicamente una intensidad uniforme" (la zona marcada con la línea de guiones en la Figura 2.1(d)); y "la diferencia entre los valores de gris promedio de las zonas correspondientes a las dos reglas anteriores es significativa". La imagen de menor resolución (nivel 1) se analiza en busca de regiones candidatas y a éstas se les aplica una ecualización por histograma seguida por un proceso de detección de bordes (nivel 2). Las regiones candidatas sobrevivientes se examinan con otro conjunto de reglas que responden a características faciales como los ojos y la boca (nivel 3).

2.1.2. Enfoques de Características Invariantes

Otro de los caminos seguidos por los investigadores para resolver el problema de detección ha sido tratar de encontrar características faciales que no varíen. La suposición

(a) $n = 1$ **(b)** $n = 4$ **(c)** $n = 8$ **(d)** ***Detección Facial***

Figura 2.1: Imagen original e imágenes de baja resolución correspondientes. Cada celda cuadrada consiste de $n \times n$ pixeles en la cual se ha reemplazado la intensidad de cada píxel por la intensidad promedio de los pixeles en esa celda.

en la cual se basan es la observación de que los seres humanos pueden detectar la cara de otras personas sin ningún esfuerzo aparente, aún en diferentes posiciones y condiciones de iluminación por lo que, de hecho, deben existir propiedades invariantes aún bajo estos cambios. Comúnmente los métodos de este tipo extraen los rasgos faciales (como cejas, ojos, nariz, boca y línea del cabello) utilizando algoritmos detectores de bordes. Basándose en las características extraídas se construye un modelo estadístico que describa las relaciones existentes entre ellas y que permita verificar las regiones donde puede haber un rostro. Un problema con estas metodologías consiste en que las características de la imagen se pueden corromper severamente debido a la iluminación, el ruido o las oclusiones. Los límites de estas características se pueden suavizar dentro del rostro mientras que las sombras pueden causar bordes muy marcados que, en conjunto, vuelven inútil la aplicación de los algoritmos de agrupamiento visual [Yang02].

Sirohey utilizó en [Sirohey93] un mapa de bordes (*detector Canny*) así como algunas heurísticas propias para eliminar y agrupar bordes, de tal forma que sólo los bordes existentes en el contorno del rostro se preserven. Posteriormente ajustó una elipse al límite entre la región facial y el fondo de la imagen. Estas elipses resultantes son las regiones que su metodología propone como rostros.

Otra de las características del rostro humano que se han tratado como invariantes es su textura distintiva, la cual se puede utilizar para diferenciarla de otros objetos. Igualmente su color se ha utilizado como rasgo diferenciador en muchas aplicaciones, desde seguimiento del movimiento de las manos a la propia detección facial [Yang02]. Sobottka y Pitas propusieron en [Sobottka96] un método para localización de rostros y extracción de características faciales utilizando forma y color. Primeramente se aplica segmentación por

color en el *espacio HSV* (modelo de color Matiz - Saturación - Valor – *Hue* - *Saturation* - *Value* –) para ubicar regiones de piel. Luego se determinan los componentes conectados a través de algoritmos de dilatación aplicados a la imagen en menor resolución. Para cada componente conectado se calcula la elipse que mejor le ajuste, utilizando momentos geométricos, aquellos objetos que estén aproximados adecuadamente por la elipse respectiva se seleccionan como candidatos a contener un rostro humano. Finalmente estos candidatos se verifican buscando características faciales dentro de los componentes conectados. Las características, como ojos y boca, se extraen con base en el hecho de que son más obscuras que el resto de la cara.

2.1.3. Métodos de Empate de Plantillas

En el empate de plantillas se predefine o parametriza manualmente un patrón *estándar* de rostro (usualmente frontal) por medio de una función. Luego, dada una imagen de entrada, se calculan independientemente los valores de correlación de los patrones estándar con respecto al contorno del rostro, los ojos, la nariz y la boca. Dichos valores determinarán la posible existencia de un rostro en el área analizada. Aunque este enfoque es fácil de implementar resulta muy poco exitoso en el manejo de variaciones en cuanto a escala, posición y forma [Yang02].

Sinha en [Sinha94] y [Sinha95] utilizó un pequeño conjunto de invariantes espaciales para describir el espacio de patrones faciales. El hecho clave para la definición de estos invariantes es que, mientras las variaciones en la iluminación cambian el brillo individual de las diferentes partes del rostro (como los ojos, las mejillas y la frente), el brillo relativo de las mismas se mantiene notoriamente. Así pues, las restricciones en el brillo entre las diferentes partes de la cara se capturan mediante un conjunto apropiado de relaciones binarias *claro-obscuro* entre subregiones de la misma. Se dirá que se ha localizado un rostro si una imagen satisface todas las restricciones establecidas. Este método de Sinha también ha sido extendido por Scassellati en [Scassellati98] y aplicado a la localización de rostros en un sistema de visión robotizado. La Figura 2.2 muestra la plantilla mejorada del proceso, conformada por 23 relaciones definidas y clasificadas más especializadamente en 11 relaciones esenciales (flechas solidas) y 12 relaciones de confirmación (flechas punteadas). Cada flecha señala hacia la zona más obscura en la relación; misma que se satisface si la diferencia entre la zona clara y la obscura supera un umbral preestablecido y, si el número de relaciones

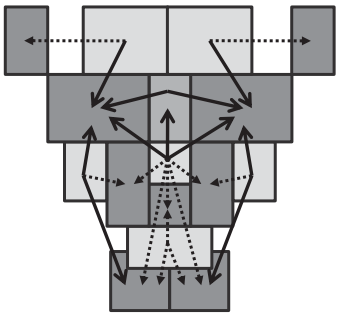


Figura 2.2: Plantilla para localización de rostros basada en el método de Sinha. La plantilla está compuesta de 16 regiones (cuadros) y 23 relaciones (flechas) [Scassellati98].

satisfechas supera su propio límite definido *a priori*, se ha localizado un rostro.

2.1.4. Métodos Basados en Apariencia

En contraste con los métodos de empate de plantillas, en los cuales a éstas las define un experto, las *plantillas* en los métodos basados en la apariencia se aprenden de una gran cantidad de imágenes de ejemplo. Por lo general este tipo de métodos se basan en técnicas de análisis estadístico y aprendizaje automático para encontrar las características relevantes de las imágenes que presentan rostros humanos y de aquéllas que no. Las características citadas se presentan en forma de modelos de distribución estadística o funciones discriminantes que, consecuentemente, se utilizan para el proceso de detección. Como estos métodos utilizan la mayor cantidad de información proporcionada por las imágenes de entrenamiento y de prueba (color de la piel, contraste entre las intensidades de los pixeles de las diferentes regiones del rostro, textura de la piel, borde de los ojos, la nariz, la boca y el cabello, etc.); presentan un desempeño aplastantemente superior a las otras 3 categorías descritas, motivo por el cual la investigación en los últimos años se ha desarrollado en torno a ellos casi exclusivamente [Yang02].

Kirby y Sirovich en [Kirby90] demostraron que las imágenes de rostros humanos se pueden codificar linealmente utilizando un número modesto de imágenes *base*. Esta demostración está basada en la *transformación de Karhunen – Loève* (expuesta en [Karhunen46] y [Loève55]); que también es conocida como *análisis de componentes principales* (*PCA - Principal Component Analysis*, véase [Joliffe86]); y como *transformación Hotelling* (ver [Hotelling33]). La idea es propuesta primeramente por Pearson en 1901 en [Pearson01] y

luego por Hotelling en 1933; consiste en que dada una colección de imágenes de entrenamiento, de $m \times n$ pixeles, representadas como vectores de tamaño $m \times n$; se puede determinar un conjunto de vectores que formen la base de un subespacio óptimo. Lo anterior se realiza de tal manera que se minimice el error cuadrado promedio entre las imágenes de entrenamiento originales y su proyección en este subespacio. A este conjunto de vectores de la base óptima se denominó *imágenes características* (*eigenpictures*), dado que son simplemente los vectores característicos (*eigenvectors*) de la matriz de covarianzas de las imágenes de rostros en el conjunto de entrenamiento, en su forma vectorizada.

En forma similar a lo anterior Turk y Pentland en [Turk91] aplicaron el análisis de componentes principales a la detección y el reconocimiento de rostros humanos. En su propuesta denominaron *rostros característicos* (*eigenfaces*) al conjunto de vectores de la base del subespacio óptimo y, a éste, *espacio de rostros*. De sus experimentos observaron que las imágenes faciales no cambian muy radicalmente al ser proyectadas[1] en el espacio de rostros (sin proporcionar una demostración analítica del porqué) mientras que las que no lo son sí lo hacen. Basados en este hecho, para detectar la presencia de una cara en una escena cualquiera se analizan todas las ubicaciones posibles en dicha imagen, se proyecta cada subregión en el espacio de rostros y se calcula su distancia con respecto al mismo. Esta distancia proporciona un indicador de *qué tan rostro* es la imagen en análisis, por lo que al resultado de este proceso de "medición" se le designó como *mapa de rostros*. El mínimo local del mapa de rostros indica la región donde existe la mayor factibilidad de que se haya detectado una cara. Cabe señalar que antes de realizar el análisis de componentes principales sobre el conjunto de entrenamiento se calcula la imagen promedio de éste (obteniendo la media de cada componente correspondiente – píxel – en las imágenes del propio conjunto). Posteriormente este promedio se substrae de cada imagen del conjunto con la finalidad de normalizar el proceso respecto a una media 0. La Figura 2.3 ejemplifica la cara promedio de un conjunto de entrenamiento y los siete primeros rostros característicos del análisis de componentes principales (los correspondientes a los valores propios – *eigenvalues* – de mayor magnitud).

Otro tipo de métodos clasificados como basados en la apariencia son las redes neuronales, las cuales se han aplicado satisfactoriamente en muchos de los problemas de

[1] La proyección de una imagen en el espacio de rostros es un vector de k coordenadas. Cada una de estas k coordenadas resulta de la sumatoria del producto, componente a componente, de la propia imagen y cada uno de los k rostros característicos.

Figura 2.3: Rostros característicos (la primera imagen es la cara *promedio* del conjunto de entrenamiento y las restantes son los primeros siete rostros característicos) [Turk91].

reconocimiento de patrones. Como la detección facial se puede plantear como un problema de patrones de dos clases (la imagen contiene o no contiene un rostro humano), se han propuesto varias arquitecturas de redes neuronales para solucionarlo. Al respecto los trabajos más significativos en el área los realizaron Rowley et al. (expuestos en [Rowley96] y [Rowley98]), empleando una red neuronal de varias capas para aprender los patrones de rostro/no rostro de imágenes de ejemplo (esto es, las intensidades y relaciones espaciales entre los pixeles). El sistema cuenta con dos módulos principales: las múltiples redes neuronales (para detectar patrones de rostro) y un módulo para toma de decisiones (para arbitrar resultados de detección posiblemente duplicados). Como se muestra en la Figura 2.4, el primer componente del proceso es una red neuronal que recibe una región de 20×20 pixeles de una imagen de entrada y produce un puntaje de salida con valores entre -1 (no hay rostro) y +1 (sí lo hay). Para detectar las caras en una escena cualquiera se aplica esta red neuronal en todas las ubicaciones posibles dentro de ella. A las escenas superiores a los 20×20 pixeles se les aplica repetidamente submuestreo y se les somete de nueva cuenta al proceso descrito de detección (en cada escala del submuestreo). Cada imagen de entrenamiento (cerca de $1,050$ en diferentes tamaños, posiciones e iluminación) se normalizó a una misma escala, posición y orientación; tomando como fundamento la ubicación de los ojos, la de la punta de la nariz y la de los extremos y el centro de la boca (etiquetados manualmente). El segundo componente realiza el arbitraje de la salida de las diferentes redes neuronales (simplemente aplicar los operadores booleanos Y/O) así como la mezcla de las detecciones que se solapan. Desafortunadamente una de las limitaciones de este procedimiento es que sólo se manejan apropiadamente rostros en posición vertical de frente; para resolver dicho inconveniente Rowley propuso en [Rowley99] añadir una red neuronal de redirección, la cual preprocesa cada ventana de entrada a fin de determinar la posible orientación del rostro, aplicar el giro conveniente que la coloque vertical y luego pasarla al sistema ya descrito.

A diferencia de los métodos que utilizan el valor particular de los pixeles de las

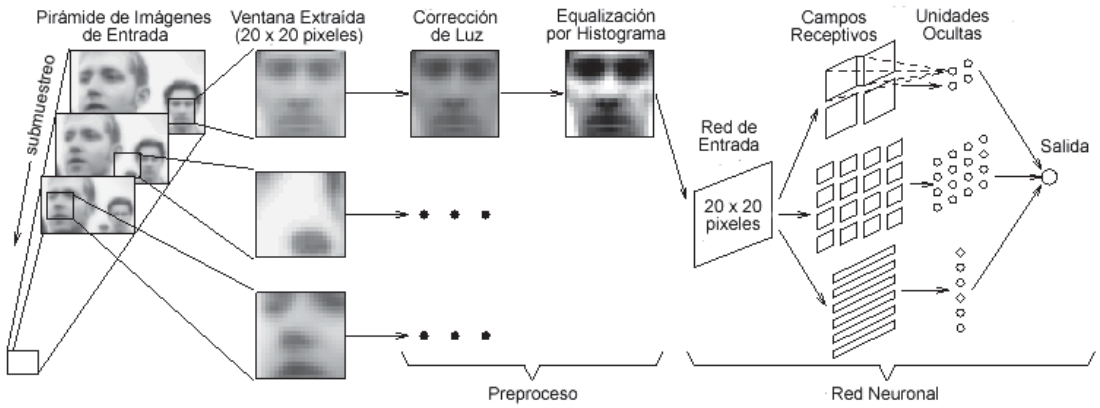


Figura 2.4: Diagrama del sistema del método de Rowley [Rowley98]. Cada rostro se preprocesa antes de pasarlo a las redes neuronales. Los sistemas de arbitraje determinarán si existe un rostro en la imagen con base en la salida de estas redes.

imágenes de entrada, el proceso de detección se puede basar en simples rasgos (*features*) tomados de éstas. El fundamento más común para tal acción es que los rasgos codifican el conocimiento específico del dominio que interesa, el cual resulta muy difícil de aprender utilizando una cantidad finita de datos de entrenamiento. Esta ventaja es el caso de los *rasgos Haar*, los cuales codifican la existencia de contrastes orientados entre regiones de una imagen así como sus relaciones espaciales. Se les denomina rasgos Haar debido a que se calculan de manera similar a los coeficientes de una transformación de *kernels de Haar* investigados por Chui en [Chui92]. El método que emplea este tipo de rasgos (denominado *A boosted cascade of simple features – Una cascada de rasgos simples con impulso*) fue propuesto por Viola y Jones en [Viola01, Viola04]). En la primera fase del proceso se entrena un *clasificador* con un cierto número de imágenes de muestra que presenten rostros humanos (ejemplos positivos) y de imágenes que no los presenten (ejemplos negativos). Estas imágenes deben estar escaladas a un mismo tamaño; en el trabajo de Viola y Jones la escala corresponde a 24×24 pixeles.

Una vez que se ha entrenado el clasificador, se aplica a la región de interés en una imagen de entrada (una ventana de la escena completa, del mismo tamaño que las imágenes del entrenamiento); si se presume que la imagen contiene un rostro humano, el clasificador devuelve el valor 1, si no, devuelve el valor 0. A fin de realizar la búsqueda en la imagen entera, se desliza la ventana de exploración a cada ubicación posible dentro de ella y, para manejar diferentes tamaños de rostros, el clasificador se diseña de tal forma que se pueda

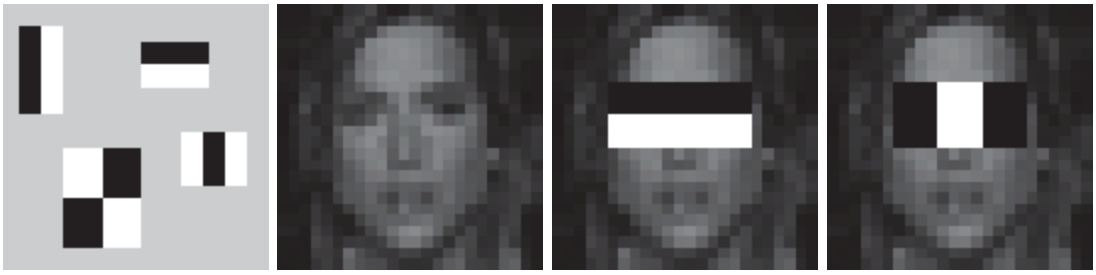
escalar muy fácilmente (lo cual es mejor que reescalar la imagen de entrada completa), lo que permite analizar repetidamente la muestra de entrada y detectar las caras de mayor o menor tamaño al de la ventana de exploración.

La palabra *cascada* en el nombre del método significa que el clasificador resultante consiste de varios clasificadores de mayor simpleza (etapas) que son aplicados a la región de interés en forma secuencial, hasta que en alguna de ellas se rechaza o hasta que finalmente se acepta atravesando toda la cascada. La palabra *impulso* significa que los clasificadores de cada etapa de la cascada son en sí mismos complejos. Es decir, constituidos de clasificadores más simples que se han mejorado utilizando alguna de las variantes del algoritmo *AdaBoost* [Freund95]. El hecho de que el método de Viola y Jones descarte las posibles regiones candidatas a contener un rostro humano al momento de fallar en una sola de las etapas es la base de su velocidad de proceso superior a los otros métodos presentados hasta la fecha.

La Figura 2.5(a) presenta algunos de los rasgos Haar empleados por las etapas del clasificador. La Figura 2.5(b) presenta una imagen de muestra, de 24×24 pixeles, que sirve como entrada para la etapa de entrenamiento o para la detección. Las Figuras 2.5(c) y 2.5(d) ejemplifican los dos principales rasgos seleccionados por el algoritmo *Adaboost* (los más determinantes para la detección de un rostro humano). Esto significa que la diferencia de intensidades entre la región de los ojos y la región superior de las mejillas es el principal fundamento para localizar un rostro humano. El segundo rasgo de mayor caracterización facial compara las intensidades entre la región de los ojos y la del puente de la nariz. En sí, el rasgo como tal consiste en el valor resultante de substraer la suma de las intensidades de los pixeles dentro de los rectángulos blancos de la suma de intensidades de los pixeles dentro de los rectángulos obscuros. Para optimizar el cálculo de la enorme cantidad de sumatorias que se tienen que realizar, el procedimiento utiliza el concepto de *imagen integral*, misma que se define como la sumatoria de las intensidades de los pixeles por encima y a la izquierda de una posición x, y cualquiera, tal y como se representa en la Figura 2.6(a). La imagen integral se calcula con la fórmula siguiente:

$$ii(x, y) = \sum_{x' \leq x, y' \leq y} i(x', y')$$

donde $ii(x, y)$ es la imagen integral e $i(x, y)$ es la imagen original. Utilizando el siguiente



(a) **(b)** **(c)** **(d)**

Figura 2.5: Ejemplos de las imágenes utilizadas en el método propuesto por Viola y Jones en [Viola01, Viola04]. a) Rasgos Haar. b) Ventana de 24×24 pixeles de entrada para los procesos de entrenamiento y detección. c) Rasgo principal seleccionado por al algoritmo *Adaboost* (el más determinante para la detección de un rostro humano). d) Segundo rasgo de mayor caracterización facial.

par de recurrencias:

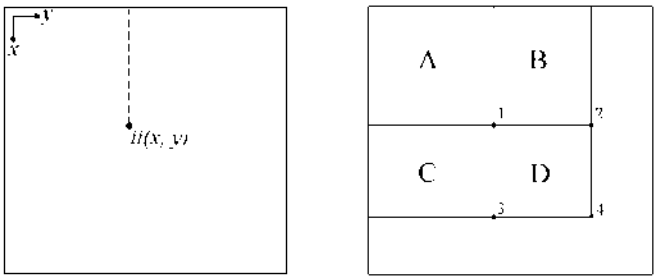
$$\begin{aligned} s(x,y) &= s(x,y-1) + i(x,y) \\ ii(x,y) &= ii(x-1,y) + s(x,y) \end{aligned}$$

(donde $s(x,y)$ es la suma acumulada de las filas, $s(x,-1) = 0$ e $ii(-1,y) = 0$) la imagen integral se puede calcular en una sola pasada sobre toda la imagen original.

Como se indica en la Figura 2.6(b), utilizando la imagen integral cualquier suma rectangular se puede calcular con sólo cuatro referencias.

El clasificador final utilizado en el trabajo de Viola y Jones consiste de 38 etapas. El número de rasgos utilizados en las primeras cinco son: 1, 10, 25, 25 y 50, respectivamente. Las capas restantes incrementan el número de rasgos empleados hasta alcanzar un total de $6,061$ en todas ellas. Se reporta que el clasificador puede trabajar a una velocidad de 15 fotos por segundo, analizando imágenes de 384×288 pixeles en una computadora personal Pentium© III a 700 MHz.

El método empleado para la detección de rostros en esta tesis se clasifica en esta última categoría. Las secciones que conforman el resto del capítulo se dedican a la descripción detallada del mismo.



(a) **(b)**

Figura 2.6: Imagen Integral. a) Representación. b) La suma de los niveles de gris correspondientes a los pixeles dentro del área D se puede calcular con cuatro referencias. El valor de la imagen integral en el punto 1 es la suma de los pixeles en el área A. El valor en el punto 2 es $A + B$, en el punto 3 es $A + C$ y, en el punto 4, es $A + B + C + D$. La suma dentro del área D se calcula como 4+1-(2+3).

2.2. Nomenclatura

A continuación se establece la nomenclatura formal de los componentes que conforman el problema (y la solución) de la detección y el reconocimiento facial.

La detección, la localización, la verificación y el reconocimiento de rostros humanos son parte de un tipo de problemas más general que consiste en el reconocimiento automatizado de un individuo en específico, basándose en características biológicas (anatómicas o físicas) y/o de conducta. Al proceso que involucra estos métodos automáticos se le denomina *biometría* [Dunn07].

Se denomina *sistema biométrico* o, simplemente, *sistema*; al sistema operacional automatizado conformado por múltiples componentes individuales (sensores, algoritmos de empate, dispositivo de despliegue de resultados, etc.) que permite capturar muestras biométricas de un usuario, procesarlas, almacenar la información obtenida de este procesamiento, comparar dicha información con la obtenida de muestras de referencia y, finalmente, decidir qué tan bien concuerdan. Este proceso se realiza con la finalidad de indicar si se ha logrado confirmar la identidad del sujeto al que pertenece la muestra de entrada. En este contexto una *muestra biométrica* es la información o los datos obtenidos de un dispositivo o componente de hardware (*sensor biométrico*) que transforma la entrada biométrica (imágenes de un rostro, de huellas dactilares, de la palma de la mano, del iris o la retina del ojo, etc.) a una señal digital, misma que se transmite al dispositivo de procesamiento,

el cual generalmente es una computadora [Dunn07].

Para los procesos de detección y, sobretodo, de reconocimiento facial, se emplea una *galería* y una *prueba*. La galería es el conjunto de muestras biométricas obtenidas de individuos conocidos o base de datos del sistema biométrico, utilizadas para una evaluación en específico o para una cierta implementación del propio sistema. La prueba consiste en una muestra biométrica adicional que se emplea como entrada al sistema a fin de compararla contra la galería para determinar su existencia en la misma (y obtener su identidad) o para indicar que el sujeto del que se obtuvo la prueba es un individuo ajeno a la galería o, simplemente, desconocido. La identificación será de *conjunto cerrado* si se sabe que la prueba de entrada existe en la base de datos, mientras que será de *conjunto abierto* si ésto no se garantiza [Dunn07].

2.3. Análisis en el Subespacio de Rostros

Las técnicas de análisis en subespacios para la detección y el reconocimiento de caras humanas están basadas en el hecho de que la clase de patrones que nos interesan, es decir las imágenes faciales, pertenecen a un subconjunto del conjunto de imágenes de entrada. Por ejemplo, una imagen pequeña de 64 pixeles de alto por 64 pixeles de ancho tiene $4,096$ pixeles en total. Estos pixeles expresan un gran número de clases de patrones como casas, autos y rostros humanos. Sin embargo, entre las $256^{4,096} > 10^{9,864}$ posibles combinaciones de los tonos de gris en estos $4,096$ pixeles (considerando imágenes en 256 niveles de gris), sólo unas pocas corresponden a los rostros. Así pues, si sólo nos interesan las imágenes correspondientes a rostros humanos, la representación original de las imágenes (al considerar todos los pixeles que las conforman) es muy redundante y su dimensión se puede reducir enormemente [Jain05].

Como se describió en la sección dedicada a los métodos basados en apariencia, con el enfoque de rostros característicos o análisis de componentes principales (ver apéndice A) una imagen facial se representa eficientemente como un vector de características de baja dimensión, esto es, un vector de *pesos*. Las características en tal subespacio proporcionan información más importante en cuanto al proceso de detección o reconocimiento que la imagen original. Este uso de técnicas de modelado de subespacios ha impulsado considerablemente el avance de la tecnología que permite solucionar estos problemas [Jain05].

Con la finalidad de establecer de manera más formal el problema de la detección y

el reconocimiento facial, salvo la definición en específico de lo que es un rostro (la cual resulta ser el obstáculo real para la resolución definitiva de ambos problemas); sea E_I la colección de imágenes I de m pixeles de alto por n pixeles de ancho, en 256 niveles diferentes de gris. Se tiene entonces que la *variedad de rostros* V_R se refiere a la subcolección de la colección de imágenes E_I, $\{V_R \subset E_I\}$, que engloba las variaciones en la apariencia del rostro. Esto es: $V_R = \{I \in E_I \mid I \text{ es un rostro}\}$. Mientras tanto, la *variedad de no rostro* V_N se refiere a la subcolección de E_I, complementaria de la anterior, que engloba las imágenes de cualquier otra cosa que no sea rostro. Esto es: $V_N = \{I \in E_I \mid I \text{ no es un rostro}\}$ teniendo, además, que $E_I = \{V_R \bigcup V_N\}$ y $\{V_R \bigcap V_N = \emptyset\}$.

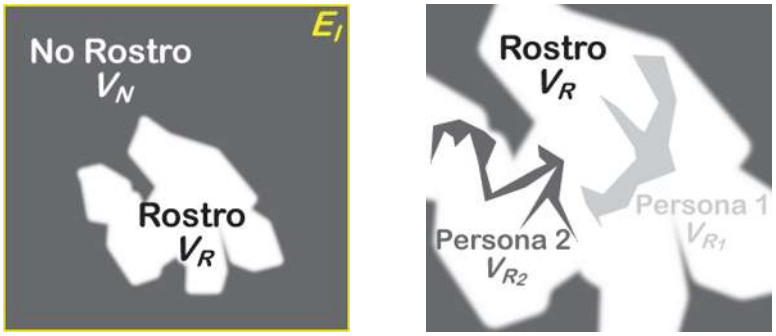
Si observamos estas *variedades* V_R y V_N inmersas en la colección de imágenes E_I, encontraremos que no son ni lineales ni convexas. Para ejemplificar este hecho la Figura 2.7(a) ilustra la variedad de rostros V_R contra la variedad de no rostros V_N como colecciones complementarias dentro de E_I y, la Figura 2.7(b), muestra gráficamente las variedades V_{R_1} y V_{R_2}, correspondientes al rostro en específico de dos individuos diferentes, inmersas en la variedad de rostros V_R completa.

La detección de rostros se puede considerar como una tarea de *distinguir entre las variedades de rostro y no rostro* en la colección de imágenes entera y, el reconocimiento, como la tarea de *distinguir entre las variedades específicas de rostros* correspondientes a diferentes individuos, inmersas en la variedad de rostros completa [Jain05].

Se debe notar que, en teoría y de acuerdo con el modelo descrito, cualquier imagen de un rostro debería pertenecer a V_R. En la práctica, debido al ruido en los sensores de las cámaras, usualmente la señal contiene un componente diferente de cero fuera de esta subcolección[2]. Esto introduce incertidumbre en el modelo y requiere técnicas algebraicas y estadísticas capaces de extraer la esencia matemática de la variedad de rostros aún en presencia del citado ruido [Jain05].

En la sección siguiente se detalla el método estadístico implementado para la detección de rostros en el presente trabajo de tesis, mismo que permita superar gran parte de los problemas descritos.

[2] Esto implica que las fronteras en la Figura 2.7(a) entre las variedades de rostro y no rostro sean en realidad borrosas. Asimismo implica que las variedades correspondientes a diferentes individuos, como V_{R_1} y V_{R_2} en la Figura 2.7(b), se pueden intersecar. Este efecto de intersección se puede presentar debido a alteraciones en las imágenes, por ejemplo debidas a cambios de iluminación; o por semejanza notoria entre sujetos, por ejemplo hermanos gemelos o padres e hijos.



(a) ***Rostro vs. No Rostro*** **(b)** ***Diferentes individuos***

Figura 2.7: Espacio de imágenes y principales variedades dentro de él.

2.4. Estimación de Densidad[3]

Como se estableció en los alcances del presente trabajo, la solución que se plantea a los problemas de detección y reconocimiento de rostros parte de la suposición de que las imágenes faciales a emplear presentan en conjunto una distribución de densidad gaussiana. Ésto es, deben ser imágenes frontales del rostro con expresión neutra, sin oclusiones y con iluminación natural o semejante a ella. Para determinar los parámetros que definen completamente esta distribución, que siendo del tipo gaussiana son la media y la varianza, se emplea un conjunto de imágenes $\{I_1, I_2, \ldots, I_k\}$ recolectado *a priori*. Con este conjunto de imágenes de $N = m \times n$ pixeles, en 256 niveles de gris, se puede formar un conjunto de vectores de entrenamiento $\mathbf{x}_k$, donde el i-ésimo componente del vector $\mathbf{x}_k$ corresponde al píxel $I_k(fila, columna)$ de la k-ésima imagen I_k con $i = (fila \times n) + columna$ y n es el ancho o número de columnas de I_k. Este conjunto de vectores $\mathbf{x}_k$ deben pertenecer a una misma clase de objetos, la cual denominaremos Ω (para fines de ejemplificación se supondrá que los objetos de la clase Ω son imágenes de "rostros humanos"). En tal sentido se construye la matriz $\mathbf{X}$ de $N \times k$ componentes, cuyas columnas corresponden a los k vectores de entrenamiento $\mathbf{x}_k$. La media de la distribución, $\overline{\mathbf{x}}$, se *estima* promediando los valores en cada una de las filas de la matriz $\mathbf{X}$ (lo que equivale a promediar, coordenada a coordenada, los vectores $\mathbf{x}_k$). De igual manera se estima la matriz de covarianzas con la

[3] Esta sección está fundamentada en su totalidad en las referencias [Moghaddam95a] y [Moghaddam96].

fórmula siguiente:

$$\mathbf{\Sigma} = \frac{1}{k}\mathbf{X}\mathbf{X}^T$$

Habiéndose estimado (robustamente) la media $\overline{\mathbf{x}}$ y la matriz de covarianzas $\mathbf{\Sigma}$ de la distribución[4], la verosimilitud de un patrón de entrada $\mathbf{x}$ está dado por:

$$P(\mathbf{x}|\Omega) = \frac{\exp[-\frac{1}{2}(\mathbf{x}-\overline{\mathbf{x}})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\overline{\mathbf{x}})]}{(2\pi)^{N/2}|\Sigma|^{1/2}} \tag{2.1}$$

La estadística suficiente para caracterizar esta verosimilitud es la *distancia de Mahalanobis*:

$$d(\mathbf{x}) = \widetilde{\mathbf{x}}^T\mathbf{\Sigma}^{-1}\widetilde{\mathbf{x}} \tag{2.2}$$

donde $\widetilde{\mathbf{x}} = \mathbf{x} - \overline{\mathbf{x}}$. Sin embargo dadas, por ejemplo, imágenes de entrenamiento de 150 pixeles de alto por 100 pixeles de ancho, la longitud o dimensión de los vectores es de $N = 150 \times 100 = 15,000$ componentes y si, además, se tiene un número $k = 1,000$ imágenes de entrenamiento, también sólo para ejemplificar; la matriz de covarianzas tendrá un tamaño de $N \times N = 15,000 \times 15,000 = \mathbf{225,000,000\ componentes}$. Esto es un problema, computacionalmente hablando, poco práctico de resolver. Más aún si se considera que se debe calcular la matriz *inversa* de la matriz de covarianzas $\mathbf{\Sigma}$.

En lugar de evaluar esta multiplicación cuadrática explícitamente, se puede realizar un cálculo mucho más eficiente y robusto, especialmente con relación a $\mathbf{\Sigma}^{-1}$. Este cálculo se introduce en el apartado siguiente.

2.4.1. Imágenes de Componentes Principales

La base vectorial resultante de una transformación de Karhunen - Loève (*KLT – Karhunen - Loève Transform–*) [Karhunen46, Loève55], se obtiene resolviendo el problema de valores característicos:

$$\mathbf{\Lambda} = \mathbf{\Phi}^T\mathbf{\Sigma}\mathbf{\Phi} \tag{2.3}$$

donde $\mathbf{\Sigma}$ es la matriz de covarianzas de los datos, $\mathbf{\Phi}$ es la matriz de vectores característicos de $\mathbf{\Sigma}$ y $\mathbf{\Lambda}$ es la matriz correspondiente de valores característicos. $\mathbf{\Phi}$ es una matriz ortonormal que define una rotación que elimina la correlación de los datos. En el *PCA*, se realiza una

[4] En la práctica, no se puede estimar una matriz de covarianzas $\mathbf{\Sigma}$ de rango completo N a partir de k observaciones independientes cuando $k < N$. Pero, como se verá, el estimador que se pretende no requiere la matriz de covarianzas completa sino solamente sus primeros M vectores característicos principales, donde $M < k$.

KLT parcial a fin de identificar los M vectores característicos correspondientes a los M valores característicos de mayor magnitud y obtener un vector de rasgos de componentes principales $\mathbf{y} = \mathbf{\Phi}_M^T \widetilde{\mathbf{x}}$, donde $\widetilde{\mathbf{x}} = \mathbf{x} - \overline{\mathbf{x}}$ es la imagen vectorizada normalizada con respecto a la media y $\mathbf{\Phi}_M$ es una submatriz de $\mathbf{\Phi}$ que contiene los M vectores característicos principales. El *PCA* se puede ver como una transformación lineal $\mathbf{y} = \mathcal{T}(\mathbf{x}) : \mathbb{R}^N \rightarrow \mathbb{R}^M$, la cual extrae un subespacio de menor dimensión de la base vectorial *KL* correspondiente a los valores característicos máximos. Estos componentes principales preservan las correlaciones lineales máximas en los datos y descartan las más pequeñas[5].

Ordenando los vectores característicos de la *KLT* con respecto a sus valores característicos y seleccionando los primeros M componentes principales, se forma una descomposición ortogonal del espacio vectorial $\mathbb{R}^N$ en dos subespacios complementarios y mutuamente excluyentes: el subespacio principal (o *espacio de características*[6] *–feature space–*) $F = \{\mathbf{\Phi}_i\}_{i=1}^{M}$ que contiene los componentes principales y, su complemento ortogonal, $\bar{F} = \{\mathbf{\Phi}_i\}_{i=M+1}^{N}$. Esta descomposición ortogonal se ilustra en la Figura 2.8(a), donde se tiene un ejemplo típico de una distribución que se encuentra completamente dentro de F. En la práctica siempre existe un componente de la señal dentro de $\bar{F}$ debido a pequeñas variaciones estadísticas en los datos o debido al ruido en las observaciones, el cual afecta a cada elemento de $\mathbf{x}$.

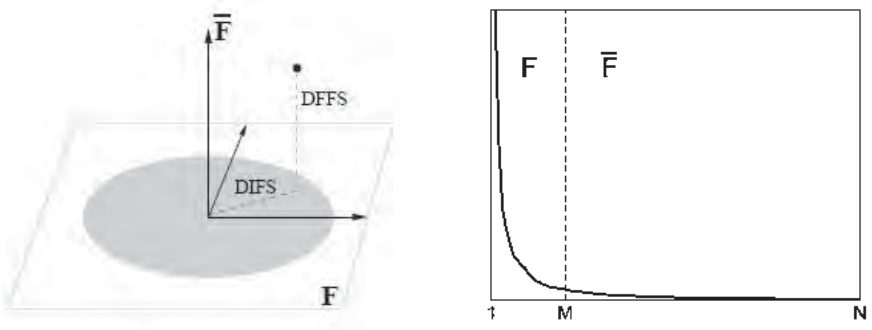
En una *KLT* parcial, el error residual de reconstrucción se define como:

$$\epsilon^2(\mathbf{x}) = \sum_{i=M+1}^{N} \mathbf{y}_i^2 = \|\widetilde{\mathbf{x}}\|^2 - \sum_{i=1}^{M} \mathbf{y}_i^2 \tag{2.4}$$

y fácilmente se puede calcular a partir de los primeros M componentes principales y la norma L_2 de la imagen $\widetilde{\mathbf{x}}$ normalizada con respecto a la media. Consecuentemente la norma L_2 de cada elemento $\mathbf{x} \in \mathbb{R}^N$ se puede descomponer en términos de sus proyecciones en estos dos subespacios. Se hace referencia al componente en el subespacio ortogonal $\bar{F}$ como la *distancia al espacio de características* (*DFFS – distance from feature space*), la cual es una simple distancia Euclidiana que equivale al error residual $\epsilon^2(\mathbf{x})$ en la Ecuación 2.4. El componente de $\mathbf{x}$ que cae *en* el espacio F es referido como la *distancia en el espacio de*

[5] En la práctica, el número de imágenes de entrenamiento k es mucho menor que la dimensión de las imágenes N, consecuentemente, la matriz de covarianzas $\mathbf{\Sigma}$ es singular. Sin embargo, los primeros $M < k$ vectores característicos siempre se pueden calcular (estimar) a partir de k muestras utilizando, por ejemplo, una descomposición de valor singular (*SVD – Singular Value Decomposition*) [Poole04].

[6] Para el interés de la presente investigación, este espacio de características es el *espacio de rostros –face space–*.



(a) *Subespacios F y $\bar{F}$* **(b) *Valores característicos***

Figura 2.8: Descomposición en subespacios de una densidad Gaussiana y del espectro típico de sus valores característicos.

características (*DIFS – distance in feature space*), pero generalmente no es una distancia basada en alguna norma, más bien, se puede interpretar en términos de la distribución de probabilidad de $\mathbf{y}$ en F.

2.4.2. Densidades Gaussianas en el Espacio F

Habiéndose determinado una forma simplificada de expresar la matriz inversa de la matriz de covarianzas, se puede reescribir $\mathbf{\Sigma}^{-1}$ utilizando los valores y vectores característicos de $\mathbf{\Sigma}$ en la forma diagonalizada:

$$\begin{aligned} d(\mathbf{x}) &= \widetilde{\mathbf{x}}^T \mathbf{\Sigma}^{-1} \widetilde{\mathbf{x}} \\ &= \widetilde{\mathbf{x}}^T [\mathbf{\Phi} \mathbf{\Lambda}^{-1} \mathbf{\Phi}^T] \widetilde{\mathbf{x}} \\ &= \mathbf{y}^T \Lambda^{-1} \mathbf{y} \end{aligned} \tag{2.5}$$

considerando que $\mathbf{\Phi}$ es ortogonal y que $\mathbf{y} = \mathbf{\Phi}^T \widetilde{\mathbf{x}}$ son las variables nuevas obtenidas con el cambio de coordenadas en una *KLT*. Debido a la forma diagonalizada, la distancia de Mahalanobis también se puede expresar en términos de la suma:

$$d(\mathbf{x}) = \sum_{i=1}^{N} \frac{\mathbf{y}_i^2}{\lambda_i} \tag{2.6}$$

En la base vectorial obtenida por la *KLT*, la distancia de Mahalanobis se encuentra convenientemente *desacoplada* en una suma ponderada de energías componentes no relacionadas. Más aún, la verosimilitud se transforma en un producto de densidades Gaussianas

separables e independientes. A pesar de su simpleza, la evaluación de la Ecuación 2.6 todavía no es computacionalmente factible debido a la dimensión tan alta. Por eso se busca entonces estimar $d(\mathbf{x})$ utilizando **únicamente las M proyecciones principales**. Intuitivamente, una elección obvia para la representación de menor dimensión es el subespacio principal indicado por el *PCA*, el cual captura el máximo grado de variación estadística de los datos[7]. Luego entonces, se divide la suma de la Ecuación 2.6 en dos partes independientes correspondientes al espacio principal $F = \{\mathbf{\Phi}_i\}_{i=1}^{M}$ y su complemento ortogonal $\bar{F} = \{\mathbf{\Phi}_i\}_{i=M+1}^{N}$:

$$d(\mathbf{x}) = \sum_{i=1}^{M} \frac{\mathbf{y}_i^2}{\lambda_i} \quad + \quad \sum_{i=M+1}^{N} \frac{\mathbf{y}_i^2}{\lambda_i} \tag{2.7}$$

Nótese que los términos en la primera suma se pueden calcular proyectando $\mathbf{x}$ en el subespacio principal F de dimensión M. Los términos restantes en la segunda suma $\{\mathbf{y}_i\}_{i=M+1}^{N}$, sin embargo, no pueden ser calculados explícitamente en la práctica debido a su alta dimensionalidad. A pesar de lo anterior, la suma de estos términos se encuentra disponible y, de hecho, *es* la cantidad *DFFS* $\epsilon^2(\mathbf{x})$; misma que se puede calcular con la Ecuación 2.4. Por lo tanto, basados en los términos disponibles, se puede formular un estimador para $d(\mathbf{x})$ como sigue:

$$\begin{aligned} \hat{d}(\mathbf{x}) &= \sum_{i=1}^{M} \frac{\mathbf{y}_i^2}{\lambda_i} \quad + \quad \frac{1}{\rho}\left[\sum_{i=M+1}^{N} \mathbf{y}_i^2\right] \\ &= \sum_{i=1}^{M} \frac{\mathbf{y}_i^2}{\lambda_i} \quad + \quad \frac{\epsilon^2(\mathbf{x})}{\rho} \end{aligned} \tag{2.8}$$

donde el término dentro de los paréntesis cuadrados es $\epsilon^2(\mathbf{x})$, mismo que se puede calcular utilizando los primeros M componentes principales. Entonces se puede escribir la forma del estimador de verosimilitud con base en $\hat{d}(\mathbf{x})$ como el producto de dos densidades Gaussianas marginales e independientes:

$$\begin{aligned} \hat{P}(\mathbf{x}|\Omega) &= \left[\frac{\exp\left(-\frac{1}{2}\sum_{i=1}^{M}\frac{\mathbf{y}_i^2}{\lambda_i}\right)}{(2\pi)^{M/2}\prod_{i=1}^{M}\lambda_i^{1/2}}\right] \cdot \left[\frac{\exp\left(-\frac{\epsilon^2(\mathbf{x})}{2\rho}\right)}{(2\pi\rho)^{(N-M)/2}}\right] \\ &= P_F(\mathbf{x}|\Omega)\ \hat{P}_{\bar{F}}(\mathbf{x}|\Omega) \end{aligned} \tag{2.9}$$

[7] En breve se verá que, dados los espectros típicos de los valores característicos observados en la práctica – Figura 2.8(b) –, esta elección es óptima por una razón diferente: Desde el punto de vista de la teoría de la información minimiza la *divergencia* entre la densidad verdadera y la estimación propuesta.

donde $P_F(\mathbf{x}|\Omega)$ es la densidad marginal real en el espacio F y $\hat{P}_{\bar{F}}(\mathbf{x}|\Omega)$ es la densidad marginal estimada en el espacio complementario ortogonal $\bar{F}$. El valor óptimo de ρ se puede determinar minimizando una función de costo apropiada $J(\rho)$. Desde el punto de vista de la teoría de la información, esta función de costo debe ser la *divergencia Kullback - Leibler* o *entropía relativa* [Cover94] entre la densidad real $P(\mathbf{x}|\Omega)$ y su estimación $\hat{P}(\mathbf{x}|\Omega)$:

$$J(\rho) = \int P(\mathbf{x}|\Omega) \log \frac{P(\mathbf{x}|\Omega)}{\hat{P}(\mathbf{x}|\Omega)} d\mathbf{x} = \mathrm{E}\left[\log \frac{P(\mathbf{x}|\Omega)}{\hat{P}(\mathbf{x}|\Omega)}\right] \tag{2.10}$$

Utilizando las formas diagonalizadas de la distancia de Mahalanobis $d(\mathbf{x})$ y de su estimación $\hat{d}(\mathbf{x})$ así como el hecho de que $\mathrm{E}[\mathbf{y}_i^2] = \lambda_i$, se tiene que (apéndice B):

$$J(\rho) = \frac{1}{2} \sum_{i=M+1}^{N} \left(\frac{\lambda_i}{\rho} \; - \; 1 \; + \; \log \frac{\rho}{\lambda_i} \right) \tag{2.11}$$

Entonces, el peso óptimo ρ se puede localizar minimizando esta función de costo con respecto a ρ. Una vez resuelta tenemos (apéndice B):

$$\rho = \frac{1}{N-M} \sum_{i=M+1}^{N} \lambda_i \tag{2.12}$$

lo cual es simplemente el promedio aritmético de los valores característicos en el subespacio ortogonal complementario $\bar{F}$. Además de su optimalidad, ρ también resulta ser un estimador insesgado de la distancia de Mahalanobis, i. e. $\mathrm{E}[\hat{d}(\mathbf{x};\rho)] = \mathrm{E}[d(\mathbf{x})]$. Esta derivación muestra que una vez seleccionado el subespacio principal F de dimensión M (como lo indica, por ejemplo, el *PCA*), el estimador óptimo de la estadística suficiente $\hat{d}(\mathbf{x})$ tendrá la forma de la Ecuación 2.8 con ρ dado por la Ecuación 2.12. En la práctica, por supuesto, sólo se tienen los primeros $k-1$ valores característicos de la Figura 2.8(b) y, en consecuencia, el resto del espectro se debe estimar ajustando una función (típicamente $1/f$) a los valores característicos disponibles, utilizando el hecho de que el último de éstos es simplemente el ruido introducido en los pixeles de la imagen. El valor de ρ se puede estimar entonces a partir de la porción extrapolada del espectro.

2.5. Detección de Máxima Verosimilitud[8]

La densidad estimada $\hat{P}(\mathbf{x}|\Omega)$ se puede utilizar para calcular una medida local de notabilidad objetivo en cada posición espacial (i, j) de una imagen de entrada basada en

[8] Esta sección está fundamentada en su totalidad en las referencias [Moghaddam95a] y [Moghaddam96].

el vector $\mathbf{x}$ obtenido por el ordenamiento lexicográfico de los valores de los pixeles en una vecindad local R:

$$S(i, j; \Omega) = \hat{P}(\mathbf{x}|\Omega)$$

para $\mathbf{x} =\downarrow [\{I(i + r, j + c) : (r, c) \in R\}]$, donde $\downarrow [\bullet]$ es el operador que convierte una subimagen en un vector. La estimación de máxima verosimilitud (*ML – Maximum Likelihood* –) de la posición del objetivo Ω está dada por:

$$(i, j)^{ML} = \text{argmax } S(i, j; \Omega)$$

Esta formulación *ML* se puede extender para estimar la escala del objeto con mapas de notabilidad *multiescala*. El cálculo de la verosimilitud se realiza (en paralelo) sobre versiones escaladas linealmente de la imagen de entrada $I^{(s)}$ correspondientes a un conjunto predeterminado de escalas (linealmente espaciadas) $\{s_1, s_2, \ldots, s_n\}$

$$S(i, j, s; \Omega) = \hat{P}(\mathbf{x}^{ijs}|\Omega)$$

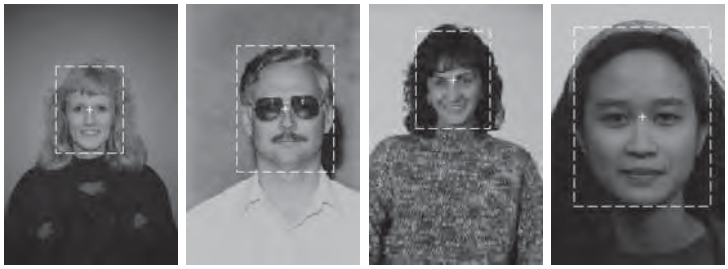
donde $\mathbf{x}^{ijs}$ es el vector obtenido de una subimagen local en la representación multiescala. Por lo tanto, el estimador *ML* de los índices espaciales y de escala se define como:

$$(i, j, s)^{ML} = \text{argmax } S(i, j, s; \Omega)$$

El enfoque anterior es el fundamento en el que se basa el método de detección de rostros aplicado en el presente trabajo de tesis y que se detalla en la siguiente sección.

2.6. Aplicación a la Detección de Rostros

El detector multiescala descrito en la sección anterior se aplica a la muestra de entrada al sistema automatizado de detección de rostros. La Figura 2.9 representa gráficamente algunos de los resultados en el proceso de detección de dicho detector, indicando la escala del rostro localizado por medio de los rectángulos de líneas blancas de guiones así como su posición, por medio de la cruz blanca en el centro del mismo. Los mapas multiescala de notabilidad $S(i, j, s; \Omega)$ se calcularon con base en el estimador de verosimilitud $\hat{P}(\mathbf{x}|\Omega)$ en un subespacio principal F de dimensión $M = 10$ utilizando el modelo Gaussiano tratado en la sección 2.4.1. Para evitar la variación debida a la iluminación, se normalizó la subimagen de entrada $\mathbf{x}$ convirtiéndola en un vector unitario con media cero.



(a) **(b)** **(c)** **(d)**

Figura 2.9: Ejemplos de detección multiescala de rostros. (Imágenes tomadas de [Moghaddam95a].)

El proceso completo de detección consiste de los siguientes pasos. En primer lugar el detector *ML* estima la posición y la escala del rostro tal y como se indica en la imagen de ejemplo 2.10(b) por medio del rectángulo límite y la cruz en el centro del mismo. Una vez que esta región se ha localizado, la escala y posición estimadas se utilizan para normalizar el rostro geométricamente en cuanto a traslación y escala, produciendo una imagen en un formato estándar que se pasará a la siguiente etapa – Figura 2.10(c) –. Un segundo proceso de detección de rasgos se aplica en esta escala fija para estimar la posición de cuatro de las características faciales: los dos ojos, la punta de la nariz y el centro de la boca – Figura 2.10(d) –. Ya que se localizaron los rasgos faciales, la imagen se modifica para alinear su geometría y su forma con un modelo canónico. Luego, la zona facial se extrae (aplicando una máscara fija) y, subsecuentemente, se normaliza con respecto al contraste a través de una ecualización por histograma. La imagen resultante será enviada al proceso de reconocimiento de rostros humanos descrito en el capítulo siguiente.

Para la detección de los cuatro rasgos faciales (ojos, punta de la nariz y centro de la boca) en la segunda etapa de detección, se utiliza un subespacio de dimensión $M = 5$, con una escala fija dado que la normalización geométrica que se realiza con base en la escala detectada para el rostro completo simplifica el trabajo. La Figura 2.11 presenta ejemplos de las imágenes utilizadas para entrenar los detectores de los rasgos faciales citados. Cabe aclarar que cada uno de los cuatro detectores utilizan únicamente como entrada para su entrenamiento (el cálculo de su *PCA* específico), el cuadro correspondiente al ojo derecho, al izquierdo, a la punta de la nariz o el rectángulo del centro de la boca, respectivamente.

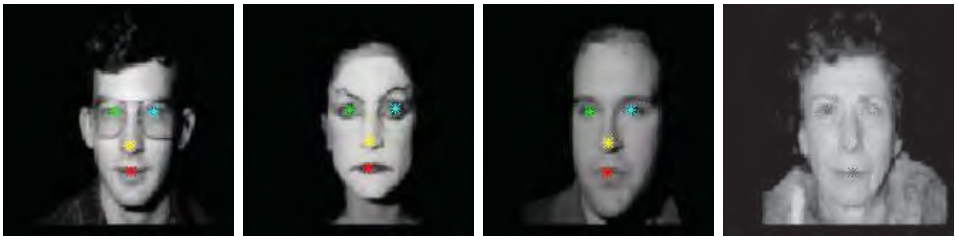
La Figura 2.12 muestra gráficamente algunas de las detecciones típicas resultantes

(a) **(b)** **(c)** **(d)** **(e)**

Figura 2.10: Etapas del proceso de detección de rostros humanos. a) Imagen original. b) Estimación de escala y posición. c) Escala y posición normalizadas geométricamente. d) Posición estimada de los rasgos faciales. e) Alineación y enmascaramiento final. (Imágenes tomadas de [Moghaddam95a].)

(a) **(b)** **(c)** **(d)**

Figura 2.11: Ejemplo de imágenes de entrenamiento para la detección de rasgos faciales. (Imágenes tomadas de [Moghaddam95a].)



(a) **(b)** **(c)** **(d)**

Figura 2.12: Detecciones típicas de rasgos faciales. (Imágenes tomadas de [Moghaddam95a].)

de los detectores de rasgos faciales ya entrenados.

Por último, se indica que las imágenes faciales mostradas en la presente sección pertenecen a la base de datos de rostros humanos en la referencia [NIST02], misma que fundamenta los trabajos de los que se tomó el concepto para implementar el método de detección facial aplicado en la presente tesis (principalmente [Moghaddam95a]).

2.7. Modificaciones a la Implementación Original

El proceso de detección de rostros descrito en la sección anterior corresponde a la implementación original de Moghaddam y Pentland en [Moghaddam95a]. Para el presente trabajo de tesis se realizaron ciertas simplificaciones a fin de mejorar la velocidad o el desempeño del sistema desarrollado (dado que los experimentos, con el proceso tal y como muestra el apartado anterior, presentaron un porcentaje de detección mucho menor que aplicando estas adecuaciones).

La primera de las simplificaciones introducidas corresponde al número de rasgos que se detectan dentro del rostro ubicado por el detector multiescala en la primera fase de detección. Dado que los rostros que se están analizando se encuentran en posición frontal, con poca o ninguna rotación hacia fuera o dentro del plano, no se requiere calcular una transformación afín completa que alinee el rostro geométricamente con el formato establecido *a priori*. En realidad sólo se requiere ubicar el centro de ambos ojos para tal fin. Esto es, se detectan los ojos, se alinean horizontalmente mediante una rotación calculada con base en su posición relativa (primeramente se traslada la posición media entre los ojos al origen de coordenadas para que la rotación se realice respecto a este último punto [Zisserman03]) y, finalmente, la distancia entre los ojos permite escalar el rostro al formato predefinido. La

(a) **(b)** **(c)** **(d)**

Figura 2.13: Normalización geométrica del rostro mediante una transformación de similaridad. a) Rostro detectado original. b) Detección de ambos ojos. c) Alineación. d) Escalamiento.

Figura 2.13 ejemplifica el proceso descrito.

El segundo cambio aplicado a la metodología del Dr. Moghaddam, consiste en que los mapas multiescala de notabilidad $S(i, j, s; \Omega)$ para detección del rostro, se calcularon empleando un subespacio principal F de dimensión $M = 31$, en lugar de $M = 10$, utilizando el modelo Gaussiano tratado en la sección 2.4.1. Asimismo la dimensión del subespacio principal F empleado en la detección de los ojos izquierdo y derecho fue de $M = 31$ componentes principales, en lugar de $M = 5$.

La tercera y última simplificación realizada al sistema consiste en que para la detección de ambos ojos se empleó, al igual que para el rostro, el proceso de búsqueda multiescala descrito en la sección 2.5 (aunque con tres escalas solamente).

Los dos últimos cambios descritos se especifican como simplificaciones al método original, dado que *disminuyen* la dificultad del problema a resolver (generando un mejor desempeño en la detección). Sin embargo, cabe señalar que imponen un mayor número de cálculos matemáticos a realizar.

2.8. Conclusiones

La detección es el primer paso para todo sistema de reconocimiento de rostros automatizado. Se han establecido muchas soluciones al problema y de formas muy variadas. Los métodos que mayor éxito han tenido son aquéllos que se basan en la apariencia del rostro humano tal y como se presenta en una escena dada, es decir, se toma una gran cantidad de ejemplos de imágenes faciales y se extrae, por medios estadísticos y matemáticos, modelos

de la esencia de lo que constituye un rostro humano. De esta manera, tales modelos se aplicarán en la identificación facial en las escenas de prueba subsecuentes.

Los modelos aprendidos resultan ser computacionalmente inviables si se pretende trabajar en el espacio de imágenes, por tal motivo, se han empleado técnicas de reducción o de análisis en subespacios mediante las cuales se obtenga la mayor parte del conocimiento específico del dominio que interesa (el rostro humano).

El método que se propone en la sección 2.4 del presente capítulo para solucionar el problema de la detección de rostros considera el análisis en el subespacio de rostros pero, a la vez, le da la importancia necesaria al componente *menor* de la solución que permanece fuera de él y que resulta vital en la formulación matemática de la distribución de probabilidad *completa* de la apariencia facial en el espacio de imágenes.

El planteamiento formulado sólo aplica para objetos de datos con distribución Gaussiana (como es el caso de las imágenes de rostros humanos en posición vertical de frente sin variaciones de iluminación que se tratan en esta investigación). Para generalizarlo a distribuciones de densidad más complejas (v. gr. imágenes con variación en la iluminación o en la pose del sujeto respecto al eje de la cámara), el propio Moghaddam et al. plantean en [Moghaddam95a] el modelo de densidad en términos de una mezcla de densidades Gaussianas, misma que permita estimar (aunque no de manera óptima como para el caso Gaussiano unimodal) la distribución de densidad de probabilidad completa de los objetos en el espacio de características.

Como punto final de las conclusiones del presente capítulo se señala que el proceso de detección del rostro, desde que se recibe la imagen de entrada (de 640 pixeles de ancho por 480 pixeles de alto) hasta que se entrega la imagen del rostro alineada, recortada y ecualizada; emplea poco más de 3 segundos en una computadora personal Intel Core© 2 Duo a 2.0 GHz.

Estimado el modelo de máxima verosimilitud propuesto y una vez aplicado a las imágenes de prueba, se obtendrá una imagen con el rostro *detectado* en un formato muy *ad-hoc* para el proceso de reconocimiento que se pretende realizar como objetivo global final del sistema. Esta imagen será el punto de entrada para los métodos de la parte complementaria de reconocimiento facial que se trata de manera muy detallada en el siguiente capítulo.

Capítulo 3

Reconocimiento de Rostros

Después de segmentar el rostro de la imagen de entrada al sistema, se procede a la determinación de la identidad del sujeto. Como se menciona en la sección 2.3, el proceso de reconocimiento facial se reduce a diferenciar, dentro del subespacio de rostros del espacio de imágenes, las diferentes variedades (o subespacios) correspondientes a cada uno de los individuos en las muestras de referencia que se tienen (base de datos de rostros) [Jain05].

3.1. Introducción

La clasificación de los métodos para detección facial ya fue descrita (sección 2.1), se hizo hincapié en el hecho de que los métodos basados en la apariencia son muy superiores al resto de los demás por lo que, de manera natural, se han extrapolado al proceso de reconocimiento. Dado que el objetivo en la detección es dividir un conjunto de imágenes en rostros o en cosas que no lo sean y viendo que la meta del reconocimiento es la división de tales rostros en imágenes del sujeto A, del sujeto B, etc.; se entiende pues, la aplicación subsecuente de los métodos citados.

3.1.1. El Subespacio de Rostros no es Lineal ni es Convexo

La Figura 3.1 muestra gráficamente la dificultad que se presenta para identificar la división del subespacio de rostros correspondiente a cada individuo dentro de un conjunto determinado. En el subespacio generado por los tres componentes principales (determinados mediante el *PCA* de las imágenes del conjunto de ejemplo tomadas en su forma vectorial) de un conjunto de rostros de ejemplo, se puede ver que éste no es lineal ni es convexo. De

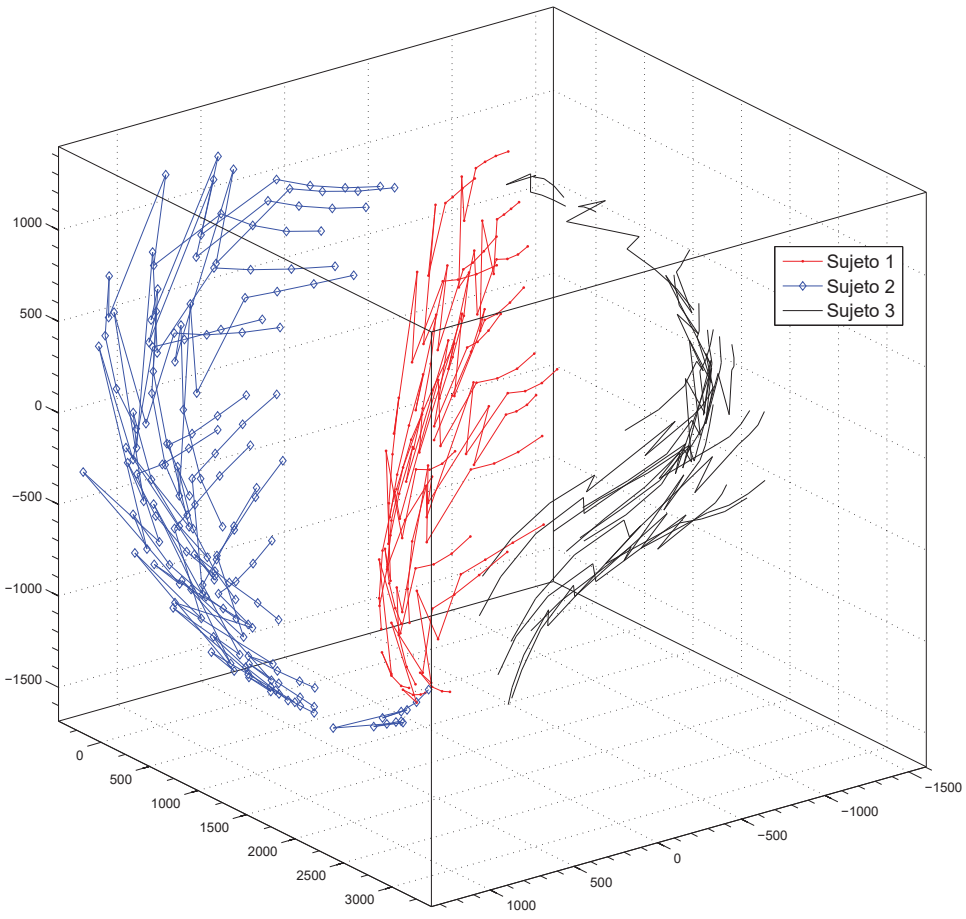


Figura 3.1: El subespacio de rostros no es lineal ni es convexo. Representación de las rotaciones mostradas en la Figura 3.2 por medio de curvas en el espacio 3D generado por los tres primeros componentes principales de un conjunto de rostros de ejemplo.

(a) *Sujeto 1.*

(b) *Sujeto 2.*

(c) *Sujeto 3.*

Figura 3.2: Transformación de rostros mediante rotaciones consecutivas.

lo anterior se deduce que los métodos *lineales* que tradicionalmente se han utilizado para resolver el problema, a pesar de sus grandes logros, no son suficientemente robustos [Jain05].

En la Figura 3.1 se presentan las divisiones del subespacio de rostros que pertenecen a cada uno de tres individuos diferentes. Se tienen 16 imágenes frontales del rostro de cada uno de estos individuos, mismas que se transforman aplicando una rotación variando 11 veces el ángulo aplicado (dos grados a la vez). Las 11 imágenes conforman una secuencia (como se ejemplifica en las tres imágenes de la Figura 3.2) y cada curva en la Figura 3.1 es la representación de dicha secuencia en el espacio de los tres primeros componentes principales. Por tanto se tienen 16 curvas por individuo. Se puede observar la no linealidad (y no convexidad) de las trayectorias.

A lo anterior cabe aplicar dos anotaciones: la primera, mientras que estos ejemplos se presentan en el subespacio generado por el *PCA*, es de esperarse que en el espacio original no modificado las curvas sean más (no lineales y no convexas) complejas. Segunda, aunque estos ejemplos son solamente rotaciones aplicadas a las imágenes, en las aplicaciones reales se presentan cambios de mayor complejidad en cuanto a cambios de iluminación, oclusiones, rotaciones dentro y fuera del plano, etc. [Jain05]

3.1.2. Maldición de la Dimensionalidad

Al igual que en el proceso de la detección, para tratar con el problema de la división del subespacio de rostros en el área correspondiente a cada individuo que ya se tiene identificado, se debe tratar con el problema de la dimensión de las imágenes vistas como vectores. Para especificar una imagen arbitraria en el espacio de rostros se necesita especificar el valor de cada píxel. Debido a ésto, la dimensión *nominal* del espacio, dictada por la representación de pixeles, es la multiplicación del número de pixeles del largo de la imagen por el de su ancho. Un número muy alto aún para imágenes de tamaño pequeño. Los métodos de reconocimiento facial que utilizan esta representación sufren de un gran número de desventajas (costo computacional muy alto, carencia del número apropiado de imágenes de entrenamiento – lo que generalmente produce matrices singulares –, etc.), la mayoría inmersas en la llamada *maldición de la dimensionalidad* [Jain05].

Para solucionar estos problemas, la investigación en el campo del reconocimiento facial ha empleado métodos lineales (y más recientemente no lineales) que permitan *extraer* la dimensión intrínseca del subespacio de rostros. Basta observar que la mayor parte del rostro humano es de superficie suave y de textura regular. Por esto el muestreo basado en pixeles es innecesariamente denso: el valor de un píxel típicamente está correlacionado enormemente con los pixeles a su alrededor. También se tiene que una vista frontal del rostro es eminentemente simétrica, lo que respalda el hecho de que la dimensión *real* del subespacio de rostros es mucho menor [Jain05].

3.1.3. Técnicas Lineales de Reducción de Dimensión

El más común de los métodos *lineales* empleados para reducir la dimensión del subespacio de rostros es el análisis de componentes principales descrito en la sección 2.3 y en el apéndice A. El *PCA* es una técnica que extrae el número deseado de *componentes principales* de los datos multidimensionales. El primer componente principal es la combinación lineal de las dimensiones originales que tiene la máxima variación, restringiendo a que los restantes *n-1* componentes principales sean ortogonales a éste [Gerbrands81].

Cuando se presentan cambios substanciales en la iluminación y la expresión del rostro la variación en los datos se debe en su mayoría a éstos. El *PCA* selecciona esencialmente el subespacio que retenga la mayor parte de la variación y, consecuentemente, la similitud entre individuos no necesariamente la determina su identidad [Jain05].

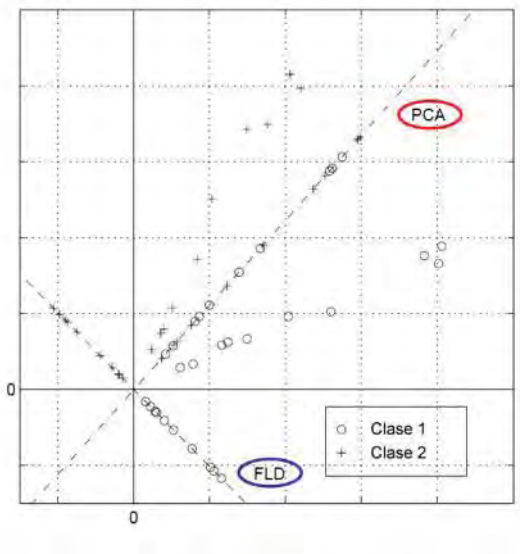


Figura 3.3: Fisherfaces (*FLD*) vs. Eigenfaces (*PCA*) [Belhumeur97].

Para resolver el problema anterior Belhumeur et al. en [Belhumeur97] emplearon los *rostros de Fisher* (*Fisherfaces*), una aplicación del discriminante lineal de Fisher (*FLD*). El *FLD* selecciona el subespacio lineal que maximiza la proporción de la variación interna de las imágenes de un mismo individuo con respecto a la variación entre las imágenes de diferentes individuos. Esto es, encuentra la proyección de los datos en la cual las *clases* (identidades) de individuos son linealmente más separables.

La Figura 3.3 presenta un ejemplo sencillo del subespacio seleccionado por la máxima variación en el *PCA* (lo que aglutina mas que separar las dos clases de objetos) contra el subespacio lineal de mayor separabilidad del *FLD* (lo que efectivamente diferencia las clases de objetos dados).

El *PCA* considera los pixeles que conforman las imágenes faciales como variables aleatorias con distribución Gaussiana y estadísticas de segundo orden minimizadas. Claramente, para cualquier distribución no Gaussiana (como es el caso de imágenes bajo diferente iluminación o pose), la variación mayor no corresponderá a los componentes principales. Dado este inconveniente otra de las propuestas más populares para reducir la dimensión involucrada es el *análisis de componentes independientes* (*ICA – Independent Component Analysis* descrito en [Bartlett98]). Este análisis minimiza las dependencias entre los datos

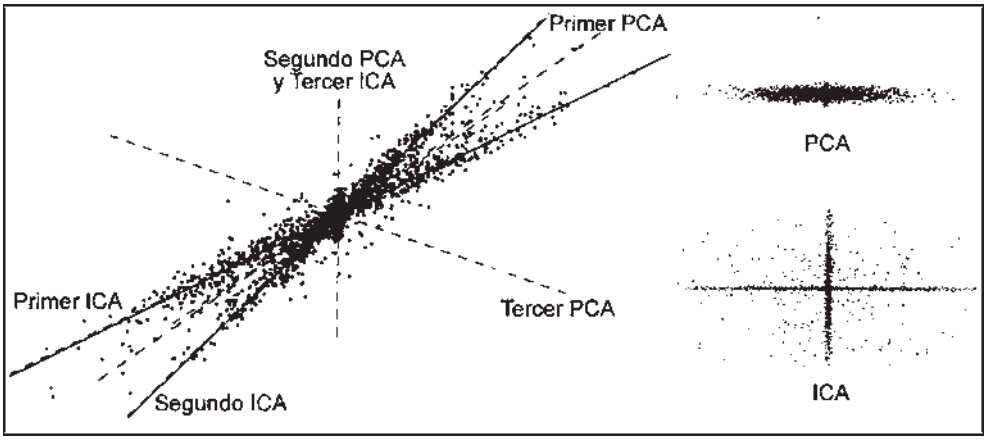


Figura 3.4: *ICA* vs. *PCA*. [Bartlett98].

de segundo y órdenes superiores, intentando encontrar la base vectorial a lo largo de la cual los datos (cuando se proyectan en ella) son *independientes estadísticamente*.

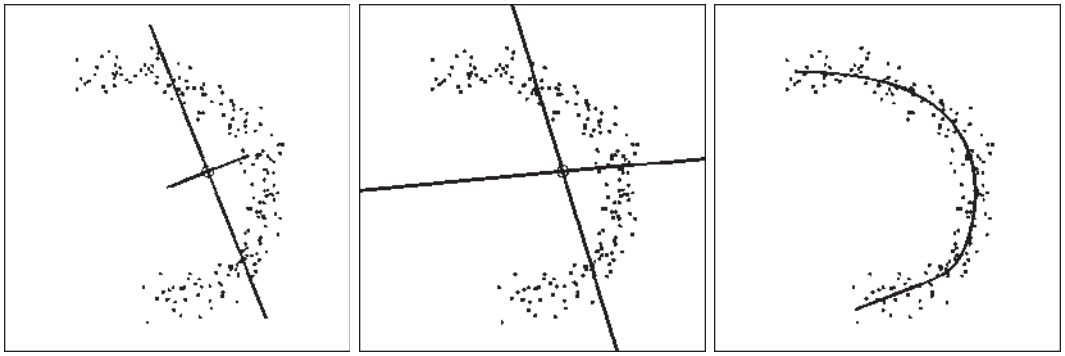
Como el *PCA*, el *ICA* realiza una transformación lineal en los datos pero con dos enormes diferencias: los vectores que conforman la nueva base *no* son ortogonales y se calculan de tal forma que su distribución *siempre* será no Gaussiana.

La Figura 3.4 ejemplifica la base vectorial determinada por el *ICA*. Del conjunto de puntos tridimensionales en la parte izquierda de la figura se calculan los tres primeros componentes principales y los tres primeros componentes independientes. Descartando el tercer vector de cada base, se construyen las proyecciones bidimensionales mostradas en la parte derecha (arriba la proyección del *PCA* y, abajo, la del *ICA*). El subespacio 2D recuperado por el *ICA* parece reflejar mucho mejor la distribución de los datos que el subespacio calculado por el *PCA*.

3.1.4. Técnicas No Lineales de Reducción de Dimensión

La propiedad que define a las técnicas no lineales de reducción de dimensión es que la *imagen inversa* del subespacio de rostros original es una superficie no lineal (curva) en el espacio de menor dimensión calculado por estos métodos, misma que *pasa por el centro mismo de los datos* a la vez que minimiza la suma total de las distancias entre los puntos de datos y su proyección en dicha superficie [Jain05].

La más conocida de estas metodologías se denomina *curvas principales* y su for-



(a) **(b)** **(c)**

Figura 3.5: Bases vectoriales calculadas. a) PCA (lineal, ordenada y ortogonal). b) ICA (lineal, no ordenada y no ortogonal). c) Curva principal (subespacio parametrizado no lineal). El círculo muestra la media global de los datos [Jain05].

mulación es básicamente una regresión no lineal. La Figura 3.5 presenta un ejemplo de esta curva principal en contraste con la base vectorial determinada por el *PCA* e *ICA*. Se observa el mucho mejor ajuste del método no lineal por sobre los dos lineales. La forma más simple de generar la curva principal es mediante redes neuronales multicapa de autocodificación que reciben el nombre de *PCA no lineal* y que implementan una función de proyección (no lineal, por supuesto) mediante una suma de sigmoides ponderada [Jain05].

La técnica del *PCA* también ha sido extendida utilizando *funciones generadoras de kernels* (funciones continuas que generan matrices simétricas definidas semi-positivas, para el ámbito de este trabajo de tesis). Aplicando el teorema de Mercer que, en forma muy resumida, establece que cualquier función generadora de kernels con las características antes citadas se puede expresar en forma de producto punto en un espacio de alta dimensión (posiblemente infinito), se transforma la no linealidad del problema de reconocimiento de rostros en un problema de solución claramente lineal en el nuevo espacio. Así, donde sea requerido un producto punto en el espacio original (como en el cálculo de la matriz de covarianzas de la que pretendemos sacar los componentes principales) se substituirá por el kernel, con el efecto adicional de no requerir el cálculo explícito (y posiblemente prohibitivo) del espacio generado por el mapeo del teorema [Jain05]. A éste método se le denomina *análisis de componentes principales con kernel* (*KPCA – Kernel PCA –*) y también ha sido extendido al discriminante lineal de Fisher; denominándose, en este caso, *Fisherfaces con kernel*. Los kernels típicos empleados incluyen funciones Gaussianas ($\exp(-\|\mathbf{x_i} - \mathbf{x_j}\|)^2/\sigma^2$),

polinomios $(\mathbf{x_i} \cdot \mathbf{x_j})^d$ y sigmoides $\tanh(a(\mathbf{x_i} \cdot \mathbf{x_j}) + b)$, todos los cuales satisfacen el mencionado teorema de Mercer [Jain05].

Las ventajas del *KPCA* sobre las redes neuronales y curvas principales son que no requiere optimización (no lineal), no está sujeto a sobreajuste de parámetros y no requiere conocimiento previo del número de dimensiones o de la arquitectura de la red. A diferencia del *PCA* tradicional, debido al mapeo a un espacio de mayor dimensión, se pueden utilizar más proyecciones de vectores característicos que la dimensión original de los datos. El inconveniente con este método resulta ser la inexistencia de una solución analítica demostrada para seleccionar el kernel y sus parámetros asociados óptimos [Jain05].

Una vez que las imágenes de rostros se han proyectado en el nuevo subespacio determinado por alguno de los métodos lineales o no lineales vistos hasta ahora, la tarea de reconocer a un individuo pasa por definir cuáles imágenes son más parecidas. Existen dos formas de realizar tal definición. La primera consiste en *medir* la distancia entre las imágenes en su representación vectorial en este subespacio. La segunda forma *mide* qué tan similares son las dos imágenes a comparar. Cuando se miden distancias se desea minimizar este valor, de tal manera que dos imágenes parecidas producen distancias pequeñas. Cuando se mide similitud se intenta maximizar su valor, por lo tanto, dos imágenes que son parecidas producirán un valor de similitud alto [Yambor00].

Obtenida la medida de distancia entre la imagen del sujeto que se pretende verificar y las imágenes de todos los individuos que se tienen en la base de datos de rostros, se selecciona la menor de ellas y se reporta la identidad correspondiente. En el proceso de identificación general, en el cual el sujeto en análisis puede no estar en la base, se define un umbral para el valor de la distancia. Si la distancia mínima calculada es menor al umbral, se actúa como en el caso anterior. Si la distancia mínima supera el umbral, se dictamina un individuo desconocido [Zhao03].

En la siguiente sección se describe en forma detallada la medida de distancia que se aplica en esta tesis.

3.2. Método Bayesiano para Cálculo de Distancia[1]

Considérese ahora un espacio de vectores $\mathbf{\Delta}$ correspondiente a diferencias, pixel a pixel, entre dos imágenes de rostros humanos ($\mathbf{\Delta} = \mathbf{I}_j - \mathbf{I}_k$). Se pueden definir dos clases

[1] Esta sección está fundamentada en su totalidad en las referencias [Moghaddam98] y [Moghaddam02].

de variaciones en las imágenes faciales: variaciones *intrapersonales* Ω_I (correspondientes, por ejemplo, a diferentes expresiones e iluminación del *mismo* individuo) y variaciones *interpersonales* Ω_E (correspondientes a variaciones entre *diferentes* individuos). La medida de similitud $S'(\mathbf{\Delta})$ se puede expresar en términos de la verosimilitud del vector de diferencias $\mathbf{\Delta}$ dada la clase de diferencias intrapersonales, esto es:

$$S'(\mathbf{\Delta}) = P(\mathbf{\Delta}|\Omega_I) \tag{3.1}$$

Aunado a lo anterior, también es posible plantear la medida de similitud de la probabilidad intrapersonal *a posteriori* de $\mathbf{\Delta}$ dada la clase de variaciones Ω_I y Ω_E mediante el teorema de Bayes:

$$S(\mathbf{\Delta}) = P(\Omega_I|\mathbf{\Delta}) = \frac{P(\mathbf{\Delta}|\Omega_I)P(\Omega_I)}{P(\mathbf{\Delta}|\Omega_I)P(\Omega_I) + P(\mathbf{\Delta}|\Omega_E)P(\Omega_E)} \tag{3.2}$$

Se observa que esta formulación Bayesiana con Ω_I y Ω_E transforma la tarea del reconocimiento facial (esencialmente un problema de clasificación n-ario) en un problema de clasificación binaria de patrones.

Las densidades de probabilidad de ambas clases se modelan como funciones Gaussianas de alta dimensión, utilizando el mismo método basado en el *PCA* empleado para la detección del rostro, los ojos y la nariz (descrito en la sección 2.4):

$$\begin{aligned} P(\mathbf{\Delta}|\Omega_E) &= \frac{e^{-\frac{1}{2}\mathbf{\Delta}^T\Sigma_E^{-1}\mathbf{\Delta}}}{(2\pi)^{D/2}|\Sigma_E|^{1/2}} \\ P(\mathbf{\Delta}|\Omega_I) &= \frac{e^{-\frac{1}{2}\mathbf{\Delta}^T\Sigma_I^{-1}\mathbf{\Delta}}}{(2\pi)^{D/2}|\Sigma_I|^{1/2}} \end{aligned} \tag{3.3}$$

Estas densidades están centradas respecto a la media dado que, para cada $\mathbf{\Delta} = \mathbf{I}_j - \mathbf{I}_i$, existe un $\mathbf{\Delta} = \mathbf{I}_i - \mathbf{I}_j$.

Debido al *PCA*, las densidades Gaussianas ocupan solamente un subespacio del espacio de imágenes (el subespacio de rostros); así, sólo los vectores característicos más importantes son relevantes para el modelado de la densidad de probabilidad. Estas densidades se emplean para evaluar las funciones de similitud en las Ecuaciones 3.1 y 3.2. El cálculo de la similitud consiste, pues, en primero substraer una imagen candidata $\mathbf{I}_{Prueba}$ de un rostro $\mathbf{I}_j$ en la base de datos. La imagen resultante $\mathbf{\Delta}$ se proyecta en los vectores característicos de la densidad Gaussiana intrapersonal para el estimador de verosimilitud en la Ecuación

3.1 (definido en la segunda parte de la Ecuación 3.3). Para el estimador Bayesiano $S(\mathbf{\Delta})$, este vector de diferencias $\mathbf{\Delta}$ también se proyecta en los vectores característicos de la densidad Gaussiana interpersonal, se calculan y normalizan los exponenciales y, finalmente, se combinan de acuerdo a la propia Ecuación 3.2. Este proceso se repite para cada rostro en la base de datos. La imagen que logre el máximo valor de similitud será considerada como la identidad del individuo que se encuentra bajo análisis.

El estimador de *máxima verosimilitud ML* (*Maximum Likelihood*) en la Ecuación 3.1 se define formalmente como:

$$ML = \arg\max[P(\mathbf{\Delta}_1|\Omega_I), P(\mathbf{\Delta}_2|\Omega_I), \ldots, P(\mathbf{\Delta}_{N_C}|\Omega_I)] \tag{3.4}$$

donde $\mathbf{\Delta}_i = I_i - I_{Prueba}$, con $i = 1, 2, \ldots, N_C$ y N_C es el número de imágenes de rostros en la base de datos. Complementariamente, el estimador bayesiano de probabilidad *máxima a posteriori MAP* en la Ecuación 3.2 se define formalmente como:

$$MAP = \arg\max[P(\Omega_I|\mathbf{\Delta}_1), P(\Omega_I|\mathbf{\Delta}_2), \ldots, P(\Omega_I|\mathbf{\Delta}_{N_C})] \tag{3.5}$$

Para grandes bases de datos estos cálculos son muy caros en cuanto a poder de cómputo se trata, por eso, es deseable simplificarlos con transformaciones previas a la realización del proceso de reconocimiento *en vivo*. Sobre tales transformaciones trata el apartado siguiente.

3.2.1. Cálculo Eficiente

Para calcular eficientemente las verosimilitudes $P(\mathbf{\Delta}|\Omega_I)$ y $P(\mathbf{\Delta}|\Omega_E)$, las imágenes $\mathbf{I}_j$ en la base de datos se preprocesan con transformaciones de *blanqueo*. Cada imagen se convierte y almacena como dos coeficientes blanqueados del subespacio de rostros: $\mathbf{y}_{\Phi_I}$ para el espacio intrapersonal y $\mathbf{y}_{\Phi_E}$ para el espacio interpersonal:

$$\mathbf{y}^j_{\Phi_I} = \mathbf{\Lambda}_I^{-\frac{1}{2}}\mathbf{V}_I\mathbf{I}_j, \qquad \mathbf{y}^j_{\Phi_E} = \mathbf{\Lambda}_E^{-\frac{1}{2}}\mathbf{V}_E\mathbf{I}_j \tag{3.6}$$

donde $\mathbf{\Lambda}_X$ y $\mathbf{V}_X$ son las matrices de los valores y vectores característicos más grandes, respectivamente, de $\mathbf{\Sigma}_X$ (substituyendo X por el símbolo I o E).

Después de este preprocesamiento la evaluación de las densidades Gaussianas se reduce a las simples distancias Euclidianas de la Ecuación 3.7. Los denominadores se precalculan, se evalúan las verosimilitudes y, con éstas, se calcula la similitud $S(\mathbf{\Delta})$ de la Ecuación

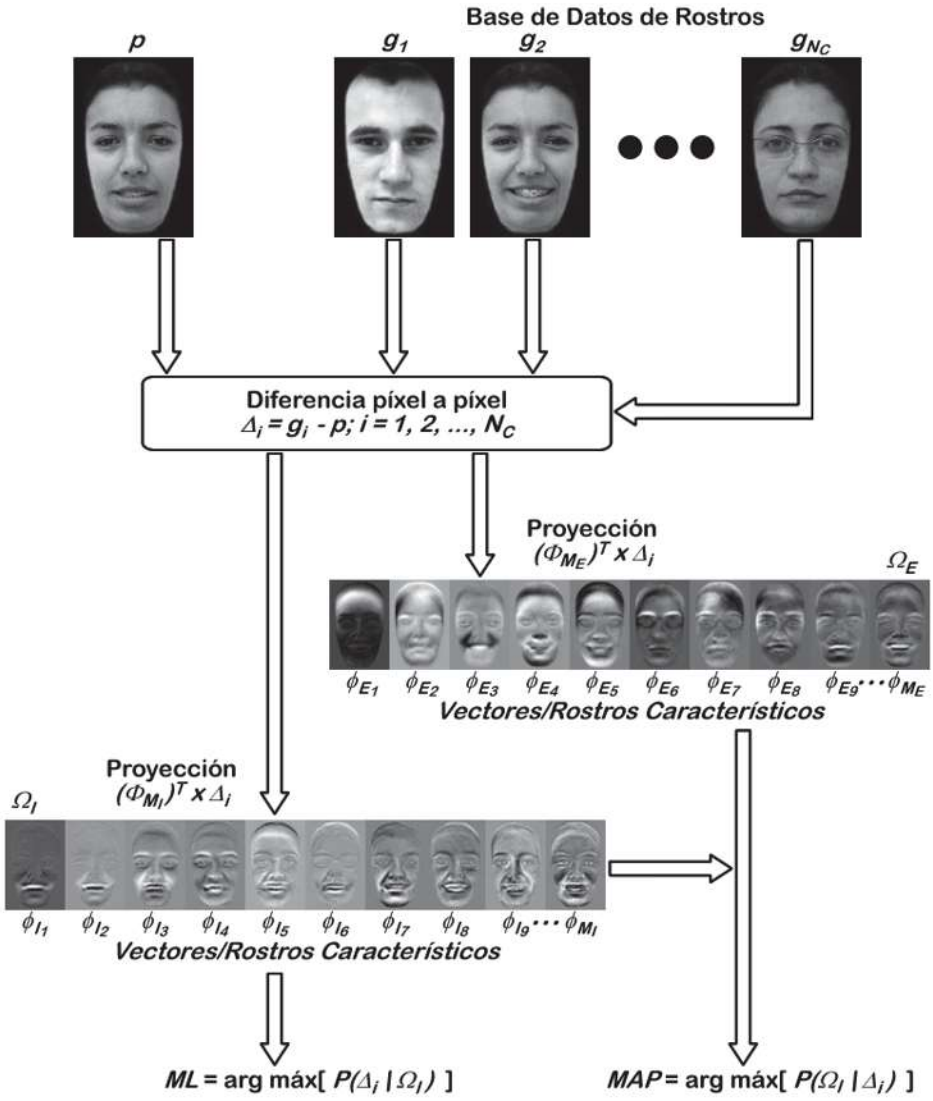


Figura 3.6: Cálculo de la similitud entre dos imágenes. La imagen de diferencias se proyecta en los dos conjuntos de rostros característicos (intrapersonal/interpersonal) para obtener las dos verosimilitudes del estimador MAP [Moghaddam02].

3.2 (como base del cómputo de la probabilidad *MAP* de la Ecuación 3.5). Las distancias Euclidianas se calculan entre los vectores $\mathbf{y}_{\Phi_I}$ de dimensión k_I así como entre los vectores $\mathbf{y}_{\Phi_E}$ de dimensión k_E. De esta forma, para cada evaluación de similitud se requieren someramente $2 \times (k_E + k_I)$ operaciones aritméticas, evitando tener que repetir la resta y proyección de las imágenes.

$$\begin{aligned} P(\mathbf{\Delta}|\Omega_I) = P(\mathbf{I} - \mathbf{I}_j|\Omega_I) &= \frac{e^{-\|\mathbf{y}_{\Phi_I} - \mathbf{y}^j_{\Phi_I}\|^2/2}}{(2\pi)^{k_I/2}|\Sigma_I|^{1/2}} \\ P(\mathbf{\Delta}|\Omega_E) = P(\mathbf{I} - \mathbf{I}_j|\Omega_E) &= \frac{e^{-\|\mathbf{y}_{\Phi_E} - \mathbf{y}^j_{\Phi_E}\|^2/2}}{(2\pi)^{k_E/2}|\Sigma_E|^{1/2}} \end{aligned} \tag{3.7}$$

La similitud $S'(\mathbf{\Delta})$ es todavía más simple de calcular, ya que solamente se evalúa la clase intrapersonal, teniendo la forma modificada siguiente (también como base del cómputo de la probabilidad *ML* de la Ecuación 3.4):

$$S'(\mathbf{\Delta}) = P(\mathbf{\Delta}|\Omega_I) = \frac{e^{-\|\mathbf{y}_{\Phi_I} - \mathbf{y}^j_{\Phi_I}\|^2/2}}{(2\pi)^{k_I/2}|\Sigma_I|^{1/2}} \tag{3.8}$$

El enfoque anterior requiere dos proyecciones del vector de diferencias $\mathbf{\Delta}$ a fin de calcular las verosimilitudes de la medida de similitud Bayesiana. El diagrama del cálculo se ilustra en la Figura 3.6. Los pasos de proyección son lineales mientras que la evaluación posterior es no lineal. Debido a las dos proyecciones *PCA* requeridas, a este método se le ha llamado una técnica de *espacio característico dual*. Se observa la proyección del vector de diferencias $\mathbf{\Delta}$ en los *rostros característicos duales* (Ω_I y Ω_E) para el cálculo posterior en la Ecuación 3.5.

Antes de pasar a las conclusiones del presente capítulo, cabe realizar el señalamiento de que el presente apartado sobre la optimización del cálculo de la similitud entre dos imágenes se incluye por completitud de la tesis. Pero tal optimización no se implementa en el sistema desarrollado debido a que no forma parte de la propuesta original con que Moghaddam y Pentland participaron en la competencia *FERET* de 1996 y 1997, la cual es el fundamento de la presente investigación.

3.3. Conclusiones

El reconocimiento de rostros humanos consiste básicamente en tomar el subespacio de rostros que el proceso de detección facial extrajo del espacio completo de las imágenes y

dividirlo en subespacios más pequeños, los cuales correspondan a la identidad de cada uno de los individuos que se tienen registrados en la base de datos de rostros que se prepara como parte de un sistema de reconocimiento. En el caso general de la identificación, se establece una división adicional etiquetada para aquellos individuos que no se encuentren en la base y, por lo tanto, sean desconocidos hasta el momento (al tiempo de concluir que el sujeto es desconocido se pueden agregar a la base de datos a fin de enriquecer el proceso de identificación con el tiempo).

A pesar de que el cálculo de un subespacio de rostros disminuye la dificultad del problema original, permanece la cuestión del manejo de las imágenes en su forma vectorial caracterizadas por una dimensión muy alta. Por lo anterior, se aplican técnicas matemáticas que transforman este problema, de carácter eminentemente no lineal, en planteamientos computacionalmente solubles (hablando en términos prácticos, por supuesto).

Las técnicas matemáticas que tradicionalmente se han aplicado a la reducción de dimensión son de tipo lineal y tienen como punto de partida el *PCA*. En etapas posteriores buscan dividir el subespacio de rostros basándose en independencia estadística, separabilidad lineal de *clases* o densidades de probabilidad.

Los métodos de reducción (dimensional) más recientes, utilizan planteamientos no lineales de tal forma que se ajustan de mejor manera al problema original. Redes neuronales, regresión no lineal, kernels e, incluso, tensores; proporcionan formas de mapeo del subespacio de rostros no lineal a espacios lineales o formas lineales que, mediante simples medidas de distancia o similitud basadas en estadísticas de primer y – a lo más – de segundo orden, clasifiquen las imágenes de prueba en identidades asociadas a sujetos.

El método Bayesiano propuesto para calcular una medida de similitud facial (sección 3.2), fundamentado en las diferencias intrapersonales de un mismo individuo así como en las diferencias interpersonales entre diferentes sujetos, resulta un planteamiento óptimo. Dado que el proceso automatizado de detección de rostros que se expone en la sección 2.4 segmenta, escala y alinea los rostros a comparar; las proyecciones en el espacio característico *dual* de estas diferencias – aún y cuando no son linealmente separables – son aproximadas óptimamente en términos de sus densidades de probabilidad *a posteriori* (bajo la suposición Gaussiana correspondiente a las imágenes de frente y en condiciones de iluminación semejantes a las que se restringe el presente trabajo de tesis).

Presentada la teoría correspondiente a los métodos de detección y reconocimiento de rostros humanos que se implementan, resta exponer los resultados experimentales que

demuestren su desempeño en la práctica. El siguiente capítulo presenta dichos resultados como soporte de todo lo visto hasta el momento.

Capítulo 4

Evaluación del Sistema

Se mencionó en capítulos anteriores la importancia trascendental del proyecto *FERET* en diversas vertientes del campo de la detección y del reconocimiento de rostros. Para el presente apartado la vertiente que interesa consiste en el protocolo desarrollado para evaluar los sistemas de detección y reconocimiento facial desde un punto de vista imparcial [Phillips98, Phillips00]. Aunque a la fecha se han desarrollado variantes y mejoras al protocolo citado, el protocolo original sigue siendo la línea base de comparación entre metodologías desarrolladas por los investigadores en el área alrededor del mundo [Jain05].

4.1. Introducción[1]

En la biometría y el reconocimiento de rostros el rendimiento se reporta sobre tres tareas comunes: la verificación, la identificación de conjunto abierto y la de conjunto cerrado. Cada tarea tiene su propio conjunto de medidas de rendimiento. Todas estas tareas tienen relación cercana, siendo la identificación de conjunto abierto el caso general.

Un sistema trabaja procesando muestras biométricas. Estas *muestras biométricas* son registros de los rasgos de una persona que permitan reconocerla, idealmente, sin equivocación (por ejemplo imágenes faciales o huellas dactilares). Una muestra biométrica puede consistir de registros múltiples, por ejemplo, cinco imágenes de la persona adquiridas al mismo tiempo o una imagen facial y una huella dactilar.

Para calcular el desempeño del sistema en las tareas señaladas se requieren tres

[1] Esta sección está fundamentada en su totalidad en las referencias [Jain05], [Phillips98], [Phillips00], [Blackburn01] y [Phillips03].

conjuntos de imágenes. El primero se denomina *galería* ($\mathcal{G}$), el cual contiene muestras de la gente conocida por el sistema. Los otros dos son *conjuntos de pruebas*. Una *prueba* es una muestra que se presenta al sistema para reconocimiento, donde el reconocimiento puede ser identificación o verificación. El primer conjunto de pruebas es $\mathcal{P}_{\mathcal{G}}$, el cual contiene muestras de gente en la galería (estas muestras son diferentes de aquéllas en la propia galería). El segundo conjunto de pruebas es $\mathcal{P}_{\mathcal{N}}$, el cual contiene muestras de gente que no está en la galería.

La identificación de conjunto cerrado es la medida de rendimiento clásica utilizada en la comunidad del reconocimiento automático de rostros, donde se le conoce como identificación. En la identificación de conjunto cerrado la pregunta básica es ¿de quién es este rostro?. Esta pregunta conlleva sentido para este tipo de identificación dado que una muestra de prueba siempre pertenece a alguien en la galería. Para el caso general de la identificación de conjunto de abierto, dado que la persona a la que corresponde la prueba no es forzosamente alguien en la galería, la pregunta básica es ¿se conoce este rostro?. En este caso el sistema debe decidir si la prueba contiene una imagen perteneciente a alguien en la galería y, si es así, debe reportar su identidad. La identificación de conjunto abierto y la de conjunto cerrado se denominan también *empate uno a muchos* o *empate* $1:N$.

En la tarea de verificación se presenta una muestra al sistema presumiendo cierta identidad y, éste, debe decidir si la muestra corresponde a la identidad pretendida. En este proceso la pregunta básica es ¿esta persona es quien dice ser?. Al procedimiento de verificación se le denomina también *autentificación* o *empate uno a uno*.

4.1.1. Identificación de Conjunto Abierto

La identificación de conjunto abierto es el caso general del reconocimiento de rostros. En esta tarea el sistema determina si una prueba p_j corresponde a una persona en la galería $\mathcal{G}$. Si se determina que la prueba se encuentra en la galería, el algoritmo debe identificar a la persona de la correspondencia.

Una galería $\mathcal{G}$ consiste de un conjunto de muestras $\{g_1, \ldots, g_{N_C}\}$ con $N_C = |\mathcal{G}|$, donde existe una muestra única por persona. Cuando una imagen de prueba p_j (perteneciente al conjunto de pruebas $\mathcal{P}_{\mathcal{G}}$ o al conjunto de pruebas $\mathcal{P}_{\mathcal{N}}$) se presenta al sistema, se compara con la galería completa. La comparación entre una prueba p_j y cada una de las muestras g_i en la galería, produce un valor de similitud s_{ij}. Entre mayor sea este valor, mayor será el

parecido entre las dos muestras (si la medida entre muestras es de *distancia* –como la *Euclidiana*– se puede convertir a similitud multiplicando su valor por -1). Un valor de similitud s_{ij} es un *empate* si g_i y p_j son muestras de la misma persona. Un valor de similitud *no es un empate* si son muestras de diferentes personas. Si p_j es una muestra de una persona en la galería, se designa como g_* su correspondencia única (en la galería). El valor de similitud entre p_j y g_* se designa como s_{*j}. La función *id()* devuelve la identidad de una muestra biométrica, con $\mathrm{id}(p_j) = \mathrm{id}(g_*)$. Para la identificación se examinan y ordenan todos los valores de similitud entre una prueba p_j y una galería. Se dice que la prueba p_j tiene rango n si s_{*j} es el n-ésimo valor de similitud más grande. Esto se designa como $\mathrm{rango}(p_j) = n$. Por ejemplo, rango 1 significa que el valor s_{*j} es el máximo de los valores de similitud calculados y, rango 5, indica que existen cuatro valores de similitud s_{ij} superiores a s_{*j}, esto es, s_{*j} es el quinto valor de similitud más grande de los calculados. Al rango 1 se le denomina también *mejor empate*.

El rendimiento para la identificación de conjunto abierto se caracteriza con dos estadísticas de desempeño: la *tasa de detección e identificación* y la *tasa de falsas alarmas*. Una prueba es detectada e identificada si la prueba es identificada correctamente y su correspondiente valor de similitud se encuentra por encima de un cierto umbral de operación τ. Estas condiciones corresponden formalmente a:

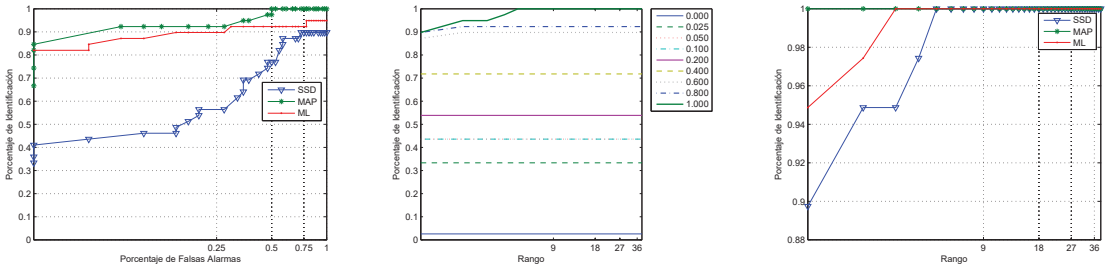
- $rango(p_j) = 1$
- $s_{*j} \geq \tau$ para el valor de similitud donde $id(p_j) = id(g_*)$

para el umbral de operación τ. La tasa de detección e identificación es la fracción de pruebas en el conjunto $\mathcal{P}_\mathcal{G}$ que son detectadas e identificadas correctamente. Esta tasa es una función del umbral de operación τ, la cual se define como:

$$P_{DI}(\tau, 1) = \frac{|\{p_j : rango(p_j) = 1, \text{ y } s_{*j} \geq \tau\}|}{|\mathcal{P}_\mathcal{G}|} \tag{4.1}$$

La tasa de falsas alarmas proporciona información del desempeño cuando una prueba no se encuentra en la galería (esto es, $p_j \in \mathcal{P}_\mathcal{N}$). Este tipo de prueba también es referida como *impostor*. Ocurre una falsa alarma cuando el valor de similitud del mejor empate supera el umbral de operación τ. Formalmente, ocurre una falsa alarma cuando:

$$\max_i s_{ij} \geq \tau$$



(a) ROC **(b) ROC *basada en rangos*** **(c) CMC**

Figura 4.1: Ejemplos de curvas *ROC*, *ROC* basada en rangos y *CMC* para la evaluación de un sistema de detección y reconocimiento facial. $|\mathcal{G}| = 40$, $|\mathcal{P}_\mathcal{G}| = 40$ y $|\mathcal{P}_\mathcal{N}| = 40$.

La tasa de falsas alarmas es la fracción de pruebas $p_j \in \mathcal{P}_\mathcal{N}$ que son falsas alarmas. Esto se calcula con la siguiente ecuación:

$$P_{FA}(\tau) = \frac{|\{p_j : \text{máx}_i\, s_{ij} \geq \tau\}|}{|\mathcal{P}_\mathcal{N}|} \tag{4.2}$$

El sistema ideal tendría una tasa de detección e identificación de 1.0 y una tasa de falsas alarmas de 0.0; toda la gente en la galería se detecta e identifica y no hay falsas alarmas. Sin embargo, en sistemas reales existe un intercambio entre ambas tasas. Modificando el umbral de operación se cambian las tasas de rendimiento. Al incrementar el umbral de operación disminuyen ambas tasas. No se pueden maximizar ambas estadísticas a la vez. Este intercambio se muestra gráficamente en una curva *característica recibida de operación* (*ROC – receiver operator characteristic –*). La Figura 4.1(a) presenta ejemplos de curvas *ROC* (la simbología *SSD*, *MAP* y *ML* se explicará en breve). El eje horizontal es la tasa de falsas alarmas (escalado logarítmicamente). El eje logarítmico enfatiza tasas pequeñas, las cuales son los puntos de operación de interés en los sistemas reales. El eje vertical es la tasa de detección e identificación. Cuando se presenta el rendimiento de un algoritmo, se deben señalar el tamaño de la galería y el de ambos conjuntos de pruebas.

En el caso general de la identificación de conjunto abierto, el sistema examina los n primeros empates entre una prueba y la galería. Se detecta e identifica una prueba de una persona en la galería con rango n si la prueba es de rango n o menor y el empate correcto se encuentra por encima del umbral de operación. Formalmente estas condiciones se expresan de la siguiente manera:

- $rango(p_j) \leq n$
- $s_{*j} \geq \tau$ para el valor de similitud donde $id(p_j) = id(g_*)$

La tasa de detección e identificación con rango n es la fracción de pruebas en el conjunto $\mathcal{P}_\mathcal{G}$ que son detectadas e identificadas correctamente con rango n. Esta tasa para el umbral de operación τ se define como:

$$P_{DI}(\tau, n) = \frac{|\{p_j : rango(p_j) \leq n, \text{ y } s_{*j} \geq \tau\}|}{|\mathcal{P}_\mathcal{G}|} \tag{4.3}$$

el cálculo de $P_{FA}(\tau)$ con rango n es igual que en el caso del rango 1.

El rendimiento de la identificación de conjunto abierto se puede graficar a lo largo de tres ejes: la tasa de detección e identificación, la tasa de falsas alarmas y el rango. Aunque el rendimiento se representa como una superficie en este espacio tridimensional, en la práctica se grafica como dos planos bidimensionales. La Figura 4.1(a) presenta ejemplos en los cuales el rango se mantiene en 1 y se muestra el intercambio entre las tasas de detección e identificación y de falsas alarmas. La Figura 4.1(b) presenta el otro formato en el que se reporta el desempeño. El eje vertical es la tasa de detección e identificación y el eje horizontal es el rango escalado logarítmicamente. Cada curva corresponde al rendimiento del mismo sistema a una tasa de falsas alarmas diferente.

4.1.2. Identificación de Conjunto Cerrado

Para el desempeño en la identificación de conjunto cerrado la pregunta no siempre es si el mejor empate es correcto, sino más bien, ¿la respuesta correcta se encuentra dentro de los primeros n empates?.

El primer paso para calcular el desempeño de la identificación de conjunto cerrado consiste en ordenar los valores de similitud entre p_j y la galería $\mathcal{G}$ y calcular los rango(p_j). La tasa de identificación para el rango n, $P_I(n)$, es la fracción de pruebas con rango n o menor. Para el rango n, sea:

$$C(n) = |\{p_j : rango(p_j) \leq n\}| \tag{4.4}$$

la cuenta acumulada del número de pruebas con rango n o menor. La tasa de identificación de rango n es:

$$P_I(n) = \frac{|C(n)|}{|\mathcal{P}_\mathcal{G}|} \tag{4.5}$$

Las funciones $C(n)$ y $P_I(n)$ son crecientes en n. La tasa de identificación de rango 1, $P_I(1)$, también se denomina *tasa de identificación correcta*, *tasa de mejor empate* o *tasa de mejor puntuación*.

El rendimiento de la identificación de conjunto cerrado se reporta en una curva *característica de empate acumulado* (*CMC –Cumulative Match Characteristic–*). Una *CMC* grafica $P_I(n)$ como una función del rango n. La Figura 4.1(c) muestra varias *CMC*. El eje horizontal es el rango en una escala logarítmica y el eje vertical es $P_I(n)$. Nótese que la curva correspondiente a la simbología *SSD* en la Figura 4.1(c) proviene de la evaluación del mismo sistema que generó las curvas *ROC* basadas en rangos de la Figura 4.1(b).

El rango 1 resume en la mayoría de los casos el rendimiento de este tipo de identificación aunque también se utilizan otros puntos como 5, 10 y 20. Al mismo tiempo la fortaleza y debilidad de una *CMC* es su dependencia en el tamaño de la galería, $|\mathcal{G}|$. Para observar esto se puede graficar el rendimiento de rango 1 contra la cardinalidad de la galería. Si se desea eliminar este efecto indeseado se grafica la tasa de identificación como porcentaje del rango, es decir, el desempeño cuando la respuesta correcta se encuentra dentro del 10 %, por ejemplo.

La identificación de conjunto cerrado es un caso particular de la identificación de conjunto abierto cuando el conjunto de pruebas $\mathcal{P}_\mathcal{N}$ se encuentra vacío y el umbral de operación $\tau = -\infty$. Este umbral de operación corresponde a una tasa de falsas alarmas de 1.0. Esto significa que $s_{*j} \geq \tau$ para todos los valores de empate y todos estos valores se reportan como falsas alarmas. Así, para cualquier n, $P_{DI}(-\infty, n) = P_I(n)$. Obsérvese que la curva en la Figura 4.1(b) con una tasa de falsas alarmas de 1.0 (la curva superior) es la *CMC* para la versión de conjunto cerrado de ese experimento (mostrada en la Figura 4.1(c) con la simbología *SSD*).

Una vez que se han definido y establecido los parámetros para la evaluación de un sistema biométrico, en los siguientes apartados se presenta tal evaluación aplicada al sistema desarrollado en el presente trabajo de tesis.

4.2. Experimentos de Detección y Reconocimiento

A fin de realizar los experimentos de detección y reconocimiento facial, es necesario contar con los conjuntos de imágenes que conformarán las galerías, ambos conjuntos de prueba (detallados en la sección anterior) y aquéllas que servirán de *entrenamiento* para los procesos automatizados dentro del sistema biométrico. También se debe considerar que el número de ejemplares influye en el *peso* estadístico de los resultados obtenidos. Por lo anterior y a fin de solventar de manera apropiada los requerimientos comentados se recurrió a dos bases de datos disponibles públicamente en internet, la base de datos de rostros del Centro Universitario de la *FEI* y la base de datos *CVL*; mismas que en el apéndice C se detallan en profundidad. Por dicha razón, en las secciones subsecuentes sólo se hará referencia a los citados conjuntos.

4.2.1. Experimentos de Reconocimiento con Alineación Manual

Con la finalidad de determinar separadamente el desempeño del sistema en los procesos de detección y reconocimiento, en los primeros experimentos realizados se alineó manualmente cada imagen facial, esto es, se indicó en forma manual la posición central de los ojos izquierdo y derecho en todas y cada una de las imágenes analizadas y, con esto, se realizó el procedimiento de reconocimiento completo descrito en el capítulo 3. Así, se obtuvo una estimación del desempeño del sistema en cuanto al reconocimiento *puro* se refiere.

Como se describe en el capítulo referido en el párrafo anterior, el proceso de reconocimiento consiste en la alineación geométrica, el enmascaramiento/recorte y la normalización de contraste de cada imagen de los rostros analizados. Luego se calculan las diferencias intrapersonales y interpersonales del conjunto de entrenamiento para crear cada subespacio correspondiente. El subespacio intrapersonal se crea con las diferencias entre las imágenes correspondientes a un mismo individuo dentro de todo el conjunto de entrenamiento (dos para la presente investigación). Mientras tanto, el subespacio interpersonal se crea con las diferencias entre imágenes correspondientes a diferentes individuos en el conjunto de entrenamiento (elegidas al azar). Finalmente, se proyectan en estos subespacios las imágenes (en su forma vectorial) de la galería, de los conjuntos de pruebas $\mathcal{P}_{\mathcal{G}}$ y $\mathcal{P}_{\mathcal{N}}$ y se calculan las curvas *ROC*, *CMC* y *ROC* basada en rangos con las fórmulas de los apartados 4.1.1 y 4.1.2.

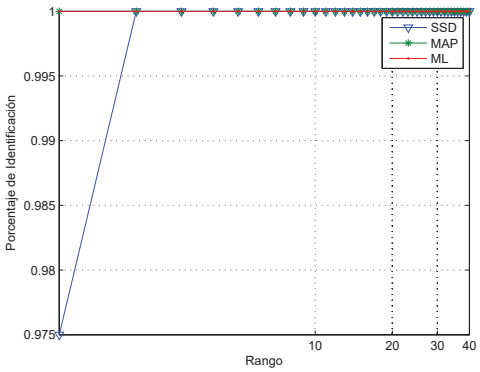
Los experimentos planteados en esta etapa se realizaron con un análisis de vali-

dación cruzada triple. Cada uno de los tres experimentos efectuados consistió en empatar correctamente las imágenes de los rostros colocados en el conjunto de prueba $\mathcal{P}_\mathcal{G}$ contra aquéllas colocadas en la galería $\mathcal{G}$, para obtener la capacidad de identificación del sistema; y en empatar los rostros colocados en el conjunto $\mathcal{P}_\mathcal{N}$ contra aquéllos en la galería, para obtener el porcentaje de falsas alarmas del sistema, puesto que las imágenes en $\mathcal{P}_\mathcal{N}$ no existen en $\mathcal{G}$. No hay que olvidar que existe también un subconjunto de imágenes que sirven de entrenamiento al sistema.

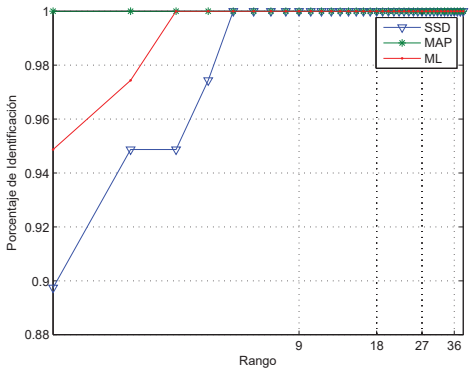
Para armar cada uno de los tres experimentos del análisis de validación cruzada triple, se repitió el siguiente procedimiento hasta por tres veces seguidas. Se seleccionó, en primer lugar, una de las dos imágenes correspondientes a cada individuo en la base de datos de rostros de la *FEI*, conformándose el subconjunto *FA* (en general se denomina de esta manera conforme a la nomenclatura utilizada en el protocolo de evaluación *FERET* [Phillips98, Phillips00]). Todas las imágenes restantes del par correspondiente a un mismo individuo conforman el subconjunto *FB* (también denominado de esta manera por el protocolo *FERET*). Como siguiente paso se crearon en forma *aleatoria* 5 particiones en ambos subconjuntos (sin solapamientos entre ellas y con correspondencia entre los elementos de las particiones de *FA* y las de *FB*, puesto que el objetivo final es empatarlos). Luego, se asignaron tres de estas particiones al proceso de entrenamiento, es decir, para la creación de los subespacios de diferencias intra y interpersonales (120 imágenes *FA* y 120 *FB*); una partición para la galería $\mathcal{G}$ (40 imágenes *FA*) y el conjunto de prueba $\mathcal{P}_\mathcal{G}$ (40 imágenes *FB*); y, la partición restante (40 imágenes *FB* dejando sin utilización las 40 *FA*), al conjunto de prueba $\mathcal{P}_\mathcal{N}$.

Para los tres experimentos, cuyos resultados se muestran en las Figuras 4.2, 4.3 y 4.4; se utilizan 35 vectores principales para definir el subespacio intrapersonal y tres veces más (105) para el subespacio interpersonal (ambos valores obtenidos mediante prueba y error, como balance entre eficacia y desempeño del sistema).

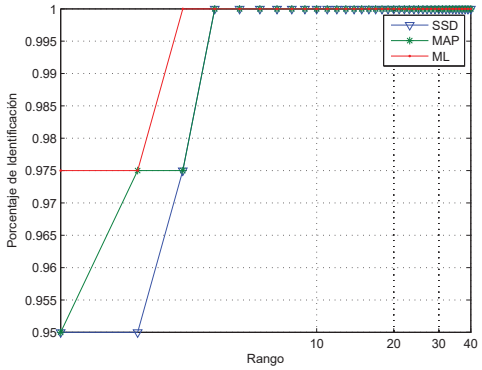
La Figura 4.2 presenta las curvas *CMC* resultantes de los tres experimentos de reconocimiento con alineación manual. Las tres subfiguras presentan curvas correspondientes a los estimadores de similitud bayesiana *máxima a posteriori* (*MAP*) –definido según la Ecuación 3.5– y de máxima verosimilitud (*ML – Maximum Likelihood*) –definido según la Ecuación 3.4–; ambos descritos en la sección 3.2. Como forma de calibración o comparación, se incluye un estimador de similitud de rostros basado en la suma de diferencias al cuadrado (*SSD – Sum of Squared Differences*); conocido también como vecino más cercano, distancia



(a) *Experimento 1*



(b) *Experimento 2*



(c) *Experimento 3*

Figura 4.2: Curvas *CMC* de los experimentos de reconocimiento con alineación manual. $|\mathcal{G}| = 40$, $|\mathcal{P}_{\mathcal{G}}| = 40$ y $|\mathcal{P}_{\mathcal{N}}| = 40$.

Euclidiana o correlación simple. Este estimador se define formalmente como:

$$SSD = \arg\max[-||\mathbf{\Delta}_i||^2] \tag{4.6}$$

donde, al igual que para los otros dos estimadores, $\mathbf{\Delta}_i = I_i - I_{Prueba}$, con $i = 1, 2, \ldots, N_C$ y N_C es el número de imágenes de rostros en la galería $\mathcal{G}$ (o base de datos).

El otro formato en que se presentan los resultados de las curvas *CMC* (las curvas *ROC* basadas en rangos) se muestran individualmente en la Figura 4.3 para cada uno de los estimadores *SSD*, *MAP* y *ML*, respectivamente; del segundo de los experimentos (el cual sirve para ejemplificar la proporción de resultados de cada estimador en los tres experimentos). En cada una de dichas gráficas se muestran diferentes curvas, mismas que

Experimento	*SSD*		*MAP*		*ML*	
1	0.60	0.60	0.88	0.88	0.78	0.80
2	0.41	0.44	0.85	0.89	0.82	0.85
3	0.73	0.75	0.93	0.95	0.93	0.93
Media	0.580	0.597	0.887	0.907	0.843	0.860
Desv. Estándar	0.161	0.155	0.040	0.038	0.078	0.066

Tabla 4.1: Resumen del desempeño en el reconocimiento con alineación manual. Se muestran los resultados a una tasa de falsas alarmas del 0 y del 5 %, respectivamente.

corresponden a diferentes tasas de falsas alarmas.

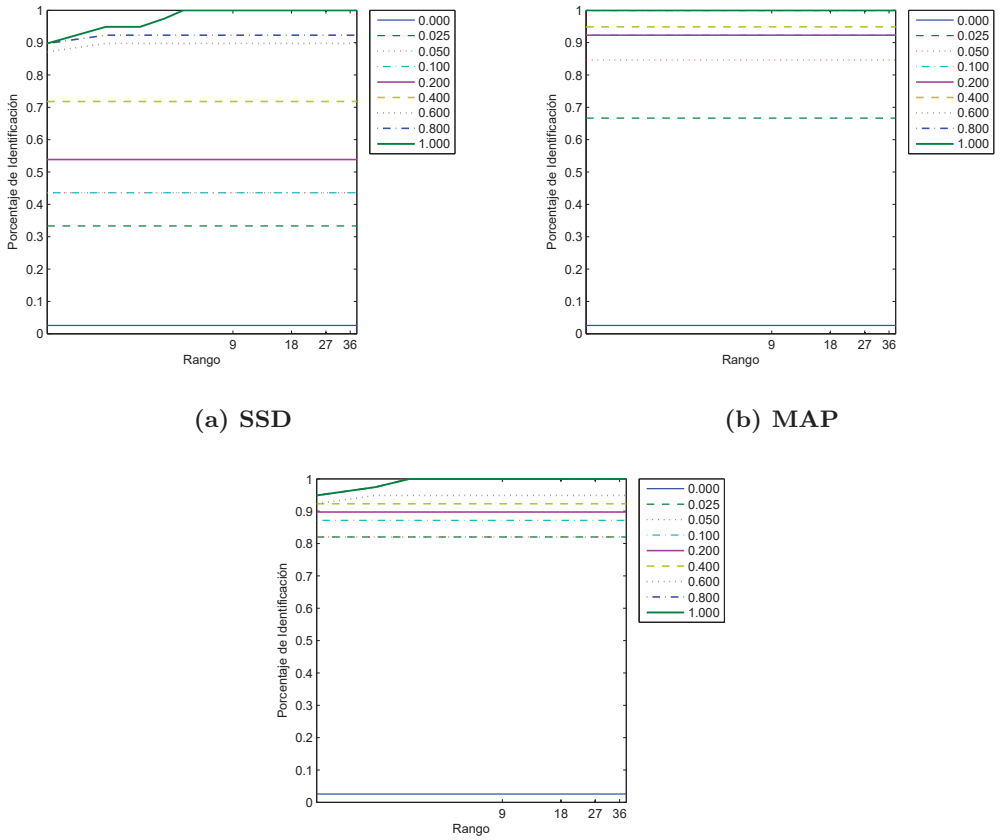
La Figura final de esta sección, 4.4, presenta las curvas *ROC* del desempeño del algoritmo de reconocimiento con los estimadores de similitud *SSD*, *MAP* y *ML* que corresponden a los tres experimentos. En el apartado siguiente se discuten brevemente los resultados de los experimentos hasta ahora descritos para continuar, en la siguiente subsección, con la experimentación de la parte de detección automática del sistema desarrollado.

Discusión

En primera instancia se observa de la Figura 4.2 que, efectivamente, el rendimiento del sistema desarrollado durante la presente investigación replica el desempeño reportado por Moghaddam y Pentland en [Moghaddam00] (para la base *FERET* completa), con un valor superior al 95 % de identificación de conjunto cerrado o *verificación* correcta de rostros de rango 1 para el método bayesiano con el estimador de similitud *MAP*. Como comparativa adicional, la Figura 4.5 presenta la gráfica *CMC* original de la evaluación en la competencia *FERET* de 1996 y 1997 (tomada de [Phillips98]).

Para resumir de mejor manera los resultados obtenidos de los experimentos anteriores, la Tabla 4.1 presenta la interpretación de las curvas *ROC* en cuanto al porcentaje de detección a una tasa de falsas alarmas del 0 y del 5 %, respectivamente; para el desempeño de los tres estimadores de similitud.

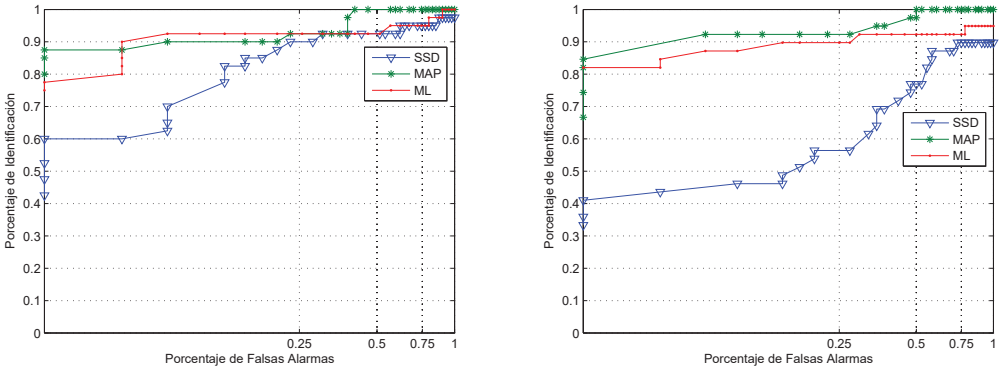
De la tabla referida se destaca que el método bayesiano con el estimador *MAP* supera al estimador de similitud *ML* y, significativamente, al *SSD*; contrario a lo expuesto por Teixeira en [Teixeira03] así como Wang y Tang en [Wang03]; los cuales reportan el efecto inverso (el estimador *ML* siempre supera al estimador *MAP*). A primera vista dicha contradicción se pudiera atribuir a la utilización, por parte del Dr. Moghaddam, de un subconjunto de aproximadamente 2,000 imágenes de rostros de la base de datos *FERET*



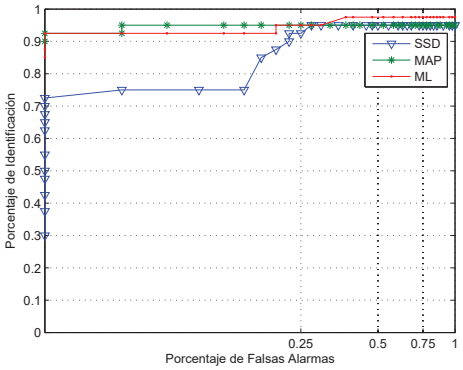
(a) SSD **(b) MAP**

(c) ML

Figura 4.3: Curvas *ROC* basadas en rangos correspondientes a los resultados del segundo experimento de reconocimiento con alineación manual (para los estimadores *SSD*, *MAP* y *ML*; respectivamente). Nótese que las curvas con una tasa de falsas alarmas de 1.0 (las curvas superiores) son las curvas *CMC* para la versión de conjunto cerrado de estos experimentos, mostradas en la Figura 4.2(b). $|\mathcal{G}| = 40$, $|\mathcal{P}_{\mathcal{G}}| = 40$ y $|\mathcal{P}_{\mathcal{N}}| = 40$.



(a) ***Experimento 1*** **(b)** ***Experimento 2***



(c) ***Experimento 3***

Figura 4.4: Curvas *ROC* correspondientes a los resultados de los tres experimentos de reconocimiento con alineación manual. $|\mathcal{G}| = 40$, $|\mathcal{P}_{\mathcal{G}}| = 40$ y $|\mathcal{P}_{\mathcal{N}}| = 40$.

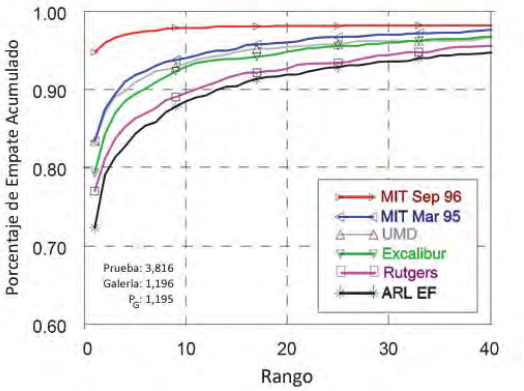


Figura 4.5: Curva *CMC* para las vistas frontales *FA/FB* en la competencia *FERET* de 1996. La curva superior etiquetada "MIT Sep 96" corresponde al método bayesiano de reconocimiento facial de Moghaddam y Pentland ([Phillips98]).

para la realización de los experimentos. Sin embargo, en las referencias citadas, Teixeira reporta la experimentación con el mismo conjunto de imágenes aunque Wang y Tang indican la utilización de la colección *FERET* pero diferentes subconjuntos. El otro motivo al que se le pudiera atribuir dicho comportamiento proviene del hecho de que Teixeira emplea el mismo método bayesiano pero con una *simplificación* del estimador *MAP* que, afirma, es equivalente al original.

Con relación al estimador de calibración, el *SSD*, su menor rendimiento se entiende del hecho de que crea divisiones *lineales* del subespacio de rostros en torno a las imágenes en la galería. Por tanto y con el fin de empatar apropiadamente las imágenes correspondientes a un mismo individuo, debe trabajar a una tasa muy alta de falsas alarmas. En contraposición, los otros dos estimadores consideran la forma *real* del espacio de rostros (no lineal y no convexa) al valorar la distribución de densidad completa en el espacio de los rostros característicos F y su complemento ortogonal $\bar{F}$ [Jain05].

Un señalamiento final importante que se debe realizar, es el hecho de que en la evaluación *FERET* original de 1996 y 1997 el enfoque principal de desempeño se atribuyó a la tasa de reconocimiento acumulado (las curvas *CMC* que, por esta exacta razón, se denominan también *curvas FERET*) y que los conceptos de falsas alarmas y curvas *ROC* se definieron/evaluaron a partir de la *Facial Recognition Vendor Test 2000 –Prueba de Vendedores de Tecnología de Reconocimiento Facial 2000–* (como se describe en [Blackburn01] y [Phillips03]) durante el año 2000. Por este motivo, el desempeño del método bayesiano

con el estimador de similitud *MAP*, propuesto por Moghaddam y Pentland, puede verse superado por propuestas más recientes diseñadas con esta evaluación en mente.

Establecidos los límites del rendimiento en el proceso de reconocimiento facial, es tiempo de hacer lo propio con el proceso de detección. El apartado siguiente detalla los experimentos relativos a dicho proceso.

4.2.2. Experimentos de Detección

Considerando que el efecto y desempeño real del proceso de detección automatizado se conoce solamente cuando se complementa con el proceso de reconocimiento, la experimentación de detección se reduce a determinar las curvas *ROC* correspondientes a la ubicación del rostro completo y del centro de ambos ojos. Como se detalló en la sección 2.7, los ojos son los únicos rasgos que se detectan (en este trabajo) y que, en realidad, se necesitan para alinear el rostro.

La Figura 4.6 presenta en una sola imagen las tres curvas citadas (la indicación de la posición del ojo –izquierdo o derecho– es relativa al sujeto, no al punto de vista del observador).

Para la creación de las curvas *ROC* correspondientes a los ojos, se debe definir primeramente un valor de distancia o de estimación de su posición dentro de la imagen (en el presente caso se emplea el estimador de máxima verosimilitud descrito en la sección 2.5). Calculado el valor del estimador para cada posible ubicación en la imagen, se establece un umbral τ en el que opere el sistema. Para los experimentos se define una detección como correcta si la ubicación señalada por el proceso automatizado genera un valor de distancia menor o igual al umbral, además de que dicha ubicación se localice dentro de un radio de n pixeles (en el presente trabajo 6 pixeles) tomando como punto de referencia la posición de los ojos señalada manualmente (tal y como sugieren Moghaddam et al. en [Moghaddam95a, Moghaddam95b]). Si la posición supera el límite de distancia impuesto se declara una falsa alarma, pero si es el valor del umbral el que se supera, no se define ni detección ni falsa alarma. Las curvas *ROC* mostradas en la Figura 4.6 se generan modificando el umbral τ. Cabe señalar que puesto que el protocolo *FERET* se enfoca más en el proceso de reconocimiento que en el de detección, estas curvas *ROC* están menos estandarizadas.

Complementariamente, para el rostro se emplea el mismo estimador de posición de máxima verosimilitud citado. La diferencia consiste en la forma de definir una detección

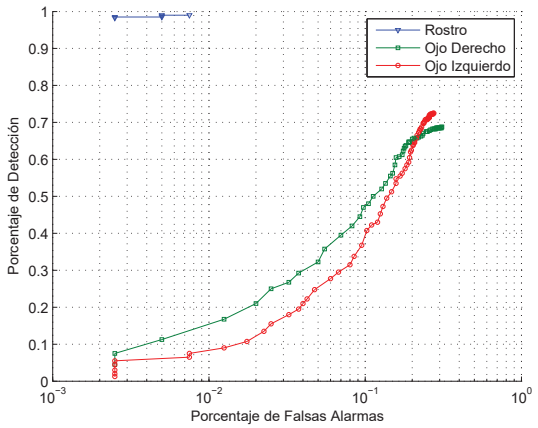


Figura 4.6: Curvas *ROC* correspondientes al proceso de detección automática del rostro y del centro de ambos ojos. Se muestran los resultados acumulados del proceso realizado para cada una de las 400 imágenes que conforman la base de datos de la *FEI*.

correcta. Tomando la subimagen rectangular que el proceso de detección señale como rostro, se definirá como apropiada la detección si ésta abarca la ubicación del centro de ambos ojos y la de la punta de la nariz (aunque la detección de la punta de la nariz no forme parte de los experimentos). Esta última coordenada también debe haberse indicado manualmente de forma previa. Las zonas que se deben abarcar incluyen no sólo la coordenada central de los ojos y de la punta de la nariz, sino también las subzonas rectangulares de cada uno de los rasgos. Para este trabajo de tesis el área correspondiente a cada uno de los ojos es de 40 pixeles de ancho por 40 de alto y para la nariz es de 24 por 24 pixeles; dimensiones que se calcularon manualmente para una escala determinada y establecida *a priori*.

Para la creación de las curvas en la Figura 4.6 se utilizaron como entrenamiento todas las imágenes de la base de datos *CVL*, es decir, se utilizaron para crear los subespacios de los *rostros característicos* (*eigenfaces*) así como de los *ojos característicos* (*eigeneyes*) que el estimador de máxima verosimilitud de la sección 2.5 requiere. Para la definición del subespacio de rostros se emplearon 21 valores y vectores característicos y, para el subespacio de ojos, se emplearon 31. De manera posterior, se estimó la pertenencia a la clase de objeto en específico de que se trata (el rostro, el ojo derecho o el ojo izquierdo) para *cada una* de las 400 imágenes originales de la colección completa de la *FEI* Brasileña. Dicho en otras palabras, con el proceso de búsqueda multiescala descrito en la sección 2.5 se detecta el rostro en el total de las imágenes de la base de datos de la *FEI* Brasileña para luego, dentro

de éste y con el mismo procedimiento, detectar el centro correspondiente a cada uno de los ojos (izquierdo y derecho). Teniendo el rostro segmentado a la par que su tamaño, se procede a normalizar la imagen hacia una escala preestablecida. Finalmente, utilizando la ubicación de los ojos se alinea verticalmente el rostro localizado. El resultado se enmascara, ecualiza y se envía al proceso de reconocimiento.

También de la Figura 4.6 se observa que la detección del rostro llega a un máximo porcentaje del 99 %, acompañado del 73 y 69 % para el ojo izquierdo y derecho, respectivamente. Lo anterior haría pensar que el proceso automático completo de detección y reconocimiento facial no debiera decrementar su desempeño grandemente en comparación a los resultados mostrados en la sección 4.2.1 para el reconocimiento con alineación manual en específico pero, como se detalla en la siguiente sección, los experimentos correspondientes muestran lo contrario.

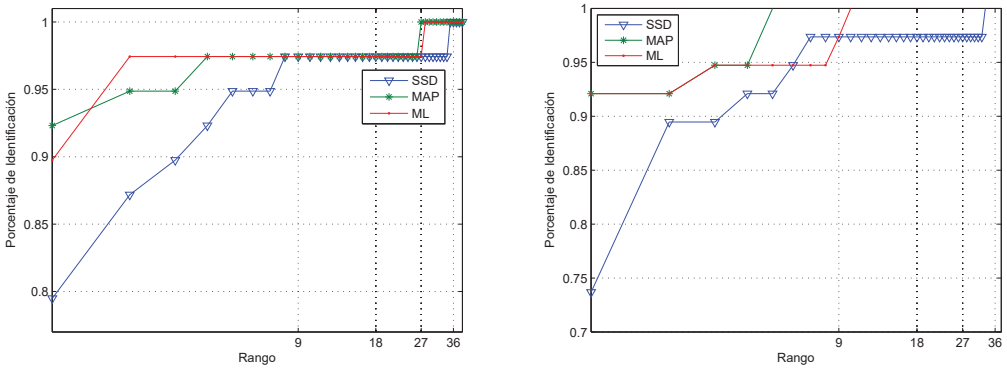
4.2.3. Experimentos de Detección y Reconocimiento

Para evaluar el desempeño completo del sistema desarrollado (detección + reconocimiento automáticos), repetiremos los tres mismos experimentos del reconocimiento con alineación manual; con la obvia modificación de que *todas* las imágenes de la colección de la *FEI* se pasarán primeramente por el detector de rostros. Como se menciona en el apartado anterior, se emplean 21 vectores característicos para la detección del rostro y 31 para la detección de los ojos.

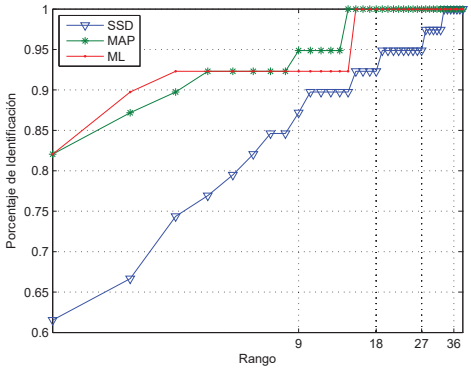
Teniendo las imágenes resultantes del procedimiento de detección (con el rostro segmentado, escalado, alineado, enmascarado y ecualizado), se procede a crear los nuevos conjuntos FA/FB y sus cinco particiones aleatorias. Con las particiones creadas, se reparten en los nuevos conjuntos de entrenamiento, las nuevas galerías y los también nuevos conjuntos de pruebas $\mathcal{P}_{\mathcal{G}}$ y $\mathcal{P}_{\mathcal{N}}$. Estos escenarios corresponden a cada uno de los tres nuevos experimentos.

Dado lo anterior, se presentan los resultados de estas nuevas pruebas en forma de curvas *CMC* en la Figura 4.7 y, en segunda instancia, las correspondientes curvas *ROC* en la Figura 4.8.

Se analiza en primer lugar que el porcentaje de identificación de rango 1, en las curvas *CMC*, se decrementa en un promedio del 10 % para los estimadores *MAP* y *ML* (nuevamente superando por muy poco margen el primero al segundo) y, para el estimador *SSD*

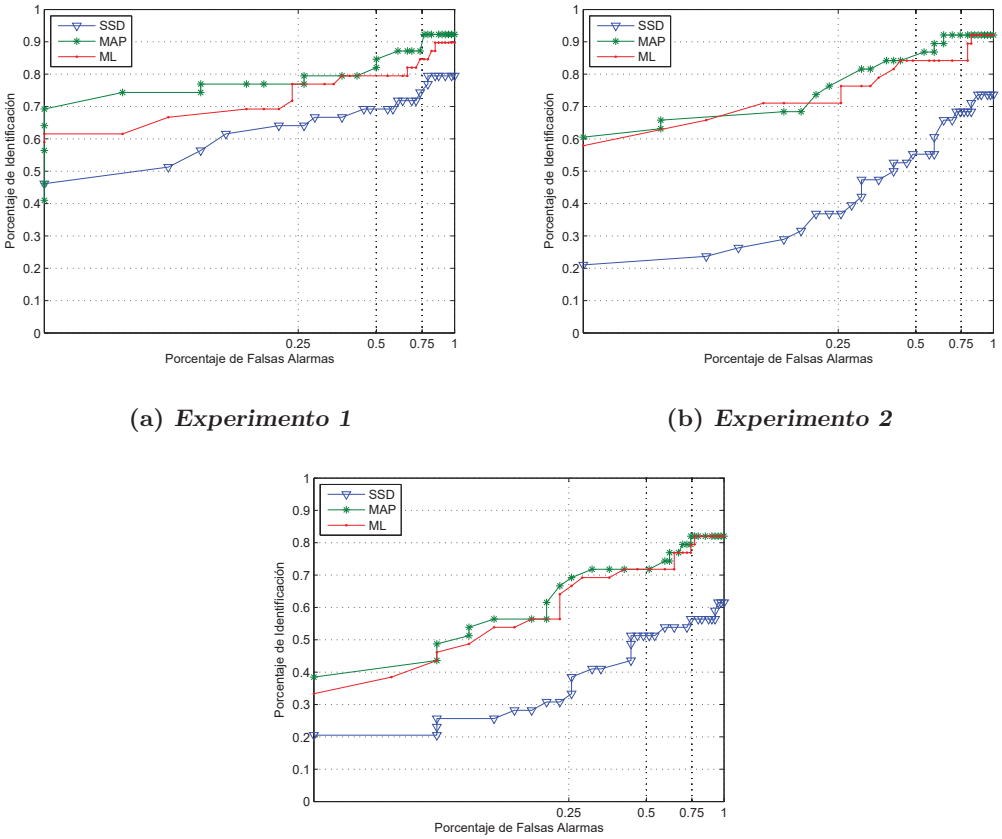


(a) ***Experimento 1*** **(b)** ***Experimento 2***



(c) ***Experimento 3***

Figura 4.7: Curvas *CMC* de los experimentos de detección y reconocimiento automáticos. $|\mathcal{G}| = 40$, $|\mathcal{P}_{\mathcal{G}}| = 40$ y $|\mathcal{P}_{\mathcal{N}}| = 40$.

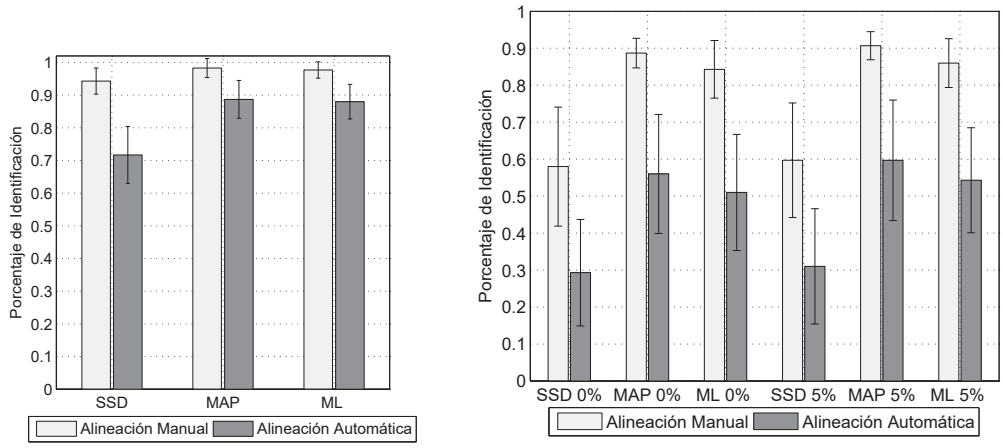


(a) ***Experimento 1***

(b) ***Experimento 2***

(c) ***Experimento 3***

Figura 4.8: Curvas *ROC* de los experimentos de detección y reconocimiento automáticos. $|\mathcal{G}| = 40$, $|\mathcal{P}_{\mathcal{G}}| = 40$ y $|\mathcal{P}_{\mathcal{N}}| = 40$.



(a) ***Comparativa de curvas*** **CMC** **(b)** ***Comparativa de curvas*** **ROC**

Figura 4.9: Comparación de resultados en las curvas *CMC* y *ROC* de los experimentos de detección y reconocimiento con alineación manual y completamente automáticos realizados. La comparativa de las curvas *ROC* se presenta para las tasas de falsas alarmas del 0 y del 5 %, respectivamente.

de comparación, poco más del 20 %. Esto se desvía bastante de lo reportado por Moghaddam y Pentland en [Moghaddam95a], donde informan que el rendimiento sólo debiera bajar, como máximo, en un 3 % de la tasa de identificación. Considerando que los conjuntos de imágenes empleados en este trabajo son diferentes a los empleados en la referencia citada, hasta cierto punto esto resultaría en cierta forma aceptable, pero el detrimento importante proviene de la evaluación de la identificación de conjunto abierto, es decir, de las curvas *ROC*. Para una mucha mejor apreciación de este efecto, se presenta el resumen de tasas de identificación y falsas alarmas en la Figura 4.9, tanto para las curvas *CMC* como para las *ROC*. Se debe observar que los valores presentados en la Figura corresponden al análisis de validación cruzada triple, esto es, se presentan los valores promedio de los tres experimentos de reconocimiento con alineación manual así como de los tres experimentos con detección y reconocimiento automáticos (la Figura también presenta los errores/desviación estándar para cada uno de los seis experimentos).

Conforme a lo dicho, el desempeño de los algoritmos automatizados disminuye un promedio de poco más del 30 % con relación a sus contrapartes con el alineamiento manual. Como la disminución es constante para los tres estimadores, la relación de superioridad $MAP \rightarrow ML \rightarrow SSD$ se mantiene. Se consideran varias explicaciones para este resultado

tan negativo. La primera proviene del hecho de que Moghaddam y Pentland reportan sus resultados ([Moghaddam95a]) considerando en sus experimentos 2,000 imágenes de la base de datos de rostros *FERET*, pero sólo indican el resultado de un 97 % de detección correcta *para el rostro* que, si se observa la Figura 4.6, resulta análogo al casi 99 % de detección correcta del rostro para todas las imágenes de la colección de la *FEI*, empleada en este trabajo de tesis. La segunda explicación proviene del hecho de que en el experimento en que sí se reporta la detección automática agregada de los ojos, la nariz y el centro de la boca; se utilizaron exclusivamente imágenes correspondientes a 155 individuos, como subconjunto del total de 1,199 sujetos cuyas imágenes conforman la base *FERET*, lo que pudiera indicar un *mínimo local* con relación al desempeño del algoritmo en este subconjunto en específico.

Dado que las curvas *ROC* se definieron y evaluaron por el *NIST* para este tipo de experimentos a partir del año 2000 ([Blackburn01]), no se tienen las curvas de dicho tipo generadas por los experimentos de Moghaddam y Pentland, lo cual impide comprobar en realidad si el efecto negativo es del algoritmo o de la implementación realizada durante la presente investigación (recuérdese lo relativo al detrimento de un solo 10 % de los resultados de las curvas *CMC*, que sí reportan Moghaddam y Pentland).

La última explicación, que se estima como la más probable, consiste en la dependencia del proceso de búsqueda multiescala en cuanto al número y separación lineal de las escalas en las que se buscan el rostro y ambos ojos. En los experimentos efectuados, probando con 100 escalas diferentes, la detección correcta de los ojos llega a subir hasta en un 10 %; pero con el gran problema de incrementar hasta en 10 veces el tiempo de cálculo requerido para el proceso. Esto indica que una mayor indagación en esta dirección permitiría obtener mejores resultados (tratando con el problema de la velocidad de ejecución, por supuesto). La próxima sección presenta un breve resumen del desempeño del sistema desarrollado, con relación a los tiempos de ejecución del proceso de detección y reconocimiento automático.

4.2.4. Evaluación de Tiempos de Ejecución

Se completa el capítulo de evaluación del sistema presentando las mediciones de tiempo de ejecución observadas durante los experimentos de detección y reconocimiento facial automatizados. Estas mediciones se realizaron en un equipo Intel Core© 2 Duo a 2.0 GHz., utilizando como ambiente de programación el software Matlab© versión 7.6.0.

En segundo término se señala que las imágenes de entrada al sistema son de 480

Rasgo Buscado	Imágenes de Entrenamiento	
	226	452
Rostro	5.5 segundos	10 segundos
Ojo Izquierdo	1.5 segundos	3 segundos
Ojo Derecho	1.5 segundos	3 segundos

Tabla 4.2: Tiempo de cálculo de los *PCA* correspondientes a la detección del rostro, del ojo izquierdo y del ojo derecho.

Escalas del Rostro	Escalas de Ambos Ojos	
	3	30
5	3 segundos	8.5 segundos
50	23.5 segundos	28.5 segundos

Tabla 4.3: Tiempos de duración del proceso de detección facial automática.

pixeles de alto por 640 pixeles de ancho. Dentro de éstas se intenta localizar un rostro cuyas proporciones se presumen de 150 pixeles de alto por 100 pixeles de ancho. Determinado el rostro se procede a detectar ambos ojos dentro del recuadro facial encontrado. La imagen correspondiente a cada uno de los ojos se presume de 40 pixeles de alto por 40 pixeles de ancho.

Aunque el número de componentes principales (rostros u ojos característicos) empleados en los experimentos del presente capítulo para la detección del rostro y de cada uno de los ojos es de 31, la Tabla 4.2 presenta los tiempos necesarios para realizar el *PCA* correspondiente a las imágenes del conjunto de entrenamiento de cada objeto citado (rostro, ojo izquierdo u ojo derecho), considerando el cálculo fijo de 256 componentes principales.

A continuación la Tabla 4.3 presenta los tiempos de duración del proceso automático de detección facial. Se muestra el número de escalas empleadas para la localización del rostro y las empleadas para la ubicación de ambos ojos. Los tiempos presentados corresponden al periodo de tiempo que transcurre desde que ingresa la imagen de entrada al sistema (de 480 por 640 pixeles) hasta que se entrega la imagen del rostro normalizada geométricamente, normalizada con respecto al contraste y, finalmente, enmascarada (según se ha descrito a lo largo de esta tesis).

Con relación al proceso de reconocimiento, la Tabla 4.4 presenta los tiempos necesarios para realizar los *PCA* de las diferencias intrapersonales e interpersonales empleados en el cálculo de los estimadores de similitud *MAP* y *ML*. El número de pares de imágenes

Imágenes de Entrenamiento	Diferencias Intrapersonales	Diferencias Interpersonales	Tiempo de Cálculo Total
60 pares	60	240	80 segundos
300 pares	300	$1,200$	445 segundos

Tabla 4.4: Tiempo de cálculo de los *PCA* correspondientes al proceso de reconocimiento.

Imágenes a Comparar	Estimador		
	SSD	*MAP*	*ML*
80 pares	1.5 segundos	7.5 segundos	3 segundos
200 pares	4.5 segundos	46 segundos	15 segundos

Tabla 4.5: Tiempos de duración del proceso de reconocimiento de rostros.

del rostro de un mismo individuo implica el mismo número de imágenes de diferencias intrapersonales, pero implica que el número de imágenes de diferencias interpersonales, como se maneja en los experimentos, es del cuádruple. Asimismo esta Tabla 4.4 presenta los tiempos necesarios para calcular, de forma fija, 256 componentes principales para cada uno de los tipos de diferencias.

Finalmente, en la Tabla 4.5 se indican los lapsos de tiempo requeridos para realizar el proceso de reconocimiento, comparando las imágenes en los conjuntos de pruebas $\mathcal{P}_{\mathcal{G}}$ y $\mathcal{P}_{\mathcal{N}}$ contra las imágenes faciales en la galería $\mathcal{G}$.

Merece la pena hacer la observación de que los número de imágenes en las Tablas 4.2, 4.3, 4.4 y 4.5 son consecuencia del diseño de los experimentos de la validación cruzada triple descrita en la sección 4.2.1. Las escalas empleadas en el proceso de detección del rostro así como de los ojos se establecieron heurísticamente, como balance entre eficacia y desempeño del sistema.

La parte final del presente capítulo presenta las conclusiones relativas a la experimentación realizada.

4.3. Conclusiones

De los experimentos realizados en el presente capítulo se puede deducir que los estimadores de probabilidad *máxima a posteriori* y de máxima verosimilitud, como se definen en los capítulos 2 y 3, modelan en forma no lineal (cuadrática) la distribución de densidad de probabilidad de las imágenes de rostros de los conjuntos de prueba utilizados; resultando

ser la forma óptima de empate de rostros, bajo la suposición Gaussiana de la distribución de densidad de las imágenes empleadas. Por tal motivo, el desempeño del reconocimiento fundamentado en dichos estimadores supera, con mucho, las posibilidades de empatamiento que la correlación simple pudiera tener (el estimador *SSD*). Incluso aún y cuando las imágenes estén perfectamente alineadas, escaladas, enmascaradas para resaltar la zona de comparación y normalizadas respecto al contraste.

Por otra parte, el soporte que el subespacio de diferencias interpersonales proporciona adicionalmente al estimador *MAP*, por sobre el subespacio de diferencias intrapersonales del estimador *ML*, le permiten un mayor discernimiento entre diferentes identidades que, con todo el preprocesamiento realizado, lucen de una forma más que similar. Por este motivo el resultado en el desempeño entre ambos métodos de comparación conlleva una diferencia generalizada del 5 % (también reportada por Moghaddam y Pentland en [Jain05]), con la salvedad de que el cálculo del estimador *MAP* presenta un costo computacional del doble con relación al cálculo del estimador *ML* (por la proyección de imágenes en ambos subespacios).

En cuanto a la detección automática del rostro y de sus rasgos internos (los ojos), también experimentalmente se pudo comprobar la gran dependencia que presenta el proceso de búsqueda multiescala en cuanto al número y separación lineal de éstas (las escalas), mismas que generan las diferentes versiones de las imágenes de entrada sobre las que se calcula el estimador de posición de máxima verosimilitud (sección 2.5). Desafortunadamente, para optimizar dicha búsqueda generando imágenes en un mayor número de escalas, el incremento en el costo computacional del cálculo puede llegar a ser prohibitivo (efecto que se agrava si el proceso se realiza para el rostro y todos los rasgos internos que se empleen – ojos, boca, nariz, línea del cabello, etc. –).

Finalmente, una conclusión aparentemente obvia es la vinculación del desempeño de un sistema referido al conjunto de imágenes utilizado. Para tal fin, las curvas de las *características recibidas de operación – ROC –*, resultan de una mayor utilidad en la evaluación del desempeño del sistema que las curvas de *empate acumulado – CMC –*. Éstas últimas ya parecen haber sido sobrepasadas por el desempeño de los sistemas de detección y reconocimiento de rostros más recientes.

Capítulo 5

Conclusiones

En este trabajo de tesis se ha implementado un sistema para la identificación de individuos basado en la detección y el reconocimiento automáticos de sus rostros, según la metodología propuesta por Moghaddam et al. en [Moghaddam95a] y [Moghaddam96]. Para tal fin, la densidad de probabilidad del rostro y de ambos ojos se modelan utilizando una distribución de densidad Gaussiana multivariada en un espacio de alta dimensión. Descomponiendo este espacio en dos subespacios complementarios, generados por los vectores característicos de la matriz de covarianzas de un conjunto de datos de entrenamiento, se puede calcular un estimador de máxima verosimilitud de pertenencia a la clase de objeto buscado; ya sea el rostro, el ojo derecho o el izquierdo. El primer subespacio se genera con los primeros M vectores característicos de la matriz, correspondientes a los M valores característicos de mayor magnitud. El segundo subespacio se genera con los vectores característicos restantes pero, dado el número tan elevado de éstos que resultaría aún para imágenes pequeñas, se plantea su estimación a través de los valores y vectores característicos del subespacio principal.

Con la dependencia del estimador de detección de máxima verosimilitud en la escala de los objetos de entrenamiento, la imagen de entrada/búsqueda se escala a un determinado conjunto de tamaños predefinidos y espaciados linealmente para, individualmente, realizar la búsqueda dentro de cada versión de la misma. El valor mínimo global del estimador en todas estas imágenes indicará el objeto *detectado*. Un gran inconveniente encontrado con esta búsqueda multiescala, consiste en la relación directamente proporcional entre el número de escalas de búsqueda, el tiempo de cálculo y su efectividad de localización correcta.

Determinada la ubicación del rostro y la de ambos ojos, la subimagen facial correspondiente se extrae de la escena original, se normaliza geométricamente de acuerdo a la escala encontrada y tomando como ángulo de alineación horizontal/vertical el establecido por la posición de los ojos. Posteriormente se enmascara la imagen para eliminar las zonas periféricas de menor importancia para el proceso de identificación y, finalmente, se le realiza una normalización de contraste. La imagen resultante será enviada al procedimiento de reconocimiento facial.

Con relación al proceso de reconocimiento facial, el estimador de máxima verosimilitud se emplea para calcular una medida de similitud basada en un análisis Bayesiano de diferencias de imágenes. Se modelan dos clases mutuamente exclusivas de variación entre dos imágenes de rostros: *intrapersonales* (variaciones en la apariencia de un individuo, debidas a cambios en la expresión o en la iluminación) y *interpersonales* (variaciones en la apariencia debidas a diferencias en la identidad). Las funciones de densidad de probabilidad Gaussiana de alta dimensión para cada una de estas clases se obtienen de un conjunto de datos de entrenamiento, empleando la descomposición en espacios característicos antes señalada y, contando con ambas densidades, se calcula una medida de similitud Bayesiana de la probabilidad *a posteriori* de pertenencia a la clase de diferencias *intrapersonales*. Esta medida de similitud permite empatar el rostro de prueba contra aquellos existentes en una base de datos previamente recolectada.

Los algoritmos anteriores se eligieron para su implementación por ser la aplicación que presentó los mejores resultados en la competencia *FERET* de los años 1996 y 1997, realizada por el Instituto Estadounidense de Normas y Tecnología (*NIST*), tal y como se describe en [Phillips98]. Aún cuando los trabajos de investigación y experimentación se iniciaron empleando el método de las *cascadas de Haar*, propuesto por Viola y Jones [Viola01], tal método presentó un rendimiento muy pobre de detección correcta del rostro menor al 80 %, claramente superado por el 99 % del método de Moghaddam.

Desafortunadamente el desempeño reportado en la prueba *FERET* se realizó considerando únicamente la identificación de conjunto cerrado, es decir, el proceso de verificación de identidad. Para el proceso general de identificación de conjunto abierto el desempeño global del sistema, en nuestra implementación, decayó muy notoriamente. No obstante lo anterior, si el área en la que se desea aplicar el algoritmo de identificación basado en el estimador de similitud Bayesiana no depende de la detección automática del rostro en tiempo real, los resultados obtenidos serán muy buenos. Ejemplo de esta aplicación puede

ser la búsqueda en una base de datos de fotografías de familiares, personas extraviadas, antecedentes criminales, etc., en la cual se alinee manualmente el rostro (por ejemplo indicando el centro de ambos ojos como se realiza en nuestra implementación) como primer paso del proceso y, luego, se efectúe el proceso de búsqueda/reconocimiento como tal.

Se tiene conocimiento de métodos desarrollados en forma posterior a 1997 que superan el desempeño del método de Moghaddam. Ejemplo claro de esto son los algoritmos participantes en las competencias *FERET* de los años 2000 y 2002, denominadas *Prueba de Vendedores de Tecnología de Reconocimiento Facial* (*FRVT –Facial Recognition Vendor Test–*), descritas en [Blackburn01] y [Phillips03]. De forma lamentable y como la parte de *vendedores* implica, dichos algoritmos se han desarrollado bajo secreto comercial sin posibilidad de acceso público, lo cual no permitió su estudio y prueba durante nuestros trabajos de investigación.

5.1. Trabajo Futuro

Se tienen varias líneas de investigación para exploración futura. Como se aprecia claramente de la Figura 4.9, el proceso de detección automática es el área de oportunidad para mejorar el desempeño del sistema. En tal sentido y observando también el buen desempeño del estimador Bayesiano de probabilidad posterior, resulta natural extender el estimador de posición (detector) basado en la verosimilitud empleando la noción de una clase "no rostro" $\bar{\Omega}$, obteniendo mapas de notabilidad *a posteriori* de la forma:

$$P(\Omega|\mathbf{x}) = \frac{P(\mathbf{x}|\Omega)P(\Omega)}{P(\mathbf{x}|\bar{\Omega})P(\bar{\Omega}) + P(\mathbf{x}|\Omega)P(\Omega)}$$

donde ahora se utiliza una regla *MAP* para estimar la posición y la escala del rostro (o de los ojos, en su caso). Esto se puede ver como un enfoque probabilístico que emplea ejemplos *positivos* y *negativos*.

Otra forma de mejorar el pobre rendimiento de la parte de detección es investigar el comportamiento de los subespacios de características (del rostro, del ojo derecho y/o del ojo izquierdo) con conjuntos de entrenamiento a diferentes escalas. Si el análisis de componentes principales se puede hacer *piramidal* o *bidimensional* en cuanto a la clase de pertenencia de objeto buscado y a sus diferentes escalas, el proceso de detección se podría realizar de forma efectiva sin un costo computacional tan alto.

Aunque de una dificultad analítica considerable, los algoritmos de detección y de reconocimiento facial se mejorarían bastante al momento de determinar los valores óptimos para la gran cantidad de parámetros cuyos valores afectan el proceso y que deben elegirse de manera eminentemente heurística: el número de valores y vectores característicos del subespacio principal; el número, la resolución y la escala de las imágenes de los diferentes conjuntos de entrenamiento; el tamaño comparativo del subespacio de diferencias interpersonales con relación al de diferencias intrapersonales, etc.

Dado que el problema principal que se encontró durante la experimentación resultó ser la determinación de la posición exacta del centro de ambos ojos, derivando en una alineación incorrecta del rostro y, por lo tanto, en fallas en el reconocimiento, se puede modelar la distribución de densidad de probabilidad de la zona ocular completa que, en escala de grises, siempre será un centro obscuro sobre fondo claro. Esto generará una mayor precisión de la ubicación relativa entre ambos ojos y, al final, del alineamiento buscado.

Aunado a las posibilidades de investigación anterior, cabe señalar que para mejorar el desempeño global del sistema así como también para extenderlo a imágenes de rostros cuya distribución de densidad de probabilidad no sea Gaussiana, se puede implementar el método que el propio Moghaddam et al. plantean en la referencia [Moghaddam95a], base de esta investigación de tesis. En dicha propuesta se plantea estimar (aunque no de manera óptima como para el caso Gaussiano unimodal) la distribución de densidad de probabilidad completa de los objetos en el espacio de características (la denominada *DIFS*) como un modelo de densidad de probabilidad en términos de una mezcla de distribuciones de densidad Gaussianas, calculadas empleando el algoritmo de *Expectación - Maximización – Expectation - Maximization –* o *EM*.

Por último y tomando como punto de partida la detección correcta del rostro (cuyo desempeño, ya se ha visto, es del orden del 99 % para los experimentos realizados), se puede plantear el mecanismo de alineación como un problema de *registro de imágenes*. En tal sentido el registro se realizaría entre el rostro detectado y la cara promedio del conjunto de entrenamiento. Más aún, se puede establecer el cálculo de la verosimilitud como guía del proceso de registro, a fin de considerar variaciones en la escala y en la rotación del rostro.

Apéndice A

Análisis de Componentes Principales

Para el desarrollo del presente apéndice se presuponen conocimientos básicos de álgebra lineal; estudio previo de los conceptos de vectores, matrices, valores y vectores característicos, espacios y subespacios (vectoriales), dimensión, ortogonalidad, las operaciones básicas de vectores y matrices, en fin, la base primordial del álgebra lineal. Para una introducción a estos temas se pueden consultar las referencias [Román93] y [Poole04]. El material que a continuación se expone se fundamenta principalmente en [Jain05] y [Shlens05], con contribuciones menores de [Smith02] y [Yambor00].

A.1. Introducción

El análisis de componentes principales (*PCA - Principal Component Analysis*) [Joliffe86] es una técnica de reducción de dimensión basada en el cálculo del número deseado de *componentes principales* de una serie de datos multidimensionales. El primer componente principal se conforma de la combinación lineal de las dimensiones originales que tiene la máxima variación; el n-ésimo componente principal se conforma de la combinación lineal con la máxima variación *sujeta* a que sea ortogonal a los $n-1$ primeros componentes principales.

La idea del *PCA* se ilustra en la Figura A.1(c); el eje etiquetado como ϕ_1 corresponde a la dirección de máxima variación y se selecciona como el primer componente principal. En el caso de dos dimensiones, el segundo componente se determina entonces de

manera unívoca por las restricciones de ortogonalidad; en un espacio de mayor dimensión el proceso de selección continuaría, guiado por la variación de las proyecciones.

El *PCA* se relaciona de manera cercana con la transformación de Karhunen - Loève (*KLT – Karhunen - Loève Transform*) [Loève55], la cual se obtuvo en el contexto de procesamiento de señales como la transformación ortogonal con la base (vectorial) $\mathbf{\Phi} = [\phi_1, \ldots, \phi_N]^T$ que para cualquier $k \leq N$ minimiza el *error de reconstrucción* L_2 promedio para los puntos de datos $\mathbf{x}$:

$$\epsilon(\mathbf{x}) = \left\| \mathbf{x} - \sum_{i=1}^{k} (\phi_i^T \mathbf{x}) \phi_i \right\| \tag{A.1}$$

Se puede demostrar [Gerbrands81] que, bajo la suposición de que los datos están centrados respecto a la media (que tienen media 0), la formulación del *PCA* y de la *KLT* son idénticas. Sin pérdida de generalidad, a partir de este instante se asumirá que los datos en realidad están centrados respecto a la media, esto es, que el rostro promedio $\overline{\mathbf{x}}$ siempre se substrae de los datos.

Los vectores que conforman la base en la *KLT* se pueden calcular de la siguiente manera: Sea $\mathbf{X}$ la matriz de datos de $N \times k$ elementos cuyas columnas $\mathbf{x}_1, \ldots, \mathbf{x}_k$ son *observaciones* de una señal perteneciente a $\mathbb{R}^N$; en el contexto del reconocimiento facial, k es el número disponible de imágenes de rostros y $N = mn$ es el número de pixeles en una imagen. La base *KLT* $\mathbf{\Phi}$ se obtiene resolviendo el problema de valores característicos $\mathbf{\Lambda} = \mathbf{\Phi}^T \mathbf{\Sigma} \mathbf{\Phi}$, donde $\mathbf{\Sigma}$ es la matriz de covarianza de los datos:

$$\mathbf{\Sigma} = \frac{1}{k} \sum_{i=1}^{k} \mathbf{x}_i \mathbf{x}_i^T = \frac{1}{k} \mathbf{X} \mathbf{X}^T \tag{A.2}$$

$\mathbf{\Phi} = [\phi_1, \ldots, \phi_k]^T$ es la matriz de vectores característicos de $\mathbf{\Sigma}$ y $\mathbf{\Lambda}$ es la matriz diagonal con los valores característicos $\lambda_1 \geq \ldots \geq \lambda_N$ de $\mathbf{\Sigma}$ en su diagonal principal, de tal forma que ϕ_j es el vector característico correspondiente al j-ésimo valor característico más grande. Luego entonces se puede demostrar que el valor característico λ_i es la variación de los datos proyectados en ϕ_i.

Así pues, para realizar el *PCA* y extraer los M componentes principales de los datos, éstos se deben proyectar al subespacio $\mathbf{\Phi}_M$, las primeras M columnas de la base *KLT* $\mathbf{\Phi}$ que corresponden a los M valores característicos de mayor magnitud de $\mathbf{\Sigma}$ (a fin de reducir la dimensión N del espacio original a la dimensión M del subespacio calculado

Originales		Centrados	
x	y	x'	y'
2.50	2.40	0.69	0.49
0.50	0.70	-1.31	-1.21
2.20	2.90	0.39	0.99
1.90	2.20	0.09	0.29
3.10	3.00	1.29	1.09
2.30	2.70	0.49	0.79
2.00	1.60	0.19	-0.31
1.00	1.10	-0.81	-0.81
1.50	1.60	-0.31	-0.31
1.10	0.90	-0.71	-1.01
$\mu_x = 1.81;\ \mu_y = 1.91$			

Tabla A.1: Datos para el análisis de componentes principales de ejemplo.

por el proceso *KLT/PCA*). Esto se puede ver como una proyección lineal $\mathbb{R}^N \rightarrow \mathbb{R}^M$, la cual retiene la máxima energía (i.e., variación) de la señal. Otra propiedad importante del *PCA* es que *elimina la correlación* de los datos: la matriz de covarianza de $\mathbf{\Phi}_M^T\mathbf{X}$ *siempre* es diagonal.

Las propiedades principales del *PCA* se resumen en lo siguiente:

$$\mathbf{x} \approx \mathbf{\Phi}_M\mathbf{y},\ \mathbf{\Phi}_M^T\mathbf{\Phi}_M = \mathbf{I},\ E\{y_i y_j\}_{i\neq j} = 0 \tag{A.3}$$

a saber, reconstrucción aproximada, ortonormalidad de la base $\mathbf{\Phi}_M$ y componentes principales sin correlación $y_i = \phi_i^T\mathbf{x}$, respectivamente.

A.1.1. Ejemplo de Aplicación del *PCA*

Enseguida se desarrolla un ejemplo del análisis de componentes principales utilizando los datos minimalistas de la Tabla A.1. El primer paso consiste en substraer la media de cada dimensión de los correspondientes datos que la conforman, esto es, restar la media de las x de cada dato en la primera columna de la tabla y la media de las y de cada dato en la segunda columna. El resultado se muestra en la tercera y cuarta columna de la misma tabla. Así mismo se ilustran en forma gráfica tanto los datos originales como los centrados con respecto a la media en las Figuras A.1(a) y A.1(b), respectivamente.

El segundo paso consiste en calcular la matriz de covarianzas de los datos centrados

la cual, sin sorpresa alguna y dado que se están empleando dos dimensiones, tiene un tamaño de 2×2 elementos y es simétrica. El resultado pues es:

$$Covarianza = \begin{bmatrix} 0.6166 & 0.6154 \\ 0.6154 & 0.7166 \end{bmatrix}$$

Tercer paso, calcular los valores y vectores característicos de la propia matriz de covarianza, ordenando los vectores de acuerdo a la magnitud de su correspondiente valor característico:

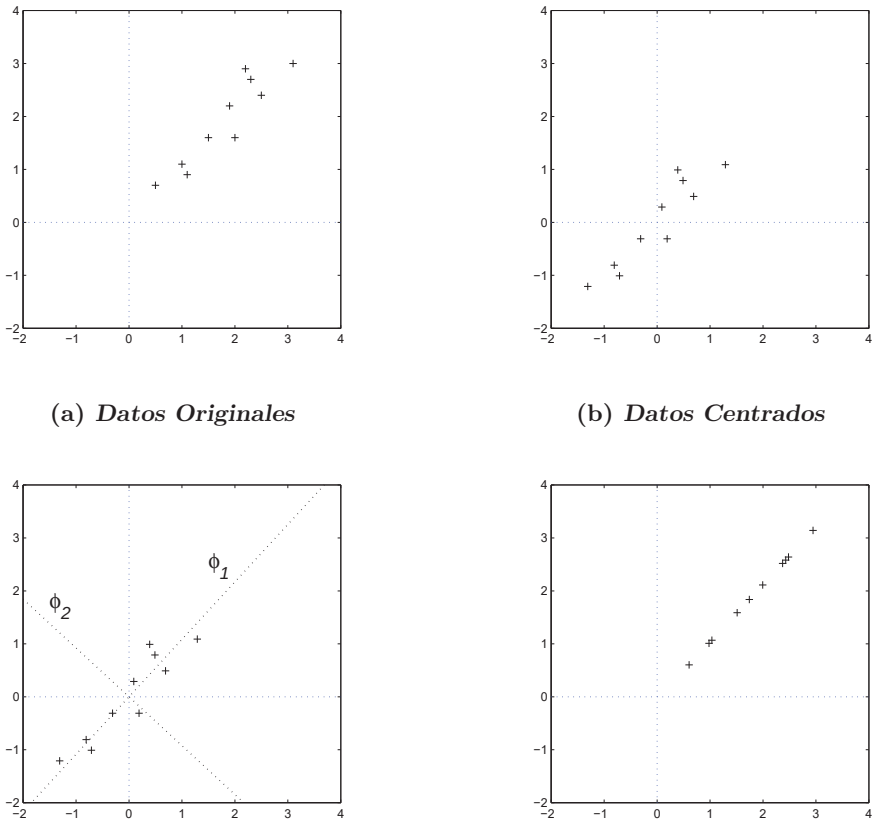
$$Eigenvalores = \begin{bmatrix} 1.2840 \\ 0.0491 \end{bmatrix}$$

$$Eigenvectores = \begin{bmatrix} 0.6779 & -0.7352 \\ 0.7352 & 0.6779 \end{bmatrix}$$

Es muy importante volver a señalar que la base obtenida del proceso es *ortonormal*, es decir, todos los vectores que la conforman son ortogonales entre sí y todos tienen norma unitaria. Esto tiene la ventaja de poder capturar de una mejor manera la dirección *lineal* de variación de los datos analizados, como se ejemplifica en la Figura A.1(c), donde se han sobrepuesto los vectores característicos a los datos. El eje etiquetado como ϕ_1 es el vector correspondiente al valor característico de mayor magnitud (y por lo tanto es el más *influyente*). La desventaja consiste en que para el caso de la detección y reconocimiento facial esta técnica intenta encontrar una variación lineal en el espacio de rostros (ver capítulo 2) aún y cuando éste, eminentemente, *no* es lineal ni convexo.

El cuarto paso consiste en calcular el nuevo conjunto de datos, esto es, aplicar la reducción de dimensiones. Se elige un cierto número M de los vectores encontrados, los cuales *deben* corresponder a los M mayores valores característicos, a fin de crear el subespacio que con un menor número de dimensiones (computacionalmente más manejable) capture la mayor variación de los datos. Esto no tiene gran importancia en el ejemplo de juguete que se está tratando, pero cuando se habla de imágenes de 64×64 pixeles con un espacio de $4,096$ dimensiones, reducidas por ejemplo a sólo 100, resaltan claramente las ventajas.

Para el ejemplo se elige $M = 1$, por lo que tomamos el vector característico respectivo y proyectamos los datos en este *nuevo* subespacio. Lo anterior se logra simplemente multiplicando la matriz transpuesta de la matriz con los vectores característicos elegidos,



(a) ***Datos Originales*** **(b)** ***Datos Centrados***

(c) ***Vectores Característicos*** **(d)** ***Datos Reconstruidos***

Figura A.1: Etapas del análisis de componentes principales (*PCA*).

por la matriz conformada de los datos centrados respecto a la media (*no* los originales). Esta multiplicación en realidad lo que efectúa es una rotación de los datos, de tal manera que los vectores característicos elegidos serán los nuevos ejes de coordenadas del subespacio, eliminando la correlación entre ellos. Esto equivaldría a tomar la Figura A.1(c) y hacerla girar en el sentido de las manecillas del reloj hasta hacer concordar los ejes ϕ_1 y ϕ_2 con los ejes horizontal y vertical originales (lo que permite ver que, efectivamente, los datos ya no tendrían correlación). Entonces se tiene que los datos proyectados son:

$$Proyección = \begin{bmatrix} 0.8280 \\ -1.7776 \\ 0.9922 \\ 0.2742 \\ 1.6758 \\ 0.9129 \\ -0.0991 \\ -1.1446 \\ -0.4380 \\ -1.2238 \end{bmatrix}$$

y como es de esperarse, estos datos sólo tienen una dimensión.

Un último paso del proceso, aunque en realidad no del *PCA*, consiste en reconstruir los datos originales a partir de la matriz *reducida* de vectores característicos. Esto es, teniendo el planteamiento verbal antes descrito:

$$Proyección = MVectores^T \times DatosCentrados$$

donde *Proyección* es el vector de datos proyectados en el nuevo subespacio, $MVectores$ es la matriz de los M vectores característicos elegidos y *DatosCentrados* es la matriz de los datos centrados respecto a la media; simplemente se realiza el despeje de la matriz *DatosCentrados* y se invierte la substracción de las medias respectivas que se realizó en el primer paso.

$$\begin{aligned} (MVectores^T)^{-1} \times Proyección &= DatosCentrados \\ MVectores \times Proyección &= DatosCentrados \\ DatosCentrados &= MVectores \times Proyección \\ DatosReconstruidos &= DatosCentrados + Medias \end{aligned}$$

considerando que por ser $MVectores$ una matriz ortogonal, su matriz inversa equivale a su matriz transpuesta y que, $Medias$, es el vector de valores promedio calculado para cada dimensión en el primer paso del *PCA*.

La matriz final de los datos reconstruidos utilizando solamente un vector característico (y por lo tanto *diferente* a la matriz de datos originales) se ilustra en forma gráfica en la Figura A.1(d). Se puede observar que los datos han perdido la información correspondiente al segundo vector característico y, por lo tanto, se ubican sobre el primero de ellos.

A.2. Descomposición de Valor Singular (*SVD – Singular Value Decomposition*)

El *PCA* involucra invariablemente el cálculo de la matriz de covarianzas $\mathbf{\Sigma}$. Esto, en el contexto de la detección y el reconocimiento facial, resulta ser una gran desventaja dado que para imágenes de 64×64 pixeles se está hablando de una matriz de covarianzas de $4,096 \times 4,096$ elementos, esto es, más de 16 millones de elementos. Lo anterior resulta computacionalmente costoso inclusive para la tecnología actual.

Existe otro medio para el cálculo de los componentes principales de la matriz de datos $\mathbf{X}$, el cual recibe el nombre de *descomposición de valor singular* (*SVD – Singular Value Decomposition*). Las secciones siguientes introducen primeramente los valores singulares de una matriz (base de la descomposición) para luego proceder a la *SVD*.

A.2.1. Valores Singulares de una Matriz

Para cualquier matriz A de $m \times n$ elementos, la matriz $A^T A$ de $n \times n$ será simétrica y, por lo tanto, puede ser diagonalizada ortogonalmente, según el *teorema espectral* [Román93]. No sólo los valores característicos de $A^T A$ son todos reales [Poole04], sino que también todos son *no negativos*. Para demostrar esta última afirmación, sea λ un valor

característico de A^TA con el vector característico unitario correspondiente $\mathbf{v}$. Entonces:

$$\begin{aligned} 0 &\leq \|A\mathbf{v}\|^2 \\ &= (A\mathbf{v}) \cdot (A\mathbf{v}) \\ &= (A\mathbf{v})^T A\mathbf{v} \\ &= \mathbf{v}^T A^T A\mathbf{v} \\ &= \mathbf{v}^T \lambda\mathbf{v} \\ &= \lambda(\mathbf{v} \cdot \mathbf{v}) \\ &= \lambda\|\mathbf{v}\|^2 \\ &= \lambda \end{aligned}$$

por tanto, tiene sentido tomar raíces cuadradas (positivas) de estos valores característicos. A estos valores resultantes de tomar las raíces cuadradas positivas de los valores característicos de la matriz A^TA de una matriz A de $m \times n$ elementos se les denomina *valores singulares* de A y se denotan mediante $\sigma_1, \ldots, \sigma_n$. Es convención acomodar estos valores singulares de tal manera que $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$. Cabe señalar que de forma análoga a la anterior se puede demostrar que los valores singulares también se pueden obtener como raíces cuadradas positivas de los valores característicos de la matriz AA^T, con la salvedad de que si $m \neq n$ el número de valores singulares obtenidos para A^TA y AA^T no es el mismo [Poole04].

A.2.2. *SVD*

Tomando como base lo visto en la sección anterior se puede decir entonces que cualquier matriz A de $m \times n$ elementos tiene la descomposición o puede ser factorizada como el producto de matrices:

$$A = U\Sigma V^T$$

donde U es una matriz ortogonal de $m \times m$ elementos, Σ una matriz "diagonal" de $m \times n$ y V una matriz ortogonal de $n \times n$. Si los valores singulares *distintos de cero* de A son:

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$$

y $\sigma_{r+1} = \sigma_{r+2} = \cdots = \sigma_n = 0$, entonces Σ tendrá la forma de bloques:

$$\Sigma = \left[\begin{array}{c|c} \overbrace{D}^{r} & \overbrace{O}^{n-r} \\ \hline O & O \end{array}\right] \begin{array}{l} \}r \\ \}m\text{-}r \end{array} \quad , \quad \text{donde} \quad D = \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r \end{bmatrix}$$

y cada matriz O es una matriz cero del tamaño apropiado (si $r = m$ o $r = n$, alguna de éstas no aparecerá).

Para construir la matriz ortogonal V, primero se determina una base ortonormal $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ de $\mathbb{R}^n$ compuesta por vectores característicos de la matriz simétrica $A^T A$ de $n \times n$ (*teorema espectral* [Román93]). Entonces

$$V = [\mathbf{v}_1 \ \cdots \ \mathbf{v}_n]$$

es una matriz ortogonal de $n \times n$.

Con respecto a la matriz ortogonal U, primero se advierte que $\{A\mathbf{v}_1, \ldots, A\mathbf{v}_n\}$ es un conjunto ortogonal de vectores de $\mathbb{R}^m$. Para ver esto, sea $\mathbf{v}_i$ el vector característico de $A^T A$ correspondiente al valor característico λ_i. Entonces, para $i \neq j$, se tiene que:

$$\begin{aligned} (A\mathbf{v}_i) \cdot (A\mathbf{v}_j) &= (A\mathbf{v}_i)^T A\mathbf{v}_j \\ &= \mathbf{v}_i^T A^T A\mathbf{v}_j \\ &= \mathbf{v}_i^T \lambda_j \mathbf{v}_j \\ &= \lambda_j (\mathbf{v}_i \cdot \mathbf{v}_j) \\ &= 0 \end{aligned}$$

debido a que los vectores característicos $\mathbf{v}_i$ son ortogonales. Ahora recuérdese que los valores singulares satisfacen $\sigma_i = \|A\mathbf{v}_i\|$ y que los primeros r de éstos son distintos de cero. Por consiguiente, podemos normalizar $A\mathbf{v}_1, \ldots, A\mathbf{v}_r$, al establecer:

$$\mathbf{u}_i = \frac{1}{\sigma_i} A\mathbf{v}_i \quad \text{para} \quad i = 1, \ldots, r$$

Esto garantiza que $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ es un conjunto ortonormal de $\mathbb{R}^m$, pero si $r < m$ no será una base de $\mathbb{R}^m$. En este caso, se extiende el conjunto $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ a una base ortonormal $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ de $\mathbb{R}^m$ (mediante el *proceso de Gram - Schmidt* [Román93], por ejemplo). Entonces se establece

$$U = [\mathbf{u}_1 \ \cdots \ \mathbf{u}_m]$$

Todo lo que queda por demostrar es que este procedimiento funciona; es decir, se necesita verificar que con U, Σ y V como se describen, se tendrá que $A = U\Sigma V^T$. Debido a que $V^T = V^{-1}$, esto es equivalente a demostrar que:

$$AV = U\Sigma$$

Se sabe que $A\mathbf{v}_i = \sigma_i \mathbf{u}_i$ para $i = 1, \ldots, r$ y que $\|A\mathbf{v}_i\| = \sigma_i = 0$ para $i = r+1, \ldots, n$. Por tanto, $A\mathbf{v}_i = \mathbf{0}$ para $i = r+1, \ldots, n$. Por consiguiente:

$$\begin{aligned}
AV &= A[\mathbf{v}_1 \ \cdots \ \mathbf{v}_n] \\
&= [A\mathbf{v}_1 \ \cdots \ A\mathbf{v}_n] \\
&= [A\mathbf{v}_1 \ \cdots \ A\mathbf{v}_r \ \mathbf{0} \ \cdots \ \mathbf{0}] \\
&= [\sigma_1\mathbf{u}_1 \ \cdots \ \sigma_r\mathbf{u}_r \ \mathbf{0} \ \cdots \ \mathbf{0}] \\
&= [\mathbf{u}_1 \ \cdots \ \mathbf{u}_m] \left[\begin{array}{ccc|c} \sigma_1 & \cdots & 0 & \\ \vdots & \ddots & \vdots & O \\ 0 & \cdots & \sigma_r & \\ \hline & O & & O \end{array}\right] \\
&= U\Sigma
\end{aligned}$$

como se requiere.

A.2.3. *SVD* y el *PCA*

Volvamos ahora al planteamiento original del *PCA* donde para la matriz $\mathbf{X}$ de $N \times k$ elementos la base *KLT* $\mathbf{\Phi}$ se obtiene resolviendo el problema de valores característicos $\mathbf{\Lambda} = \mathbf{\Phi}^T \mathbf{\Sigma} \mathbf{\Phi}$, donde $\mathbf{\Sigma}$ es la matriz de covarianza de los datos según la Ecuación A.2.

Ahora se puede definir una nueva matriz $\mathbf{Y}$ como una matriz de $M \times k$ elementos:

$$\mathbf{Y} = \frac{1}{\sqrt{k}} \mathbf{X}^T$$

donde cada *columna* de $\mathbf{Y}$ está centrada respecto a la media. La definición de $\mathbf{Y}$ se clarifica

analizando $\mathbf{Y}^T\mathbf{Y}$:

$$\begin{aligned}\mathbf{Y}^T\mathbf{Y} &= (\frac{1}{\sqrt{k}}\mathbf{X}^T)^T(\frac{1}{\sqrt{k}}\mathbf{X}^T) \\ &= \frac{1}{k}(\mathbf{X}^T)^T\mathbf{X}^T \\ &= \frac{1}{k}\mathbf{X}\mathbf{X}^T \\ \mathbf{Y}^T\mathbf{Y} &= \mathbf{\Sigma}\end{aligned}$$

es decir, $\mathbf{Y}^T\mathbf{Y}$ es la matriz de covarianzas de $\mathbf{X}$. Luego, según lo visto hasta ahora, como la base $\mathbf{\Phi}$ está conformada por los vectores característicos de $\mathbf{\Sigma}$, si se calcula la *SVD* de $\mathbf{Y}$ las columnas de la matriz V contendrán los vectores característicos de $\mathbf{Y}^T\mathbf{Y} = \mathbf{\Sigma}$. Por tanto, las columnas de V *son* los componentes principales de $\mathbf{X}$.

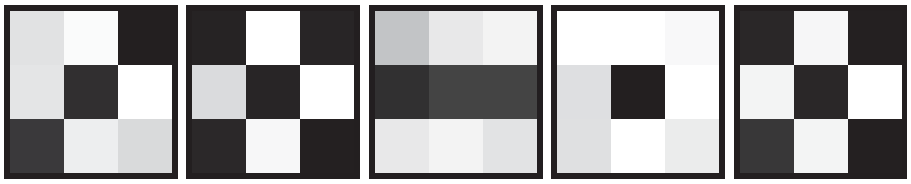
Lo que esto significa es que V genera el *espacio fila* de $\mathbf{Y} \equiv \frac{1}{\sqrt{k}}\mathbf{X}^T$. Por tanto, V también debe generar el *espacio columna* de $\frac{1}{\sqrt{k}}\mathbf{X}$. Como el objetivo final es encontrar una base ortonormal que *exprese* los vectores que conforman las columnas de $\mathbf{X}$ (el *espacio columna* de $\mathbf{X}$); ésta se puede calcular directamente sin necesidad de construir $\mathbf{Y}$. Por simetría, las columnas de U producidas por la *SVD* de $\frac{1}{\sqrt{k}}\mathbf{X}$ también *deben ser* los componentes principales [Shlens05].

A.2.4. Ejemplo de Aplicación de la *SVD* para Calcular el *PCA*

A continuación se desarrolla un ejemplo más elaborado para demostrar la *SVD*. En primer término se tienen las cuatro imágenes de entrenamiento que se muestran en la Figura A.2 y que están marcadas como $x_1, \ldots, x_4$. La imagen restante, y_1, se introduce como una prueba de identificación a una escala muy reducida dado que las imágenes cuentan con solamente nueve pixeles.

La Tabla A.2 muestra los datos correspondientes a los valores de gris de las imágenes de entrenamiento (por ordenamiento lexicográfico de sus pixeles), la media correspondiente a cada una de las "dimensiones" y, finalmente, la matriz $\mathbf{X}$ de los datos centrados respecto a la *imagen promedio*. Se incluyen también los datos respectivos a la imagen de prueba y_1, tanto originales como centrados.

Los valores característicos (como el valor cuadrático de los valores singulares)



(a) x_1 **(b)** x_2 **(c)** x_3 **(d)** x_4 **(e)** y_1

Figura A.2: Imágenes de entrenamiento y de prueba para la *SVD* de ejemplo.

Originales					Medias	Centrados **X**				
x_1	x_2	x_3	x_4	y_1	μ's	x'_1	x'_2	x'_3	x'_4	y'_1
225	10	196	255	20	171.50	53.50	-161.50	24.50	83.50	-151.50
229	219	35	223	244	176.50	52.50	42.50	-141.50	46.50	67.50
48	24	234	224	44	132.50	-84.50	-108.50	101.50	91.50	-88.50
251	255	232	255	246	248.25	2.75	6.75	-16.25	6.75	-2.25
33	18	59	0	21	27.50	5.50	-9.50	31.50	-27.50	-6.50
238	247	244	255	244	246.00	-8.00	1.00	-2.00	9.00	-2.00
0	17	243	249	4	127.25	-127.25	-110.25	115.75	121.75	-123.25
255	255	57	255	255	205.50	49.50	49.50	-148.50	49.50	49.50
217	2	226	235	2	170.00	47.00	-168.00	56.00	65.00	-168.00

Tabla A.2: Datos para la descomposición de valor singular de ejemplo.

Proyecciones				
x_1''	x_2''	x_3''	x_4''	y_1''
-103.0851	-265.9231	229.7636	139.2447	-266.6479
-117.3060	98.2879	125.9002	-106.8821	80.7476

Tabla A.3: Datos proyectados al subespacio calculado por la *SVD* de ejemplo.

obtenidos por la *SVD* de la matriz $\mathbf{X}$ son:

$$\lambda_1 = 38,381 \ \geq \ \lambda_2 = 12,674 \ \geq \ \lambda_3 = 5,695 \ \geq \ \lambda_4 = 0$$

y sus correspondientes vectores característicos unitarios son (no se incluye ν_4 por corresponder a un valor característico 0):

$$\nu_1 = \begin{bmatrix} 0.3562 \\ -0.2785 \\ 0.4796 \\ -0.0317 \\ 0.0350 \\ 0.0088 \\ 0.5601 \\ -0.2963 \\ 0.4022 \end{bmatrix} \quad \nu_2 = \begin{bmatrix} -0.5521 \\ -0.4885 \\ 0.0443 \\ -0.0479 \\ 0.1051 \\ -0.0035 \\ 0.1115 \\ -0.4917 \\ -0.4324 \end{bmatrix} \quad \nu_3 = \begin{bmatrix} -0.2636 \\ 0.3470 \\ 0.3092 \\ 0.0635 \\ -0.2219 \\ 0.0777 \\ 0.5845 \\ 0.4011 \\ -0.3908 \end{bmatrix}$$

Para el presente caso demostrativo se implementa la reducción de dimensionalidad eligiendo los $M = 2$ vectores característicos que corresponden a los valores de mayor magnitud. Posteriormente se proyectan los datos de entrenamiento y los de prueba en este nuevo *subespacio*, resultando los datos mostrados en la Tabla A.3.

Finalmente y a fin de resolver el problema planteado de la identificación, calculamos la *distancia euclideana* o *norma* L_2 entre las imágenes de entrenamiento y la imagen de prueba (se pueden utilizar otras medidas de distancia, esto es sólo para completar el ejemplo). La distancia más corta nos indicará a cuál de las cuatro identidades "pertenece" la imagen de prueba, el resultado es:

Norma L_2 entre y_1 y x_i			
x_1	x_2	x_3	x_4
65,977.9968	308.1871	248,463.1354	199,953.7113

Así pues, la imagen más parecida a y_1 es la imagen de entrenamiento x_2. Lo cual se puede comprobar por inspección visual de la Figura A.2.

El proceso seguido en el presente ejemplo es el método original implementado por Turk y Pentland [Turk91], en uno de los trabajos más influyentes de los últimos años en el área.

Apéndice B

Determinación del Valor Óptimo de ρ para el Estimador $\hat{P}(\mathbf{x}|\Omega)$

Para obtener el valor óptimo del estimador $\hat{P}(\mathbf{x}|\Omega)$ en base al parámetro ρ de la Ecuación 2.9, se debe resolver la Ecuación 2.10. Para tal efecto se utilizan las formas diagonalizadas de la distancia de Mahalanobis $d(\mathbf{x})$ y su estimación $\hat{d}(\mathbf{x})$ (Ecuaciones 2.7 y 2.8, respectivamente). Por tanto, se tiene que:

$$
\begin{aligned}
J(\rho) &= \mathrm{E}\left[\log \frac{P(\mathbf{x}|\Omega)}{\hat{P}(\mathbf{x}|\Omega)}\right] \\
&= E\left\{ \log \frac{\dfrac{\exp\left(-\frac{1}{2}\sum\limits_{i=1}^{M}\frac{\mathbf{y}_i^2}{\lambda_i} - \frac{1}{2}\sum\limits_{i=M+1}^{N}\frac{\mathbf{y}_i^2}{\lambda_i}\right)}{(2\pi)^{N/2}\prod\limits_{i=1}^{N}\lambda_i^{1/2}}}{\left[\dfrac{\exp\left(-\frac{1}{2}\sum\limits_{i=1}^{M}\frac{\mathbf{y}_i^2}{\lambda_i}\right)}{(2\pi)^{M/2}\prod\limits_{i=1}^{M}\lambda_i^{1/2}}\right] \cdot \left[\dfrac{\exp\left(-\frac{\sum\limits_{i=M+1}^{N}\mathbf{y}_i^2}{2\rho}\right)}{(2\pi\rho)^{(N-M)/2}}\right]} \right\}
\end{aligned}
$$

$$
\begin{aligned}
J(\rho) &= \mathrm{E}\left\{ \log \frac{(2\pi\rho)^{\frac{(N-M)}{2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{M}\frac{\mathbf{y}_i^2}{\lambda_i} - \frac{1}{2}\sum_{i=M+1}^{N}\frac{\mathbf{y}_i^2}{\lambda_i} - \left(-\frac{1}{2}\sum_{i=1}^{M}\frac{\mathbf{y}_i^2}{\lambda_i} \right) - \left(-\frac{\sum_{i=M+1}^{N}\mathbf{y}_i^2}{2\rho} \right) \right]}{(2\pi)^{\frac{(N-M)}{2}} \prod_{i=M+1}^{N}\lambda_i^{1/2}} \right\} \\
&= \mathrm{E}\left[\log \frac{\rho^{\frac{(N-M)}{2}} \exp\left(-\frac{1}{2}\sum_{i=M+1}^{N}\frac{\mathbf{y}_i^2}{\lambda_i} + \frac{\sum_{i=M+1}^{N}\mathbf{y}_i^2}{2\rho} \right)}{\prod_{i=M+1}^{N}\lambda_i^{1/2}} \right] \\
&= \mathrm{E}\left\{ \log\left[\frac{\prod_{i=M+1}^{N}\rho^{1/2}}{\prod_{i=M+1}^{N}\lambda_i^{1/2}} \right] + \log\left[\exp\left(-\frac{1}{2}\sum_{i=M+1}^{N}\frac{\mathbf{y}_i^2}{\lambda_i} + \frac{\sum_{i=M+1}^{N}\mathbf{y}_i^2}{2\rho} \right) \right] \right\} \\
&= \mathrm{E}\left\{ \log\left[\prod_{i=M+1}^{N}\left(\frac{\rho}{\lambda_i}\right)^{1/2} \right] - \frac{1}{2}\sum_{i=M+1}^{N}\frac{\mathbf{y}_i^2}{\lambda_i} + \frac{1}{2\rho}\sum_{i=M+1}^{N}\mathbf{y}_i^2 \right\} \\
&= \mathrm{E}\left[\frac{1}{2}\sum_{i=M+1}^{N}\left(\log\frac{\rho}{\lambda_i}\right) - \frac{1}{2}\sum_{i=M+1}^{N}\frac{\mathbf{y}_i^2}{\lambda_i} + \frac{1}{2}\sum_{i=M+1}^{N}\frac{\mathbf{y}_i^2}{\rho} \right]
\end{aligned}
$$

y, finalmente, aprovechando que $\mathrm{E}[\mathbf{y}_i^2] = \lambda_i$, se tiene:

$$
\begin{aligned}
J(\rho) &= \frac{1}{2}\sum_{i=M+1}^{N}\left[\log\frac{\rho}{\lambda_i} - \frac{\mathrm{E}(\mathbf{y}_i^2)}{\lambda_i} + \frac{\mathrm{E}(\mathbf{y}_i^2)}{\rho} \right] \\
&= \frac{1}{2}\sum_{i=M+1}^{N}\left(\log\frac{\rho}{\lambda_i} - 1 + \frac{\lambda_i}{\rho} \right) \\
&= \frac{1}{2}\sum_{i=M+1}^{N}\left(\frac{\lambda_i}{\rho} - 1 + \log\frac{\rho}{\lambda_i} \right)
\end{aligned}
$$

El peso óptimo ρ se puede calcular minimizando esta función de costo con respecto a ρ. Resolviendo la ecuación $\frac{\partial J}{\partial \rho} = 0$ conlleva a:

$$
0 = \frac{\partial}{\partial\rho}\left[\frac{1}{2}\sum_{i=M+1}^{N}\left(\frac{\lambda_i}{\rho} - 1 + \log\frac{\rho}{\lambda_i} \right) \right]
$$

$$
\begin{aligned}
0 &= \sum_{i=M+1}^{N} \lambda_i(-1)\rho^{-2} \;+\; 0 \;+\; \frac{\partial}{\partial\rho}\left[\sum_{i=M+1}^{N}(\log\rho - \log\lambda_i)\right] \\
\sum_{i=M+1}^{N}\frac{\lambda_i}{\rho^2} &= \sum_{i=M+1}^{N}\frac{1}{\rho} \;+\; 0 \\
\frac{1}{\rho}\sum_{i=M+1}^{N}\lambda_i &= N - M \\
\rho &= \frac{1}{N-M}\sum_{i=M+1}^{N}\lambda_i
\end{aligned}
$$

lo cual es simplemente el promedio aritmético de los valores característicos en el subespacio ortogonal complementario $\bar{F}$.

Apéndice C

Bases de Datos de Rostros Utilizadas en los Experimentos

Debido a su estructura tridimensional, flexible y compleja, la apariencia de un rostro se ve afectada por un gran número de factores que incluyen la identidad, la pose, la iluminación, la expresión facial, la edad, las oclusiones e, incluso, el cabello, la barba y el bigote. El desarrollo de algoritmos robustos ante estas variaciones requiere bases de datos de tamaño suficiente que incluyan variaciones de estos factores cuidadosamente controladas. Más aún, se necesitan bases de datos comunes para la evaluación comparativa de los algoritmos. La recolección de bases de datos de gran calidad consume grandes recursos, pero la disponibilidad pública de bases de datos de rostros es importante para el avance de las investigaciones en el área [Jain05].

En este apartado se describen detalladamente las bases de datos de rostros empleadas en los experimentos desarrollados durante el presente trabajo de investigación[1].

[1] Lamentablemente y pese a los esfuerzos realizados, no fue posible tener acceso con el tiempo suficiente a la base de datos *idónea*, es decir, a la base de datos de imágenes faciales del programa *FERET*, disponible en [NIST93]. El Instituto Nacional de Estándares y Tecnología de los Estados Unidos (*NIST*), responsable de la misma, nos proporcionó la autorización de acceso a la base de datos casi simultáneamente con la finalización de los experimentos realizados durante la presente investigación, por lo que se decidió no considerarla en este documento.

C.1. Base de Datos de Rostros del Centro Universitario de la *FEI* en Brasil

En cuanto a la parte correspondiente de los experimentos de reconocimiento facial se utilizó la base de datos de rostros de la *FEI* (referida en [FEI05]). Esta base es una colección de datos brasileña que contiene imágenes tomadas entre los meses de junio del año 2005 y marzo del año 2006 en el Laboratorio de Inteligencia Artificial del Centro Universitario de la *FEI* (*Fundação Educacional Inaciana*) en San Bernardo del Campo, San Pablo, Brasil. Contiene 14 imágenes de cada uno de los 200 individuos que la conforman (100 hombres y 100 mujeres); un total de 2,800 imágenes a color tomadas contra un fondo homogéneo blanco con los sujetos en una posición vertical de frente con rotaciones laterales hasta los 180 grados. La escala varía alrededor de un 10 %, siendo el tamaño original de cada imagen de 640 pixeles de ancho por 480 de alto. Todos los rostros pertenecen a estudiantes o personal de la *FEI* entre los 19 y los 40 años de edad, con diferencias en la apariencia, cabello y adornos.

La Figura C.1 presenta un ejemplo de las 14 variaciones individuales en esta base de datos. Debido a las restricciones impuestas en el acceso a la misma, no es posible ejemplificar una mayor cantidad de miembros. Por lo anterior, sólo se señala que en el presente trabajo de investigación se emplearon las dos imágenes correspondientes a la última fila de la Figura C.1 de cada uno de los 200 individuos que conforman el conjunto y, claramente, en forma posterior a la conversión de las fotografías a escala de grises.

C.2. Base de Datos de Rostros del Laboratorio de Visión por Computadora de la Universidad de Ljubljana en Slovenia

La segunda y última colección de fotografías empleada en esta investigación, en específico para el proceso de detección automatizada, corresponde a un conjunto de imágenes en color (accesible en [Peer03a]) recolectado en el año 2003 por el Dr. Peter Peer de la Universidad de Ljubljana en Slovenia, conjunto descrito en [Peer03b]. Se conforma de un total de 114 sujetos diferentes, con 7 imágenes individuales por cada uno de ellos, tomadas a una resolución de 640 pixeles de ancho por 480 de alto. La mayoría de individuos tienen

Figura C.1: Imágenes de ejemplo correspondientes a la base de datos de rostros del Centro Universitario de la *FEI* en Brasil [FEI05]. En la presente investigación se emplearon las dos imágenes de la última fila correspondientes a cada uno de los 200 individuos que conforman el conjunto.

Figura C.2: Imágenes de rostros de ejemplo correspondientes a la base de datos de rostros del Laboratorio de Visión por Computadora de la Universidad de Ljubljana en Slovenia [Peer03a]. En el presente trabajo se emplean las dos últimas imágenes mostradas, relativas a cada uno de los 114 individuos que conforman la base.

alrededor de los 18 años por tratarse de alumnos y profesores del propio Laboratorio de Visión por Computadora. Del total de la base, aproximadamente, el 90 % son hombres.

La Figura C.2 presenta un ejemplo de las 7 diferentes imágenes individuales en esta base (denominada *CVL – Computer Vision Laboratory*). Al igual que para la base de datos *FEI* y en concordancia con el acuerdo de licencia/entrega, no se deben reproducir o publicar las imágenes que conforman el conjunto (fuera del ejemplo descrito, por supuesto). Por lo anterior, sólo se menciona que en el presente trabajo se emplean las dos últimas imágenes mostradas en la Figura C.2, correspondientes a cada uno de los 114 individuos congregados en la citada base de datos.

Referencias

[Bartlett98] Bartlett, M., Lades, H., y Sejnowski, T. Independent component representations for face recognition. *En Proceedings of the SPIE: Conference on Human Vision and Electronic Imaging III*, tomo 3299, págs. 528–539. 1998.

[Belhumeur97] Belhumeur, P., Hespanha, J. P., y Kriegman, D. J. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, July 1997.

[Blackburn01] Blackburn, D. M., Bone, M., y Phillips, P. J. Face recognition vendor test 2000. Inf. téc., National Institute of Standards and Technology, February 2001. http://www.frvt.org.

[Chui92] Chui, C. K. *An Introduction to Wavelets*. Academic Press Professional, Inc., San Diego, CA, USA, 1992.

[Cover94] Cover, T. M. y Thomas, J. A. *Elements of Information Theory*. John Wiley & Sons, New York, 1994.

[Dunn07] Dunn, J. S. y Podio, F. The Biometric Consortium, 2007. Recurso en internet: http://www.biometrics.org/.

[FEI05] FEI. FEI face database, 2005. Recurso en internet: http://www.fei.edu.br/~cet/facedatabase.html.

[Freund95] Freund, Y. y Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *En Eurocolt'95: European Conference on Computational Learning Theory*, págs. 23–37. 1995.

[Gerbrands81] Gerbrands, J. J. On the relationships between SVD, KLT and PCA. *Pattern Recognition*, 14(1-6):375–381, 1981.

[Grgic05] Grgic, M. y Delac, K. Face recognition homepage, 2005. Recurso en internet: http://www.face-rec.org/.

[Hotelling33] Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441,498–520, 1933.

[Intel Corp.00] Intel Corp. y SourceForge.Net. The open computer vision library, 2000. Recurso en internet: http://opencvlibrary.sourceforge.net/.

[Jain05] Jain, A. K. y Li, S. Z. *Handbook of Face Recognition*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[Joliffe86] Joliffe, I. T. *Principal Components Analysis*. New York: Springer–Verlag, 1986.

[Karhunen46] Karhunen, K. Über lineare methoden in der wahrscheinlichkeitsrechnung. *En Series AI: Mathematica–Physica*, tomo 37, págs. 3–79. Annales Academiae Scientiarum Fennicae, 1946. (Traducido por RAND Corp., Santa Mónica, Calif., Reporte Técnico T-131, Agosto 1960).

[Kirby90] Kirby, M. y Sirovich, L. Application of the Karhunen-Loève procedure for the characterization of human faces. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(1):103–108, 1990.

[Loève55] Loève, M. *Probability Theory*. Princeton, N. J.:Van Nostrand, 1955.

[Moghaddam95a] Moghaddam, B. y Pentland, A. Probabilistic visual learning for object detection. *En ICCV'95: Proc. of the International Conference on Computer Vision*, págs. 786–793. June 1995.

[Moghaddam95b] Moghaddam, B. y Pentland, A. A subspace method for maximum-likelihood target detection. *En ICIP'95: Proc. of the International Conference on Image Processing*. October 1995.

[Moghaddam96] Moghaddam, B., Nastar, C., y Pentland, A. A bayesian similarity measure for direct image matching. Inf. Téc. 393, Massachussets Institute of Technology, Media Laboratory Perceptual Computing Section, 1996.

[Moghaddam98] Moghaddam, B., Jebara, T., y Pentland, A. Efficient MAP / ML similarity matching for visual recognition. *En ICPR'98: Proceedings of the 14th International Conference on Pattern Recognition*, tomo 1, pág. 876. IEEE Computer Society, Washington, DC, USA, 1998.

[Moghaddam00] Moghaddam, B., Jebara, T., y Pentland, A. Bayesian face recognition. *Pattern Recognition*, 33(11):1771–1782, November 2000.

[Moghaddam02] Moghaddam, B. Principal manifolds and probabilistic subspaces for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(6):780–788, 2002.

[NIST93] NIST. The FERET database, 1993. Recurso en internet: http://www.nist.gov/humanid/feret/.

[NIST02] NIST. The NIST mugshot identification database (MID), 2002. Recurso en internet: http://www.nist.gov/srd/nistsd18.htm.

[NIST03] NIST. The color FERET database, 2003. Recurso en internet: http://www.nist.gov/humanid/colorferet/.

[Pearson01] Pearson, K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572, 1901.

[Peer03a] Peer, P. The CVL face database, 2003. Recurso en internet: http://www.lrv.fri.uni-lj.si/facedb.html.

[Peer03b] Peer, P., Batagelj, B., Kovac, J., y Solina, F. Color-based face detection in the "15 Seconds of Fame" art installation. *En Mirage 2003, Conference on Computer Vision / Computer Graphics Collaboration for Model-based Imaging, Rendering, Image Analysis and Graphical Special Effects*, págs. 38–47. March 2003.

[Phillips98] Phillips, P. J., Wechsler, H., Huang, J., y Rauss, P. J. The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, 16(5):295–306, 1998.

[Phillips00] Phillips, P. J., Moon, H., Rizvi, S. A., y Rauss, P. J. The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.

[Phillips03] Phillips, P. J., Grother, P., Micheals, R. J., Blackburn, D. M., Tabassi, E., y Bone, M. Face recognition vendor test 2002: Evaluation report. NISTIR 6965, National Institute of Standards and Technology, March 2003. http://www.frvt.org.

[Poole04] Poole, D. *Algebra Lineal: Una Introducción Moderna*. International Thomson Editores, S. A. de C. V., D. F., México, 2004.

[Román93] Román, J. B. *Algebra Lineal*. McGraw - Hill/Interamericana de España, S. A., Madrid, España, 1993.

[Rowley96] Rowley, H. A., Baluja, S., y Kanade, T. Human face detection in visual scenes. *En* D. S. Touretzky, M. C. Mozer, y M. E. Hasselmo, eds., *Advances in Neural Information Processing Systems*, tomo 8, págs. 875–881. The MIT Press, 1996.

[Rowley98] Rowley, H. A., Baluja, S., y Kanade, T. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, 1998.

[Rowley99] Rowley, H. A. *Neural Network - Based Face Detection*. Tesis Doctoral, School of Computer Sciencie, Computer Science Department, Carnegie Mellon University, May 1999.

[Scassellati98] Scassellati, B. Eye finding via face detection for a foveated active vision system. *En AAAI/IAAI*, págs. 969–976. 1998.

[Shlens05] Shlens, J. A tutorial on Principal Components Analysis, 2005. Recurso en internet: http://www.snl.salk.edu/~shlens/pub/notes/pca.pdf.

[Sinha94] Sinha, P. Object recognition via image invariants: A case study. *Investigative Ophthalmology and Visual Science*, 35:1735–1740, May 1994.

[Sinha95] Sinha, P. *Perceiving and Recognizing Three-Dimensional Forms.* Tesis Doctoral, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA., 1995.

[Sirohey93] Sirohey, S. A. Human face segmentation and identification. Inf. Téc. CS-TR-3176, University of Maryland, USA, 1993.

[Smith02] Smith, L. I. A tutorial on Principal Components Analysis, 2002. Recurso en internet: http://www.cs.cmu.edu/˜elaw/papers/pca.pdf.

[Sobottka96] Sobottka, K. y Pitas, I. Face localization and facial feature extraction based on shape and color information. *En ICIP'96: Proceedings of the 1996 IEEE International Conference on Image Processing*, tomo III, págs. 483–486. September 1996.

[Teixeira03] Teixeira, M. *The Bayesian Intrapersonal/Extrapersonal Classifier.* Proyecto Fin de Carrera, CSU Computer Science Department, July 2003.

[Turk91] Turk, M. y Pentland, A. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[Viola01] Viola, P. y Jones, M. J. Rapid object detection using a boosted cascade of simple features. *En CVPR'01: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, tomo 1, págs. I–511–I–518. 2001.

[Viola04] Viola, P. y Jones, M. J. Robust real-time face detection. *International Journal on Computer Vision*, 57(2):137–154, 2004.

[Wang03] Wang, X. y Tang, X. Unified subspace analysis for face recognition. *En Ninth IEEE International Conference on Computer Vision (ICCV'03)*, tomo 1, pág. 679. 2003.

[Yambor00] Yambor, W. S. Analysis of PCA-based and Fisher discriminant-based image recognition algorithms. Inf. Téc. CS-00-103, Colorado State University, July 2000.

[Yang94] Yang, G. y Huang, T. S. Human face detection in a complex background. *Pattern Recognition*, 27(1):53–63, 1994.

[Yang02] Yang, M.-H., Kriegman, D. J., y Ahuja, N. Detecting faces in images: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):34–58, 2002.

[Zhao03] Zhao, W., Chellappa, R., Phillips, P. J., y Rosenfeld, A. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458, 2003.

[Zisserman03] Zisserman, A. y Hartley, R. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge, United Kingdom, 2ª ed$^{\underline{\text{ón}}}$., 2003.

Glosario

Algoritmo AdaBoost- Algoritmo utilizado para aprendizaje rápido de una función de clasificación entre (y dado) un conjunto de imágenes de entrenamiento de ejemplos positivos y negativos así como un conjunto de rasgos de imagen que caractericen a cada uno de estos conjuntos.

Análisis de componentes principales (PCA)- Véase *transformación de Karhunen – Loève*.

Biometría (como característica)- Característica biológica (anatómica o física) y de conducta que es medible y que se puede utilizar para reconocimiento automatizado.

Biometría (como proceso)- Métodos automatizados de reconocimiento de un individuo basados en características biológicas (anatómicas o físicas) y de conducta.

Correlación- Mide la proporción de cambio entre los pixeles de dos imágenes. Es una medida de similitud entre -1 (las imágenes son totalmente opuestas entre sí) y $+1$ (las imágenes son idénticas).

Coseno del ángulo- Medida de similitud que multiplicada por -1, negando su valor, se convierte en medida de distancia; la cual calcula el coseno del ángulo entre dos vectores normalizados con base en su producto punto.

Curvas principales- Técnica no lineal de reducción de dimensión que consiste, básicamente, en una formulación de regresión no lineal sobre los datos o vectores del espacio original.

Detección de rostros- Proceso en el cual un sistema biométrico determina la posición de todos los rostros humanos existentes en una imagen de entrada.

Distancia al espacio de rostros- Distancia existente entre una imagen de prueba proyectada al subespacio óptimo calculado por una transformación de Karhunen – Loève y su correspondiente espacio de rostros.

Distancia de Mahalanobis- Calcula la suma del producto de las coordenadas correspondientes entre dos vectores, ponderando su valor con respecto a la magnitud del valor característico asociado (a cada dimensión).

Espacio característico dual- Conjunción de los espacios vectoriales generados por los vectores característicos principales del conjunto de diferencias entre las imágenes de un mismo individuo y el generado por los vectores respectivos al conjunto de diferencias entre las imágenes de diferentes individuos.

Espacio de rostros- Subespacio óptimo calculado por medio de una transformación de Karhunen – Loève.

FERET – *FacE REcognition Technology program-* Programa de desarrollo y evaluación de tecnología de reconocimiento de rostros auspiciado por el gobierno de los Estados Unidos entre 1993 y 1997.

Fisherfaces- Técnica de reducción de dimensión de un espacio vectorial que selecciona el subespacio lineal que maximiza la proporción de la variación (su dispersión en el espacio) de los vectores correspondientes a una misma clase, con respecto a la variación entre vectores de diferentes clases (en reconocimiento de rostros los vectores son imágenes faciales y las clases son identidades correspondientes a varios individuos).

FLD- Véase *Fisherfaces*.

Galería- Conjunto de individuos conocidos o base de datos de un sistema biométrico, utilizados para una evaluación en específico o una cierta implementación.

ICA – Independent Component Analysis- Otra técnica de reducción de dimensión que elige los componentes (nueva base vectorial) que minimice las dependencias de segundo y superiores órdenes entre los vectores del espacio original, con las características obligatorias de ser independientes estadísticamente y tener una distribución no Gaussiana además de, generalmente, no ser ortogonales.

Identificación de rostros- Proceso en el cual un sistema biométrico busca dentro de una base de datos una referencia que empate con una muestra biométrica de entrada y, si se encuentra, devuelve la identidad correspondiente. La identificación es de *conjunto cerrado* si se sabe que la muestra de entrada existe en la base de datos, mientras que será de *conjunto abierto* si esto no se garantiza. El sistema debe determinar si la muestra de entrada existe o no en la base de datos y devolver su identidad, en caso afirmativo.

Imagen integral- Representación intermedia de una imagen utilizada para calcular en forma muy optimizada rasgos (de Haar) rectangulares de ésta.

KPCA – Kernel PCA- Técnica de reducción de dimensionalidad que extiende el *PCA* al utilizar un kernel para el cálculo de los componentes principales del espacio original, mapeando entre dicho espacio y uno nuevo de mayor dimensión (posiblemente infinito). El kernel utilizado debe cumplir con el teorema de Mercer para que donde se requiera un producto punto en el espacio de origen (como en el cálculo de la matriz de covarianzas de la que se obtienen los componentes principales) se substituya por el kernel, sin la necesidad de realizar el cálculo explícito (y posiblemente prohibitivo) del espacio generado por el mapeo del teorema.

LDA – Linear Discrimant Analysis- Véase *Fisherfaces*.

Localización de rostros- Proceso en el cual un sistema biométrico determina la posición de un solo rostro humano de todos los posibles existentes en una imagen de entrada.

Mapa de rostros- Imagen creada con los valores de distancia al espacio de rostros calculados para cada ubicación posible en una imagen de prueba en la que se pretende realizar un proceso de localización de rostros humanos. El valor mínimo de este mapa indica la región con mayor factibilidad de contener una cara.

Métodos basados en apariencia- Métodos de detección y reconocimiento de rostros que *aprenden* los patrones o plantillas que caracterizan una imagen como rostro humano a través de un conjunto de imágenes de entrenamiento, mismas que deben capturar la variabilidad representativa de la apariencia facial (para lograr un buen desempeño del sistema biométrico). Posteriormente los patrones aprendidos se utilizarán para los procesos de detección y/o reconocimiento facial.

Métodos basados en conocimiento- Métodos de detección y reconocimiento de rostros basados en reglas, que aplican el conocimiento humano de lo que constituye un rostro. Usualmente las reglas capturan las relaciones entre rasgos faciales.

Métodos basados en enfoques de características invariantes- Métodos de detección y reconocimiento de rostros que tratan de encontrar los elementos estructurales que existen aún y cuando la posición, el punto de vista o la iluminación varíen y, una vez determinados, los utilizan para localizar los rostros.

Métodos de empate de plantillas- Métodos de detección y reconocimiento de rostros que almacenan una gran cantidad de patrones estándares de rostros para describir a una cara como un todo o a sus rasgos separadamente. La correlación calculada entre una imagen de entrada y los patrones almacenados permitirá el proceso de detección y/o el de reconocimiento facial.

Modelo- Representación utilizada para caracterizar a un individuo.

Muestra biométrica- Información o datos computarizados obtenidos de un dispositivo sensor biométrico. Las imágenes de un rostro o de alguna huella dactilar son ejemplos de muestras biométricas.

Norma L_1 – *Norma rectilínea*- Medida de distancia que suma la diferencia absoluta entre las coordenadas correspondientes de dos vectores.

Norma L_2 – *Norma Euclidiana*- Medida de distancia que suma la diferencia cuadrada entre las coordenadas correspondientes de dos vectores.

PCA- Véase *transformación de Karhunen – Loève.*

Plantilla- Representación digital de las características distintivas de un individuo como medidas tomadas de una muestra biométrica en específico.

Pose- Posición de un individuo con respecto al eje óptico de un observador o de la cámara en una imagen fotográfica.

Prueba- Muestra biométrica que se utiliza como entrada en un sistema biométrico a fin de compararla contra una o más referencias en la galería.

Rasgos Haar- Características de una imagen que codifican la existencia de contrastes orientados entre regiones de la misma así como sus relaciones espaciales. Se les denomina rasgos Haar debido a que se calculan de manera similar a los coeficientes de una transformación de *kernels de Haar*.

Reconocimiento de rostros- Una modalidad biométrica que utiliza una imagen de la estructura física visible del rostro de un individuo con propósitos de reconocimiento.

Rostros característicos (*eigenfaces*)- Conjunto de vectores que conforman la base (vectorial) del subespacio óptimo determinado por una transformación de Karhunen – Loève.

Rostros característicos duales- Rostros característicos (base vectorial) del espacio característico dual.

Sistema biométrico- Sistema operacional automatizado conformado por múltiples componentes individuales (sensores, algoritmos de empate, dispositivo de despliegue de resultados, etc.), que permite capturar muestras biométricas de un usuario, procesarlas, almacenar la información obtenida de este procesamiento, comparar dicha información con la obtenida de muestras de referencia y, finalmente, decidir qué tan bien empatan para indicar si se ha logrado confirmar la identidad del sujeto al que pertenece la muestra de entrada.

***Template*-** Véase plantilla.

Teorema de Mercer- Establece que cualquier función generadora de kernels definidos semi-positivos (para el alcance del presente trabajo de tesis) se puede expresar en forma de producto punto en un espacio de alta dimensión (posiblemente infinito).

Transformación de Karhunen – Loève- Determinación de la base de un subespacio óptimo dado un conjunto de imágenes de entrenamiento de mxn pixeles representadas como vectores de mxn componentes, minimizando el error cuadrado promedio entre las imágenes originales y sus proyecciones en dicho espacio. Cuando las imágenes del conjunto de entrenamiento se centran con respecto a la media del mismo, la transformación de Karhunen – Loève, la transformación Hotelling y el análisis de componentes principales (PCA) son equivalentes.

Transformación Hotelling- Véase *transformación de Karhunen – Loève.*

Transformaciones de blanqueo – *Whitening transformations*- El blanqueo sobre un conjunto de vectores consiste en centrar (al conjunto) respecto a la media (media cero) así como transformar su varianza en uno.

Valores singulares- Valores resultantes de tomar las raíces cuadradas positivas de los valores característicos de la matriz $A^T A$ (o AA^T) de una matriz A de $m \times n$ elementos.

Variaciones interpersonales- Variaciones en un conjunto de imágenes debido a que están asociadas a diferentes individuos.

Variaciones intrapersonales- Variaciones en un conjunto de imágenes asociadas a un mismo individuo y que corresponden a, por ejemplo, cambios e la iluminación o en la pose.

Variedad de no rostros- Subespacio, del espacio de las imágenes de un tamaño específico, que corresponde a todas aquellas imágenes que no representan a un rostro humano.

Variedad de rostros- Subespacio, del espacio de las imágenes de un tamaño específico, que corresponde a todas aquellas imágenes que representan a un rostro humano.

Verificación- Proceso en el cual un sistema biométrico intenta confirmar la identidad pretendida de una muestra biométrica de entrada comparándola contra una o más muestras que con antelación se ingresaron en una base de datos.