Tesis defendida por

Susana Garduño Massieu

y aprobada por el siguiente comité

Dr. Vitaly Kober
Director del Comité

Dr. Hugo Homero Hidalgo Silva
Miembro del Comité

Dr. Josué Álvarez Borrego
Miembro del Comité

Dr. José Antonio García Macías
Coordinador del Programa de Posgrado en Maestría en Ciencias en Ciencias de la Computación

Dr. David Hilario Covarrubias Rosales
Director de la Dirección de Estudios de Posgrado

11 de enero de 2013.

CENTRO DE INVESTIGACIÓN CIENTÍFICA Y DE EDUCACIÓN SUPERIOR DE ENSENADA

PROGRAMA DE POSGRADO EN CIENCIAS EN CIENCIAS DE LA COMPUTACIÓN

Reconocimiento confiable de rostros bajo condiciones del mundo real

Tesis

que para cubrir parcialmente los requisitos necesarios para obtener el grado de Maestro en Ciencias

Presenta:

Susana Garduño Massieu

Ensenada, Baja California, México.
2013

Resumen de la tesis de Susana Garduño Massieu, presentada como requisito parcial para la obtención del grado de Maestro en Ciencias en Ciencias de la Computación. Ensenada, Baja California, México.

Reconocimiento confiable de rostros bajo condiciones del mundo real.

Resumen aprobado por:

_____

Dr. Vitaly Kober

El hombre se desenvuelve en las actividades cotidianas principalmente a través de su sentido visual; depende de él para llevar a cabo tareas naturales como la identificación y clasificación, es decir, el reconocimiento de individuos y objetos a su alrededor. Durante los últimos años se ha dado un auge en las investigaciones sobre el reconocimiento de rostros automático en el área de reconocimiento de patrones y visión por computadora.

Existen varias aplicaciones de reconocimiento de rostros en la actualidad, que van desde la seguridad social hasta aplicaciones recreativas como en sistemas de realidad virtual. Sin embargo, el desempeño de los sistemas en estas aplicaciones es limitado debido a variaciones inherentes en las imágenes utilizadas para el reconocimiento, como iluminación no uniforme, cambios de postura, oclusión, e incluso ruido de los sensores de adquisición. Aunque estos factores se han estudiado para el desarrollo de nuevos métodos, sólo se han atacado los problemas en casos específicos incluyendo pocos factores a la vez.

Existen numerosas técnicas para el reconocimiento de rostros en imágenes estáticas. Dentro de las técnicas basadas en imagen, las cuales usan toda la región del rostro como entrada al sistema, los siguientes métodos estadísticos se han utilizado ampliamente: análisis de componentes principales para Eigenfaces, análisis discriminanate lineal para Fisherfaces y análisis de componentes independientes para dos arquitecturas. Sin embargo estas técnicas se basan en el reconocimiento de rostros en imágenes previamente segmentadas. Otra técnica de esta categoría se basa en métodos de filtrado lineal, que usan la correlación para caracterizar la similitud entre un patrón de referencia y uno de prueba. A pesar de los logros alcanzados actualmente en esta área, el problema de reconocimiento confiable en escenas reales bajo condiciones no controladas sigue abierto.

En este trabajo se estudiaron las técnicas de reconocimiento de rostros estadísticas mencionadas así como las técnicas de correlación y se propone un algoritmo basado en filtros de correlación compuestos que no requiere de segmentación previa del rostro a reconocer. Se utiliza el criterio de capacidad de discriminación para sintetizar un filtro compuesto que, para reconocer el rostro de un individuo dado, incorpora información a rechazar de otros individuos y se adapta a la escena de entrada. Además se introduce una técnica adaptativa para maximizar el desempeño del filtro propuesto, logrando así tener una técnica de reconocimiento automático de rostros que no depende de una etapa de preprocesamiento para la detección y segmentación del rostro.

Palabras Clave: **reconocimiento de rostros, filtros de correlación, reconocimiento de patrones.**

Abstract of the thesis presented by Susana Garduño Massieu as a partial requirement to obtain the Master of Science degree in Computer Science. Ensenada, Baja California, México.

Reliable face recognition under real world conditions.

Abstract approved by:

_____

Dr. Vitaly Kober

Humans perform daily tasks, mainly, thanks to their visual system; they depend on it to perform natural tasks such as identification and classification, e.g., recognition of individuals and objects around them. In pattern recognition and computer vision domains, there has been an increasing interest in automatic face recognition.

Nowadays there are several face recognition applications in a number of broad areas, from social security to entertainment with virtual reality systems. Facial recognition systems face challenging problems due to inherent variations on different factors at image acquisition, such as nonuniform illumination, pose changes, occlusion and even sensor noise in acquisition, resulting in a poor performance of these systems. Even though these factors have been studied to develop new robust recognition methods, they have been approached under constrained scenarios, considering few factors at the same time.

Numerous techniques have been proposed for face recognition in still images. Among image based methods, which use the whole face region as the input to the system, the following statistical methods are widely used: principal component analysis for Eigenfaces, linear discriminant analysis for Fisherfaces, and independent component analysis with two architectures. Another technique of this type is based on linear filtering methods, which uses correlation to characterize similarity between a reference and test patterns. Despite recent achievements in this area, the problem of reliable facial recognition in a real, uncontrolled scene still remains open.

In this work, the above statistical based face recognition techniques were studied, as well as correlation techniques, and an algorithm is proposed based on composite correlation filters. The algorithm does not require prior input face segmentation. A single composite filter is synthesized using the discrimination capability criterion which, in order to recognize a given face of an individual, incorporates information to reject from other individuals and adapts to the input scene. Moreover an adaptive technique is introduced to maximize the proposed filter performance thus achieving an automatic face recognition technique that does not depend on a preprocessing step for detection and segmentation of the face.

Keywords: **face recognition, correlation filters, pattern recognition.**

Dedicatorias

A mis padres,
a mi hermano
y a Pablo.

Agradecimientos

Por su apoyo, sus consejos y amor en todo momento y etapa de mi vida, le agradezco a mi familia, que son mi ejemplo a seguir.

A Pablo, gracias por estar incondicionalmente y siempre ahí, sin importar los kilómetros de distancia, por todo el apoyo y cariño.

Gracias a mi asesor, el Dr. Vitaly Kober, por su paciencia y conocimiento transmitido durante este tiempo para realizar este proyecto. También agradezco a los miembros del comité el Dr. Hugo Hidalgo y el Dr. Josué Álvarez, por su apoyo, sus comentarios y observaciones valiosas.

Un especial agradecimiento a Pablo Aguilar, por su disposición al contestar las múltiples preguntas teóricas y técnicas que me surgieron durante y fuera de su estancia en CICESE.

Agradezco a mis compañeros de generación por todos los momentos memorables a lo largo de estos años de estudiantes. En especial a Daniel-san y Marcus, por apoyarme en los momentos más difíciles al realizar este trabajo, darme ánimos y por su amistad. Gracias, Daniel, por las tradicionales “casitas felices”.

Agradezco a las chicas; Valeria, Maythé y Gio, por la organización de eventos, convivios y las palmaditas en la espalda que a veces se necesitan.

A los ex-miembros del cubo, gracias por las pláticas amenas y “ñoñas”, la fruta, los panqués, el helado y demás. Y a los compañeros de los otros cubos que, con todo y quejas, han hecho de esta etapa toda una experiencia.

Gracias a Javier por no dejar oxidar esa amistad y porque desde lejos me “echó porras” y ánimos.

A todos los que de una u otra manera me ayudaron y me apoyaron durante este periodo, muchas gracias.

Agradezco al CICESE, a CONACYT y a México por haberme dado la oportunidad de realizar mis estudios de posgrado con apoyo económico.

No. Registro: 242906.

Contenido

Lista de figuras

Lista de tablas

Capítulo 1. Introducción

El rostro humano es uno de los patrones más comunes que aparecen ante la visión del hombre. Su habilidad innata para detectar y reconocer rostros vuelve estas actividades en tareas que no requieren de esfuerzo; es su principal método de identificación. El reconocimiento automático del rostro humano es un problema importante en el desarrollo y aplicación del reconocimiento de patrones. En la actualidad hay muchas aplicaciones que utilizan reconocimiento de rostros, ya sea para fines de seguridad social, como la vigilancia y la identificación de individuos sospechosos, o en el sector de entretenimiento para juegos de video, realidad aumentada, cámaras fotográficas, e interacción hombre-máquina.

En las últimas décadas se han incrementado los problemas sociales que atentan contra la integridad física de los individuos alrededor del mundo. Los sistemas de seguridad han mejorado con la adquisición de nuevas tecnologías; los gobiernos han sido motivados a utilizar sistemas basados en características corporales o de comportamiento, comúnmente llamados sistemas biométricos. Éstos son atractivos debido a que pueden integrarse a cualquier aplicación que requiera de un control de acceso o de seguridad, eliminando los riesgos asociados de las tecnologías que se basan en lo que portan o en lo que saben los individuos, en vez de su verdadera identidad (Abate, Nappi, Riccio, & Sabatino, 2007).

Entre las biometrías más comunes se encuentran el iris y las huellas digitales, aunque otras se han estudiado, como la geometría de los dedos y las palmas de las manos, voz, firma y rostro (Abate et al., 2007). Las biometrías presentan algunas desventajas a considerar. Por ejemplo, el reconocimiento de iris es muy preciso, pero es muy costoso implementarlo y no es tan aceptado por la gente; por otro lado, el reconocimiento de huellas digitales es bastante confiable y no incomoda, pero se necesita cooperación de los individuos con el procedimiento. Por el contrario, el reconocimiento de rostros es una alternativa que presenta un equilibro de compromiso entre la precisión y la aceptación (Abate et al., 2007). Además, presenta la ventaja de que puede usarse en lugares grandes y concurridos de gente que no necesariamente está consciente de su participación en los sistemas. Inclusive, de las biometrías mencionadas, el rostro es la que proporciona mejor

información para la correspondencia, según los factores de evaluación de sistemas de documentos de viaje legibles por máquinas (Machine Readable Travel Documents, MRTD) (Li & Jain, 2005).

1.1 Sistemas de reconocimiento de rostros

Un Sistema de Reconocimiento de Rostros (SRR) por lo general consta de una etapa precedente de entrenamiento (ver Figura 1). En ésta, se cuenta con un conjunto de imágenes de rostros de distintas personas al que se le llama base de datos o conjunto de rostros de entrenamiento. De este conjunto se extrae información relevante para el reconocimiento (como de representación o características), que posteriormente se almacena en otra base de datos que se utiliza para los resultados del SRR. Cuando se desea reconocer un rostro dado, éste se alimenta al sistema y pasa por el mismo proceso de extracción de información que el conjunto de rostros de entrenamiento. Una vez extraída la información relevante del rostro a reconocer, se compara con aquella almacenada de los rostros de entrenamiento. Un proceso de clasificación es el encargado de esta comparación para que posteriormente el sistema decida si el rostro se reconoce o no.

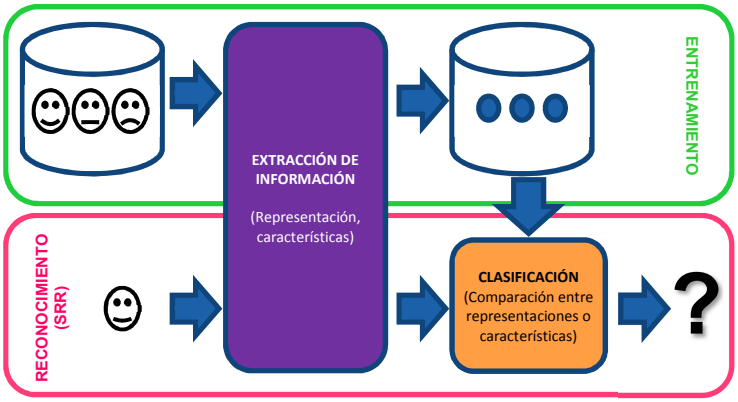


Figura 1. Sistema de reconocimiento de rostros.

Como el reconocimiento de rostros se considera un problema de reconocimiento de patrones visuales, el rostro se puede ver como un objeto tridimensional a identificar por su proyección en dos dimensiones sujeta a variaciones de iluminación, postura y expresiones, entre otras. Un SRR comúnmente tiene cuatro grandes etapas: detección, alineamiento, extracción de características o información y correspondencia. Las primeras dos etapas corresponden a los procesamientos de localización y normalización previos al

reconocimiento como tal (Li & Jain, 2005). La detección segmenta el área del rostro del fondo de la imagen, mientras que la alineación lo localiza con mayor precisión y normaliza los rostros; la detección sólo proporciona estimaciones burdas de localización y escala de cada rostro. Para normalizar la imagen, se pueden localizar características faciales como los ojos, nariz, boca y contorno del rostro, así como utilizar transformaciones geométricas, respecto al tamaño y postura, basándose en los puntos localizados.

La extracción de características o información provee datos útiles para distinguir entre diferentes rostros, e idealmente, que permita la tolerancia a variaciones de apariencia. En la correspondencia se compara la información extraída de un rostro con la de imágenes registradas en una base de datos (incluida en la clasificación en la Figura 1). Esta última etapa proporciona como salida la identidad del rostro, si se encuentra una correspondencia con un alto nivel de confiabilidad o, en caso contrario, simplemente indica que no se encontró la identidad.

1.1.1 Modos de operación y tareas de los sistemas de reconocimiento de rostros

Un SRR puede operar en dos modos: verificación (autenticación) o identificación (referida simplemente como reconocimiento). La verificación de un rostro, vista como una tarea a realizar, es una correspondencia uno a uno que compara una imagen de rostro de entrada al sistema (o de prueba) con la imagen del rostro registrada en el sistema que se presume es la identidad afirmada. Por el contrario, la identificación de rostros es una correspondencia de uno a muchos, en la que se compara una imagen de rostro con los múltiples rostros registrados en una base de datos, para asociar la identidad del rostro de prueba a una de aquellos registrados. En este modo, se tienen dos tipos de tareas: identificación en conjunto abierto e identificación en conjunto cerrado. En la segunda, se suponen los rostros de prueba pertenecientes a uno de los individuos registrados en la base de datos, mientras que en la primera tarea, los rostros de prueba no necesariamente deben pertenecer a alguno de los individuos registrados (Li & Jain, 2005; Wechsler, 2007).

1.1.2 Problemas de los sistemas de reconocimiento de rostros

Los SRR actuales se enfrentan a problemas debido a algunos factores inherentes a las condiciones de adquisición de las imágenes. Los principales y que ya han sido objeto de

estudio, son las variaciones de iluminación, los cambios en postura de los individuos, las variaciones físicas del rostro debido al paso del tiempo, expresiones faciales de los individuos y, por último, las oclusiones que se presentan debido a la perspectiva o existencia de otros individuos u objetos en la imagen, o por portar accesorios como bufandas, lentes (oscuros o no), o por presencia de vello facial.

Resulta evidente la necesidad de estudiar y desarrollar técnicas que permitan a los SRR tener un desempeño satisfactorio independientemente de las condiciones presentes a la hora de adquirir la información; la iluminación del día o el clima no deben ser importantes, así como la posición y tamaño del rostro en el cuadro de adquisición del sensor, por ejemplo. Las técnicas de reconocimiento actuales normalmente están diseñadas para trabajar con imágenes de rostros de entrenamiento y de prueba capturadas en condiciones ideales; es decir, las imágenes, si bien presentan algunas variaciones de las mencionadas arriba, es de manera muy moderada. Además, su desempeño depende mucho de las etapas previas de detección y alineamiento. De aquí la necesidad de tener técnicas que funcionen con imágenes de rostros de prueba que presenten variaciones en condiciones reales.

1.2 Clasificación de técnicas de reconocimiento de rostros

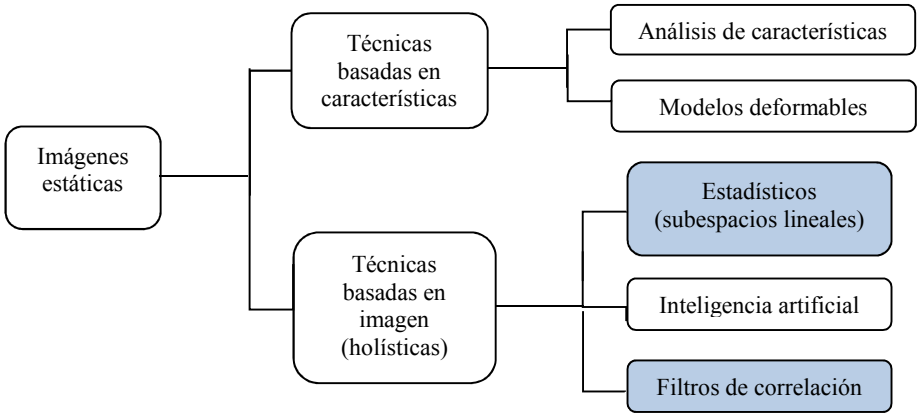


Figura 2. Clasificación de enfoques de reconocimiento de rostros.

Un SRR puede trabajar con imágenes estáticas o con video. Enfocándose en las imágenes estáticas, las diferentes técnicas existentes para el reconocimiento de rostros se pueden clasificar en dos grandes grupos (véase Figura 2), con base en la clasificación de Zhao, Chellappa, Phillips y Rosenfeld (2003) y de Jafri y Arabnia (2009).

1.2.1. Técnicas basadas en características

Estas técnicas requieren de un preprocesamiento de la imagen de entrada para identificar, extraer y medir características distintivas como los ojos, nariz, boca, etc., (Jafri & Arabnia, 2009). Posteriormente, se calculan relaciones geométricas entre ellas y se utilizan técnicas estándares de reconocimiento de patrones para la correspondencia de los rostros basada en estas medidas. En esta categoría se encuentran métodos puramente geométricos o basados en el análisis de características, el cual organiza las características visuales en el concepto global de rostro y características faciales a partir de la información geométrica del rostro (Hjelmås & Low, 2001). Entre los primeros trabajos, el de Kanade (1973) calculaba un vector de 16 parámetros faciales (razones, áreas, distancias y ángulos) y usaba distancia euclidiana; posteriormente Brunelli y Poggio (1993) extendieron el vector a 35 parámetros. Cox, Ghosn y Yianilos (1996) propusieron un método que obtenía tasas altas de reconocimiento, pero la extracción de características era manual. Los modelos deformables tienen como propósito la extracción de características no rígidas y complejas (Hjelmås & Low, 2001). Sin embargo, como los modelos no encajan a la perfección en las estructuras de la imagen, se requiere establecer un compromiso entre tolerancia y precisión de reconocimiento (Jafri & Arabnia, 2009). A pesar de que el conocido método Elastic Bunch Graph Matching (EBGM) (Wiskott, Fellous, Krüger, & von der Malsburg, 1997), que genera un grafo por cada rostro, haya tenido buen desempeño en una de las evaluaciones más renombradas, la versión presentada requería de una configuración manual del grafo para los primeros 70 rostros antes de que fuera confiable (Jafri & Arabnia, 2009).

Aunque estas técnicas son relativamente robustas a variaciones de la posición en la imagen de entrada y, en principio, pueden hacerse invariantes a tamaño, orientación e iluminación (Jafri & Arabnia, 2009), además de brindar una representación compacta agilizando el proceso de correspondencia, su principal desventaja es la dificultad de detección automática de características; no tienen un nivel de precisión muy alto y requieren una capacidad de cómputo considerable. Además, la implementación de algunas de ellas requiere tomar decisiones arbitrarias sobre la relevancia de las características (Jafri & Arabnia, 2009).

1.2.2. Técnicas basadas en imagen

Las técnicas de esta categoría identifican rostros usando representaciones globales; utilizan toda la región del rostro como entrada de datos crudos al sistema de reconocimiento (Zhao et al., 2003). Algunas se basan en métodos estadísticos (o de subespacios lineales), en los que la representación de la región del rostro más utilizada está basada en proyecciones del análisis de componentes principales (PCA), debido a su éxito en la reconstrucción de rostros en espacios de menor dimensión (Zhao et al., 2003). Dentro de este grupo se encuentran las técnicas: Eigenfaces, que emplea un clasificador de vecinos cercanos; Fisherfaces, que emplea tanto PCA como el análisis discriminante lineal (LDA); análisis de componentes independientes (ICA), que es una generalización de PCA; así como técnicas que usan métodos Bayesianos mediante medidas probabilísticas de similitud, e incluso técnicas que utilizan algoritmos genéticos, como EP (Evolutionary Pursuit).

Pertenecen también a esta categoría las técnicas basadas en inteligencia artificial que utilizan herramientas como redes neuronales y técnicas de aprendizaje máquina para reconocer rostros (Jafri & Arabnia, 2009). La idea consiste en tener una red con una neurona por pixel de la imagen, aunque no se construye directamente a partir de la imagen de entrada, sino de la resultante de alguna técnica de reducción de dimensión. El mapa de auto organización (SOM) es una red neuronal invariante respecto a cambios menores en una imagen de muestra (Jafri & Arabnia, 2009), entrenada mediante aprendizaje no supervisado; no necesita de datos de imagen preclasificados. Se utiliza para tener una representación discreta de baja dimensión de las muestras de entrenamiento, llamada mapa. Las redes neuronales basadas en decisión probabilística se han modelado para diferentes aplicaciones, como detector de rostros, localizador de ojos y reconocedor de rostros (Abate et al., 2007). Lin, Kung y Lin (1997) propusieron una red particular que busca regiones como cejas, ojos y nariz, pero no la boca, con el objeto de evitar la influencia de variaciones faciales por expresiones y lograr construir un sistema robusto (Zhao et al., 2003). La estructura de las redes neuronales depende mucho del área de aplicación, por lo que diferentes contextos resultan en redes bastante diferentes y, en general, presentan problemas cuando incrementa el número de clases. Además, tampoco son adecuados para reconocimiento de un solo modelo, ya que se necesitan muchas imágenes modelo por

persona para que el entrenamiento del sistema sea óptimo (Abate et al., 2007). Otros métodos utilizados han sido las máquinas de soporte vectorial (SVM) y los basados en modelos ocultos de Markov (HMM). Estos últimos han sido reportados invariantes a escala de las imágenes de rostros (Jafri & Arabnia, 2009).

Por último, las técnicas de reconocimiento de patrones mediante filtros de correlación se han aplicado exitosamente en el reconocimiento automático de objetivos (ATR) y han tenido un gran desarrollo en las últimas décadas. Recientemente también se han aplicado para reconocimiento de rostros y, aunque han tenido un auge más lento, proveen ventajas sobre los otros enfoques dentro de las categorías mencionadas. Una de ellas es la invarianza a desplazamiento que tienen los filtros, otra es que se pueden diseñar para presentar tolerancia a ruido y alta discriminación y, además, proveen expresiones de forma cerradas (Savvides, Kumar, & Pradeep, 2002). Estos métodos utilizan la transformada de Fourier para realizar la correspondencia en el dominio de frecuencia.

De las distintas técnicas de reconocimiento presentadas, las holísticas o basadas en imagen, son las que han dominado en los últimos años (Zhao, et al., 2003), en especial con el uso de PCA. De aquí la elección de trabajar con estas técnicas en este proyecto.

1.3. Objetivos

Como se estableció anteriormente, las técnicas tradicionales actuales de reconocimiento de rostros están diseñadas para trabajar con imágenes segmentadas. Estas imágenes se pueden obtener mediante una etapa previa de detección que influye en el desempeño de los métodos, restringiendo su aplicación en situaciones reales. Dado el éxito de los filtros de correlación en el reconocimiento de objetos, este trabajo se basa en ellos para aplicarlos a dichas situaciones reales.

1.3.1. Objetivo general

Proponer una técnica de reconocimiento de rostros, basada en filtros de correlación, con entrenamiento controlado y reconocimiento en imágenes del mundo real tomadas bajo condiciones no controladas.

1.3.2. Objetivos específicos

1. Estudiar e implementar las técnicas de reconocimiento de rostros basadas en los métodos de análisis de componentes principales, análisis de discriminante lineal y de análisis de componentes independientes.
2. Adaptar las técnicas estudiadas para utilizarlas en situaciones reales: con imágenes de entrada capturadas bajo condiciones no controladas, sin preprocesamiento manual (como segmentación).
3. Diseñar filtros de correlación para el reconocimiento de rostros, considerando un conjunto de imágenes de referencia capturadas en condiciones ideales y una imagen de entrada en condiciones no controladas (con fondo complejo).
4. Proponer una técnica basada en correlación y modelos matemáticos para reconocer rostros sin preprocesamiento manual de la imagen de entrada, siendo tolerante a pequeñas distorsiones.
5. Evaluar la técnica propuesta, comparar su desempeño con las técnicas estudiadas y probar su efectividad.

1.4. Limitaciones y suposiciones

En este trabajo se supone una situación realista como una en la que un SRR es capaz de detectar el rostro de entrada sin necesidad de un preprocesamiento manual, como la segmentación, por lo tanto, que pueda trabajar con imágenes que contengan fondo complejo (existencia de varios objetos en la escena) y/o ruido.

Un escenario real se considera como aquel que tiene un conjunto de imágenes de entrenamiento capturadas en condiciones ideales o controladas (sin fondo o con poco fondo uniforme) y una imagen de entrada al sistema capturada en condiciones no controladas. No se considera el caso en que las imágenes de entrenamiento estén capturadas en condiciones no controladas. Las imágenes usadas están en escala de gris. Por último, se considera la tarea principal del SRR como la de identificación en conjunto cerrado; es decir, todos los rostros de prueba pertenecen a los individuos de las imágenes de referencia.

1.5. Investigación relacionada relevante

Se mencionan a continuación trabajos del área relacionados con el desarrollo del proyecto de tesis llevado acabo.

1) Reconocimiento de rostros con métodos estadísticos. Como se mencionó anteriormente, una de las técnicas más populares que se sigue utilizando es la de Eigenfaces (Turk & Pentland, 1991), que se basa en el método de PCA para reducción de dimensiones, basado en los trabajos de Sirovich y Kirby (1987) y Kirby y Sirovich (1990) para la representación y reconstrucción de imágenes de rostros. Otra técnica muy relevante actualmente es la de Fisherfaces (Belhumeur, Hespanha, & Kriegman, 1997), la cual también se basa en la reducción de dimensiones, pero utiliza el LDA enfocándose así en la discriminación de clases. Más recientemente se propuso un método usando la generalización de PCA con ICA aplicado al reconocimiento de rostros para proveer una mejor representación, estableciendo dos arquitecturas (Bartlett, Lades, & Sejnowski, 1998; Bartlett, Movellan, & Sejnowski, 2002).

2) Reconocimiento de rostros con filtros de correlación. Entre los trabajos de reconocimiento por filtros de correlación aplicado a rostros, se tiene el presentado por Saavides, et al., (2002), en el que aplican filtros compuestos para la tarea de verificación de rostros usando pocas imágenes de entrenamiento por clase. Posteriormente Kumar, et al., (2006), propusieron dos métodos de filtros compuestos: CFA (Class-dependence feature analysis) y KCFA (Kernel CFA), aplicados a la verificación de rostros siguiendo el protocolo FRGC (Face Recognition Grand Challenge), al crear un espacio de características de manera similar a las técnicas estadísticas. Aunque se muestran mejores resultados respecto a las técnicas estadísticas, estos trabajos atacan el problema de reconocimiento de rostros a partir de imágenes segmentadas y no se aplican directamente a escenas del mundo real en fondos complejos.

3) Reconocimiento de patrones con filtros de correlación. Las escenas del mundo real se pueden modelar como un objeto incrustado en un fondo disjunto; Javidi y Wang (1994) propusieron filtros óptimos para el modelo de escena disjunto mediante la optimización de criterios de desempeño evitando las limitaciones de filtros clásicos anteriores. Estos filtros se conocen como el filtro óptimo generalizado y el filtro de correspondencia generalizado.

Existen técnicas de filtros compuestos para cuando se consideran distorsiones geométricas en las escenas, como la de las funciones discriminantes sintéticas, propuesta por Casasent, (1984), que utiliza un conjunto de imágenes de entrenamiento que capturan las posibles apariencias que puedan presentar los objetos. Aunque tienen algunas desventajas, estos filtros se han utilizado satisfactoriamente para reconocer objetos en fondos complejos mediante un algoritmo adaptativo alcanzando un nivel de discriminación deseado (González-Fraga, Kober & Álvarez-Borrego, 2006). Asimismo, recientemente se diseñaron filtros basados en los propuestos por Javidi y Wang (1994) para reconocimiento de objetos degradados y distorsionados por ruido, usando también un algoritmo iterativo de adaptación (Ramos-Michel & Kober, 2008).

1.6. Organización de la tesis

El resto de la tesis se organiza como sigue. El capítulo 2 presenta conceptos que forman la base teórica con la que se realizó este trabajo. Éstos son importantes para que el lector comprenda algunos términos que se utilizan en el resto de la tesis. Las técnicas estadísticas de reconocimiento de rostros que fueron implementadas y que se incluyen en el estudio comparativo, así como el detector de rostros con el que se adaptaron para su aplicación en situaciones reales, se describen en el capítulo 3. El capítulo 4 es la base teórica del enfoque de reconocimiento de patrones por filtros de correlación con la que se formuló la técnica propuesta de reconocimiento de rostros. Esta última se presenta en el capítulo 5. En el capítulo 6 se describen los experimentos realizados para la comparación entre las técnicas estadísticas y la técnica propuesta, así como los resultados correspondientes. Finalmente se exponen las conclusiones de este trabajo de investigación, al igual que las oportunidades de trabajo a futuro. La última parte es un apéndice que describe la base de datos de rostros utilizada en los experimentos.

Capítulo 2. Fundamentos teóricos

2.1 Introducción

Como el reconocimiento de rostros se apoya ampliamente en la teoría del procesamiento de imágenes, a continuación se presentan conceptos importantes que forman parte de la base del trabajo de investigación realizado.

Una imagen cualquiera es una función bidimensional *f*(*x,y*), donde *x* y *y* son las coordenadas espaciales y la amplitud de la función en cualquier coordenada (*x,y*), es la intensidad o nivel de gris en ese punto (González & Woods, 2008). Esta función *f* se convierte en una imagen digital a través de un muestreo y una cuantización. El muestreo corresponde a la digitalización de las coordenadas del plano y la cuantización a la digitalización de los valores de la amplitud.

2.2 Transformada de Fourier

La idea de la transformada de Fourier es representar una función como una combinación de funciones senos y/o cosenos. Además, tiene la característica de que la función puede ser reconstruida completamente con un proceso inverso. Con esto, el avance de la tecnología digital y, en particular, el algoritmo de la transformada rápida de Fourier, se tienen técnicas que permiten el estudio e implementación de distintos enfoques de procesamiento de imágenes. La transformada de Fourier de una imagen bidimensional $f(x, y)$, denotada como $\mathcal{F}\{f(x, y)\}$, está definida por:

$$\mathcal{F}\{f(x,y)\} = F(u,v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)e^{-2i\pi(ux+vy)}dxdy. \tag{1}$$

De manera análoga, dada *F*(*u, v*), se puede obtener *f*(*x*, *y*) usando la transformada inversa de Fourier, denotada como $\mathcal{F}^{-1}\{F(u, v)\}$, definida:

$$\mathcal{F}^{-1}\{F(u,v)\} = f(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(u,v)e^{2i\pi(ux+vy)}dudv, \tag{2}$$

donde *u* y *v* son las variables de frecuencia, *x* y *y* son las variables espaciales y donde *i* es el número imaginario $\sqrt{-1}$. Estas dos expresiones constituyen el par transformado de Fourier continuo bidimensional. De la ecuación (2), se puede interpretar que la función $f(x, y)$ es

una combinación lineal de funciones complejas $e^{2i\pi(ux+vy)}$ que, utilizando la fórmula de Euler

$$e^{i\phi} = \cos\phi + i\sin\phi, \tag{3}$$

están compuestas de funciones seno y coseno, donde $F(u,v)$ es la función de peso. En general, los componentes de la transformada de Fourier son complejos. De aquí que $F(u,v)$ se puede expresar en coordenadas polares, como se hace con los números complejos:

$$F(u,v) = |F(u,v)|e^{i\phi(u,v)}, \tag{4}$$

donde

$$|F(u,v)| = \sqrt{[Re\{F(u,v)\}]^2 + [Im\{F(u,v)\}]^2}, \tag{5}$$

es el espectro de amplitud o la magnitud de la transformada de Fourier y

$$\phi(u,v) = \arctan\left[\frac{Im\{F(u,v)\}}{Re\{F(u,v)\}}\right], \tag{6}$$

es el ángulo de fase o el espectro de fase de la transformada de Fourier. En las ecuaciones (5) y (6), $Re\{F(u.v)\}$ y $Im\{F(u.v)\}$ representan la parte real y la parte imaginaria de $F(u.v)$, respectivamente. En el análisis de Fourier también resulta conveniente conocer el espectro de potencia definido como:

$$\begin{aligned} P(u,v) &= |F(u,v)|^2 \\ &= [Re\{F(u,v)\}]^2 + [Im\{F(u,v)\}]^2 \\ &= F(u,v)F^*(u,v), \end{aligned} \tag{7}$$

donde el superíndice * denota el conjugado complejo.

2.2.1. Propiedades de la transformada de Fourier

Desde el punto de vista de procesamiento de imágenes, las siguientes propiedades de la transformada de Fourier son interesantes.

1) Linealidad. La transformada de Fourier es un operador lineal, por lo tanto, para dos funciones $f_1(x,y)$ y $f_2(x,y)$ y para dos constantes arbitrarias *a* y *b*, se tiene:

$$\mathcal{F}\{af_1(x,y) + bf_2(x,y)\} = aF_1(x,y) + bF_2(x,y). \tag{8}$$

Además, la transformada de Fourier es distributiva respecto a la suma, pero no respecto a la multiplicación (Gonzalez & Woods, 2002). Esto es:

$$\mathcal{F}\{f_1(x,y)\cdot f_2(x,y)\} \neq F_1(x,y)\cdot F_2(x,y).$$

2) Traslación (o desplazamiento). Para dos constantes x_0 y y_0 se tienen las expresiones para el desplazamiento en el dominio espacial y en el dominio de frecuencia:

$$\mathcal{F}\{f(x-x_0, y-y_0)\} = F(u,v)e^{-2i\pi(ux_0+vy_0)}, \tag{9}$$
$$\mathcal{F}\{f(x,y)e^{2i\pi(u_0x+v_0y)}\} = F(u-u_0, v-v_0).$$

3) Rotación. Con coordenadas polares, $x = r\cos\theta$, $y = r\sin\theta, u = \omega\, cos\varphi, v = \omega\sin\varphi$ se puede representar una función como $f(r,\theta)$ y su transformada de Fourier como $F(\omega,\varphi)$, para las que se cumple que:

$$\mathcal{F}\{f(r,\theta+\theta_0)\} = F(\omega,\varphi+\theta_0). \tag{10}$$

4) Simetría. Si se tiene el par transformado bidimensional $f(x,y)$ y $F(u,v)$ (Pratt, 2001), entonces:

$$\mathcal{F}\{f^*(x,y)\} = F^*(-u,-v). \tag{11}$$

Si $f(x,y)$ es una función real, su transformada de Fourier es simétrica conjugada (Gonzalez & Woods 2002), esto es:

$$F(u,v) = F^*(-u,-v), \tag{12}$$

y además, el espectro es simétrico respecto al origen, es decir, $|F(u,v)| = |F(-u,-v)|$.

Teorema de Parseval. Representa la conservación de la energía. La energía en el dominio espacial es igual a la energía en el dominio de frecuencia (Kumar, Mahalanobis, & Juday, 2005).

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|f(x,y)|^2dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|F(u,v)|^2dudv. \tag{13}$$

2.2.2. Transformada de Fourier discreta

Al tratar con imágenes digitales, se tiene que la función *f*(*x*, *y*) es discreta, es decir la imagen continua es muestreada $M \times N$ veces en las direcciones *x* y *y*, respectivamente, separadas por Δx y Δy unidades. La transformada de Fourier bidimensional discreta de una imagen digital $f(x,y)$ de tamaño $M \times N$ define como (Gonzalez & Woods, 2008):

$$F(u,v) = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)e^{-2i\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)}, \tag{14}$$

para $u = 0,1,2,\ldots,M-1$ y $v = 0,1,2,\ldots,N-1$. Análogamente, dada $F(u,v)$, la transformada de Fourier inversa discreta se define como:

$$f(x,y) = \frac{1}{MN}\sum_{u=0}^{M-1}\sum_{v=0}^{N-1} F(u,v)e^{2i\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)}, \tag{15}$$

para $x = 0,1,2,\dots,M-1$ y $y = 0,1,2,\dots,N-1$.

2.2.3. Convolución y el teorema de convolución

La convolución de dos funciones bidimensionales en el dominio espacial está dada por la integral (González & Woods, 2008):

$$f(x,y) * h(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(m,n)h(x-m,y-n)dmdn, \tag{16}$$

donde el operador $*$ denota la convolución.

El teorema de la convolución está dado por las expresiones (González & Woods, 2008):

$$\mathcal{F}\{f(x,y) * h(x,y)\} = F(u,v)H(u,v), \tag{17}$$
$$\mathcal{F}\{f(x,y)h(x,y)\} = F(u,v) * H(u,v),$$

donde $F(u,v)$ y $H(u,v)$ son las transformadas de Fourier de $f(x,y)$ y $h(x,y)$, respectivamente. Por lo tanto, la ecuación de la convolución también se puede expresar como:

$$f(x,y) * h(x,y) = \mathcal{F}^{-1}\{F(u,v)H(u,v)\}. \tag{18}$$

Para funciones discretas, la convolución está dada por (González & Woods, 2008):

$$f(x,y) * h(x,y) = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} f(m,n)h(x-m,y-n), \tag{19}$$

para $x = 0,1,2,\dots,M-1$ y $y = 0,1,2,\dots,N-1$. Esta expresión obtiene un periodo de una secuencia periódica bidimensional.

2.2.4. Correlación y el teorema de correlación

La correlación de dos funciones bidimensionales complejas está dada por:

$$f(x,y) \circ h(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(m,n)h^*(x+m,y+n)dmdn, \tag{20}$$

donde el operador $\circ$ denota la correlación.

El teorema de correlación establece que

$$\mathcal{F}\{f(x,y) \circ h(x,y)\} = F(u,v)H^*(u,v), \tag{21}$$
$$\mathcal{F}\{f(x,y)h^*(x,y)\} = F(u,v) \circ H(u,v).$$

Por lo tanto, se puede calcular la correlación mediante la expresión:

$$f(x,y) \circ h(x,y) = \mathcal{F}^{-1}\{F(u,v)H^*(u,v)\}. \tag{22}$$

Si $f(x,y) = h(x,y)$, es decir, son la misma función, entonces del teorema de correlación se tiene que $\mathcal{F}\{f(x,y) \circ f(x,y)\} = |F(u,v)|^2 = F(u,v)F^*(u,v)$ y $\mathcal{F}\{|f(x,y)|^2\} = F(u,v) \circ F(u,v)$, que denotan la llamada autocorrelación. De otra manera, se le llama correlación cruzada para hacer hincapié que se trata de funciones distintas. Para funciones discretas, la correlación está dada por:

$$f(x,y) \circ h(x,y) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} f(m,n)h^*(x+m, y+n), \tag{23}$$

para $x = 0,1,2,\dots,M-1$ y $y = 0,1,2,\dots,N-1$.

2.3 Sistemas lineales

Los sistemas lineales tienen varias propiedades que ofrecen ciertas ventajas en el procesamiento de imágenes. Aquí se define la linealidad en una dimensión, que es fácilmente extrapolada al caso bidimensional.

Un sistema transforma un conjunto de funciones de entrada en otro conjunto de funciones de salida. El proceso de transformación se puede representar a través de uno o varios operadores. Sea $\boldsymbol{T}$ un sistema de un operador que produce una función de salida $g(x)$ para una función de entrada $f(x)$, se tiene (Gonzalez & Woods, 2008):

$$\boldsymbol{T}[f(x)] = g(x). \tag{24}$$

La linealidad consiste en una suma ponderada de funciones de entrada que da como resultado una suma idénticamente ponderada de las funciones de salida (Kumar et al., 2005). Formalmente, si para dos funciones de entrada $f_1(x)$ y $f_2(x)$ se tienen las salidas $g(x)$ y $g_2(x)$, respectivamente y para dos constantes *a* y *b*, un sistema lineal debe cumplir con lo siguiente (Gonzalez & Woods, 2008):

$$\begin{aligned} \boldsymbol{T}[af_1(x) + bf_2(x)] &= a\boldsymbol{T}[f_1(x)] + b\boldsymbol{T}[f_2(x)] \\ &= ag_1(x) + bg_2(x). \end{aligned} \tag{25}$$

Un sistema invariante a desplazamiento es aquel en el que si a una función de entrada se le aplica un desplazamiento x', la salida resultante está desplazada por la misma cantidad (Kumar et al., 2005). Es decir:

$$\boldsymbol{T}[f(x-x')] = g(x-x'), \tag{26}$$

para toda entrada $f(x)$ y todo desplazamiento x'.

Sea el impulso $\delta(x-x')$, el término

$$h(x;x') = \boldsymbol{T}[\delta(x-x')], \tag{27}$$

denota la respuesta al impulso del sistema $\boldsymbol{T}$ en la posición x (Gonzalez & Woods, 2008). Si $\boldsymbol{T}$ es invariante a desplazamiento, entonces

$$\boldsymbol{T}[\delta(x-x')] = h(x-x'). \tag{28}$$

Por lo tanto, un sistema lineal $\boldsymbol{T}$ se puede caracterizar completamente por su respuesta al impulso $h(x)$. Si se conoce la respuesta de $\boldsymbol{T}$ a un impulso, la respuesta $g(x)$ a cualquier función $f(x)$ se puede calcular en términos de $h(x)$ a través de la convolución. Esto es (Bovik, 2005):

$$\begin{aligned} g(x) &= T[f(x)] \\ &= T\left[\int_{-\infty}^{\infty} f(x')\delta(x-x')dx'\right]; \end{aligned} \tag{29}$$

si el sistema es lineal, entonces

$$\begin{aligned} g(x) &= \int_{-\infty}^{\infty} f(x')\, T[\delta(x-x')]dx' \\ &= \int_{-\infty}^{\infty} f(x')\, h(x;x')dx'; \end{aligned} \tag{30}$$

si además el sistema es invariante a desplazamiento, entonces

$$\begin{aligned} g(x) &= \int_{-\infty}^{\infty} f(x')\, h(x-x')dx' \\ &= f(x) * h(x). \end{aligned} \tag{31}$$

Dado que la transformada de Fourier es una transformación lineal y, tomando en cuenta sus otras propiedades, se puede decir que caracteriza un sistema lineal invariante a desplazamiento (LSI, Linear Shift Invariant). La transformada de Fourier de la respuesta al impulso $h(x)$ se conoce como respuesta en frecuencia o función de transferencia del

sistema. En el procesamiento de imágenes, los filtros lineales se caracterizan en términos de su respuesta en frecuencia; las operaciones de correlación y convolución se llevan a cabo, comúnmente, en el espacio de frecuencia mediante los algoritmos de la transformada rápida de Fourier (FFT).

2.4 Distorsiones en imágenes

Las imágenes digitales registradas son el resultado de un proceso de adquisición y cuantización que producen diferentes degradaciones de la escena original. Las degradaciones son distorsiones que pueden ser causadas por movimiento del sensor, condiciones atmosféricas o errores en el sistema de adquisición. Este último es llamado ruido, el cual es información indeseable que contamina la imagen. Las variaciones geométricas, en cambio, se refieren a distorsiones que se presentan en la imagen registrada debido a la perspectiva y estado de la cámara.

2.4.1. Ruido

Un modelo sencillo de distorsión es aquel que sólo presenta ruido. Este modelo se puede expresar como una imagen $f(x,y)$ afectada por ruido aditivo $n(x,y)$, considerado independiente, produciendo la imagen distorsionada $g(x,y)$:

$$g(x,y) = f(x,y) + n(x,y), \tag{32}$$

(González & Woods, 2008). Por lo general, el ruido puede modelarse matemáticamente con algunas distribuciones de probabilidad. Los modelos comúnmente utilizados son los siguientes.

1) Ruido Gaussiano. Éste se define por la función de densidad de probabilidad gaussiana como:

$$p(z) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(z-\bar{z})^2}{2\sigma^2}}, \tag{33}$$

donde z representa la intensidad, $\bar{z}$ es el valor medio de z y σ es su desviación estándar (González & Woods, 2008).

2) Ruido Uniforme. La función de densidad de probabilidad de una variable uniforme está dada por (González & Woods, 2008):

$$p(z) = \begin{cases} \dfrac{1}{b-a}, & si\ a \le z \le b \\ 0, & de\ otra\ forma \end{cases}. \tag{34}$$

Para esta distribución, la media y la varianza están definidas como (González & Woods, 2008):

$$\bar{z} = \frac{a+b}{2} \quad \text{y} \quad \sigma^2 = \frac{(b-a)^2}{12}.$$

3) Ruido Impulsivo (sal y pimienta). Este ruido se define por la función de densidad de probabilidad:

$$p(z) = \begin{cases} P_a & para\ z = a \\ P_b & para\ z = b \\ 0 & en\ otro\ caso \end{cases}. \tag{35}$$

Si $b > a$, la intensidad b aparecerá como un punto luminoso en la imagen. Contrariamente, el nivel a aparecerá como un punto oscuro. Si P_a o P_b son cero, el ruido impulsivo se llama unipolar. Si ninguna de las probabilidades es cero, en especial si son iguales, el ruido se asemejará a granos de sal y pimienta distribuidos aleatoriamente en la imagen (González & Woods, 2008).

2.4.2. Distorsiones geométricas

Estas distorsiones se basan, generalmente, en coordenadas cartesianas, en las que el origen de la imagen continua (0,0), se encuentra en la esquina inferior izquierda, mientras que para la imagen discreta está en la esquina superior izquierda con índices (1,1). Existen relaciones para pasar de una representación a otra. A continuación se describen algunas de las distorsiones respecto a las coordenadas cartesianas (Pratt, 2001).

1) Traslación. En el espacio vectorial, esta transformación se representa como:

$$\begin{bmatrix} x_k \\ y_j \end{bmatrix} = \begin{bmatrix} u_q \\ v_p \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix}, \tag{36}$$

donde t_x y t_y son las constantes de traslación y (x_k ,y_j) es la nueva posición del pixel (u_q ,v_p) (Pratt, 2001).

2) Rotación. La rotación se puede obtener calculando:

$$\begin{bmatrix} x_k \\ y_j \end{bmatrix} = \begin{bmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{bmatrix} \begin{bmatrix} u_q \\ v_p \end{bmatrix}, \tag{37}$$

donde θ es el ángulo de la rotación con respecto al eje horizontal de la imagen de entrada y (x_k, y_j) es la nueva posición del pixel (u_q, v_p) (Pratt, 2001).

3) Escalamiento. Esta distorsión se obtiene modificando la imagen de entrada como:

$$\begin{bmatrix} x_k \\ y_j \end{bmatrix} = \begin{bmatrix} s_x & 0 \\ 0 & s_y \end{bmatrix} \begin{bmatrix} u_q \\ v_p \end{bmatrix}, \tag{38}$$

donde s_x y s_y son las constantes de escalamiento positivas, pero no necesariamente enteras, aplicadas al pixel (u_q, v_p) y (x_k, y_j) es el pixel escalado (Pratt, 2001).

2.5. Evaluación de sistemas de reconocimiento de rostros

Existen protocolos de experimentación para diseñar y evaluar SRR que indican cómo se debe hacer la evaluación y cómo calcular los resultados. Dos de los más importantes son FERET[1] (Phillips, Moon, Rizvi, & Rauss, 2000) y FRVT[2] 2002 (este último está diseñado para sistemas biométricos en general), que son la base de las evaluaciones FRVT 2006 y MBE 2010[3] (Li & Jain, 2005).

Como se mencionó en el capítulo 1, existen tres tareas que un SRR puede realizar, dentro de los dos modos de operación: verificación, identificación en conjunto abierto e identificación en conjunto cerrado. Cada una, establecidas en los protocolos mencionados, tienen sus medidas de desempeño, siendo la identificación de conjunto abierto el caso general y las otras dos casos particulares.

En general, para calcular el desempeño se requieren tres conjuntos de imágenes: el primero llamado galería ($\mathcal{G}$) contiene, en términos de sistemas biométricos, las muestras biométricas de las personas conocidas por el sistema. Los otros dos conjuntos son de prueba; una prueba es una muestra biométrica que se le presenta en la entrada al sistema ya sea para verificarla o identificarla. El primer conjunto de prueba ($\mathcal{P}_{\mathcal{G}}$) contiene las muestras biométricas de las personas pertenecientes a la galería (pertenecen a las mismas personas pero son diferentes muestras). El segundo conjunto ($\mathcal{P}_{\mathcal{N}}$) contiene muestras biométricas de personas que no están en la galería (Li & Jain, 2005).

En la Figura 3 se muestra un esquema de estos conjuntos de imágenes.

---

[1] Face recognition technology.
[2] Face recognition vendor test.
[3] Multiple biometric evaluation.

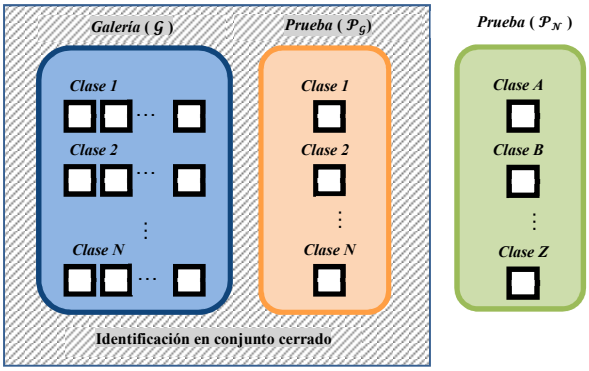


Figura 3. Conjuntos de imágenes para la evaluación de un SRR. La galería se puede ver como el conjunto de entrenamiento y el conjunto de prueba formado por dos subconjuntos $\mathcal{P}_{\mathcal{G}}$ y $\mathcal{P}_{\mathcal{N}}$.

2.5.1. Identificación de conjunto abierto

En la tarea de identificación de conjunto abierto, dada una muestra de prueba, se tiene como objetivo contestar: ¿se sabe de quién es este rostro? Como la persona en la prueba no necesariamente está en la galería, el sistema debe decidir si la prueba pertenece a una persona de la galería y, en caso de que así sea, entonces reportar la identidad de dicha persona. Cuando la galería es pequeña, al problema que resuelve esta tarea se le conoce como watch list, en inglés. Tanto la tarea de identificación en conjunto abierto como en conjunto cerrado se consideran como correspondencia 1 a *N*.

Dicho lo anterior, 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 pertenece a la galería, entonces el algoritmo de reconocimiento elegido identifica a la persona en la prueba.

La galería $\mathcal{G}$ es un conjunto de muestras biométricas $\{g_1, g_1, \dots, g_{|\mathcal{G}|}\}$, con una muestra biométrica por persona, suponiendo el caso más simple. Cuando se le presenta una prueba p_j al sistema, ésta se compara con cada muestra de la galería g_i produciendo un valor de similitud s_{ij}. Los valores de similitud más grandes indican que ambas muestras son muy similares. Comúnmente se usa una medida de distancia entre las muestras biométricas, pudiendo convertirse a un valor de similitud al negarla. Entonces, un valor de similitud s_{ij} es un valor de correspondencia si g_i y p_j son muestras biométricas de la misma persona. Si p_j es una muestra de una persona en la galería, entonces g^* es su correspondencia única en la galería y su valor de similitud se denota como s_{*j}. Sea *id*() la función que regresa la identidad de una muestra biométrica, entonces $id(p_j) = id(g^*)$.

Para la identificación, todos los valores de similitud entre la prueba p_j y una galería, se examinan y se ordenan (Li & Jain, 2005).

Rango de correspondencia. Una prueba p_j tiene rango *n* si s_{*j} es el *n-ésimo* valor de similitud más grande. Esto se denota como $rango(p_j) = n$. El rango n = 1 también es conocido como mejor correspondencia.

2.5.2. Verificación

En la tarea de verificación, considerada como una correspondencia 1 a 1, se presenta una muestra biométrica en la entrada del sistema y se afirma una identidad. El sistema debe decidir si la muestra pertenece a la identidad afirmada. Por lo tanto, esta tarea pretende responder: ¿es esta persona quien afirma ser? (Li & Jain, 2005).

2.5.3. Identificación de conjunto cerrado

La identificación de conjunto cerrado es la tarea clásica utilizada por la comunidad de reconocimiento de rostros automático, a la cual sencillamente se le conoce como identificación, para medir el desempeño de un SRR. En esta tarea, dada una muestra de prueba, se pretende responder ¿de quién es este rostro?, ya que la muestra pertenecerá necesariamente a alguna persona de la galería. Como se indicó, es un caso particular de la identificación de conjunto abierto en la que el conjunto de prueba $\mathcal{P}_{\mathcal{N}}$ está vacío.

Ahora bien, la pregunta no siempre es ¿la mejor correspondencia es correcta?, sino ¿está la respuesta correcta dentro de las mejores *n* correspondencias? Para responderla, primero se ordenan los valores de similitud entre p_j y la galería $\mathcal{G}$ y se calcula el $rango(p_j)$. La tasa de identificación para el rango *n*, $P_I(n)$, es la fracción de pruebas en rango *n* o menor. Para rango *n*, la cuenta acumulativa del número de pruebas en rango *n* o menor se define:

$$C(n) = \left|\{p_j : rango(p_j) \leq n\}\right|. \tag{39}$$

La tasa de identificación en rango *n* se calcula con:

$$P_I(n) = \frac{|C(n)|}{|\mathcal{P}_{\mathcal{G}}|}. \tag{40}$$

Las funciones $C(n)$ y $P_I(n)$ son no decrecientes respecto a n. La tasa de identificación en rango n = 1, $P_I(1)$, también se conoce como la tasa de identificación correcta o tasa de mejor correspondencia (Li & Jain, 2005). Este último criterio es el que se usa para evaluar los diferentes métodos de reconocimiento de rostros en este proyecto.

Comúnmente, el desempeño de la identificación en conjunto cerrado se reporta con una curva característica de correspondencia acumulada (CMC[4], o también CMS[5]), la cual grafica $P_I(n)$ como función del rango n. Asimismo, se resume su desempeño con el de rango n = 1; aunque también se utilizan los rangos 5, 10 ó 20. La desventaja de la CMC es la dependencia en el tamaño de la galería $|\mathcal{G}|$ (Li & Jain, 2005). Estas curvas son útiles cuando se tienen matrices de similitud.

2.6.1. Evaluación con validación cruzada

En la evaluación, desde un punto de vista de algoritmos de aprendizaje, los datos usados en el entrenamiento (o galería), frecuentemente dependen de la habilidad de especialistas, por lo que se dispone de pocos datos. Predecir el desempeño basado en una cantidad limitada de datos es un problema controversial (Witten, Frank, & Hall, 2011).

Para problemas de clasificación el desempeño de un clasificador comúnmente se mide en términos de la tasa de error. Si el clasificador predice correctamente la clase de una instancia dada, se cuenta como éxito, si no, se cuenta como error. La tasa de error es la proporción de errores sobre el total del conjunto de instancias; mide el desempeño global del clasificador. Al tratarse de reconocimiento de rostros, se emplea la tasa de identificación, definida en la sección anterior y referida como tasa de reconocimiento en las secciones siguientes. Generalmente, mientras más grande sea la muestra de entrenamiento, mejor es el clasificador, aunque los resultados comienzan a disminuir al excederse de un cierto volumen de entrenamiento (Witten et al., 2011). También, mientras más grande sea la muestra de prueba, más precisa es la estimación del error.

Prácticamente siempre se tiene el problema de poca disponibilidad de información o de datos insuficientes. Sin embargo, puede sacarse el máximo provecho de un conjunto de datos limitado, apartando una porción de éstos para las pruebas (procedimiento de reserva)

[4] Cumulative Match Characteristic
[5] Cumulative Match Score

y el resto usarlo para entrenamiento (y si es necesario, una parte usarla para validación). Se suele tomar un tercio para prueba y el resto para entrenamiento.

En ocasiones la muestra de entrenamiento o de prueba no es representativa, por lo que, idealmente, se hace un muestreo aleatorio para garantizar que cada clase esté representada apropiadamente tanto en el conjunto de entrenamiento como en el de prueba. A este procedimiento le llaman estratificación o reserva estratificada. Para evitar sesgos debidos a la elección de una muestra en particular, en el método de reserva se repite todo el proceso, entrenamiento y prueba, varias veces con diferentes muestras aleatorias. Las tasas de error en cada iteración se promedian para obtener la tasa de error total. Este es el método reserva repetitiva de estimación de tasa de error.

La validación cruzada es otra variante del método de reserva en la que se elige un número de particiones en que se dividirán los datos; una partición para prueba y el resto para entrenamiento. Por cada partición, se hace una iteración, en la que se alterna el conjunto de prueba y entrenamiento para que al final cada partición haya servido de prueba una sola vez. Este es el método de validación cruzada de "*k*" particiones; si se usa la estratificación, entonces se trata de validación cruzada de "*k*" particiones estratificada (o bien, validación cruzada de "*k*" particiones aleatoria). En la Figura 4 se muestra un esquema de cómo tomar las particiones para formar los conjuntos de entrenamiento y prueba de la validación cruzada con "*k*" particiones y cómo calcular la tasa de error promedio. La técnica de evaluación estándar cuando se dispone de datos limitados, es la validación cruzada de 10 particiones aleatoria. Aunque también se usa con 5 o con 20 particiones.

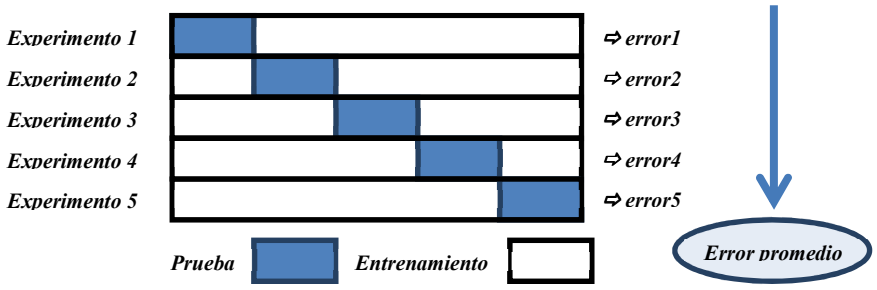


Figura 4. Esquema de validación cruzada de *k* particiones, donde *k* = 5.

Capítulo 3. Técnicas de detección y reconocimiento de rostros

El rostro humano, aunque en su forma como imagen digital es altamente dimensional, su apariencia es bastante restringida. Es decir, sus características son similares: es en general simétrico, los ojos aparecen casi siempre a la misma altura y en ambos lados, la nariz en el centro, etc. Se podría decir que están restringidos a un subespacio, llamémosle espacio de rostros, del espacio altamente dimensional de las imágenes de rostros digitales. En este capítulo se presentan tres técnicas basadas en subespacios en las que cada una tiene una representación particular, dependiendo de diferentes puntos de vista estadísticos.

3.1 Eigenfaces

Una de las técnicas más populares para tratar el problema de reconocimiento de rostros es la de Eigenfaces. Ésta surgió a partir de la idea de Sirovich y Kirby (1987) y Kirby y Sirovich (1990) de utilizar el PCA para la representación de rostros de manera económica. Ellos lo emplearon, en principio, para la compresión de información aprovechando que es un método de reducción de dimensiones. PCA es un método estadístico que transforma linealmente un conjunto de datos a otros, llamados componentes principales. Cada componente contiene la varianza de los datos originales; el primer componente contiene la máxima varianza, el segundo contiene la segunda máxima varianza y así sucesivamente. Posteriormente, Turk y Pentland (1991) propusieron su uso para el reconocimiento de rostros, dando origen a la técnica de Eigenfaces.

Supóngase que se tienen *M* imágenes de rostros de tamaño $N \times N$, pertenecientes a distintos individuos, a los que nos referiremos como clases, disponibles para entrenar un sistema de reconocimiento. Sea $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M\}$ el conjunto de imágenes de rostros de entrenamiento en su versión vector (calculado de manera lexicográfica), donde $\mathbf{x}_i \in \mathbb{R}^d$ y $d = N \times N$. De aquí se esperaría que un rostro, tratándolo como un punto, esté bien representado en este espacio $\mathbb{R}^d$ (altamente dimensional si se considera, por ejemplo, $N = 256$); es decir que todos los rostros del mismo individuo, o clase, estén aglomerados para que, con una técnica de clasificación, se puedan separar de aquellos de otras clases. Se

puede entonces definir la matriz $\mathbf{X} \in d \times M$ que agrupa al conjunto de rostros de entrenamiento en sus columnas. Ahora, sea

$$\boldsymbol{\mu} = \frac{1}{M}\sum_{i=1}^{M} \mathbf{x}_i, \tag{41}$$

la media del conjunto de imágenes de entrenamiento y sea $\boldsymbol{\varphi}_i = \mathbf{x}_i - \boldsymbol{\mu}$, el rostro i centrado, es decir, el vector que indica por cuánto difiere el rostro de la media. El objetivo es obtener un conjunto de vectores que mejor describan la distribución de los datos, en este caso, de los rostros en el espacio completo de imágenes. Estos vectores son los conocidos eigenvectores asociados a los eigenvalores de la matriz de covarianza de las imágenes de rostros de entrenamiento centrados, definida como

$$\boldsymbol{COV} = \frac{1}{\mathrm{M}}\sum_{\mathrm{i}=1}^{\mathrm{M}} \boldsymbol{\varphi}_i \boldsymbol{\varphi}_i{}^{T} = \boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}. \tag{42}$$

La idea básica de esta técnica de reconocimiento de rostros es, en la etapa de entrenamiento, calcular los eigenvectores del conjunto de imágenes de entrenamiento, los cuales caracterizan la variación entre las imágenes de rostros y son llamados eigenfaces porque al visualizarlos tienen apariencia semejante a un rostro fantasmal. Dado que la matriz $\boldsymbol{COV} \in d \times d$, se trata de determinar $d = N^2$ eigenvectores y eigenvalores, lo cual se vuelve un problema intratable considerando los tamaños típicos de las imágenes (si $N = 256$, entonces se tendría una matriz de 65536×65536). Sin embargo se pueden calcular los eigenvalores y eigenvectores de la matriz **COV** al encontrar primero los eigenvalores y eigenvectores de una matriz $M \times M$. Sea $\mathbf{L} = \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}$, considérese

$$\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{v}_{\mathrm{i}} = \lambda_{\mathrm{i}}\mathbf{v}_{\mathrm{i}}, \tag{43}$$

donde $\mathbf{v}_{\mathrm{i}}$ son los eigenvectores de **L**. Multiplicando ambos lados de (43) por $\boldsymbol{\Phi}$ se obtiene:

$$\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{v}_{\mathrm{i}} = \lambda_{\mathrm{i}}\,\boldsymbol{\Phi}\mathbf{v}_{\mathrm{i}}, \tag{44}$$

$$\mathbf{COV}\,(\boldsymbol{\Phi}\mathbf{v}_{\mathrm{i}}) = \lambda_{\mathrm{i}}\,(\boldsymbol{\Phi}\mathbf{v}_{\mathrm{i}}),$$

por lo que $\mathbf{w}_{\mathrm{i}} = \boldsymbol{\Phi}\mathbf{v}_{\mathrm{i}}$ son los eigenvectores de **COV** y, por lo tanto, son éstos los eigenfaces con los que, mediante una combinación lineal, se pueden representar todos los rostros de entrenamiento considerados.

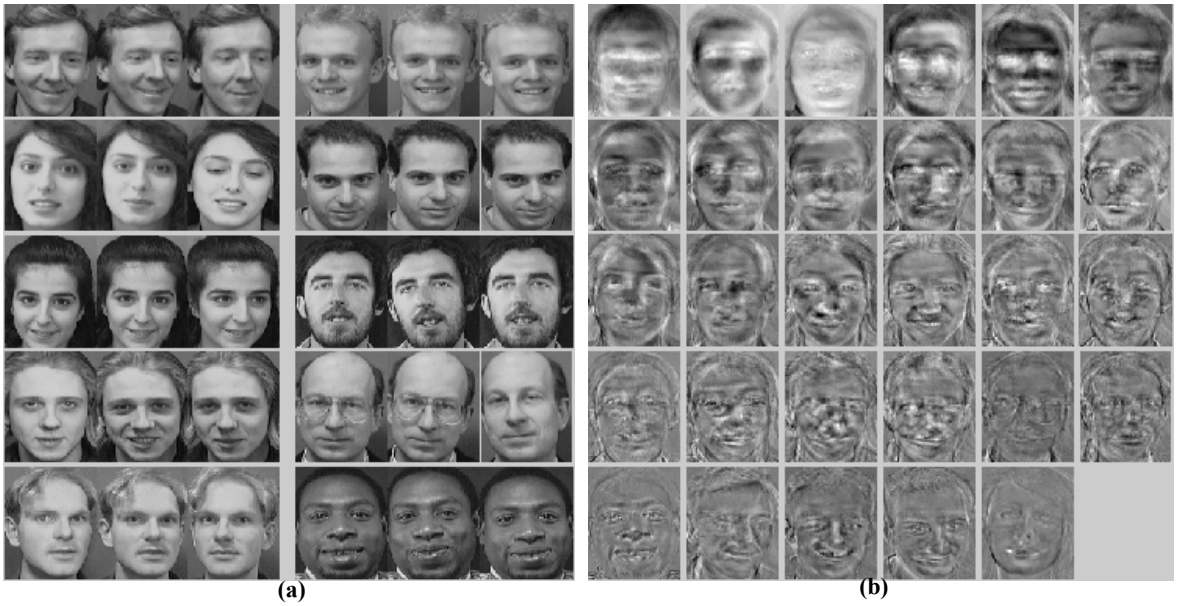


Figura 5. Rostros y eigenfaces. (a) Conjunto de imágenes de rostros de entrenamiento. (b) Primeros 29 eigenfaces correspondientes a los eigenvectores calculados de la matriz de covarianza del conjunto de entrenamiento.

Se toman sólo los mejores *K* eigenvectores, que corresponden a los *K* eigenvalores más grandes, después de ordenarlos, los cuales proveen la información más útil para caracterizar la variación entre las imágenes. De esta manera se obtiene una matriz de proyección $\mathbf{W} = [\mathbf{w}_2, \mathbf{w}_2, \dots, \mathbf{w}_\mathrm{K}]$ con la que se puede obtener una representación de las imágenes de entrenamiento en un espacio de menor dimensión. Es decir, se hace una proyección de cada rostro del conjunto de entrenamiento al espacio de eigenfaces, obteniendo un conjunto de pesos que se almacenan para su uso en la etapa de reconocimiento. De hecho, cada rostro del conjunto de entrenamiento se puede aproximar usando sólo los mejores *K* eigenfaces. Los pesos de proyección se obtienen con:

$$\mathbf{y}_i = \mathbf{W}^T \boldsymbol{\varphi}_i. \tag{45}$$

En la etapa de reconocimiento, el sistema se alimenta de una imagen de rostro a reconocer que se proyecta al espacio de eigenfaces, obteniendo una serie de pesos, de la misma manera que las imágenes de entrenamiento, previamente centrada con la media del conjunto de entrenamiento, con la ecuación (45). Para realizar la clasificación, el sistema luego compara los pesos del rostro de entrada al sistema con cada uno de aquellos

calculados y almacenados previamente en la etapa de entrenamiento, determinando si se trata de un rostro conocido o desconocido por medio de algún clasificador.

En este proyecto se utiliza el clasificador de vecinos cercanos con la distancia euclidiana o norma L2 cuadrada, definida por:

$$d_{L2}(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_2^2 = \sum_i (x_i - y_i)^2, \tag{46}$$

para dos vectores **x** y **y**.

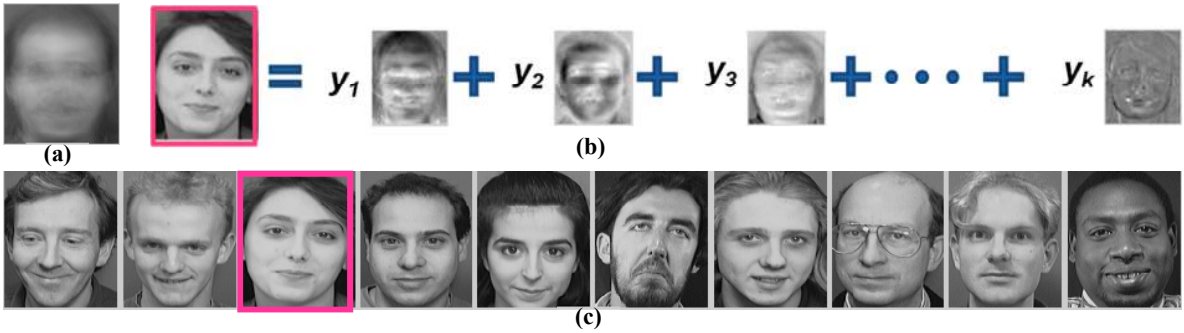


Figura 6. Proyección de un rostro al subespacio de eigenfaces. (a) Media del conjunto de entrenamiento. (b) Respresentación del rostro enmarcado mediante la combinación lineal de eigenfaces. (c) Imágenes de rostros de prueba.

La Figura 6 muestra la media del conjunto de entrenamiento de la Figura 5(a), un conjunto de imágenes de prueba y la representación del rostro de prueba enmarcado, mediante la combinación lineal de los eigenfaces calculados de la Figura 5(b). La gran ventaja de esta técnica es que provee una solución práctica del problema de reconocimiento de rostros. Por otro lado, su desempeño es influenciado por el fondo presente en las imágenes.

3.2 Fisherfaces

El LDA es una técnica estadística que utiliza el discriminante lineal de Fisher (FLD[6]) para maximizar la diferencia entre clases de un conjunto de datos y minimizar la diferencia intraclase de los mismos. Se emplea en problemas de clasificación y en la técnica de Fisherfaces (Belhumeur et al, 1997) se usa para el reconocimiento de rostros.

Como se vio en la sección anterior, PCA busca un conjunto de vectores ortonormales que maximizan el esparcimiento (o variación) de los datos. En términos de la matriz de esparcimiento total ($\mathbf{S}_T$), o matriz de covarianza, se tiene:

[6] Fisher Linear Discriminant.

$$\mathbf{S}_T = \sum_{i=1}^{M} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T. \tag{47}$$

En otras palabras, la matriz de proyección PCA ($\mathbf{W}_{PCA}$) se obtiene maximizando el esparcimiento total de los datos (ver ecuación (47)), por lo que se tiene:

$$\mathbf{W}_{PCA} = \max_{W} |\mathbf{W}^T \mathbf{S}_T \mathbf{W}| = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_K], \tag{48}$$

que se resuelve calculando los eigenvectores de la matriz $\mathbf{S}_T$: $\mathbf{S}_T \mathbf{w}_i = \lambda_i \mathbf{w}_i$, $i = 1, 2, \dots, K$.

Sin embargo, aunque PCA proyecta los datos de tal manera que se capte la mayor variación posible entre todos los datos, no permite la separabilidad de las clases pues maximiza tanto la variación debido al esparcimiento entre clases como la variación debido al esparcimiento intraclase, lo cual no es útil para la clasificación. Por ejemplo, cuando la variabilidad de los datos proviene por cambios de iluminación, el espacio creado por PCA incorporará estas variaciones y los rostros proyectados a ese espacio no estarán correctamente aglomerados, pudiendo provocar que las clases se mezclen (Belhumeur et al, 1997). Por el contrario, lo que hace el FLD es proyectar los datos de tal manera que haya la mayor separación posible entre los elementos que pertenecen a distintas clases y la menor separación posible entre los elementos de una misma clase (ver Figura 7).

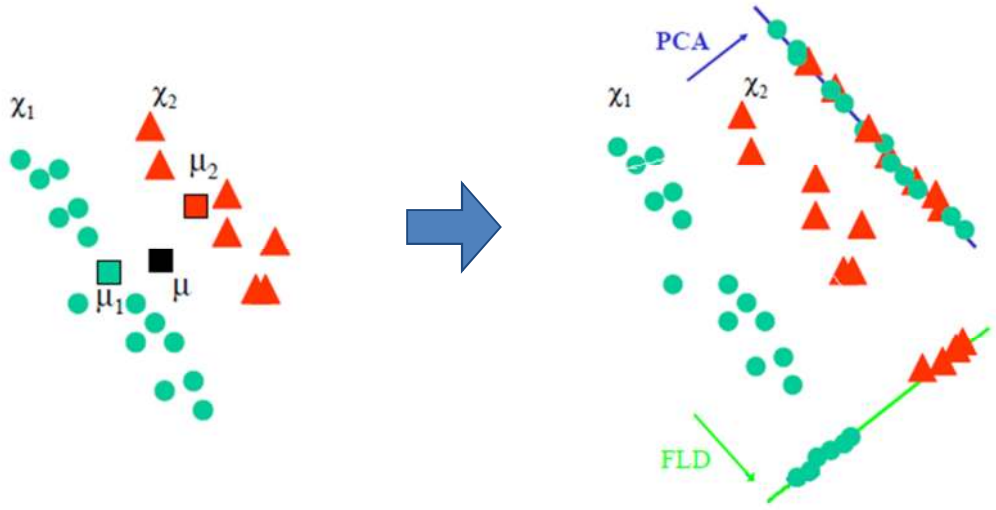


Figura 7. Diferencia entre proyecciones encontradas con PCA y con LDA, donde χ_1 y χ_2 denominan dos clases. Recuperada de http://www4.comp.polyu.edu.hk/~comp435/notes/lec08-face-recog.pdf (por Z. Li)

La técnica de Fisherfaces aprovecha que se cuenta con un conjunto de entrenamiento etiquetado, es decir, se conocen la clase a la que pertenece cada individuo.

Similarmente a como se estableció para PCA, sea $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M\}$ el conjunto de imágenes de rostros de entrenamiento en su versión vector, pero que pertenecen a una de c clases $\{\chi_1, \chi_2, \dots, \chi_c\}$. La técnica requiere del cálculo de tres matrices: la matriz de esparcimiento entre clases ($\mathbf{S}_B$), la matriz de esparcimiento intra-clases ($\mathbf{S}_W$) y la matriz de esparcimiento total ($\mathbf{S}_T$), la cual es la suma de las matrices de esparcimiento $\mathbf{S}_B$ y $\mathbf{S}_W$. Las matrices $\mathbf{S}_B$ y $\mathbf{S}_W$ se definen como:

$$\mathbf{S}_B = \sum_{i=1}^{C} |\chi_i| (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T, \tag{49}$$

y

$$\mathbf{S}_W = \sum_{i=1}^{C} \sum_{x_k \in \chi_i} (\mathbf{x}_k - \boldsymbol{\mu}_i)(\mathbf{x}_k - \boldsymbol{\mu}_i)^T, \tag{50}$$

donde $|\chi_i|$ es el número de imágenes de rostros de la clase χ_i, $\boldsymbol{\mu}_i$ es la media de la clase χ_i y $\boldsymbol{\mu}$ es la media de todo el conjunto de entrenamiento definida en (41).

La proyección del discriminante lineal de Fisher ($\mathbf{W}_{FLD}$) se obtiene maximizando la razón entre el esparcimiento entre clases y el esparcimiento intra-clases, esto es:

$$\mathbf{W}_{FLD} = \max_W \frac{|\mathbf{W}^T \mathbf{S}_B \mathbf{W}|}{|\mathbf{W}^T \mathbf{S}_W \mathbf{W}|} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_K], \tag{51}$$

lo cual es un problema de optimización que puede resolverse calculando los eigenvectores de $\mathbf{S}_B$ y $\mathbf{S}_W$; es decir $\mathbf{S}_B \mathbf{w}_i = \lambda_i \mathbf{S}_W \mathbf{w}_i$ o bien $\mathbf{S}_W{}^{-1} \mathbf{S}_B \mathbf{w}_i = \lambda_i \mathbf{w}_i, \ \ i = 1, 2, \dots, K$.

Ahora, esta proyección es óptima cuando $\mathbf{S}_W$ es no-singular; es decir, cuando tiene inversa. Sin embargo, cuando se trata con conjuntos de imágenes de rostros para reconocimiento, se tiene que $\mathbf{S}_W$ es casi siempre singular; su rango es menor a $M - c$ y por lo general la dimensión de cada imagen en versión vector es mucho mayor que el número de imágenes de entrenamiento. Para sobrellevar el problema de la singularidad, la técnica de Fisherfaces primero proyecta el conjunto de imágenes de entrenamiento a un espacio de menor dimensión, para que la matriz $\mathbf{S}_W$ sea no-singular, utilizando la técnica de PCA. Primero se proyectan a un espacio de $M - c$ dimensiones y posteriormente se aplica la proyección del FLD para reducir el espacio obtenido a $c - 1$ dimensiones.

Tomando la ecuación (48), se tiene que la matriz de proyección del FLD ahora se define como:

$$\mathbf{W}_{FLD} = \max_{W} \frac{\left|\mathbf{W}^T \mathbf{W}_{PCA}{}^T \mathbf{S}_B \mathbf{W}_{PCA} \mathbf{W}\right|}{\left|\mathbf{W}^T \mathbf{W}_{PCA}{}^T \mathbf{S}_W \mathbf{W}_{PCA} \mathbf{W}\right|}. \quad (52)$$

Por último, la proyección que se utiliza en la técnica de Fisherfaces está compuesta por la proyección $\mathbf{W}_{PCA}$ y la proyección $\mathbf{W}_{FLD}$, como sigue:

$$\mathbf{W}_{Fisher}{}^T = \mathbf{W}_{FLD}{}^T \mathbf{W}_{PCA}{}^T. \quad (53)$$

Resumiendo, en la etapa de entrenamiento de esta técnica se hace una proyección de cada rostro del conjunto de entrenamiento al espacio de eigenfaces, calculando los eigenvectores como en la técnica de Eigenfaces, pero se toman los $K = M - c$ mejores eigenvectores, donde M es el número de imágenes de entrenamiento y c es el número de clases. Se quiere obtener una representación de las imágenes de entrenamiento en un espacio de menor dimensión al mismo tiempo que se maximiza la separación entre clases y se minimiza la separación intra-clase, por lo que se hace la proyección de cada una de las imágenes del subespacio al espacio de Fisher calculando los eigenvectores de las matrices de esparcimiento entre clases e intra-clases de los datos proyectados previamente. Estos eigenvectores, llamados fisherfaces, se muestran en la Figura 8. Similarmente como en la técnica de Eigenfaces, las proyecciones resultan en pesos que se almacenan para uso en la etapa de reconocimiento. Estos pesos de proyección se obtienen de la siguiente manera:

$$\mathbf{y}_i = \mathbf{W}_{Fisher}{}^T \boldsymbol{\varphi}_i. \quad (54)$$

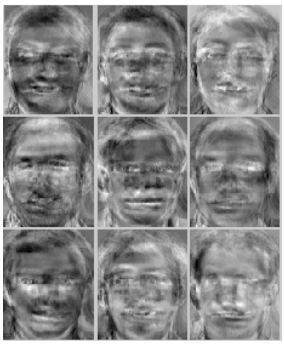


Figura 8. Ejemplo de los c-1 Fisherfaces, donde c = 10.

En la etapa de reconocimiento, el sistema se alimenta de una imagen de rostro a reconocer que se proyecta, de la misma manera que las imágenes de entrenamiento, al espacio de Fisherfaces, obteniendo una serie de pesos. El sistema luego compara los pesos del rostro de entrada al sistema con aquellos calculados y almacenados previamente en la

etapa de entrenamiento, determinando si se trata de un rostro conocido o desconocido por medio de algún clasificador.

En este proyecto se utiliza el algoritmo de vecinos cercanos utilizando la distancia euclidiana o norma L2 cuadrada, como con la técnica de Eigenfaces, de la ecuación (46).

Una de las ventajas que tiene esta técnica es que presenta cierta tolerancia a variaciones de iluminación y expresiones faciales. Sin embargo no considera cambios en postura.

3.3 Análisis de componentes independientes para el reconocimiento de rostros

Esta técnica se deriva del problema de la fiesta de coctel que se puede ejemplificar como sigue: suponga que cuatro personas hablan en cuatro puntos distintos de un salón y cuatro micrófonos captan mezclas de las cuatro señales de voz (ver Figura 9).

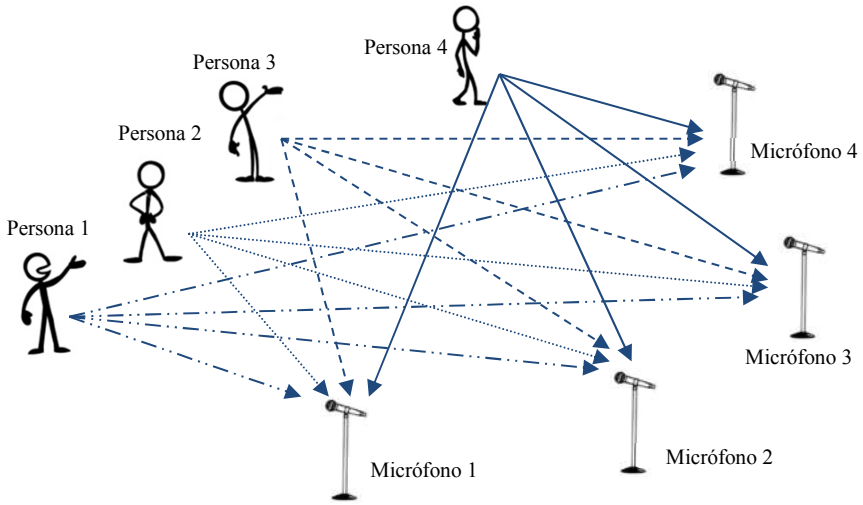


Figura 9. Problema de la fiesta de coctel.

Las señales que producen los micrófonos, defínanse como $x_1(t)$, $x_2(t)$, $x_3(t)$ y $x_4(t)$, son una combinación lineal de las señales de voz emitidas por las personas, denotadas por $s_1(t)$, $s_2(t)$, $s_3(t)$ y $s_4(t)$. Por lo tanto, las señales $x_i(t)$ se pueden expresar en forma matricial como:

$$\begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ x_4(t) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \begin{bmatrix} s_1(t) \\ s_2(t) \\ s_3(t) \\ s_4(t) \end{bmatrix},$$

donde los parámetros a_{ij} dependen de la distancia entre las personas y los micrófonos. Podría ser útil determinar las señales de voz originales $s_i(t)$ a partir de las señales producidas por los micrófonos $x_i(t)$ sin conocimiento alguno de los parámetros de mezcla a_{ij} (Hyvärinen, Karhunen, & Oja, 2001). En realidad, suponiendo que la matriz **A** (que consiste de los parámetros a_{ij}) es invertible, se pueden separar las señales originales haciendo una estimación de la matriz de mezcla **A** (o bien de la matriz de separación **W**) a partir de la mezcla de señales, tal que $\mathbf{u} = \mathbf{Wx} = \mathbf{W}(\mathbf{As})$, como se muestra en la Figura 10. En este caso, la matriz **W** es una estimación de la inversa de la matriz **A**.

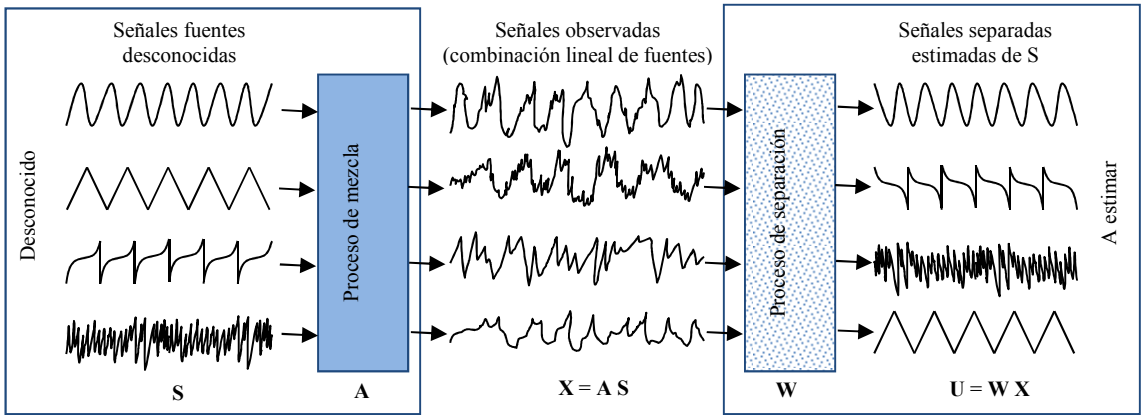


Figura 10. Separación y estimación de las señales fuente.

Originalmente se trata este problema con el ICA propuesto por Comon (1994). En la actualidad existen varios algoritmos que realizan ICA; el desarrollado por Bell y Sejnowski (1995) fue exitosamente aplicado para separar señales auditivas mezcladas aleatoriamente y también para separar señales de electroencefalograma (EEG). Más aún, este algoritmo de ICA también se ha empleado para el problema de reconocimiento de rostros. Bartlett, Lades, & Sejnowski, (1998) y Bartlett, Movellan y Sejnowski (2002), propusieron dos arquitecturas para abordar el problema de reconocimiento de rostros mediante el ICA.

La técnica del ICA básica funciona con dos grandes pasos que se ejemplifican a continuación con datos bidimensionales. Supóngase que se tiene un par de variables aleatorias (A y B) y se obtiene una combinación lineal de ellas; por lo tanto el campo que forman está correlacionado (ver Figura 11).

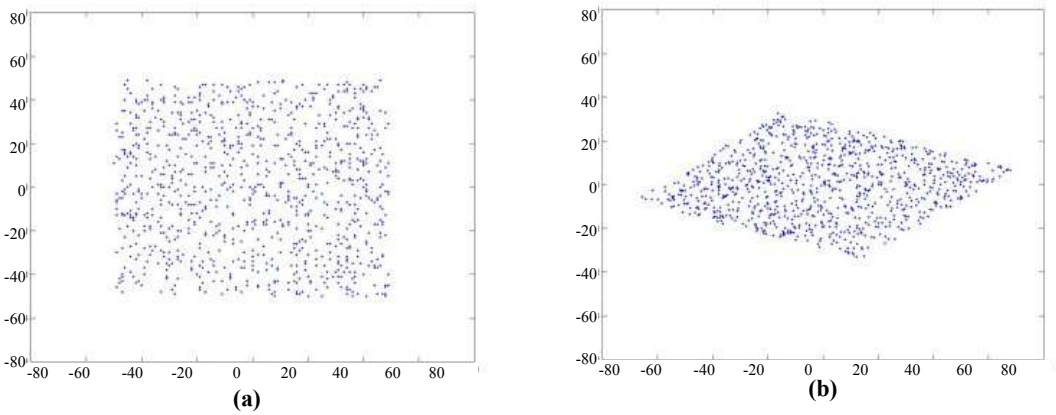


Figura 11. Campo aleatorio de variables A y B en (a) y su combinación lineal en (b). Recuperado de http://sccn.ucsd.edu/~arno/indexica.html (por Arnaud Delorme)

Cuando están correlacionados los datos, al tener información sobre una de las variables, automáticamente se tiene información de la otra variable y viceversa. El primer paso consiste en un preprocesamiento muy utilizado: el de blanqueado, el cual elimina la correlación de los datos. Se puede observar en la Figura 12 (a) que las variables tienen una distribución gaussiana. Para estimar los elementos originales o regresar al campo original, ICA calcula la rotación necesaria de los datos, minimizando la gaussianidad de los mismos, como se aprecia en la Figura 12 (b). Esta rotación corresponde al segundo paso.

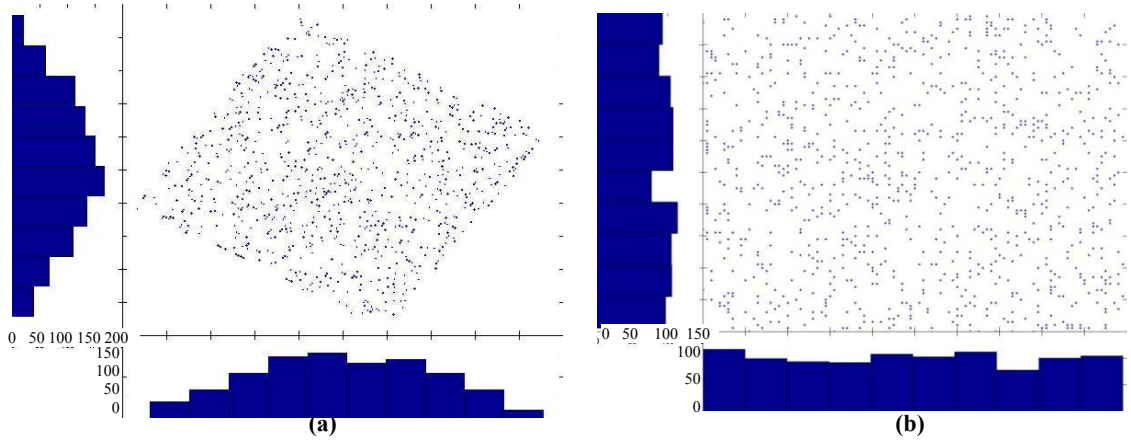


Figura 12. Campo blanqueado (a) y rotación calculada por ICA (b). Recuperado de http://sccn.ucsd.edu/~arno/indexica.html (por Arnaud Delorme)

En particular, la Arquitectura I propuesta por Bartlett et al., (1998) y Bartlett et al., (2002) supone una serie de imágenes de rostros **X** que son combinaciones lineales de otras imágenes fuente desconocidas **S**. Las combinaciones están dadas por una matriz de mezcla

desconocida **A** y, como en el problema de la fiesta de coctel, se pueden estimar los componentes (imágenes fuente) independientes (ver Figura 13).

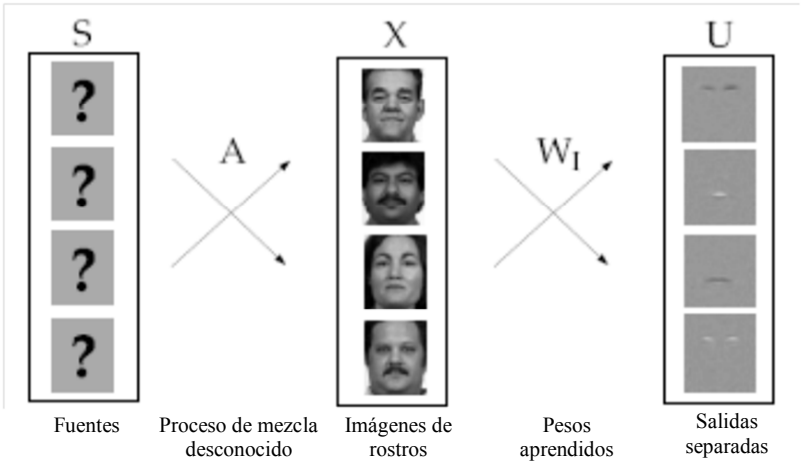


Figura 13. Modelo de síntesis de imagen para la Arquitectura I de ICA. Recuperada de Bartlett (2001), 45p.

Como se mencionó anteriormente, en el algoritmo ICA para la Arquitectura I, el objetivo es encontrar un conjunto de imágenes base, estadísticamente independientes. A diferencia de como se maneja la matriz de imágenes de rostro de entrenamiento en PCA y LDA, en esta técnica se organiza el conjunto de imágenes en sus versiones vector renglón. Es decir, sea $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M\}$ un conjunto de imágenes de rostros en su versión vector (calculado de manera lexicográfica), donde $\mathbf{x}_i \in \mathbb{R}^d$ y $d = N \times N$, la matriz que agrupa a este conjunto de rostros de entrenamiento se define ahora como $\mathbf{X} \in M \times d$, con cada imagen $\mathbf{x}_i$ forma un renglón de la matriz. Generalmente, lo que hace ICA es encontrar una matriz **W** tal que $\mathbf{U} = \mathbf{WX}$ sean tan estadísticamente independientes como sea posible, donde **U** son los componentes que se quieren estimar.

En la Arquitectura I, el proceso de blanqueado descrito anteriormente, se realiza pasando los datos **X** centrados por la matriz $\mathbf{W}_Z$**,** calculada con la covarianza de **X**. Esto es:

$$\mathbf{W}_Z = 2 \cdot (\mathbf{X}\mathbf{X}^T)^{-\frac{1}{2}}. \tag{55}$$

Este proceso elimina los estadísticos de primer y segundo orden de los datos. Tanto la media como la covarianza son ceros y las varianzas se ecualizan.

Para la estimación de los componentes independientes en **U**, el proceso de rotación descrito anteriormente tiene como objetivo maximizar la no-gaussianidad de **WX**. Existen varias medidas de no-gaussianidad; una de las más comunes es la kurtosis (considerada

como la versión normalizada del cuarto momento de una variable aleatoria) y otra es la neguentropía, la cual se basa en la cantidad de entropía teórica en términos de información. Varios algoritmos maximizan la no-gaussianidad de diferentes maneras; otros enfoques son los de maximización de información mutua y de estimación de máxima verosimilitud (Hyvärinen et al., 2001). En la Arquitectura I se realiza la estimación de los componentes independientes con el algoritmo Infomax (Bell & Sejnowski, 1995) (perteneciente al último enfoque mencionado), el cual encuentra la matriz de proyección **W**. Este algoritmo es una regla de aprendizaje no supervisado que se deriva del principio de transferencia óptima de información en neuronas a través de funciones de transferencia sigmoidales. El algoritmo supone una entrada x arbitraria y una salida y, que se pasan por una función logística g (Bartlett, 2001):

$$y = g(u) = \frac{1}{1 + e^{-u}} \quad u = wx + w_0. \tag{56}$$

"El peso óptimo w para maximizar la transferencia de información sobre x es aquel que maximiza la entropía de la salida. Este peso óptimo se encuentra aplicando un ascenso de gradiente en la entropía de la salida y respecto a w" (Bartlett, 2001, p.42). "Cuando se tienen múltiples entradas y salidas, la maximización de la entropía conjunta de la salida procura que las salidas individuales se muevan hacia la independencia estadística" (Bartlett, 2001, p.42). La regla de actualización para la matriz de pesos **W**, para múltiples entradas y salidas está dada por (Bartlett, 2001):

$$\Delta \mathbf{W} = (\mathbf{I} + y' u^T)\mathbf{W}, \tag{57}$$

donde $y' = \frac{\partial}{\partial y_i}\frac{\partial y_i}{\partial u_i} = \frac{\partial}{\partial u_i}\ln\frac{\partial y_i}{\partial u_i}$. En la implementación de Bartlett et al., (1998) y Bartlett et al., (2002) se usó $y' = (1 - 2y_i)$.

En la Arquitectura I de ICA, se obtiene finalmente una matriz de proyección (o transformación) $\mathbf{W}_I = \mathbf{W}\mathbf{W}_Z$, donde la matriz **W** es aquella aprendida o calculada por Infomax y $\mathbf{W}_Z$ es la matriz del blanqueado. El ICA no se realiza directamente sobre las entradas **X** sino que se realiza sobre una combinación lineal de ellas. Dado que las imágenes de rostro en **X** se suponen como una combinación lineal de un conjunto de fuentes independientes desconocidas, el modelo no se afecta si se usa otra combinación lineal de las imágenes. Con el propósito de reducir las dimensiones del problema, se aplica

PCA sobre las imágenes en $\mathbf{X}$, tomando sólo m componentes correspondientes a los eigenvalores más grandes obteniendo la matriz $\mathbf{P}_m$, con los componentes formando las columnas de la matriz (proceso descrito en la sección de PCA).

Luego, en la etapa de entrenamiento, se obtiene la representación (o pesos) de las imágenes de entrenamiento centradas $\mathbf{X}$ en el subespacio de PCA, que se define como $\mathbf{R}_m = \mathbf{X}\,\mathbf{P}_m$. La aproximación o reconstrucción de las imágenes de entrenamiento se puede obtener con $\widetilde{\mathbf{X}} = \mathbf{R}_m\,{\mathbf{P}_m}^T$.

Considerando la matriz de proyección $\mathbf{W}_I$, se tienen los componentes estimados

$$\mathbf{W}_I{\mathbf{P}_m}^T = \mathbf{U} \quad \Rightarrow \quad {\mathbf{P}_m}^T = {\mathbf{W}_I}^{-1}\mathbf{U}, \tag{58}$$

por lo que

$$\widetilde{\mathbf{X}} = \mathbf{R}_m\,{\mathbf{W}_I}^{-1}\mathbf{U}. \tag{59}$$

De esta manera, los coeficientes de la combinación lineal de las fuentes estadísticamente independientes $\mathbf{U}$ (o pesos de representación en el espacio ICA) para $\widetilde{\mathbf{X}}$ se obtienen de los renglones de

$$\mathbf{B} = \mathbf{R}_m\,{\mathbf{W}_I}^{-1}. \tag{60}$$

Un ejemplo de las fuentes o componentes independientes encontrados por esta técnica se muestran en la Figura 14.

En la etapa de reconocimiento, al alimentarse el sistema de una o varias imágenes de rostro a reconocer, para cada una de estas imágenes de prueba se calcula su representación (o pesos) en el espacio de ICA; es decir, se proyecta al espacio de ICA. Primeramente se calcula su representación en el subespacio de PCA, con $\mathbf{R}_{prueba} = \mathbf{X}_{prueba}\,\mathbf{P}_m$ y posteriormente se proyecta, como en la ecuación (60), como $\mathbf{B}_{prueba} = \mathbf{R}_{prueba}\,{\mathbf{W}_I}^{-1}$.

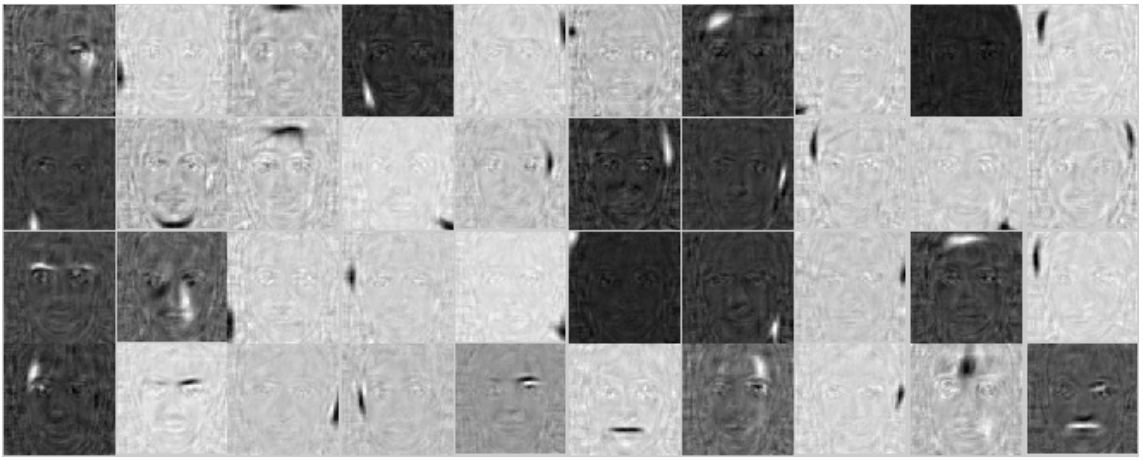


Figura 14. 40 de los 137 componentes independientes encontrados de un conjunto de entrenamiento de 342 individuos. Estos componentes se derivan del 40% de los componentes principales del conjunto de entrenamiento.

Finalmente, teniendo los pesos de proyección tanto del conjunto de imágenes de entrenamiento, previamente calculados y almacenados, como los de la(s) imagen(es) de prueba, éstos se comparan y se determina si se trata de un rostro conocido mediante un clasificador, exactamente como en las técnicas de Eigenfaces y de Fisherfaces.

En la propuesta de Bartlett et al., (1998) y Bartlett et al., (2002) se utiliza la clasificación con el algoritmo de vecinos cercanos, utilizando la distancia coseno entre los vectores de pesos **b** de entrenamiento y prueba, definida por el ángulo coseno:

$$\alpha = \frac{\boldsymbol{b}_{prueba} \cdot \boldsymbol{b}_{entrenamiento}}{\|\boldsymbol{b}_{prueba}\| \cdot \|\boldsymbol{b}_{entrenamiento}\|}. \tag{61}$$

Esta técnica presenta cierta tolerancia a variaciones por el paso del tiempo y cambios de iluminación.

3.4 Algoritmo de detección Viola & Jones

Viola y Jones (2001) propusieron una plataforma de detección de objetos, posteriormente para rostros (Viola & Jones, 2004), con una capacidad de procesamiento de imágenes muy rápida y tasas de detección altas. Las tres contribuciones clave que aportaron fueron: la extracción eficaz de características de las imágenes a través de una representación propuesta llamada imagen integral; la segunda es la utilización de clasificadores simples para la selección de características críticas o más importantes, a partir de un gran número de potenciales de ellas, con una variante de un algoritmo de

aprendizaje. Por último, la propuesta de un método de combinación de clasificadores en cascada para mejor desempeño en la detección.

3.4.1. Características

La plataforma de detección propuesta por Viola y Jones usa características similares a las funciones de base Haar. Para calcular estas características de manera rápida, se introdujo la representación de la imagen integral, la cual requiere pocas operaciones por pixel. Originalmente se usan tres tipos de características: bi-rectangular, tri-rectangular y cuatri-rectangular como se muestra en la Figura 15 (a).

Las características bi-rectangulares contienen valores que corresponden a la diferencia de la suma de pixeles en dos regiones rectangulares del mismo tamaño y forma siendo adyacentes vertical u horizontalmente (ver Figura 15). El valor de las características tri-rectangulares es la diferencia de la suma de pixeles en dos regiones rectangulares exteriores y la suma de la región rectangular interior o central. El valor de las características cuatri-rectangulares se calcula sumando las diferencias entre la suma de pixeles de pares diagonales de regiones rectangulares como se muestra en la Figura 15 (a). La suma de pixeles en la región blanca de cada característica se substrae de la suma de pixeles en la región negra.

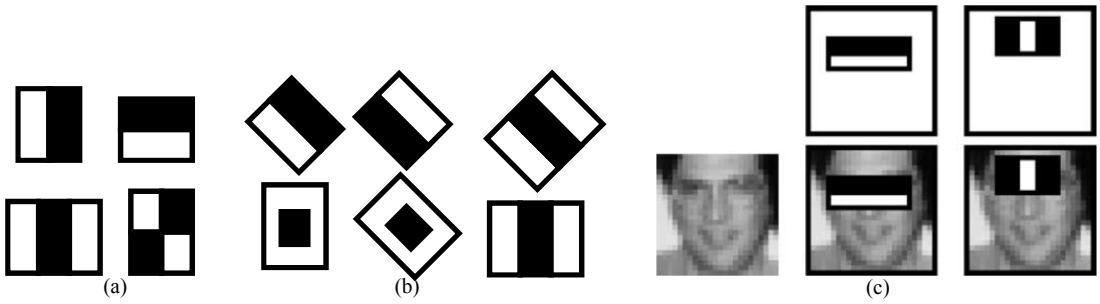


Figura 15. Características tipo Haar. (a) Características originales. (b) Características extendidas. (c) Características relevantes para detección de ojos y nariz. Imagen (c) recuperada de Viola y Jones (2004), 144p.

En algunas implementaciones recientes del detector de Viola y Jones se utilizan extensiones de las características tipo Haar (Figura 15(b)), propuestas por Lienhart y Maydt (2002), como es el caso del detector utilizado en este proyecto.

3.4.2. Imagen integral

Una de las innovaciones de esta plataforma de detección es la forma de calcular las características rectangulares rápidamente usando una representación intermediaria. A dicha representación se le llama imagen integral, la cual contiene la suma de pixeles arriba y a la izquierda de una coordenada (x,y) dada, inclusive (ver Figura 16 (a)), definida como:

$$ii(x,y) = \sum_{x' \leq x, y' \leq y} i(x',y'), \tag{62}$$

donde $ii(x,y)$ es la imagen integral e $i(x,y)$ es la imagen original. Usando el siguiente par de recurrencias:

$$s(x,y) = s(x,y-1) + i(x,y), \tag{63}$$

$$ii(x,y) = ii(x-1,y) + s(x,y), \tag{64}$$

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

Con la imagen integral, cualquier suma rectangular de pixeles de cualquier tamaño, puede calcularse con cuatro referencias a un arreglo, como se muestra en la Figura 16 (b). Por lo tanto, la diferencia entre dos sumas rectangulares se puede calcular usando ocho referencias; no obstante, como se trata de dos regiones rectangulares adyacentes se puede calcular con sólo seis referencias. Para las características tri-rectangulares se calcula con ocho referencias y para las cuatri-rectangulares con nueve referencias.

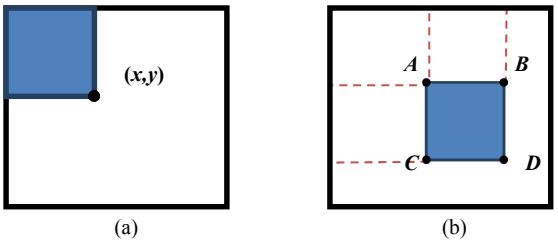


Figura 16. Imagen integral y su cálculo. (a) Imagen integral en el punto (x,y) . (b) Cálculo de la imagen integral en la región sombreada a partir de cuatro referencias: A, B, C y D. El área de la región es igual a D + A - B - C.

La gran ventaja que aporta el uso de la imagen integral se debe a que, a diferencia de los enfoques convencionales de detección que se basan en el cálculo de una pirámide de diferentes escalas de la imagen a procesar, lo que se escala es el detector mismo. En este caso, la escala base que se utiliza es de 24×24 pixeles y, en vez de usar el detector a un

tamaño fijo y escalar la imagen, se hace un escaneo con el detector en diferentes escalas sin que el cálculo de las características influya significativamente en el tiempo de procesamiento, pues se realiza con un número de operaciones constante para cualquier tamaño.

3.4.4. Clasificadores

Viola y Jones afirman que en una ventana de 24×24 pixeles se tienen aproximadamente 160,000 posibles características para calcular (de las ilustradas en la Figura 15(a)), ya que éstas se consideran en todos los posibles tamaños y posiciones dentro de la ventana. En cualquier ventana de una imagen, el número de características de tipo Haar es enorme; mucho más grande que el número de pixeles, por ejemplo, para el caso mencionado arriba se trata de 576 pixeles. Aunque las características se pueden calcular eficientemente, su cálculo completo se vuelve muy costoso. Se supone entonces que se puede combinar un número pequeño de estas características para formar un clasificador efectivo, dado que se espera que sólo algunas de ellas obtengan valores relativamente altos cuando se encuentren en la región de un rostro. Por lo tanto, para realizar una clasificación rápida, el proceso de aprendizaje del clasificador debe excluir la mayor parte de las características y concentrarse en un conjunto pequeño de características críticas (Viola & Jones, 2004).

En la propuesta del detector de Viola y Jones, un clasificador simple selecciona un número reducido de las características más importantes pertenecientes al inmenso conjunto de características potenciales (aquellas que pudieran formar parte de un rostro) a partir de una sub-ventana, usando una variante del algoritmo de aprendizaje AdaBoost (Freund & Schapire, 1995), motivados por el trabajo de Tieu y Viola (2000). Originalmente AdaBoost se usa para mejorar el desempeño de clasificación de un algoritmo simple de aprendizaje. Esto lo consigue combinando un conjunto de funciones clasificadoras débiles para formar un clasificador fuerte, como se muestra en la Figura 17. A este proceso le llaman aprendedor débil porque no se espera siquiera que la mejor función de clasificación clasifique muy bien los datos de entrenamiento.

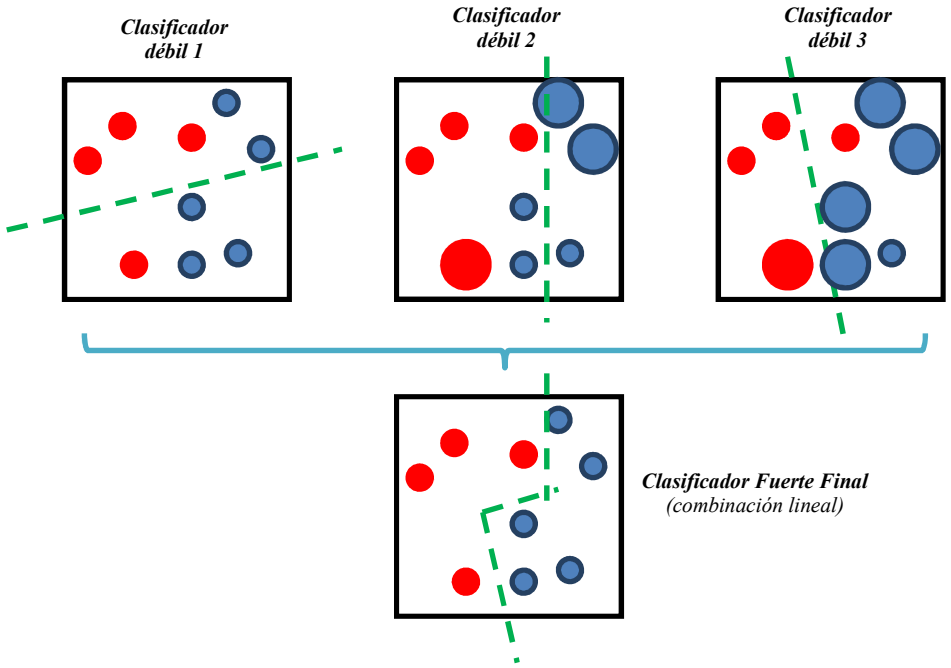


Figura 17. Proceso de *boosting* de clasificadores. De manera iterativa se pueden crear nuevos clasificadores basados en los pesos de los datos que no se clasifican correctamente en los clasificadores anteriores. Al final un clasificador fuerte se crea como combinación lineal de los clasificadores en cada iteración.

El algoritmo AdaBoost utilizado en Viola y Jones funciona de la siguiente manera. Dado un conjunto de características tipo Haar y un conjunto de entrenamiento de imágenes positivas y negativas, denotadas como x_i, con sus respectivas etiquetas de clasificación $y_i \in \{1, 0\}$, cada característica se considera un posible clasificador débil. El clasificador se puede definir matemáticamente como (Viola & Jones, 2004):

$$h(x, f, p, \theta) = \begin{cases} 1, & si\ pf(x) > p\theta \\ 0, & de\ otra\ forma' \end{cases} \tag{65}$$

donde x es la ventana de 24x24 pixeles, f es la característica considerada, p es la polaridad y θ es el umbral que decide si la imagen se clasifica como positiva (un rostro) o negativa (no es rostro). En una iteración dada, el algoritmo seleccionará una sola característica rectangular que mejor separe las imágenes en positivo y negativo, es decir que minimice el número de errores de clasificación del conjunto de entrenamiento. Esto resulta en que en cada paso del proceso de boosting se vea como un proceso de selección que se repite un número T predeterminado de veces.

Aunque el algoritmo de AdaBoost es eficiente, el conjunto de clasificadores débiles que forma puede ser muy grande; hay un clasificador débil por cada combinación de

característica/umbral, además de que depende del número de iteraciones *T* del algoritmo (Viola & Jones, 2004). Por otro lado, un solo clasificador fuerte puede rechazar una gran porción de la imagen que no es rostro, pero no es suficiente para tener niveles bajos de falsas alarmas. Por lo tanto, en la plataforma de detección de Viola y Jones se proponen un método para combinar clasificadores más complejos en una estructura de cascada que incrementa significativamente la velocidad del detector, concentrándose en regiones prometedoras de rostro de la imagen considerada.

3.4.5. Cascada de Clasificadores

La idea es que se pueden construir clasificadores boosted (clasificadores fuertes) más pequeños y por lo tanto más eficientes, que rechazan muchas ventanas negativas, mientras que detectan casi todas las instancias positivas. Este enfoque de cascada se puede ver como una clasificación por etapas en donde las primeras constan de clasificadores fuertes simples, es decir, que contienen pocas características tipo Haar, las cuales rechazan la mayoría de las ventanas analizadas. Las últimas etapas constan de clasificadores fuertes más complejos, que contienen más características, con las que se tiene mayor certeza de que la ventana contiene un rostro y entonces se obtienen tasas bajas de falsos positivos.

El proceso de clasificación en cascada del detector se muestra en la Figura 18. En éste, una sub-ventana de la imagen considerada pasa por la primera etapa en la que, si se determina que no contiene un rostro, inmediatamente se descarta. Si por el contrario, se clasifica como posible rostro, la sub-ventana pasa al clasificador de la siguiente etapa y así sucesivamente hasta que en alguna de ellas se descarte, o que al llegar a la última efectivamente se detecte el rostro. El detector propuesto en Viola y Jones (2004), consta de 38 etapas en la cascada, que en total contienen 6060 características tipo Haar. La primera etapa contiene dos características, la segunda contiene diez características y las siguientes contienen 25, 50 y hasta más características en los clasificadores.

Ahora, como se mencionó al inicio de esta sección, una vez entrenado el detector, éste se escanea a través de la imagen que se quiere procesar en varias escalas y posiciones. Recordando, el detector es el que se escala, no la imagen. Por lo tanto, para cada posición

de la sub-ventana elegida y para cada escala el detector puede arrojar múltiples detecciones. Para resolver esto se requiere de un posprocesamiento del resultado de la detección.

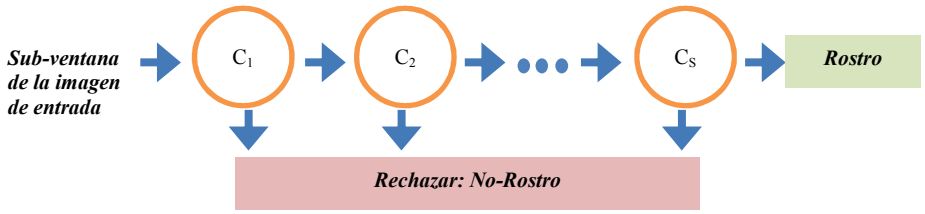


Figura 18. Esquema de clasificadores en cascada para la detección.

En este proyecto se utiliza una implementación del detector de Viola y Jones que se basa en el detector entrenado disponible en la librería de código abierto OpenCV[7].

3.5. Implementación y adaptación de algoritmos estadísticos con segmentación automática

Hasta este momento se han descrito tres de las técnicas de reconocimiento de rostros estadísticas o de subespacio líneal más conocidas y utilizadas. Estas técnicas están diseñadas originalmente para tratar el problema del reconocimiento de rostros, en particular, cuando las imágenes de prueba o de entrada al sistema están segmentadas. En otras palabras se trata sólo de correspondencia de imágenes de rostros, mientras que el problema de reconocimiento en realidad engloba, como se mencionó en el primer capítulo, tanto la detección y localización del rostro en una imagen como su correspondencia e identificación o verificación, según sea el caso. Dado que lo que se pretende con este proyecto es el reconocimiento de rostros en escenas reales sin preprocesamiento manual, la implementación de los métodos descritos en este capítulo fue adaptada mediante el uso del detector de Viola y Jones como preprocesamiento de segmentación automática.

El detector es una implementación en MATLAB realizada por Dirk-Jan Kroon[8] (Kroon, 2010) que utiliza el clasificador entrenado que provee la librería de código abierto OpenCV. El método de Eigenfaces está basado en la implementación de Delac K., Grgic M. y Grgic S. (2006)[9]; el de Fisherfaces en la implementación de Philipp Wagner[10]

[7] http://opencv.org/
[8] http://www.mathworks.com/matlabcentral/fileexchange/29437-viola-jones-object-detection
[9] http://www.face-rec.org/source-codes/
[10] http://www.bytefish.de/documents

(Wagner, 2012) y la implementación de ICA Arquitectura I es la provista por Bartlett et al, (2002)[11].

A continuación se muestra una comparación del desempeño de reconocimiento de estas técnicas, tanto para el caso cuando están segmentadas las imágenes de prueba, como cuando se hace la adaptación con el detector mencionado para imágenes no segmentadas, utilizando la base de datos de rostros Faces 1999 de Caltech, descrita en el apéndice.

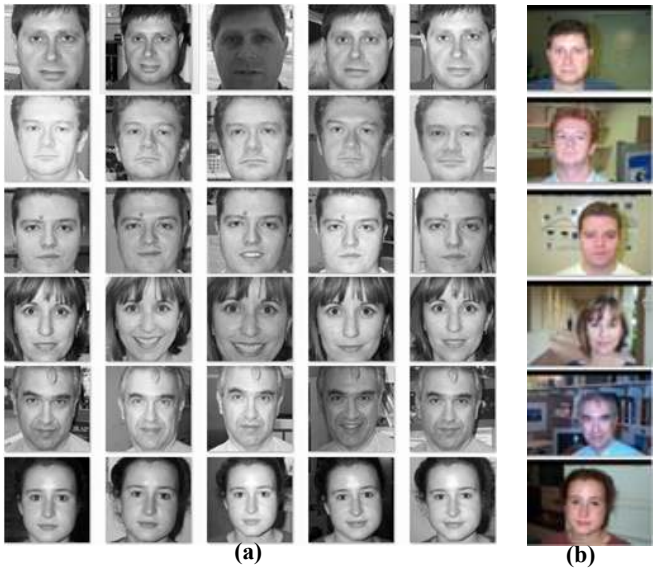


Figura 19. Ejemplos de imágenes de la base de datos de rostros Caltech. (a) Imágenes utilizadas en el entrenamiento de los métodos y como prueba para el enfoque tradicional. (b) Escenas reales utilizadas como prueba para el reconocimiento.

Para ilustrar el efecto en el desempeño de un sistema de reconocimiento de rostros, basado en métodos estadísticos, al incluir una etapa de preprocesamiento que segmenta la región del rostro en la escena de entrada al sistema, se entrenó un sistema con las técnicas de Eigenfaces, Fisherfaces y de ICA Arquitectura I con un conjunto de imágenes de rostros de entrenamiento capturadas en condiciones ideales, como se muestran en la Figura 19 (a). Se utilizaron dos conjuntos de imágenes de prueba: A y B. El conjunto A consta de imágenes capturadas en condiciones ideales o controladas, similares a las de entrenamiento, donde el rostro aparece en la mayor parte de la imagen. El conjunto B consta de escenas reales, es decir, los rostros aparecen en distintas ubicaciones y posiciones en la imagen y, a la vez, con distintos fondos como se observa en la Figura 19 (b). Para el conjunto de prueba

[11] http://mplab.ucsd.edu/~marni/code.html

B se pasó cada una de las escenas por una etapa de segmentación usando el detector descrito anteriormente y sólo aquellas que se pudieron segmentar o en las que se logró detectar un rostro, pasaron como entrada al sistema.

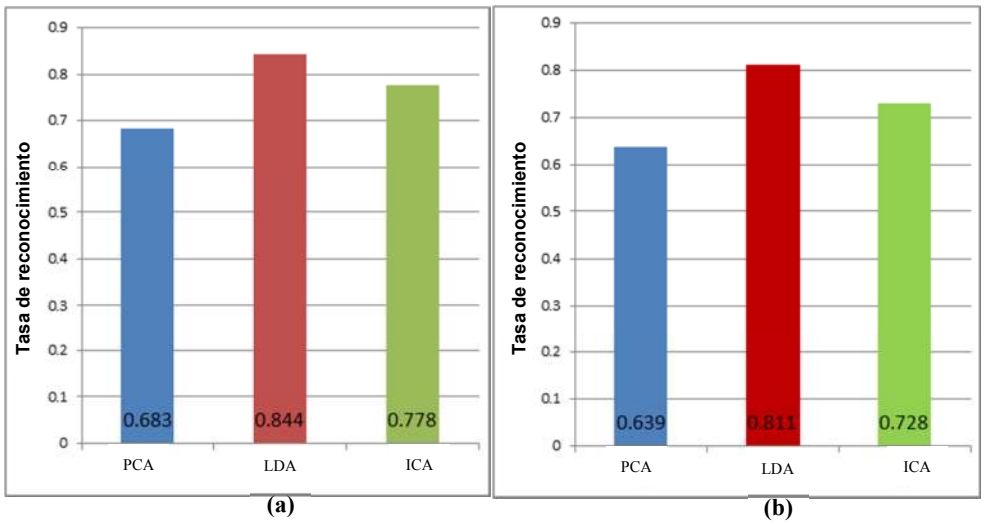


Figura 20. Desempeño respecto a la tasa de reconocimiento de las técnicas de Eigenfaces (PCA), Fisherfaces (LDA) y Análisis de componentes independientes (ICA). (a) Desempeño con imágenes de prueba en condiciones ideales (el conjunto A). (b) Desempeño con escenas reales (el conjunto B) y segmentación automática.

En la Figura 20 se muestra el desempeño, en términos de la tasa de reconocimiento, de las técnicas estadísticas en las situaciones mencionadas. En ésta se observa que al utilizar la etapa de segmentación automática (Figura 20 (b)) el desempeño de cada una de las técnicas disminuye aproximadamente entre 4 y 6%; un 6.4% para Eigenfaces, 3.95% para Fisherfaces y 6.4% para ICA Arquitectura I, respecto a su desempeño con imágenes de prueba de los mismos rostros, pero segmentadas.

Cabe mencionar que el conjunto de entrenamiento para este experimento fue de 10 imágenes por clase, las cuales fueron tomadas secuencialmente de como están registradas. Asimismo los conjuntos de prueba A y B contienen las 10 imágenes por clase restantes en la base de datos. No obstante, es importante señalar que el decremento en el desempeño es influido por varios fenómenos al incluir la etapa se segmentación automática. Si bien se trata de un detector ampliamente conocido y utilizado (el de Viola y Jones) por sus altas tasas de detección, no está libre de producir falsas alarmas. Puede suceder que el detector segmente el rostro pero dentro de una región muy grande, en donde la proporción del fondo es mayor que la del rostro. Esto depende también del posprocesamiento que se le dé a las

posibles múltiples detecciones que entregue el detector. Otra situación puede ser que el detector simplemente segmente una región que no contenga a un rostro, sino a otro objeto con el cual lo confundió, lo que a veces sucede cuando el rostro no está bien iluminado, está oscuro y hay mucho contraste con el fondo o cuando no hay mucho contraste con el fondo, como si no hubiera profundidad.

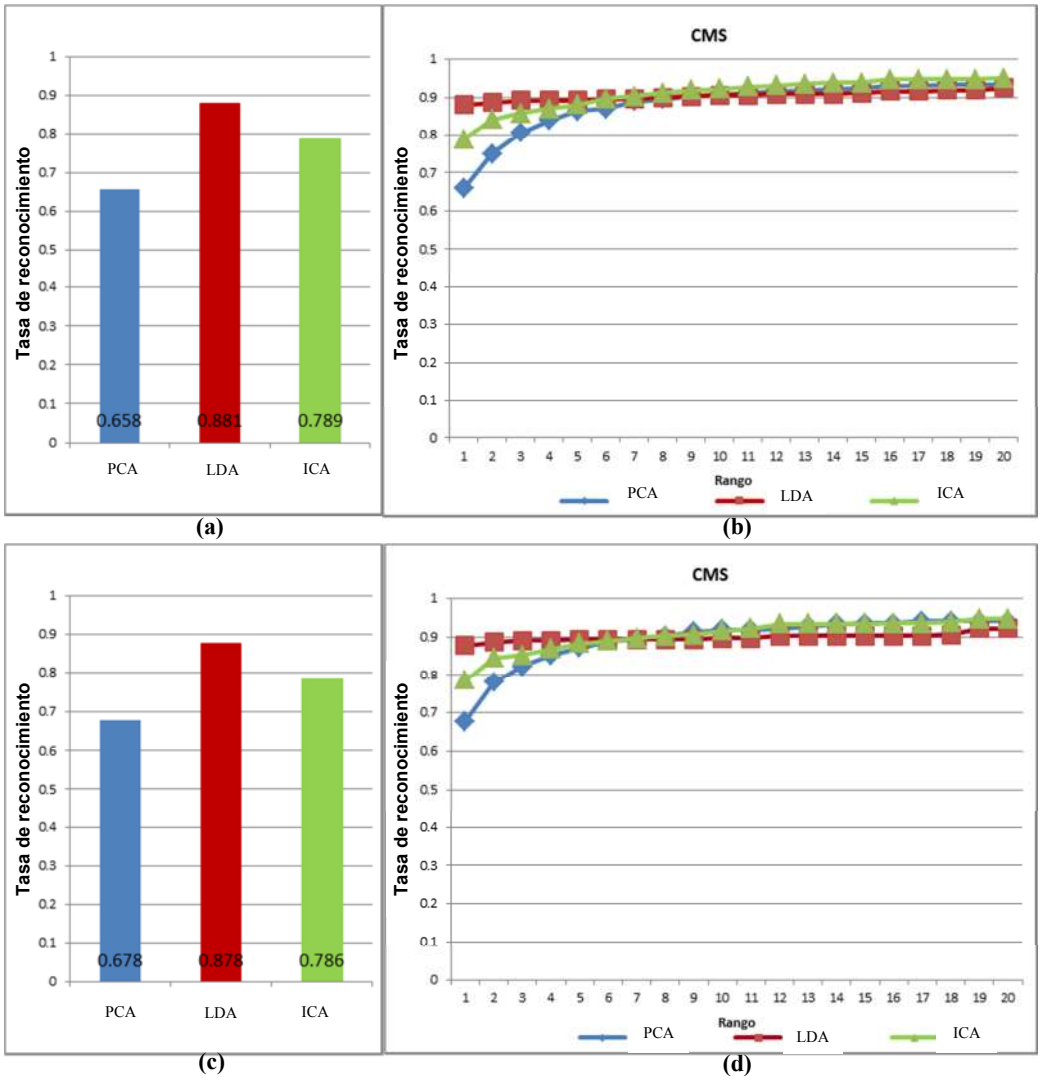


Figura 21. Tasa de reconocimiento promedio y curvas CMS (validación cruzada). (a) y (b) Validación cruzada de 4 particiones. (c) y (d) Validación cruzada de 10 particiones.

En la Figura 21 se observa el desempeño de los tres métodos como resultado de la validación cruzada, de 4 y 10 particiones. Las respectivas curvas CMS revelan que, aunque en promedio, la técnica de Fisherfaces tenga el mejor desempeño en rango 1, la técnica de

ICA Arquitectura I mejora el desempeño al subir el rango; es decir, cubre más el área de la gráfica, lo cual es lo que se espera ver en este análisis.

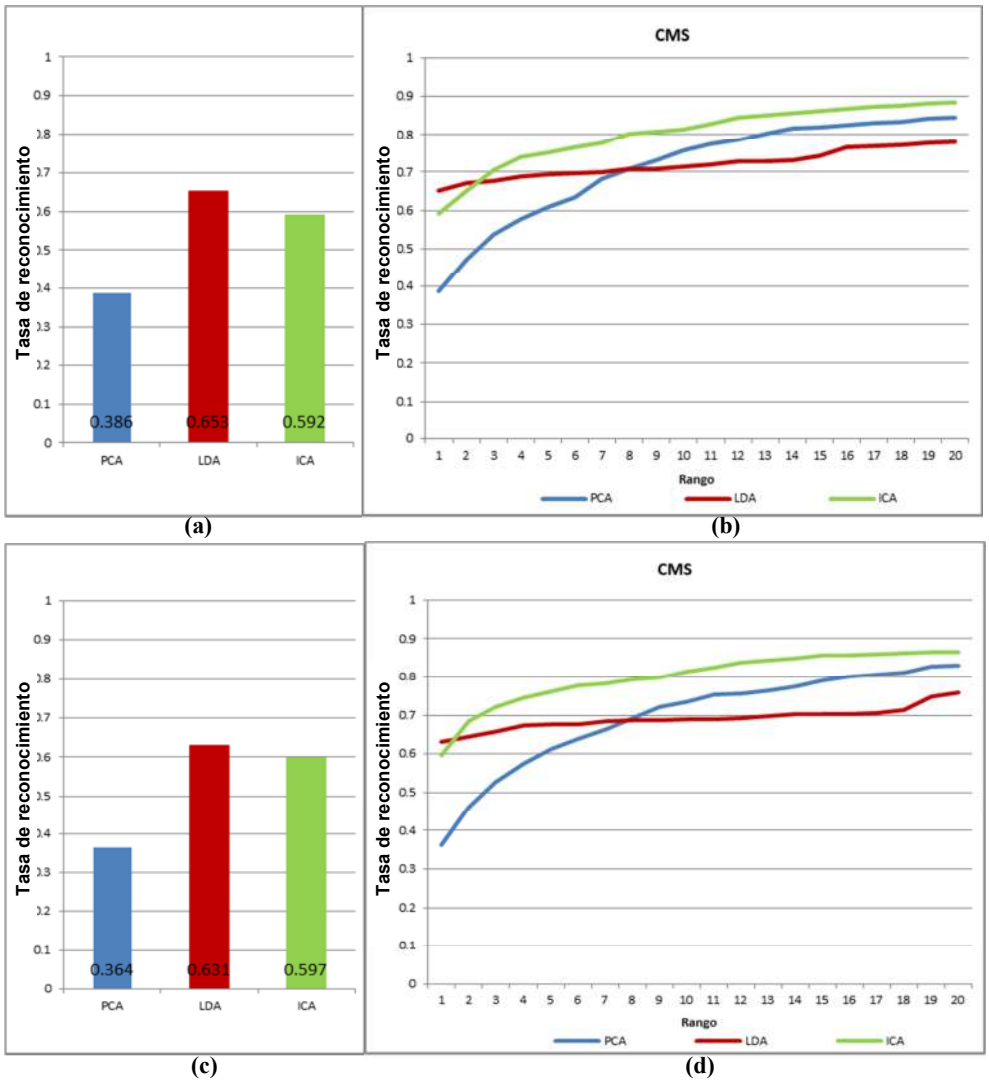


Figura 22. Tasa de reconocimiento promedio y curvas CMS (validación cruzada aleatoria). (a) y (b) Validación cruzada de 4 particiones aleatorias. (c) y (d) Validación cruzada de 10 particiones aleatorias.

Lo mismo sucede al aplicar la validación cruzada con particiones aleatorias, esto es, tomando los elementos de cada partición de manera aleatoria. La curva CMS de la Arquitectura I de ICA va incrementando a medida que sube de rango, como se muestra en la Figura 22 (b) y (d), mientras que la curva de Fisherfaces se mantiene casi constante.

3.6. Resumen

En este capítulo se describieron las técnicas tradicionales de reconocimiento de rostros, basadas en métodos estadísticos, que posteriormente se compararán con la técnica propuesta. Las tres técnicas trabajan con proyecciones a subespacios empleando la reducción de dimensiones. Asimismo se presentó uno de los detectores de rostros más utilizados actualmente, que se empleó para adaptar las técnicas descritas y aplicarlas en el reconocimiento de rostros en escenas reales. Se pudo observar cómo el desempeño del propio detector influye en el de estas técnicas, negativamente.

Capítulo 4. Teoría de reconocimiento por correlación

4.1. Introducción

El enfoque de reconocimiento de patrones por correlación establece el grado de similitud entre una señal de referencia (o, en este caso, una imagen de entrenamiento) seleccionada y la señal de entrada (o imagen de prueba) mediante la operación de correlación, definida en la sección 2.2.4. Ambas imágenes contienen un objeto, a veces referido como objetivo. Éste se desea localizar en la imagen de prueba a través de la información del objeto en la imagen de entrenamiento. La salida de la correlación se puede ver como otra imagen en la que los valores grandes indican la ubicación del objetivo de la imagen de entrenamiento en la imagen de prueba. En ocasiones la salida de correlación presenta valores grandes en ubicaciones que no corresponden al objeto de la imagen de prueba, llamados errores normales, si están cercanos a las coordenadas del objeto, o anómalos, si aparecen en cualquier parte de la salida. Los filtros de correlación, diseñados a partir de las imágenes de entrenamiento, optimizan criterios de desempeño y se basan en la selección de modelos de escena que caracterizan las imágenes usadas para poder reducir dichos errores.

4.2. Modelos de escena

Para tener mayor exactitud en la localización del objeto en la imagen de prueba es indispensable contar con un modelo matemático de escena que la represente adecuadamente. A continuación se presentan dos modelos muy utilizados en el diseño de filtros de correlación.

4.2.1. Modelo aditivo (traslapado)

En este modelo, la escena de entrada $s(x, x_0)$ contiene un objetivo $t(x - x_0)$ con ubicación desconocida, distorsionado por ruido aditivo $n(x)$ y se expresa como:

$$s(x, x_0) = t(x - x_0) + n(x). \tag{66}$$

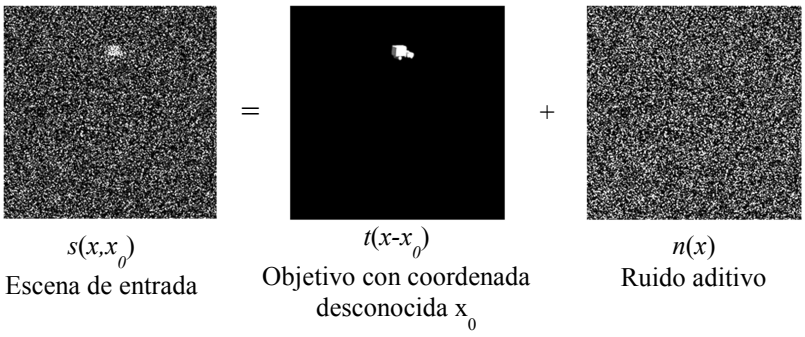


Figura 23. Modelo de escena aditivo.

En la Figura 23 se muestra el modelo con imágenes donde se aprecia el objetivo traslapado con el ruido en la escena $s(x, x_0)$. De este modelo se derivan filtros clásicos como el filtro de correspondencia (MF) de VanderLugt (1964), el filtro inverso, el filtro de sólo fase (POF) de Horner y Gianino (1984) y el filtro óptimo (OF) de Yaroslavsky (1993).

4.2.2. Modelo disjunto (no traslapado)

Este modelo caracteriza matemáticamente las imágenes del mundo real (Ramos Michel, 2008). Éstas contienen un fondo disjunto (no traslapado) del objetivo y están corrompidas por ruido traslapado producido por los sensores de adquisición. Su expresión es:

$$s(x, x_0) = t(x - x_0) + b(x)\overline{w}_t(x - x_0) + n(x), \tag{67}$$

donde la escena observada $s(x)$ contiene un objetivo $t(x)$ ubicado en x_0, como en el modelo anterior, pero que se encuentra espacialmente disjunto del fondo compuesto por $b(x)\overline{w}_t(x - x_0)$, e igualmente está corrompida por ruido aditivo $n(x)$. El fondo $b(x)$ se considera como un proceso aleatorio estacionario (con valor esperado μ_b) y $\overline{w}_t(x)$ es la región de soporte inversa del objetivo. La región de soporte del objetivo $w_t(x)$ se define como 1 en el área del objetivo y cero fuera de ella, por lo que $\overline{w}_t(x) = 1 - w_t(x)$. Entre los filtros derivados de este modelo se encuentran el generalizado óptimo y el generalizado de correspondencia, propuestos por Javidi y Wang (1994). La Figura 24 muestra un ejemplo de una escena real con este modelo.

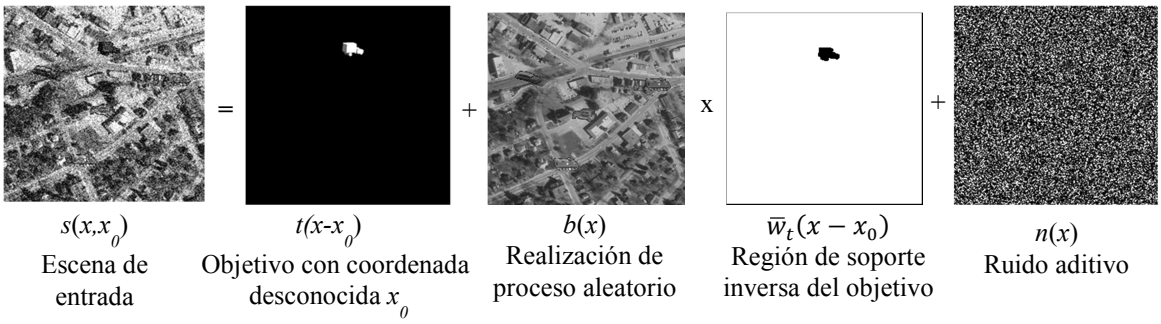


Figura 24. Modelo de escena disjunto.

4.3. Métricas de desempeño.

Para medir el desempeño de los filtros de correlación, se intenta optimizar criterios o métricas, ya sea maximizándolas o minimizándolas. En las siguientes definiciones, $E[\cdot]$ y $VAR[\cdot]$, se refieren a la esperanza y la varianza, respectivamente.

4.3.1. Razón señal a ruido

El filtro de correspondencia está diseñado para maximizar la razón de señal a ruido (SNR) de la salida del filtro (Javidi y Wang, 1994). Ésta se define como:

$$SNR = \frac{|E[y(x_0, x_0)]|^2}{VAR[y(x_0, x_0)]}, \tag{68}$$

donde $y(x_0, x_0)$ es el pico de la salida de correlación en la posición del objeto en la escena. Los valores altos de SNR indican tolerancia al ruido reduciendo la probabilidad de error de detección. El filtro de correspondencia es óptimo en términos de SNR para detectar un objetivo corrompido por ruido aditivo traslapado con media cero (Javidi y Wang, 1994).

4.3.2. Razón señal a ruido promedio

Otro criterio es la razón del cuadrado de la esperanza del pico de correlación respecto a la varianza promedio de salida, referido como SÑR y expresado como:

$$S\tilde{N}R = \frac{|E[y(x_0, x_0)]|^2}{\overline{VAR[y(x, x_0)]}}, \tag{69}$$

donde la barra superior en el denominador significa la integración normalizada (o promedio espacial) (Javidi y Wang, 1994). Este criterio es similar al SNR pero para el modelo disjunto.

4.3.3. Razón de energía del pico a la salida

El criterio de la razón de energía del pico a la salida (POE) se expresa como:

$$POE = \frac{|E[y(x_0, x_0)]|^2}{\overline{E[y(x, x_0)^2]}}, \tag{70}$$

donde $y(x,x_0)$ es la salida del filtro cuando el objetivo está en la posición x_0 en la escena de entrada y la barra superior en el denominador significa el promedio espacial sobre x (Javidi y Wang, 1994), como en el caso del criterio SÑR. El promedio se debe a que el ruido de fondo no es estacionario porque está disjunto del objetivo.

4.3.4. Capacidad de discriminación

La capacidad de discriminación (DC) está definida como la habilidad de un filtro para distinguir entre un objetivo a reconocer y cualquier otro objeto considerado como un objeto falso. Siguiendo el modelo de imagen disjunto, el objetivo está incrustado en un fondo que contiene objetos falsos y la DC (Yaroslavsky, 1993) se define como sigue:

$$DC = 1 - \frac{|C_b(0,0)|^2}{|C_t(0,0)|^2}, \tag{71}$$

donde $C_b(0,0)$ es el valor máximo del plano de correlación en el área del fondo y $C_t(0,0)$ es el valor máximo en el plano de correlación en el área del objetivo. Estas áreas son complementarias y un valor DC cercano a la unidad significa que el filtro tiene muy buena habilidad para discriminar. Los valores cero y negativos de DC significan que el filtro fracasa al distinguir un objetivo dado.

4.4. Filtros de correlación clásicos

En esta sección se presentan los filtros de correlación básicos que se han utilizado para el reconocimiento de objetos y que sirvieron como referencia para el desarrollo de este proyecto.

4.4.1. Filtro de correspondencia

Dada una escena bajo el modelo de escena aditivo, se tiene que el filtro que optimiza la SNR, es el filtro de correspondencia (MF) propuesto por VanderLugt (1967):

$$H(u) = \alpha \frac{T^*(u)}{P_n(u)}, \tag{72}$$

donde α es una constante arbitraria compleja, $T^*(u)$ es la transformada de Fourier conjugada compleja del objeto de referencia $t(x)$ a reconocer y $P_n(u)$ es el espectro de potencia del ruido aditivo $n(x)$. Cuando se trata de ruido blanco, el espectro de potencia $P_n(u)$ es constante, por lo que si $P_n(u) = 1$, se tiene:

$$H(u) = \alpha T^*(u). \tag{73}$$

Al fijar el valor de $\alpha = 1$ la respuesta al impulso del filtro, $h(x)$, en el caso de ruido blanco, resulta en una versión invertida del objeto de interés, $t(-x)$. En el caso general $h(x)$ es proporcional a $t(-x)$, razón a la que debe su nombre de filtro de correspondencia. Por lo tanto, la correlación se puede ver como la salida de un filtro de correspondencia, el cual es un filtro lineal invariante a desplazamiento cuya respuesta al impulso es la versión reflejada de la señal o imagen de referencia y que se puede demostrar que es óptima para detectar señales conocidas corrompidas por ruido blanco aditivo. (Kumar et al., 2005).

Aunque el filtro de correspondencia es óptimo en cuanto a que maximiza la SNR al detectar un objeto de referencia conocido en ruido aditivo, la limitación de éste es que es muy sensible a los cambios que pueda tener la imagen de referencia y presenta poca eficiencia de luz (Kumar, 1992). Además, produce picos de correlación muy anchos.

4.4.2. Filtro óptimo

El filtro óptimo (OF) (Yaroslavksy, 1993) es el filtro que optimiza el criterio POE; minimiza la probabilidad de errores anómalos en la localización del objeto. Cuando el objeto a reconocer es pequeño respecto a la escena de entrada, la respuesta en frecuencia del filtro óptimo se puede expresar como sigue:

$$H(u) = \frac{T^*(u)}{|T(u)|^2 + |S(u)|^2}, \tag{74}$$

donde $T(u)$ y $S(u)$ son la transformada de Fourier del objetivo $t(x)$ y de la escena de entrada $s(x)$, respectivamente.

Su principal ventaja es que minimiza la probabilidad de falsas alarmas y se adapta a la escena ya que su función de transferencia considera el espectro de energía de los objetos no deseados. Sin embargo, presenta una desventaja al tener baja eficiencia de luz, en la implementación óptica (Ramos Michel, 2008).

4.4.3. Filtro de correspondencia generalizado

El filtro de correspondencia generalizado (GMF) maximiza el criterio SÑR, similar al caso del filtro de correspondencia para el modelo aditivo y se define como (Javidi y Wang, 1994):

$$H^*(u) = \frac{E\left[S(u,x_0)e^{jux_0}\right]}{VAR[S(u,x_0)]}, \tag{75}$$

el cual, siguiendo el modelo disjunto definido anteriormente, se puede expresar como (Aguilar González, 2011):

$$H(u) = \frac{[T(u) + \mu_s \bar{W}(u)]^*}{\frac{1}{2\pi} B_s^0(u) * |\bar{W}(u)|^2 + N_s(u)}, \tag{76}$$

donde '$*$' es el operador de convolución y $B_s^0(u)$ es el espectro de potencia del fondo en la escena. Este filtro realiza la correspondencia con un objeto compuesto por el objeto original a detectar, más la región inversa de soporte del objeto, pesada con el valor promedio del fondo. En este caso son tan importantes tanto el objeto, como el fondo, para el modelo no traslapado de correspondencia.

La ventaja de este filtro es que maximiza el pico de autocorrelación y minimiza la varianza de la salida, mientras que su principal desventaja es que produce planos de correlación donde el pico de correlación puede perderse en el ruido de salida de promedio muy alto (Javidi y Wang, 1994). Por lo tanto este filtro tiene pobre desempeño respecto al criterio DC (Aguilar González, 2011).

4.4.4. Filtro óptimo generalizado

El filtro óptimo generalizado (GOF) optimiza el criterio POE en el modelo de escena disjunto, maximizando la intensidad del pico de correlación en la ubicación del objeto y minimizando la energía promedio del plano de correlación. La función de transferencia del filtro se define como (Javidi y Wang, 1994):

$$H^*(u) = \frac{E\left[S(u,x_0)e^{jux_0}\right]}{E[|S(u,x_0)|^2]}. \tag{77}$$

Aplicando la información correspondiente al modelo de escena disjunto, se obtiene la expresión (Aguilar González, 2011):

$$H(u) = \frac{[T(u) + \mu_s \bar{W}(u)]^*}{|T(u) + \mu_s \bar{W}(u)|^2 + \frac{1}{2\pi} B_s^0(u) * |\bar{W}(u)|^2 + N_s(u)}, \quad (78)$$

donde $B_s^0(u)$ es el espectro de potencia del fondo en la escena. Cuando no hay un fondo disjunto, la respuesta de frecuencia de la ecuación (78) coincide con la respuesta de frecuencia del filtro óptimo.

4.5. Filtros compuestos

Los filtros de correlación como el filtro MF presentan ciertas desventajas cuando los patrones a reconocer difieren de los patrones de referencia, ya sea por cambios de escala o por rotaciones, entre otros (Kumar et al., 2005). Para abordar esta problemática se puede aplicar un filtro de correspondencia diseñado para cada variante posible de la escena. Sin embargo, requeriría capacidades enormes de procesamiento y almacenaje. Por lo tanto, se han desarrollado filtros compuestos de correlación que se derivan de varias imágenes de entrenamiento representativas de las perspectivas que se espera tener del objeto a reconocer (Kumar et al., 2005).

Los filtros compuestos se usan cuando se dispone de varias imágenes que son perspectivas representativas de un objeto o patrón a reconocer. Estas imágenes se utilizan en el entrenamiento o diseño del filtro para que pueda reconocer al objeto aun cuando se presenta con distorsiones. Sin embargo se espera que la distorsión que se presente sea similar a alguna de las imágenes de entrenamiento. Consecuentemente, la selección de las imágenes de entrenamiento es un paso importante en el diseño de filtros compuestos (Kumar et al., 2005).

Además de reconocer a un objeto para el cual se entrenó, un filtro compuesto debe rechazar todo lo demás: debe tener una tasa de reconocimiento alta mientras mantiene una tasa de aceptación falsa, o falsas alarmas, baja (Kumar et al., 2005).

4.5.1 Funciones discriminantes sintéticas

Entre los primeros filtros compuestos se encuentran aquellos basados en funciones discriminantes sintéticas (SDF) (Casasent, 1984). Los filtros SDF se caracterizan por producir un valor específico en el origen del plano de correlación en respuesta a cada una de las imágenes con las que se entrena (o diseña) (Kumar et al., 2005). Por lo general se

utiliza, en un problema de dos clases, el valor de la unidad para las imágenes de entrenamiento de la clase 1, por ejemplo la clase que se desea aceptar (o detectar) y cero para las imágenes de entrenamiento de la clase 2, o bien la que se desea rechazar. Dicho esto, el filtro SDF se considera una combinación lineal de las imágenes de entrenamiento en la que los pesos son tales que la salida de la correlación en el origen alcanza los valores especificados previamente, en respuesta a las imágenes de entrenamiento.

Considérese un conjunto de *N* imágenes de entrenamiento. Sea u_i el valor en el origen del plano de correlación $g_i(x)$ producido por el filtro $h(x)$ en respuesta a la imagen de entrenamiento $t_i(x)$. Usando la notación vector-matriz $\mathbf{t}_i$ y $\mathbf{h}$ son las representaciones vectoriales de $t_i(x)$ y $h(x)$, respectivamente, obtenidas por un escaneo lexico-gráfico y resultando en vectores columna. Dado que el filtro SDF se diseña para obtener un valor específico a la salida de la correlación en respuesta a las imágenes de entrenamiento, se tiene entonces:

$$u_i = \mathbf{t}_i{}^{\mathrm{T}}\mathbf{h}\,, \quad 1 \le i \le N\,. \tag{79}$$

Tomando todo el conjunto de entrenamiento, se tiene:

$$\mathbf{u} = \mathbf{R}^{\mathrm{T}}\mathbf{h}, \tag{80}$$

donde el superíndice $^{\mathrm{T}}$ indica traspuesta, *d* es el número de pixeles en cada imagen $t_i(x)$, $\mathbf{h}$ es el vector $d \times 1$ del filtro, $\mathbf{R} = [\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_N]$ es una matriz $d \times N$, con los *N* vectores de imágenes de entrenamiento como columnas y $\mathbf{u} = [u_1, u_2, \dots, u_N]$ es un vector $N \times 1$ de las restricciones, o valores de pico de correlación especificados, para cada imagen de entrenamiento. Como se mencionó, el filtro $\mathbf{h}$ se compone como una combinación lineal de las imágenes de entrenamiento. Sea $\mathbf{a}$ el vector-columna de pesos que se asigna a cada imagen de entrenamiento, entonces:

$$\mathbf{h} = \mathbf{R}\mathbf{a}\,. \tag{81}$$

Sustituyendo la ecuación (81) en la (80) y despejando $\mathbf{a}$ posteriormente, el filtro SDF toma la siguiente forma:

$$\mathbf{h} = \mathbf{R}(\mathbf{R}^{\mathrm{T}}\mathbf{R})^{-1}\,\mathbf{u}, \tag{82}$$

expresada en el dominio espacial. Generalmente, como se dijo, se asigna la unidad como el valor predeterminado para los picos de correlación de las imágenes de entrenamiento. Como los SDF sólo controlan el plano de correlación en el origen, pueden aparecer lóbulos

laterales cerca del pico provocando errores de localización. Además, las restricciones sólo se cumplen cuando en la escena de entrada aparece exactamente una de las imágenes de entrenamiento. Sin embargo, si aparece una apariencia ligeramente distinta de las consideradas en el entrenamiento, se puede esperar que el filtro obtenga una salida similar a las obtenidas en el diseño; es decir, que el filtro sea capaz generalizar.

4.5.2. Filtro de mínima energía de correlación promedio

Dado que los filtros SDF sólo tienen control sobre el origen del plano, es conveniente tratar de minimizar o suprimir los lóbulos laterales que puedan ocurrir para asegurar un pico de correlación bien definido y reducir las posibilidades de error. Mahalanobis, et al., (1987) propusieron lograr esto minimizando la energía en el plano de correlación. Considérese el plano de correlación $g_i(x)$, su transformada de Fourier $G_i(u)$, la respuesta en frecuencia $H(u)$ de $h(x)$ y la transformada de Fourier de la imagen de entrenamiento $t_i(x)$ como $X_i(u)$. Le energía promedio de correlación (ACE), para las N imágenes de entrenamiento, se define como:

$$\begin{aligned} ACE &= \frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{d}|g_i(m)|^2 \\ &= \frac{1}{dN}\sum_{i=1}^{N}\sum_{k=1}^{d}|G_i(k)|^2 \\ &= \frac{1}{dN}\sum_{i=1}^{N}\sum_{k=1}^{d}|H(k)|^2|X_i(k)|^2, \end{aligned} \tag{83}$$

donde d es el número de pixeles en cada imagen $t_i(x)$. Usando notación vector-matriz, se tiene que $\mathbf{h}$ es el filtro en versión vector columna y se define la matriz diagonal $\mathbf{X}_i$ cuyos elementos en la diagonal principal son $X_i(u)$. Se reescribe la expresión de ACE como:

$$\begin{aligned} ACE &= \mathbf{h}^{+}\left[\frac{1}{dN}\sum_{i=1}^{N}|\mathbf{X}_i|^{\mathbf{2}}\right]\mathbf{h} \\ &= \mathbf{h}^{+}\mathbf{D}\mathbf{h}, \end{aligned} \tag{84}$$

donde el superíndice $^{+}$ indica traspuesta conjugada, $\mathbf{D} = \frac{1}{dN}\sum_{i=1}^{N}|\mathbf{X}_i|^{\mathbf{2}}$ es una matriz diagonal $d \times d$.

Al minimizar la ACE se espera reducir los lóbulos laterales y afilar más el pico. Ahora, así como en los filtros SDF, el pico de correlación se controla a través de restricciones. Representándolas en el dominio de frecuencia, se tiene que:

$$d\mathbf{u} = \mathbf{S}^{+}\mathbf{h}, \tag{85}$$

donde $\mathbf{S}$ es ahora la matriz cuyas columnas son las transformadas de Fourier de las imágenes de entrenamiento. Por lo tanto, el filtro de mínima energía promedio de correlación (MACE) minimiza el criterio ACE, sujeto a las restricciones de la ecuación (85). Finalmente se trata de un problema de optimización cuadrático que se resuelve usando el método de multiplicadores de Lagrange, que da como resultado la expresión para el filtro MACE en el dominio de frecuencia:

$$\mathbf{h} = \mathbf{D}^{-1}\mathbf{X}(\mathbf{X}^{+}\mathbf{D}^{-1}\mathbf{X})^{-1}\mathbf{u}. \tag{86}$$

El filtro MACE es efectivo para encontrar imágenes en fondos complejos, aunque presenta dos desventajas: como no se diseña considerando ruido, no es inmune a él y, por otro lado, es muy sensible a variaciones intraclase. Esta última se refiere a que sólo cumple con las restricciones cuando se presenta en la entrada una de las imágenes con las que se entrenó, pero no tiene mucha capacidad de generalización.

4.5.3. Otros filtros compuestos

Otro tipo de filtro compuesto es aquel que minimiza la varianza del ruido para disminuir las fluctuaciones en la salida del filtro. A este filtro se le conoce como filtro MVSDF. Por lo tanto, mientras que el filtro MACE se encarga de producir un pico de correlación más fino o definido, el filtro MVSDF provee robustez respecto al ruido, suponiendo que el ruido tiene media cero (Kumar, et al., 2005).

El filtro SDF es en realidad un caso particular del MVSDF cuando se trata del reconocimiento de las imágenes de entrenamiento en presencia de ruido aditivo blanco. Para el caso general se requiere el cálculo de la matriz de covarianza del ruido de entrada y el de su inversa, resultando complicado debido a su gran tamaño, por lo que se vuelve una desventaja de este filtro (Kumar, et al., 2005).

Aunque los filtros SDF proveen tolerancia a distorsiones controlando los picos de correlación de salida al especificarlos para las imágenes de entrenamiento, existe otra

manera de proveer dicha tolerancia sin imponer tales restricciones duras en el filtro. Se usa una métrica para distorsión definida como la variación promedio en las imágenes después del filtrado y se consideran los planos de correlación como versiones linealmente transformadas de las imágenes de entrada al aplicar el filtro. Por lo que es importante tanto el pico de correlación como el plano de correlación completo (Kumar et al., 2005).

Ahora, "además de presentar tolerancia a distorsiones, un filtro de correlación debe de obtener valores de pico altos para facilitar la detección y ubicar la posición del objeto" (Kumar et al., 2005, p. 217). Para resolver esto, se maximiza la respuesta promedio del filtro a las imágenes de entrenamiento sin utilizar restricciones duras como los filtros SDF, maximizando el criterio de la altura de correlación promedio (ACH). Por último, para reducir el efecto del ruido y fondo en la salida del filtro, se reduce la varianza del ruido de salida. El filtro que optimiza estos criterios es el llamado filtro MACH.

4.6. Implementación del reconocimiento por correlación

Para utilizar los filtros clásicos y filtros compuestos que se presentaron en las secciones anteriores, en el problema de reconocimiento de objetos, en general, se hace la implementación siguiente. Primeramente se considera una escena $s(x)$, en la que se quiere detectar y localizar un objeto, junto con un filtro a aplicar. El plano de correlación $c(x)$ se calcula como:

$$c(x) = \mathcal{F}^{-1}\{S(u)H(u)\}, \tag{87}$$

donde $S(u)$corresponde a la transformada de Fourier de la escena $s(x)$ y $H(u)$ es la función de transferencia o respuesta en frecuencia del filtro a aplicar. La multiplicación de *S* y *H* es elemento a elemento. Después se detecta y se localiza el pico más alto obtenido sobre el plano $c(x)$ y, de esta manera, se determina la ubicación del objetivo en la escena para posteriormente pasar a la clasificación. En la Figura 25 se muestra la forma general de la implementación de un sistema de reconocimiento mediante filtros de correlación.

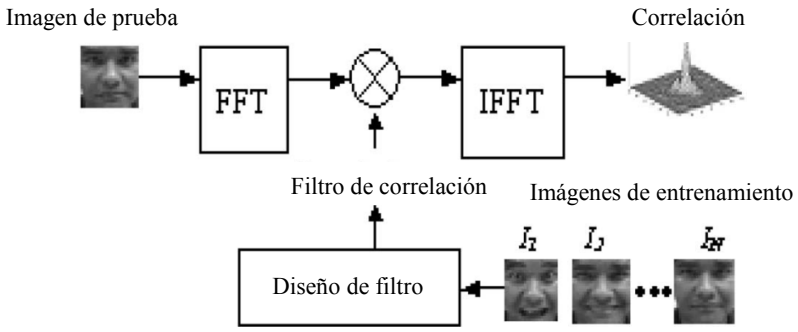


Figura 25. Esquema del reconocimiento por correlación. Recuperado de Savvides, et al., (2002), 56p.

4.7. Resumen

En este capítulo se presentaron conceptos de base que permiten establecer el modelo de escena para representar las imágenes reales que se utilizan en el problema de reconocimiento de rostros propuesto, así como filtros de correlación óptimos para estos modelos y los respectivos criterios de desempeño que optimizan. Para resolver el problema de reconocimiento de rostros contemplando los factores de variación que se presentan en las imágenes reales, se describieron técnicas de filtros compuestos que consideran un conjunto de distintas perspectivas de los objetos de interés.

Capítulo 5. Técnica de reconocimiento por correlación propuesta

En el capítulo 3 se presentaron tres de las técnicas de reconocimiento más populares, utilizadas en su forma básica, pero que fueron diseñadas con el propósito de reconocer rostros en escenas de situaciones ideales, no en escenas reales donde se requiere hacer la detección y localización del rostro, al mismo tiempo que reconocerlo (clasificarlo). Por otro lado, de las técnicas de reconocimiento de patrones por correlación, en el capítulo 4 se presentaron filtros clásicos para dos modelos de escena, así como filtros compuestos que se utilizan para tomar en cuenta múltiples apariencias de un objeto a reconocer.

Ahora bien, para abordar el problema de reconocimiento de rostros, resolviendo el problema de detección y localización, implícitamente en escenas reales, es decir, donde no se tiene el rostro segmentado, se decide utilizar el modelo de escena disjunto (de la ecuación (67)), ya que permite caracterizar matemáticamente las escenas del mundo real. Por esta razón, los filtros propuestos en este proyecto se basan en el filtro GOF, considerando también los filtros compuestos SDF para incluir la información de varias perspectivas de un individuo con las que se cuenta, para su reconocimiento.

El desempeño de los filtros utilizados se determina usando el criterio de capacidad de discriminación definido en la ecuación (71). Además, no se considera la presencia de ruido aditivo $n(x)$ del modelo de escena disjunto.

5.1. Filtro blanqueado o GOF simplificado

Como se mencionó, de acuerdo al modelo de escena disjunto, se elige el filtro GOF como base, considerando una simplificación en la cual no se requiere de conocimiento a priori del espectro de potencia del fondo disjunto de la escena de entrada (ver ecuación (78)). En la etapa de entrenamiento (o diseño) del filtro de una clase dada, se define el siguiente filtro blanqueado o filtro GOF simplificado 1 (GOFS1):

$$H_{GOFS1} = \frac{T_i(u)}{|B(u)|^2}, \tag{88}$$

donde $T_i(u)$ es la transformada de Fourier de la imagen de entrenamiento *i* del individuo de la clase dada y $|B(u)|^2$ es el espectro de potencia del fondo $b(x)$ similar al esperado en la escena de entrada y al cual se refiere como fondo típico en las secciones subsecuentes.

En la etapa de reconocimiento se crea un filtro similar en donde, para hacer una adaptación a la escena de entrada desconocida, se utiliza la información completa de la escena (su espectro de potencia), junto con la información obtenida en el proceso de entrenamiento (objetos falsos creados a partir de otras clases). El denominador de la respuesta en frecuencia del filtro se define con la estimación del espectro de potencia del fondo disjunto de la escena de entrada. El filtro blanqueado para el reconocimiento o GOF simplificado 2 (GOFS2) se expresa como:

$$H_{GOFS2} = \frac{T_i(u)}{|S(u)|^2 + |T_i(u)|^2}, \tag{89}$$

donde $|T_i(u)|^2$ es el espectro de potencia de la imagen de entrenamiento *i* y $|S(u)|^2$ es el espectro de potencia de la escena de entrada. De esta manera se permite una adaptación más dinámica a la escena real de entrada.

5.2 Algoritmo del filtro SDF adaptativo clásico

Como se mencionó en la sección 4.5.1, un filtro SDF se puede diseñar para aceptar (o bien, detectar) un conjunto de objetos y rechazar otro conjunto compuesto de objetos no deseados u objetos falsos, eligiendo la unidad como el valor del pico de correlación para el conjunto de aceptación y cero para el pico de correlación del conjunto de rechazo. De este modo el conjunto de imágenes de entrenamiento está compuesto por todos los objetos del conjunto de aceptación y todos los objetos del conjunto de rechazo. Por lo tanto, se define a $S = \{t_1(x), \dots, t_L(x), p_1(x), \dots, p_M(x)\}$ como el conjunto de imágenes de entrenamiento con *L* objetos de aceptación, denotados por $t_i(x)$, para *i* = 1...*L,* y *M* objetos de rechazo, denotados por $p_i(x)$, para *i* = 1...*M.* Se tiene entonces que el filtro SDF se compone con la siguiente combinación lineal:

$$h(x) = \sum_{i=1}^{L} a_i t_i(x) + \sum_{i=L+1}^{L+M} a_i p_{i-L}(x). \tag{90}$$

Entonces, en la ecuación (80), **R** ahora es una matriz con *L*+*M* columnas, en donde cada columna es la versión vector de cada elemento del conjunto *S* y **u** está definido como

$$\mathbf{u} = [\underbrace{1, \ldots, 1}_{K\ unos}, \underbrace{0, \ldots, 0}_{M\ ceros}]^T. \tag{91}$$

En González-Fraga et al., (2006), se propuso un algoritmo adaptativo que tiene como objetivo suprimir todos los picos falsos que puedan aparecer en el plano de correlación debidos a lóbulos laterales que no controlan los filtros SDF. En ese algoritmo, para un objetivo a reconocer dado, se pueden rechazar objetos falsos y un fondo dado de manera iterativa. En cada iteración el algoritmo suprime el pico del lóbulo lateral más alto en el plano de correlación, incrementando así la capacidad de discriminación hasta que se alcanza un valor deseado.

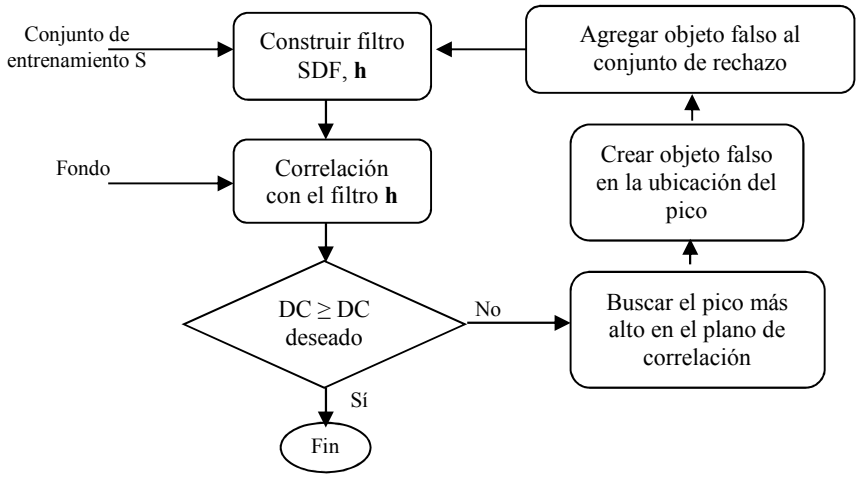


Figura 26. Algoritmo de SDF adaptativo clásico

El algoritmo del filtro SDF adaptativo clásico, ilustrado en la Figura 26, se puede resumir con los siguientes pasos:

1. Construir el filtro SDF **h** con los objetos en el conjunto de entrenamiento S.
2. Realizar la correlación del fondo conocido, llamado fondo típico, con el filtro **h** y calcular el valor DC correspondiente.
3. Si el valor DC calculado es mayor o igual al valor DC deseado, el algoritmo adaptativo termina. Si no, entonces continuar al paso 4.
4. Buscar el pico más alto en el plano de correlación.
5. Crear un objeto falso a partir del fondo en la ubicación del pico.
6. Agregar el objeto falso creado al conjunto de rechazo y actualizar el conjunto S.
7. Regresar al paso 1.

5.3 Filtro SDF blanqueado propuesto

El algoritmo propuesto se basa en el diseño de filtros SDF y en un proceso iterativo de adaptación, similar al descrito en la subsección anterior, para maximizar la discriminación entre múltiples clases. Primeramente, considerando una clase dada como la clase verdadera, se define un nuevo objetivo (o imagen de entrenamiento) compuesto $t_i(x)$, el cual resalta la importancia, tanto de la forma del objeto (a través de su región de soporte) y de la media del fondo típico, como la del propio objeto a detectar. Este nuevo objetivo (o nueva imagen de entrenamiento) se define como:

$$t_i{}^k(x) = t_i{}^k(x) + \mu_b \overline{w_i}^k(x), \tag{92}$$

donde, el superíndice *k* indica la clase verdadera, μ_b es la media del fondo típico y $\overline{w_i}^k(x)$ es la región de soporte inversa de la imagen de entrenamiento $t_i(x)$ original. Por lo tanto las *L* imágenes de entrenamiento $t_i(x)$ para la clase verdadera están ajustadas con la media del fondo típico.

En el diseño se consideran también escenas de entrenamiento compuestas por las imágenes de entrenamiento $t_i(x)$ originales del conjunto de aceptación y por la media del fondo típico que se espera ver en las escenas de entrada. Estas escenas se modelan de la misma manera que las imágenes de entrenamiento de la ecuación anterior, es decir:

$$s_i{}^k(x) = t_i{}^k(x) + \mu_b \overline{w_i}^k(x). \tag{93}$$

Asimismo se consideran escenas de rechazo que se componen de las imágenes de entrenamiento $t_j{}^l(x)$ de las clases a rechazar y de la media del fondo típico. Por lo tanto, las escenas de rechazo se modelan como:

$$\bar{s}_j{}^l(x) = t_j{}^l(x) + \mu_b \overline{w_j}^l(x), \quad l \neq k. \tag{94}$$

Retomando el conjunto de entrenamiento *S* de la subsección anterior y su notación, sean $T_i(u)$ y $P_i(u)$ el espectro de potencia del objeto de entrenamiento $t_i(x)$ y del patrón de rechazo $p_i(x)$, respectivamente, sus filtros de blanqueado (ecuación (88)) se definen:

$$\hat{T}_i(u) = \frac{T_i(u)}{|B(u)|^2}, \tag{95}$$

$$\hat{P}_i(u) = \frac{P_i(u)}{|B(u)|^2}.$$

Si $\hat{t}_i(x)$ y $\hat{p}_i(x)$ son la transformada de Fourier inversa de la respuesta en frecuencia de éstos, se define $\hat{S} = \{\hat{t}_1(x), \ldots \hat{t}_L(x), \hat{p}_1(x), \ldots, \hat{p}_M(x)\}$, como el conjunto de respuestas al impulso óptimas para cada imagen de entrenamiento $t_i(x)$ y patrones a rechazar $p_i(x)$.

Con los conjuntos $\hat{S}$ y S se puede construir un filtro SDF como:

$$h_{SDF}(x) = \sum_{i=1}^{L} a_i \hat{t}_i(x) + \sum_{i=L+1}^{L+M} a_i \hat{p}_{i-L}(x). \tag{96}$$

Usando notación vector-matriz, las ecuaciones (80), (81) y (82) se redefinen como:

$$\mathbf{u} = \mathbf{Q}^{+}\mathbf{h}_{\mathrm{SDF}}, \tag{97}$$

$$\mathbf{h}_{\mathrm{SDF}} = \mathbf{R}\mathbf{a}, \tag{98}$$

$$\mathbf{h}_{\mathrm{SDF}} = \mathbf{R}(\mathbf{Q}^{+}\mathbf{R})^{-1}\mathbf{u}, \tag{99}$$

donde **Q** es la matriz cuyas columnas corresponden a las versiones vector de los elementos del conjunto *S* y **R** es la matriz con las versiones vector de los elementos del conjunto $\hat{S}$.

Habiendo definido los componentes del nuevo filtro SDF, denominado filtro SDF blanqueado, el diseño del filtro para una clase dada comienza con la construcción de un $\mathbf{h}_{\mathrm{SDF}}$ a partir de los conjuntos *S* y $\hat{S}$ de dicha clase. En un inicio, no hay patrones de rechazo $p_i(x)$ ni $\hat{p}_i(x)$. Luego, el filtro se correlaciona con cada escena de entrenamiento de la clase verdadera. Se calculan los picos máximos en el área del objetivo de los planos de correlación resultantes y su mínimo se define como mínimo pico verdadero. Después, el filtro se correlaciona con cada escena de rechazo. Se calcula el máximo pico en el área del objetivo de los planos resultantes y el valor máximo entre ellos se define como máximo pico falso. Con estos dos valores, se define el criterio DC unificado como sigue:

$$DC\ unificado = 1 - \frac{(máximo\ pico\ falso)^2}{(mínimo\ pico\ verdadero)^2}.$$

Si el DC unificado de la iteración actual es mayor que el DC unificado de la iteración anterior (después de la primera iteración), entonces se crea un objeto falso a partir de la ubicación del pico en la escena de rechazo correspondiente al máximo pico falso. Se calcula el objeto $\hat{p}(x)$, se actualizan los conjuntos $\hat{S}$ y *S* con los patrones de rechazo y se compone un nuevo filtro $\mathbf{h}_{\mathrm{SDF}}$ con todos los elementos de los conjuntos *S* y $\hat{S}$. El área del

objeto se toma como un cuadrado que enmarca sólo el contorno del rostro en la imagen. El proceso iterativo se continúa hasta que el DC unificado sea menor o igual al DC unificado de la iteración anterior. En dado caso el último objeto falso creado, que ocasiona que no mejore el DC unificado, se elimina del filtro SDF y de los conjuntos $\hat{S}$ y S. El diagrama de la siguiente figura muestra el diseño iterativo del filtro.

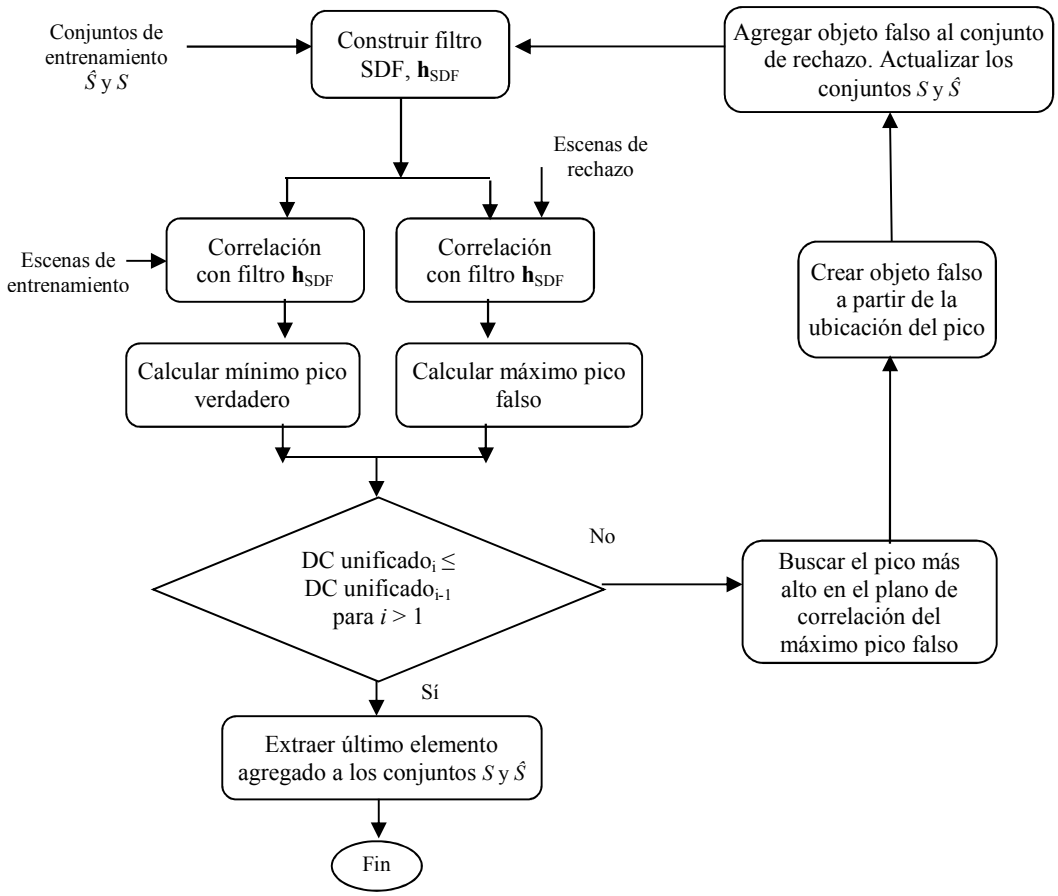


Figura 27. Algoritmo adaptativo para SDF blanqueado

Los pasos del diseño iterativo se pueden resumir como sigue:

1. Construir el filtro SDF $\mathbf{h}_{SDF}$ con los objetos en los conjuntos de entrenamiento S y Ŝ. Inicialmente todos los objetos pertenecen a la clase verdadera.

2. Realizar la correlación del filtro $\mathbf{h}_{SDF}$ con las escenas de entrenamiento definidas en la ecuación (93) y calcular mínimo pico verdadero.

3. Realizar la correlación del filtro h_{SDF} con las escenas de rechazo definidas en la ecuación (94). Calcular máximo pico falso.

4. Calcular el criterio DC unificado de la iteración actual.

5. Si el DC unificado de la iteración actual es menor o igual al de la iteración anterior, después de la primera iteración, se elimina en el filtro $\mathbf{h}_{SDF}$ el último elemento agregado a los conjuntos S y $\hat{S}$ y el algoritmo adaptativo termina. Si no, continúa al paso 6.

6. Crear objeto falso $p_i(x)$ y su $\hat{p}_i(x)$ correspondiente con la transformada de Fourier inversa de la ecuación (95) en la ubicación del máximo pico falso del plano de correlación correspondiente. Actualizar los conjuntos S y $\hat{S}$ con $p_i(x)$ y $\hat{p}_i(x)$.

7. Ir al paso 1.

Cabe mencionar que, a diferencia del algoritmo adaptativo SDF clásico, la creación del objeto falso en cada iteración se lleva a cabo de manera diferente. Dado que en el algoritmo adaptativo SDF clásico se trata de imágenes de entrenamiento que contienen un objetivo de duración finita, es decir, sólo hay información en la región de soporte del objetivo, el objeto falso se crea a partir de la imagen del fondo típico con la ubicación del pico máximo de correlación desplazado al centro y con la forma de la unión de las regiones de soporte de las imágenes de entrenamiento. Por el contrario, en esta propuesta al tratarse de imágenes de entrenamiento $\hat{t}_i(x)$ blanqueadas, éstas contienen un objetivo compuesto del tamaño de la imagen. En este caso, no es necesario extraer el objeto falso con la forma de la unión de las regiones de soporte mencionadas, sino que solamente se desplaza la ubicación del pico máximo falso al centro del nuevo patrón a rechazar.

5.3.1. Reconocimiento y clasificación

En los párrafos anteriores se describe el diseño del filtro SDF blanqueado propuesto, el cual corresponde a la etapa de entrenamiento de lo que constituye un sistema de reconocimiento. Para la etapa de reconocimiento y clasificación se realiza un proceso que se describe a continuación. La matriz **R** en la ecuación (99) se recalcula con los elementos de S actualizados en el proceso de diseño y calculando las respuestas en frecuencia del conjunto $\hat{S}$ a través de la información disponible de las escenas de entrada, con la ecuación (89), por lo que:

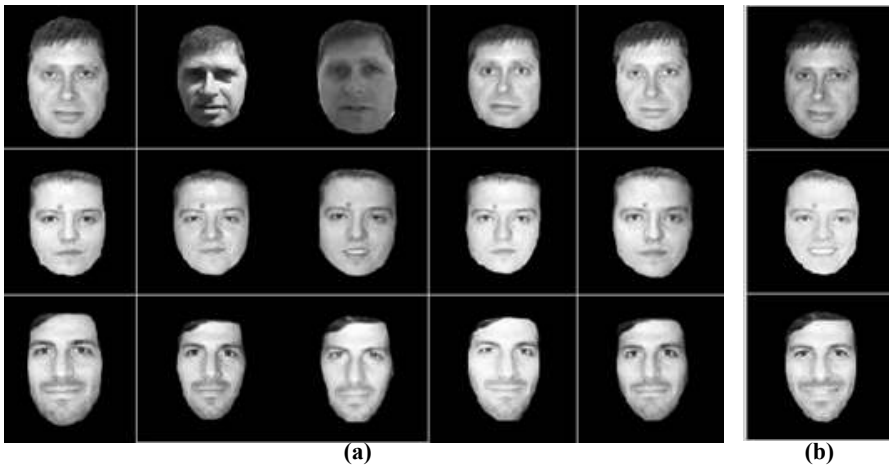
$$\hat{T}_i(u) = \frac{T_i(u)}{|S(u)|^2 + |T_i(u)|^2}, \qquad (100)$$

$$\hat{P}_i(u) = \frac{P_i(u)}{|S(u)|^2 + |P_i(u)|^2},$$

donde $T_i(u)$ y $P_i(u)$ son la transformada de Fourier de la imagen de entrenamiento *i* y del patrón de rechazo *i*, respectivamente y $|T_i(u)|^2$ y $|P_i(u)|^2$ son sus espectros de potencia.

En términos del sistema de reconocimiento, se sintetiza un filtro $\mathbf{h}_{SDF}$ para cada clase con la ecuación (99) y se correlaciona cada uno con la escena de entrada. Se calcula el criterio DC con la ecuación (71) en cada plano de correlación resultante y la clase del filtro correspondiente al máximo DC es la que define la clase del objeto en la escena de entrada.

Un ejemplo del desempeño de los filtros SDF blanqueados descritos se ilustra en la Figura 31. Las imágenes de rostro utilizadas pertenecen a la base de datos Caltech (ver apéndice). Un ejemplo de las imágenes de entrenamiento correspondientes a tres clases, así como los rostros de prueba se muestra en la Figura 28. Las escenas de prueba utilizadas corresponden a las mostradas en la Figura 29.



(a) **(b)**

Figura 28. Ejemplo de (a) rostros del conjunto de entrenamiento y (b) del conjunto de prueba de la base de datos de Caltech.

La Figura 30 muestra ejemplos del filtro GOFS1 y de objetos de aceptación y rechazo utilizados en el diseño de un filtro SDF blanqueado.

Por columnas, las gráficas de la Figura 31 corresponden a la correlación de los filtros SDF blanqueados construidos para las tres clases, con escenas de la clase 1, 2 y 3. Por renglones, los dos primeros corresponden a la correlación con escenas de prueba

sintéticas (Figura 29(a)) y los dos últimos a la correlación con escenas reales (Figura 29(c)). La Figura 31(a)-(c) y (g)-(i) muestran el desempeño de los filtros entrenados incluyendo los objetos de prueba; (d)-(f) y (j)-(l) muestran el desempeño sin incluir a los objetos de prueba en el entrenamiento. El tamaño de conjunto de entrenamiento se define por el número de objetos de cada clase incluidos en dicho conjunto. En este caso, debido a que las escenas de prueba contienen objetos en una ubicación fija, no se considera el desempeño dentro de un intervalo de confianza.

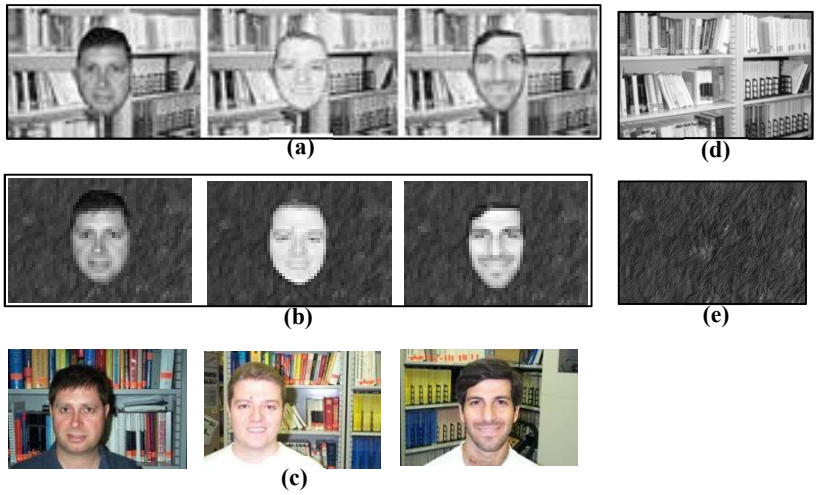


Figura 29. Escenas de prueba sintéticas y reales correspondientes a los objetos de prueba de la Figura 28(b). (a) Escenas de prueba sintéticas sobre el fondo no-homogéneo (d). (b) Escenas de prueba sintéticas sobre el fondo homogéneo (e). (c) Escenas de prueba reales; los fondos contienen estructuras similares al fondo (d).

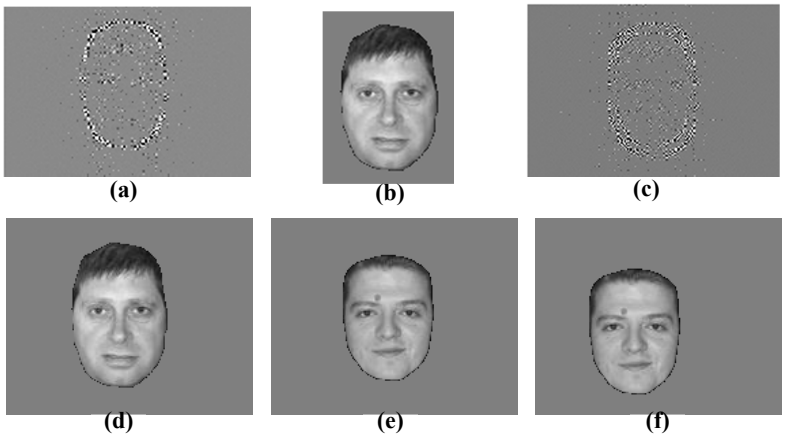


Figura 30. Ejemplo de: (a) filtro blanqueado, (b) imagen de entrenamiento, (c) filtro SDF blanqueado, (d) escena de entrenamiento, (e) escena de rechazo y (f) objeto de falso creado.

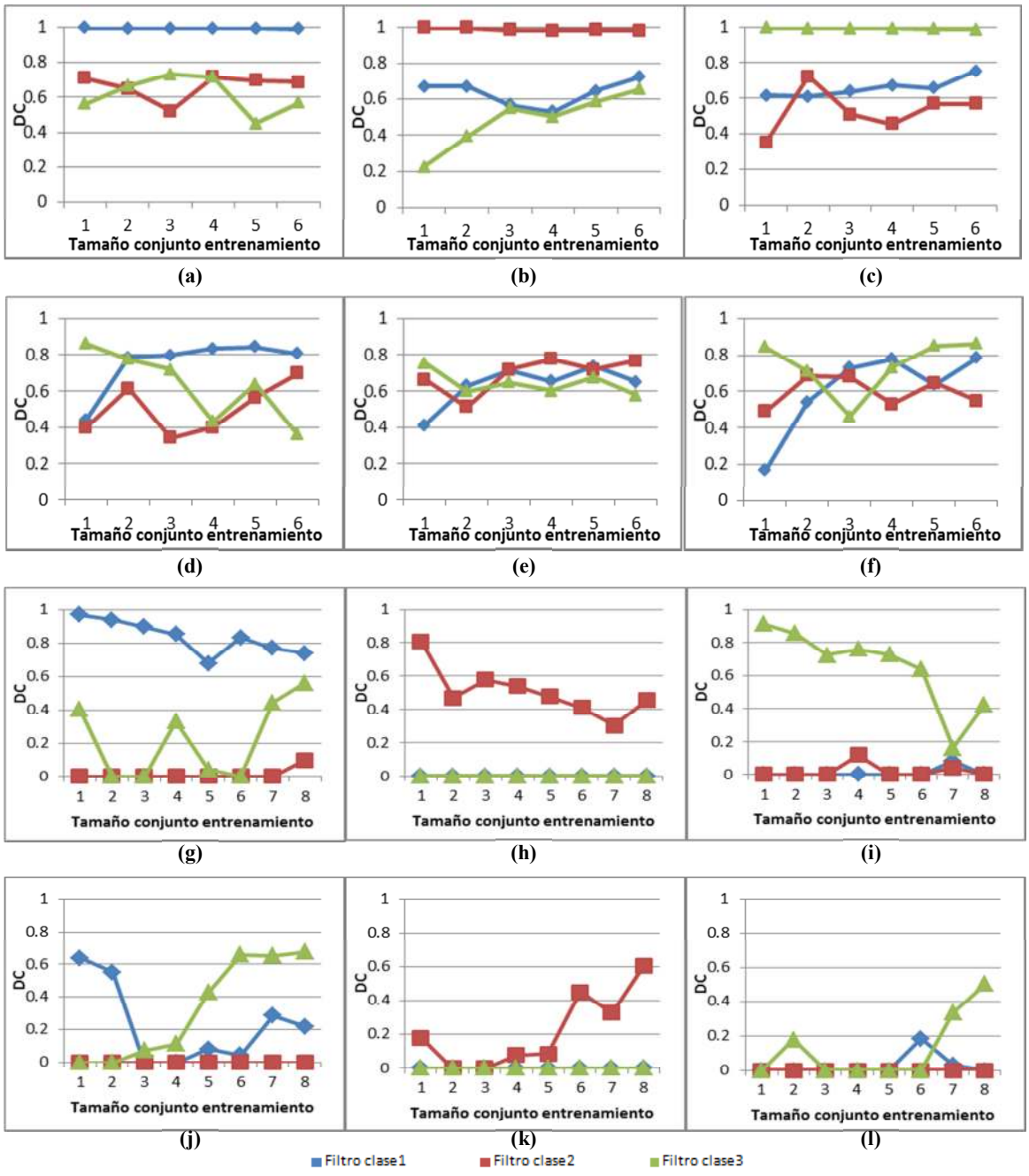


Figura 31. Desempeño en términos de DC de filtros entrenados con varios tamaños de entrenamiento, correlacionados con tres escenas sintéticas y tres escenas reales.

Se puede observar que, tanto en el caso de escenas de prueba sintéticas como reales, los filtros de la clase correcta, es decir, de la clase del objeto en la escena de prueba, obtienen siempre el mayor desempeño, por lo tanto clasifican correctamente las escenas de prueba. Sin embargo, esto sucede cuando los objetos mismos de prueba fueron incluidos en

el entrenamiento (ver Figura 31(a)-(c) y (g)-(i)). Por el contrario, para algunos conjuntos de entrenamiento, cuando no incluyen a los objetos de prueba, los filtros no siempre clasifican correctamente a las escenas de prueba (como se observa en la Figura 31(d)-(f) y (j)-(l)).

Aun cuando los filtros son entrenados con varias vistas de una clase, en ocasiones no tienen la suficiente información para generalizar todas las posibles variantes de una clase y por lo tanto, al presentarse en la prueba una apariencia que no se tomó en cuenta en el entrenamiento, se hace difícil discriminar y obtener el mayor nivel DC. De aquí que puede haber una confusión por parte de los filtros.

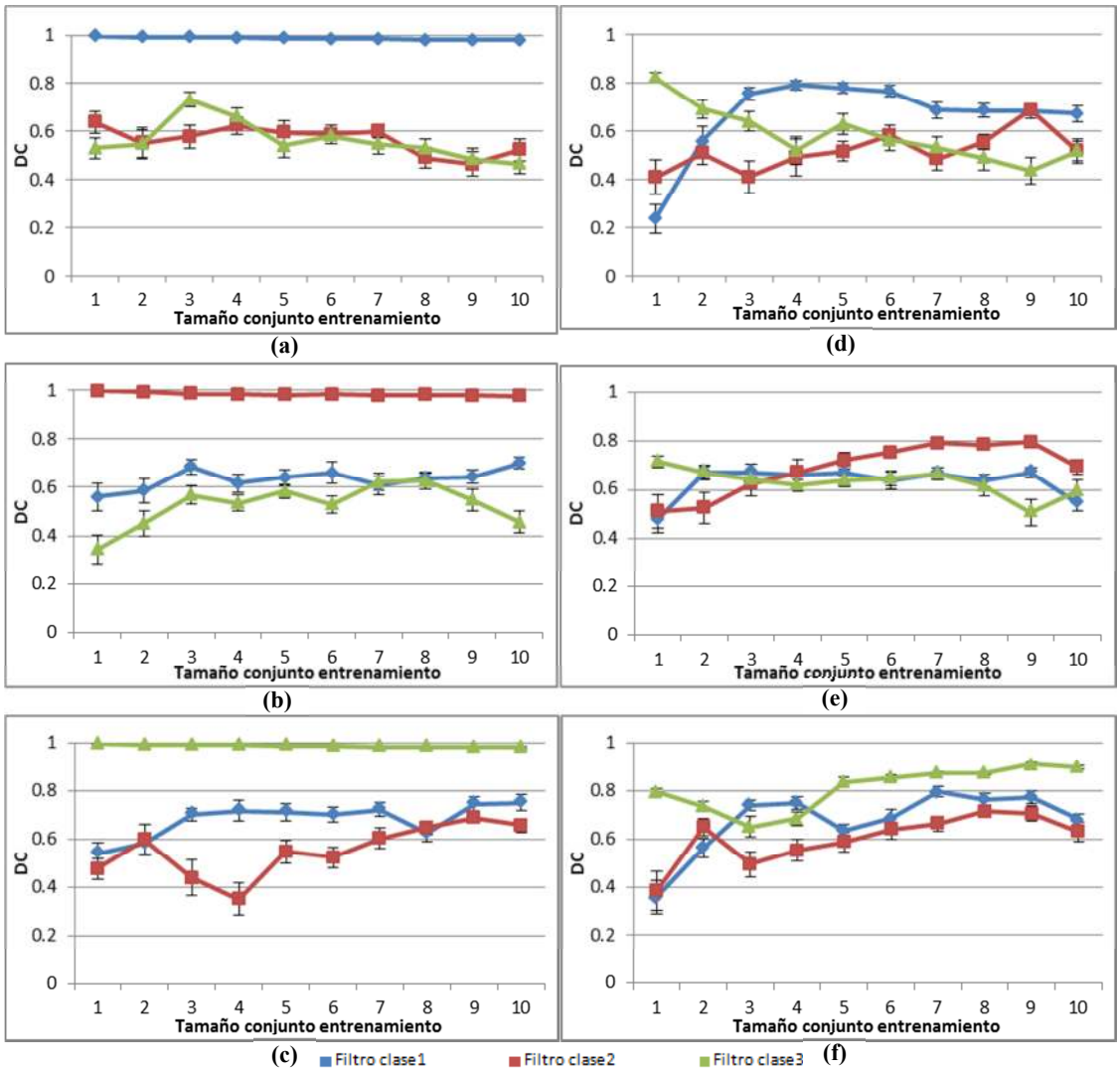


Figura 32. Desempeño con 95% de confianza en términos de DC de filtros, entrenados con varios tamaños de entrenamiento, correlacionados con 30 escenas sintéticas de las clases 1, 2 y 3, en las que la ubicación del rostro es aleatoria y utilizando el fondo típico de la Figura 29(d).

Ahora, en las gráficas de la Figura 32 y la Figura 33 se muestra el desempeño de los filtros haciendo una prueba distinta con escenas sintéticas. Se utilizaron 30 escenas de prueba por clase en las que se varía la ubicación del rostro aleatoriamente. Las 30 escenas contienen uno de los dos fondos distintos considerados; uno es muy heterogéneo, con estructuras complejas (un librero en una oficina) y el otro es muy homogéneo, con una textura de un pastizal. En dichas gráficas se muestra el desempeño en términos de DC con una confianza de 95%; los incisos (a)-(c) de la Figura 32 muestran el desempeño de filtros entrenados incluyendo los objetos de prueba y los incisos (d)-(f) cuando no se incluyen. Por renglones, las gráficas de la Figura 32 corresponden al desempeño de los filtros correlacionados con escenas de prueba de la clase 1, 2 y 3, respectivamente. En la Figura 32 se tiene una situación similar al caso anterior, en la que se obtiene mejor desempeño al incluir los objetos de prueba en el entrenamiento. Sin embargo, se puede notar que a partir de cuatro o cinco objetos de entrenamiento por clase, el valor DC promedio de cada filtro de la clase correcta comienza a incrementarse en el caso cuando no se incluyen los objetos de prueba, clasificando correctamente las escenas de prueba. Por otro lado, las gráficas de la Figura 33 muestran un comportamiento más estable; es muy similar para cuando se incluyen los objetos de prueba en el entrenamiento y para cuando no se incluyen. Esto resulta de utilizar un fondo más homogéneo, como el que se aprecia en la Figura 29(e) de un pastizal, que se aproxima más a las condiciones teóricas del diseño de los filtros GOF del modelo disjunto, al referirse al fondo como un proceso estacionario donde la media es constante en cualquier lugar. Para la Figura 33, los incisos (a)-(c) muestran el desempeño de filtros de las tres clases entrenados incluyendo los objetos de prueba y los incisos (d)-(f) cuando no se incluyen. Por renglones, las gráficas de la Figura 33 corresponden al desempeño de los filtros al correlacionarse con escenas de prueba de la clase 1, 2 y 3, respectivamente.

Cabe mencionar que las pruebas que se presentan en esta sección dependen de la selección del conjunto de entrenamiento, la cual se realizó de manera secuencial como estaban registradas las imágenes. De aquí que para algunos tamaños puede haber imágenes de entrenamiento que no aporten la información más útil para representar a su clase de manera fiel.

En las subsecciones siguientes se verán un par de técnicas que precisamente tratan de abordar las limitantes del diseño actual de los filtros SDF blanqueados presentados, evitando la confusión de los filtros.

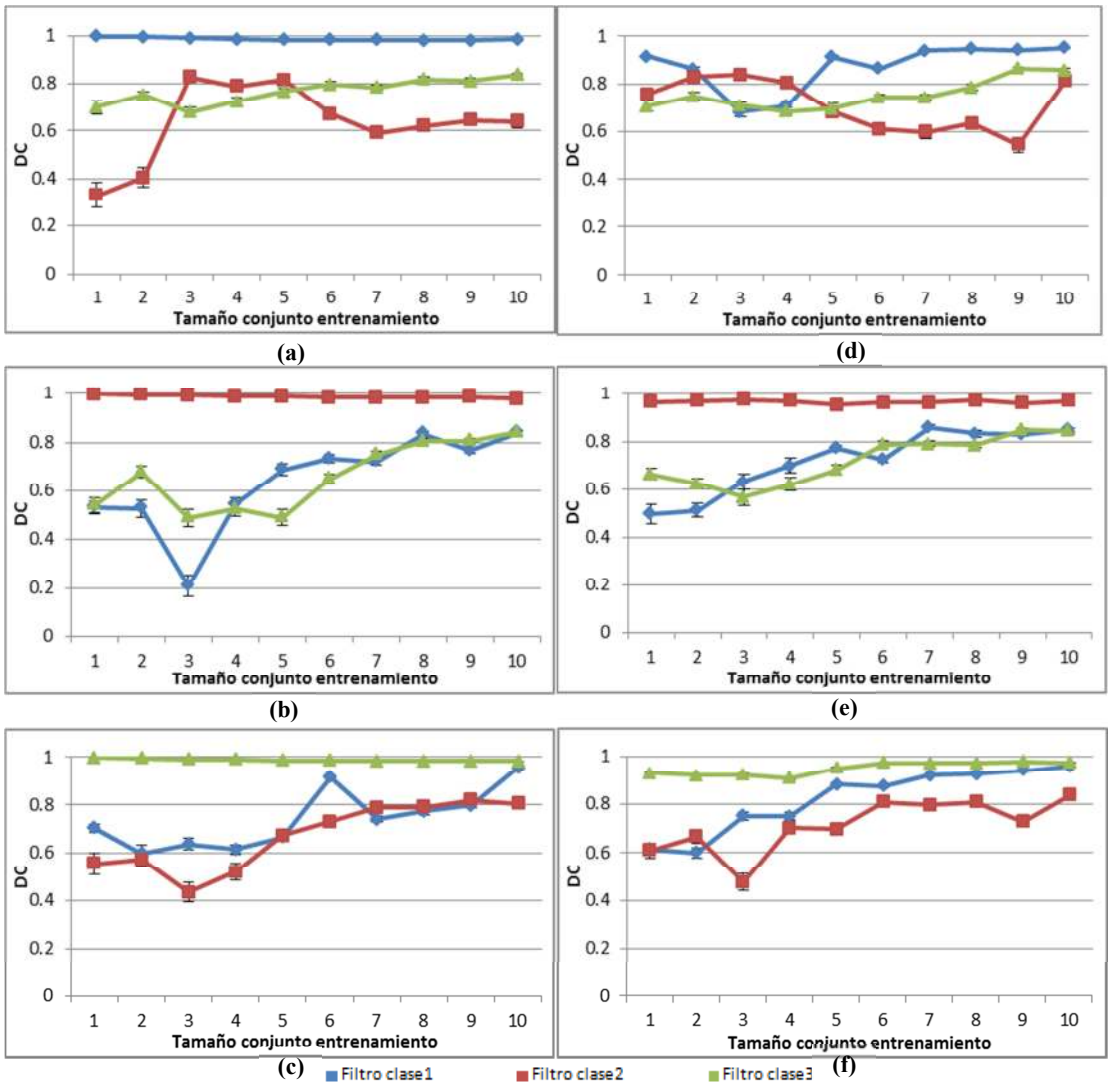


Figura 33. Desempeño al 95% de confianza en términos de DC de filtros, entrenados con varios tamaños de entrenamiento, correlacionados con 30 escenas sintéticas de las clases 1, 2 y 3, en las que la posición del rostro es aleatoria y utilizando el fondo típico, considerado homogéneo, de la Figura 29(e).

5.4. Técnicas para mejorar el desempeño de los filtros SDF blanqueados

Los trabajos previos que se han presentado sobre reconocimiento de patrones mediante filtros compuestos de correlación, en específico utilizando los filtros SDF dentro del modelo de imagen no traslapado, han dado buenos resultados considerando distorsiones del objeto a reconocer que realmente preservan una forma bien definida y, en casos de

reconocimiento multiclase, cuando los objetos de una clase tienen una forma muy distinta a los objetos de otra clase.

En el caso de reconocimiento multiclase de objetos de tipo rostro, bajo dicho modelo de escena, un problema muy grande es que se consideran objetos de distintas clases con formas muy similares, a veces con contenido muy similar cuando se trata de rostros con rasgos de una población étnica en particular. Por lo tanto en este proyecto se trabajó en la investigación de técnicas que pudieran ayudar a incrementar la diferencia entre los rostros pertenecientes a distintas clases. Se presentan a continuación dos técnicas estudiadas.

5.4.1. Expansión del conjunto de entrenamiento a través de transformaciones afines.

Como se mostró en la sección anterior, los filtros SDF blanqueados diseñados con el algoritmo adaptativo obtienen los mejores resultados cuando en la escena de entrada se presenta una de las apariencias del objeto utilizadas en la etapa de entrenamiento; sin embargo su desempeño puede bajar cuando la apariencia del objeto en la escena de entrada no se toma en cuenta en el entrenamiento. De aquí el problema fuerte de la pobre representación de cada clase al disponer de un conjunto limitado de imágenes de entrenamiento, disminuyendo la capacidad de generalización de los filtros. Para abordar esta problemática una idea es la siguiente: dado un conjunto de L imágenes de entrenamiento de una clase, se desea expandir dicho conjunto generando imágenes distorsionadas a partir de transformaciones afines (rotaciones y escalamientos), de cada uno de los elementos del conjunto original. Sea L_r el número de rotaciones y L_e el número de escalas consideradas, sean Δ_r y Δ_e los respectivos pasos de rotación y escalamiento, por cada imagen de entrenamiento, primero se generan las L_r rotaciones en incrementos de Δ_r. Se obtiene entonces un conjunto de entrenamiento de LL_r imágenes. Se prosigue a generar un nuevo conjunto redimensionando cada imagen del nuevo conjunto en incrementos de Δ_e hasta obtener L_e imágenes distorsionadas de cada una. Al final se obtiene un conjunto de entrenamiento expandido de LL_rL_e elementos por clase.

Se podrían, entonces, incluir todas las imágenes de entrenamiento del conjunto expandido en el diseño de un filtro. Si bien esta técnica resulta en un aumento en la complejidad del sistema de reconocimiento (si se considera un conjunto de entrenamiento

inicial grande o muchas clases) además de un posible decremento en el desempeño del filtro, se puede lidiar con esto utilizando las capacidades de los sistemas de cómputo actuales, como la ejecución en paralelo de trabajos. Es decir, se pueden incluir subconjuntos del conjunto de entrenamiento distorsionado en el diseño de un banco de filtros de una clase, de manera que se tomen en cuenta todas las distorsiones de todas las clases, sin incrementar demasiado el número de objetos incluidos en cada filtro y ejecutando en paralelo, tanto el algoritmo adaptativo de diseño, como la parte del reconocimiento al hacer la correlación de cada filtro con la escena de entrada. Otra posibilidad es implementar los filtros en un sistema óptico, en el que el tiempo de ejecución de una correlación no es una limitante pues se hace a la velocidad de la luz.

En este trabajo, debido al tiempo limitado, se investigó el desempeño que tienen los filtros SDF blanqueados solamente agregando, en su diseño, alguna de las imágenes distorsionadas al conjunto de entrenamiento original. Partiendo de un conjunto de entrenamiento inicial de $L \in \{1, 2, 3\}$ imágenes por clase, para $K = 3$ clases, se generaron 25 imágenes distorsionadas. Los parámetros elegidos para crear las distorsiones fueron los siguientes. Se generaron $L_r = 5$ rotaciones en incrementos de $\Delta_r = 1.5^\circ$ en un intervalo de $[-3^\circ, 3^\circ]$; y se generaron $L_e = 5$ imágenes escaladas en incrementos de $\Delta_e = 0.05$ en un intervalo de $[0.9, 1.1]$.

En las siguientes figuras se muestra el desempeño de los filtros diseñados al correlacionarse con escenas de prueba de cada clase. Se entrenaron tres filtros para cada clase, considerando tres tamaños de entrenamiento (con 1, 2 y, 3 objetos por clase). Para cada tamaño se agregó una imagen distorsionada del último elemento en cada conjunto de entrenamiento de la clase correspondiente. Los incisos (a)-(c) de la Figura 34 corresponden a la correlación con escenas prueba sintéticas de la clase 1, 2 y 3, respectivamente; los incisos (d)-(f) corresponden a la correlación con escenas prueba reales de la clase 1, 2 y 3, respectivamente. La Figura 35 muestra el desempeño, al 95% de confianza, de los filtros correlacionados con 30 escenas sintéticas en las que se varía la ubicación del objeto sobre un fondo complejo (el de la Figura 29 (d)) en los incisos (a)-(c), los cuales corresponden a escenas prueba sintéticas de la clase 1, 2 y 3; los incisos (d)-(f) corresponden a escenas

prueba sintéticas de la clase 1, 2 y 3, respectivamente, utilizando un fondo homogéneo (el de la Figura 29(e)).

En comparación con los casos problemáticos mostrados en la Figura 31(d)-(f) y (j)-(l) con tamaños de entrenamiento de 1 a 3 objetos por clase, se observa que incluyendo por lo menos una versión distorsionada de un elemento del conjunto inicial se mejora considerablemente el desempeño de los filtros. Por ejemplo en la Figura 34(a) el nivel DC alcanzado por el filtro de la clase 1 construido a partir de un objeto de entrenamiento por clase, más el mismo objeto distorsionado, sube aproximadamente a 0.8, en comparación al 0.44 que alcanza sin incluir las distorsiones, en la Figura 31(d). Similarmente ocurre con el nivel DC obtenido por los filtros de la clase 2 y 3 con dos y tres objetos de entrenamiento por clase, respectivamente, como se observa en la Figura 34(b) y (d), en comparación a los niveles de los filtros correspondientes en la Figura 31(e) y (f). Aunque la mejora de desempeño no es igual en el caso de escenas reales (ver Figura 34(d)-(f)), sí se obtiene cierto incremento en los niveles DC más bajos que se observan en la Figura 31(j)-(l).

Para el caso del desempeño en escenas sintéticas de prueba con 30 ubicaciones aleatorias del objeto, se observa igualmente una mejora del nivel DC alcanzado por los filtros para los tres primeros tamaños de entrenamiento por clase. Esto se aprecia en la Figura 35, tanto para escenas con el fondo con estructuras complejas (en los incisos (a)-(c)), como para escenas con el fondo homogéneo (en los incisos (d)-(f)). Sin embargo, la mejora que proporciona el incorporar elementos distorsionados del conjunto de entrenamiento inicial de una clase, puede resultar no tan estable como se esperaría, como se observa en algunos casos de las figuras mencionadas en esta sección.

Además, idealmente se requiere de un método para obtener, de manera automática, dichas distorsiones a incorporar al conjunto de entrenamiento, es decir, un método de selección. En esta investigación, estas distorsiones se seleccionaron de manera experimental y visual. Por esta razón se explora una segunda técnica alternativa que se presenta a continuación.

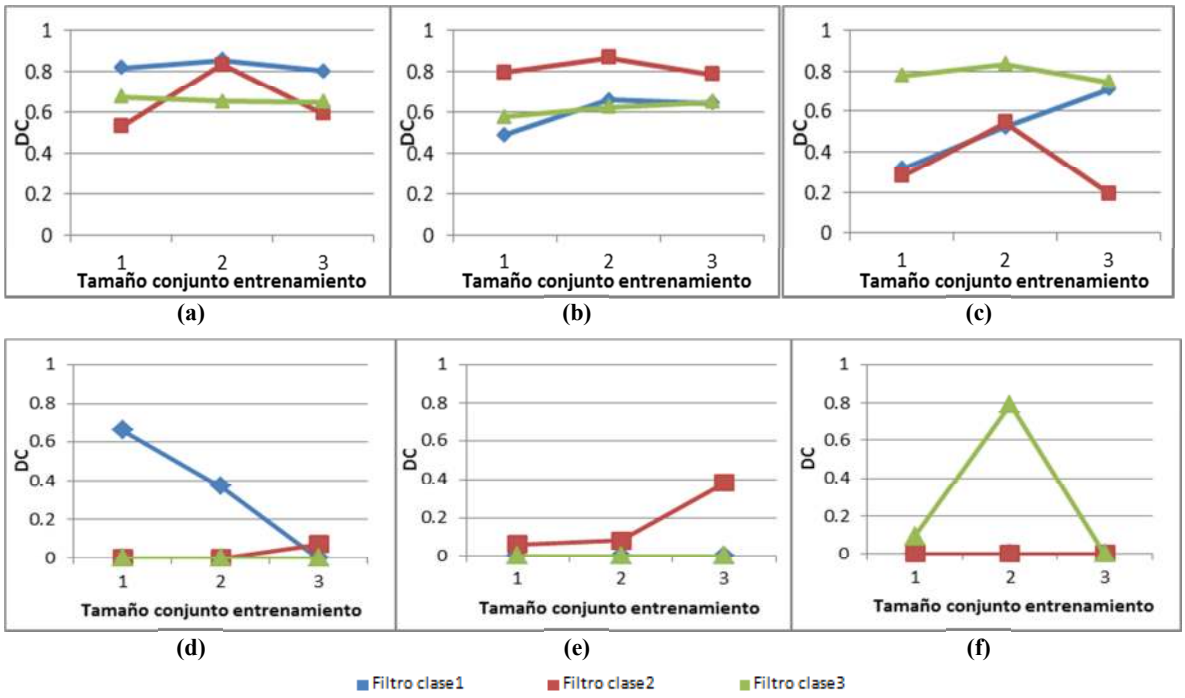


Figura 34. Desempeño en términos de DC, de filtros entrenados con varios tamaños de entrenamiento por clase incluyendo distorsiones de los objetos de entrenamiento originales, correlacionados con tres escenas sintéticas y tres escenas reales.

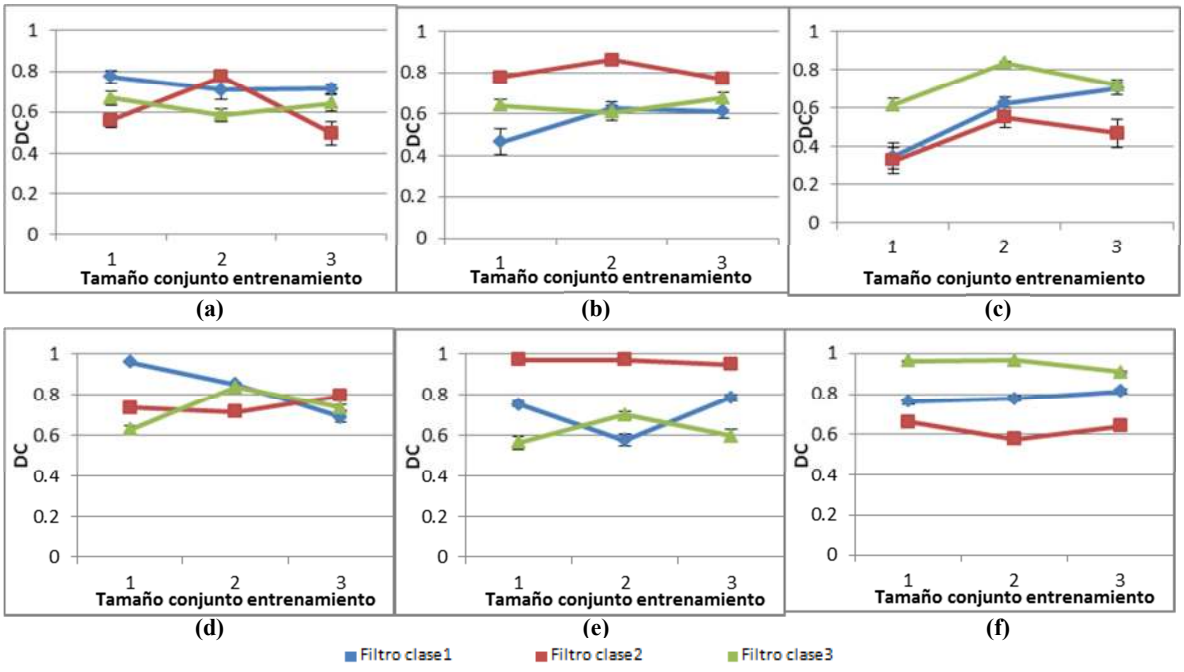


Figura 35. Desempeño con 95% de confianza en términos de DC, de filtros entrenados con varios tamaños de entrenamiento por clase incluyendo distorsiones de los objetos de entrenamiento originales, correlacionados con 30 escenas sintéticas de las en las que la posición del rostro es aleatoria.

5.4.2. Técnica basada en la generación de máscaras binarias de bloqueo de frecuencias.

El objetivo de esta técnica es separar lo más posible una clase de la otra. Es decir, debido a que los filtros SDF blanqueados prestan más atención a la información de la forma de los rostros, o bien su contorno, un filtro entrenado para reconocer una clase dada podría confundir a un objeto que pertenece a otra clase. Aunque se haga un entrenamiento del filtro de una clase con un proceso adaptativo de rechazo a los objetos de entrenamiento de otras clases, como los objetos en las escenas de entrada pueden resultar muy diferentes en cuanto a forma y contornos a la apariencia que se puede generalizar del entrenamiento, es recomendable tratar de separar lo más posible la salida de correlación de dicho filtro para objetos de su propia clase y para objetos de otras clases. Kober y Ovseyevich (2000), proponen un método para construir filtros de sólo fase (POF) y mejorar su desempeño en términos de DC, bloqueando un mínimo número de frecuencias mediante máscaras binarias. Con base en este último, se propone la siguiente metodología iterativa de creación de máscaras binarias de bloqueo de frecuencias.

Sea K el número de clases consideradas, d el número de pixeles en las imágenes y L el número de imágenes a aceptar del conjunto entrenamiento para una clase. Primero, para una clase k se correlaciona su filtro SDF blanqueado entrenado con todos sus objetos de entrenamiento correspondientes (excluyendo los patrones a rechazar calculados en el entrenamiento), se calculan los picos de correlación en el área del objeto y se toma el mínimo de ellos. Luego, se correlacionan todos los filtros de las clases restantes con los mismos objetos de entrenamiento de la clase k. Se calculan los picos de correlación en el área del objeto y se toma el máximo de ellos. El proceso de bloqueo se acciona mediante la especificación de un porcentaje de distancia que define la separación que se quiere tener entre el mínimo pico de autocorrelación (el primero calculado) y el máximo pico de correlación cruzada (el segundo calculado). Sea factor = 1/(100 - porcentaje de distancia), para cada clase se debe cumplir la siguiente desigualdad:

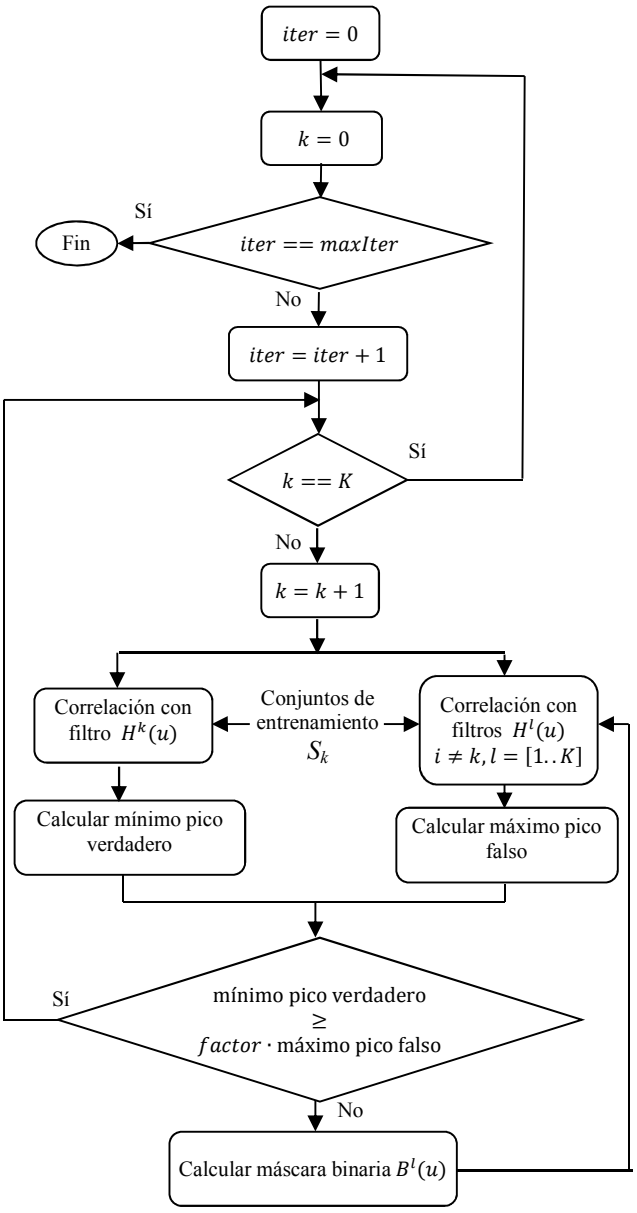
$$\min_i \left\{ \left| \sum_u H^k(u) T_i^k(u) B^k(u) \right| \right\} \geq factor \cdot \max_{i,l} \left\{ \left| \sum_u H^l(u) T_i^k(u) B^l(u) \right| \right\} \tag{101}$$

Para $i = [1 \dots L]$, $k, l \in \{1 \dots K\}$, $l \neq k$, donde $H^k(u)$ es la respuesta en frecuencia del filtro SDF blanqueado $h(x)$ de la clase *k* en cuestión, $T_i^k(u)$ es la transofrmada de Fourier de la *i*-ésima imagen de entrenamiento (a aceptar) de la clase *k* y $B^k(u) \in [0,1]$ es la máscara de bloqueo binaria de la clase en cuestión. Inicialmente, todas las máscaras constan de elementos igual a uno, es decir, no hay ningún elemento bloqueado. El argumento del primer término de la desigualdad en la ecuación (101) corresponde a los picos de correlación en el centro del plano, resultantes de la correlación del filtro de la clase *k* dada con cada elemento de entrenamiento de la misma clase y aplicando la máscara binaria correspondiente. Al hacer la suma se calcula la transformada inversa de Fourier, por lo tanto, el primer término se refiere al mínimo pico de autocorrelación, ya que se trata de elementos de una misma clase. En realidad equivale al mínimo pico verdadero, definido en la sección 5.3. Por el contrario, el argumento del segundo término de la desigualdad corresponde a los picos de correlación en el centro del plano de los filtros de las otras clases con cada elemento de entrenamiento de la clase *k* y aplicando sus respectivas máscaras binarias. Dicho de otra manera, es el máximo pico de correlación cruzada, el cual equivale al máximo pico falso, definido en la sección 5.3.

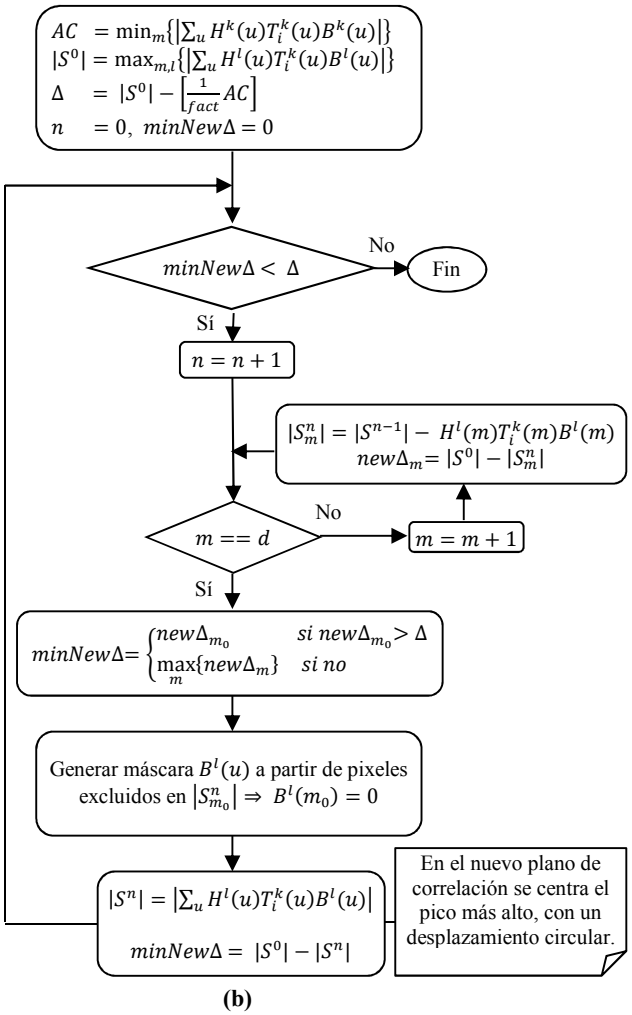
De no cumplirse la desigualdad, se prosigue a calcular (o actualizar) la máscara binaria de bloqueo para la clase *l* correspondiente al máximo pico falso. El proceso para calcular tal máscara se describe en la siguiente subsección. Una vez generada la máscara, ésta se aplica al correlacionar de nuevo los filtros de las clases restantes con las imágenes de entrenamiento de la clase *k*, verificando así que se cumpla la desigualdad (recalculando el término de la derecha). Si ésta no se cumple, se modificará la máscara de la clase *l* que resulte con el máximo pico falso debido a alguno de los objetos de entrenamiento de la clase *k*, con el mismo proceso que se describe a continuación. Cuando todas las correlaciones cruzadas cumplan con la desigualdad, el proceso se reinicia para la siguiente clase, se recalculan ambos términos de la desigualdad y así sucesivamente hasta que la desigualdad se cumpla para todas las *K* clases.

En este algoritmo, puede suceder que al generar las máscaras de las últimas clases se modifiquen los picos de autocorrelación correspondientes a las primeras clases, por lo que la desigualdad podría no cumplirse más. Debido a esto, se presenta el algoritmo de

manera iterativa, para que en iteraciones subsecuentes se logre mantener la desigualdad. El número de iteraciones está definido por *maxIter*, como se muestra en el diagrama de la Figura 36(a).



(a)



(b)

Figura 36. Diagramas del método de bloqueo de frecuencias. (a) Diagrama de la primera parte, iniciadora del bloqueo. (b) Diagrama del cálculo de máscaras binarias de bloqueo.

5.4.2.1. Cálculo de máscaras binarias de bloqueo

En este proceso se desea que el filtro de la clase *l* obtenga un pico de correlación máximo igual o menor a $(1/fact) \cdot$ mínimo pico verdadero. En el paso 3 del algoritmo de la Figura 37, se calcula la distancia del máximo pico falso al valor del pico que se desea, definida por Δ (ver Figura 38(a)). En la primera iteración se calculan *d* picos de correlación centrados, $|S_m^n|$, mediante sumas de los elementos de la correlación en el dominio de frecuencia, correspondientes al máximo pico falso, excluyendo cada elemento de la correlación, uno a la vez. Luego se calcula la distancia que se tiene del máximo pico falso a cada uno de los picos $|S_m^n|$. Los pasos 7-9 de la Figura 37 corresponden a este proceso. Una vez calculadas las distancias, se define $minNew\Delta$ como la distancia más cercana a Δ (ya

sea más grande o más pequeña), como en el paso 10 de la Figura 37. Con base en esta distancia se eligen los elementos que serán bloqueados en la máscara binaria $B^l(u)$ del paso 11. Al bloquear un nuevo elemento de la máscara, se calcula el pico de correlación de la iteración actual, $|S^n|$, del cual se irán extrayendo elementos en la siguiente iteración y además se actualiza la distancia del máximo pico falso al pico de correlación de la iteración actual, $minNew\Delta$. Por último, si la distancia actual $minNew\Delta$ es mayor o igual a Δ, se ha logrado el objetivo y el proceso termina. De lo contrario, se comienza una nueva iteración para calcular otro elemento a bloquear en la máscara binaria $B^l(u)$. Este proceso, también se muestra en manera de diagrama en la Figura 36(b).

1. $AC =$ mínimo pico verdadero $= \min_i\{|\sum_u H^k(u)T_i^k(u)B^k(u)|\}$
2. $|S^0| =$ máximo pico falso $= \max_{i,l}\{|\sum_u H^l(u)T_i^k(u)B^l(u)|\}$
3. $\Delta = |S^0| - \left[\frac{1}{fact}AC\right]$
4. $n = 0$, $minNew\Delta = 0$
5. Mientras $minNew\Delta < \Delta$
6. $\quad n = n + 1$
7. $\quad \forall m \in \{1..d\}$
8. $\quad\quad |S_m^n| = |S^{n-1}| - H^l(m)T_i^k(m)B^l(m)$
9. $\quad\quad new\Delta_m = |S^0| - |S_m^n|$
10. $minNew\Delta = \begin{cases} new\Delta_{m_0} & si\ new\Delta_{m_0} > \Delta \\ \max_m\{new\Delta_m\} & si\ no \end{cases}$
11. Generar máscara $B^l(u)$ a partir de pixeles excluidos en $|S_{m_0}^n| \Rightarrow B^l(m_0) = 0$
12. $|S^n| = |\sum_u H^l(u)T_i^k(u)B^l(u)|$
    *En el nuevo plano de correlación se centra el pico más alto, con un desplazamiento circular.
13. $minNew\Delta = |S^0| - |S^n|$

Figura 37. Algoritmo del cálculo de máscaras binarias de bloqueo.

Es importante mencionar que al modificar la máscara binaria, se debe correlacionar el filtro con el objeto de entrenamiento aplicando la máscara para actualizar el pico de correlación; éste debe centrarse en el plano en dominio espacial (si es que el máximo pico se encuentra ahora en otra ubicación), para así poder utilizar la forma de sumas en dominio de frecuencia. Se realiza de esta manera con el fin de reducir el número de correlaciones que se harían para calcular los elementos a bloquear.

Otro aspecto importante es que, para preservar la simetría del espectro de los filtros y obtener un plano de correlación lineal real, en la implementación de esta metodología se aprovechó el hecho de que la transformada de Fourier presenta simetría (ver propiedades de la transformada de Fourier del capítulo 2) y en ocasiones se bloquean dos elementos simétricos a la vez, como se muestra en la Figura 38(b). Con esto se reduce aún más el número de cálculos. En la Figura 38(c) se muestran los elementos utilizados en los cálculos, correspondientes a los pasos 7-9 de la Figura 37, en la zona sombreada (que tienen simetría) y la zona en negro (que no tienen simetría), para una imagen de 8×8 pixeles.

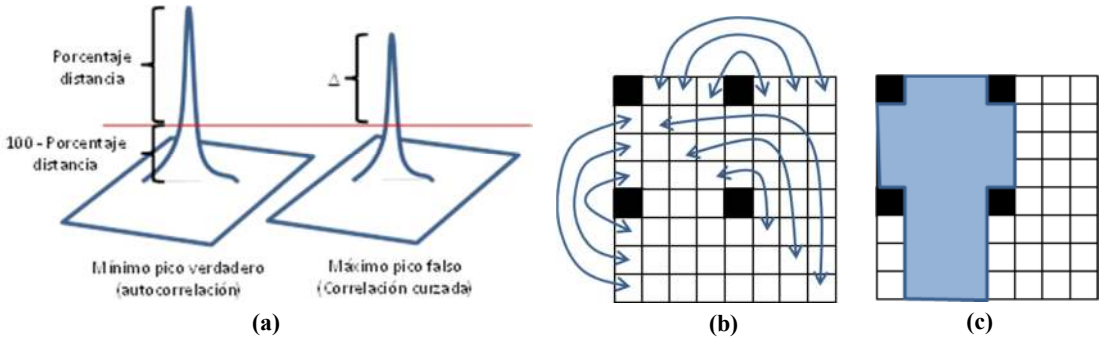


Figura 38. Método de bloqueo de frecuencias y su funcionamiento. (a) Elementos considerados para accionar el bloqueo. (b) Simetría de la transformada de Fourier discreta. (c) Zonas sombreadas de la transformada de Fourier discreta utilizadas en el cálculo de la máscara binaria.

5.5. Resumen

Se presentó en este capítulo, una técnica de reconocimiento de rostros en escenas reales, basada en filtros de correlación. Esta técnica utiliza un esquema iterativo de adaptación en el diseño de cada filtro, para incorporar información de distintas apariencias del rostro de un individuo dado y rechazar rostros de otros individuos. Además se presentan dos técnicas para maximizar la diferencia entre clases y mejorar el desempeño de los filtros diseñados.

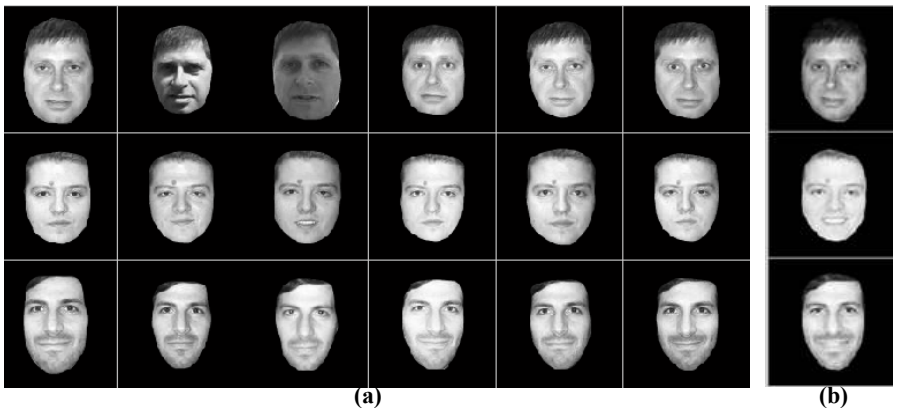
Capítulo 6. Experimentos y resultados

Con el fin de comparar las técnicas estadísticas de reconocimiento de rostros estudiadas en el capitulo 3 (adaptadas con el detector de Viola y Jones) y la técnica con filtros de correlación propuesta, se realizaron dos experimentos generales usando la base de datos Faces 1999 de Caltech, descrita en el apéndice.

Las imágenes de la base de datos se convirtieron a formato PNG y se redimensionaron a tamaño 223×147, usando la rutina *imresize()* de MATLAB con interpolación bicúbica. Además se transformaron de tres canales de color RGB a escala de gris, mediante la rutina *rgb2gray()*. Para crear los conjuntos de entrenamiento (o galería, como se definen en la sección 2.5), se hizo una segmentación manual de los rostros de la base de datos Caltech tratando de capturar la forma completa del rostro, excluyendo la región del cabello y orejas, dando como resultado imágenes como las mostradas en la Figura 39. Esto se hizo mediante el editor de gráficos GIMP; al conjunto de imágenes resultantes se le refiere como la base de datos Caltech segmentada.

Figura 39. Imágenes segmentadas para su uso en el entrenamiento de sistemas de reconocimiento de rostros.

Se utilizaron seis imágenes de tres individuos (clase 1, clase 2 y clase 3) de la base de datos Caltech segmentada para formar los conjuntos de entrenamiento y una imagen de cada uno de esta base, junto con la escena completa correspondiente de la base datos Caltech para pruebas. Éstas se pueden apreciar en la Figura 40 y en la Figura 41.



(a) (b)

Figura 40 Rostros de la base de datos Caltech segmentada, utilizadas en los experimentos. (a) Imágenes para la selección de los conjuntos de entrenamiento. (b) Rostros de prueba.

Todas las técnicas se implementaron en MATLAB R2010b. Las implementaciones de las técnicas estadísticas (referidas con las siglas PCA, LDA e ICA, en las gráficas de resultados), así como el detector de rostros, se basaron en implementaciones recuperadas de internet, como se indica en la sección 3.5. En particular, el detector de rostros utiliza el clasificador entrenado para rostros que proporciona la librería de código abierto OpenCV. Para la implementación de la técnica de Eigenfaces se tienen dos configuraciones; una en la que se toman todos los componentes principales posibles (M-1, donde M es el número de imágenes de entrenamiento) y otra en la que se toma el 40% de los componentes, cantidad que es usual tomar, según la literatura. La primera configuración se tomó en cuenta porque, como en los experimentos mostrados se utilizan conjuntos de entrenamiento muy pequeños, al desechar 60% de los componentes, se desecha una gran parte de la información significativa. Para el caso de la implementación de ICA de la Arquitectura I, se utilizó la configuración propuesta en Bartlett et al., (1998), con 40% de componentes principales. La implementación de la técnica propuesta utiliza las rutinas de MATLAB *fft2()* e *ifft2()* para poder realizar la correlación entre las imágenes. En las siguientes subsecciones se presentan resultados tanto de la técnica de correlación propuesta sin aplicar bloqueo de frecuencias como al aplicar la técnica de bloqueo de frecuencias. Para esta misma, el parámetro que acciona el mecanismo de bloqueo (el porcentaje de distancia deseada entre el pico de autocorrelación y el máximo pico de correlación cruzada de los filtros respecto a las imágenes de entrenamiento) se estableció experimentalmente en 88.89% del pico de autocorrelación.

En ambos experimentos se determina el desempeño de las técnicas respecto a la tasa de reconocimiento (definida en la sección 2.5). En particular, el desempeño de los filtros de correlación propuestos se determina mediante el criterio DC.

6.1. Experimento 1: Reconocimiento en escenas sintéticas y reales con diferente tamaño de conjunto de entrenamiento.

Este experimento tiene como objetivo reconocer (detectar y clasificar) un rostro en una posición fija de la escena de prueba, con un conjunto de entrenamiento de tamaño determinado. Se consideran dos casos: uno en el que se usa un conjunto de prueba A con escenas sintéticas y otro en las que se usa un conjunto de prueba B con escenas reales. Las escenas sintéticas del conjunto A se componen por una imagen de fondo (ver Figura 41(b)) y un rostro de entrenamiento superpuesto, como las mostradas en la Figura 41(a). Las escenas reales corresponden a las de la Figura 41(c).

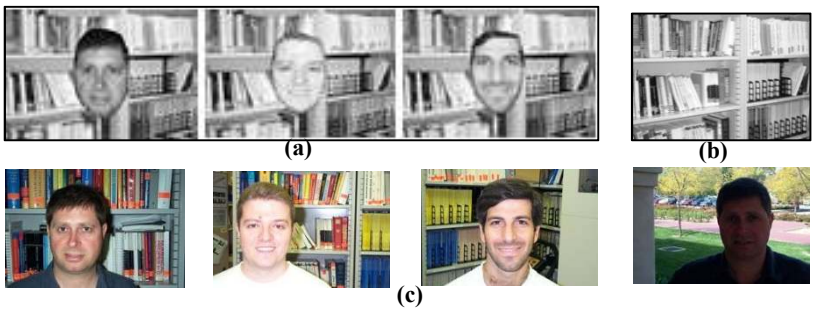


Figura 41. Escenas sintéticas y reales usadas en experimentos. (a) Escenas sintéticas del conjunto A de prueba. (b) Fondo para componer las escenas sintéticas. (c) Escenas reales del conjunto B de prueba.

Para establecer niveles de confianza del desempeño de las técnicas, se hicieron 30 selecciones aleatorias de los conjuntos de entrenamiento. Se formaron conjuntos con 2, 3 y 4 imágenes de entrenamiento por clase, por lo que las galerías tuvieron un tamaño de 6, 9 y 12, respectivamente. Los resultados del conjunto de prueba A se muestran en la Figura 42.

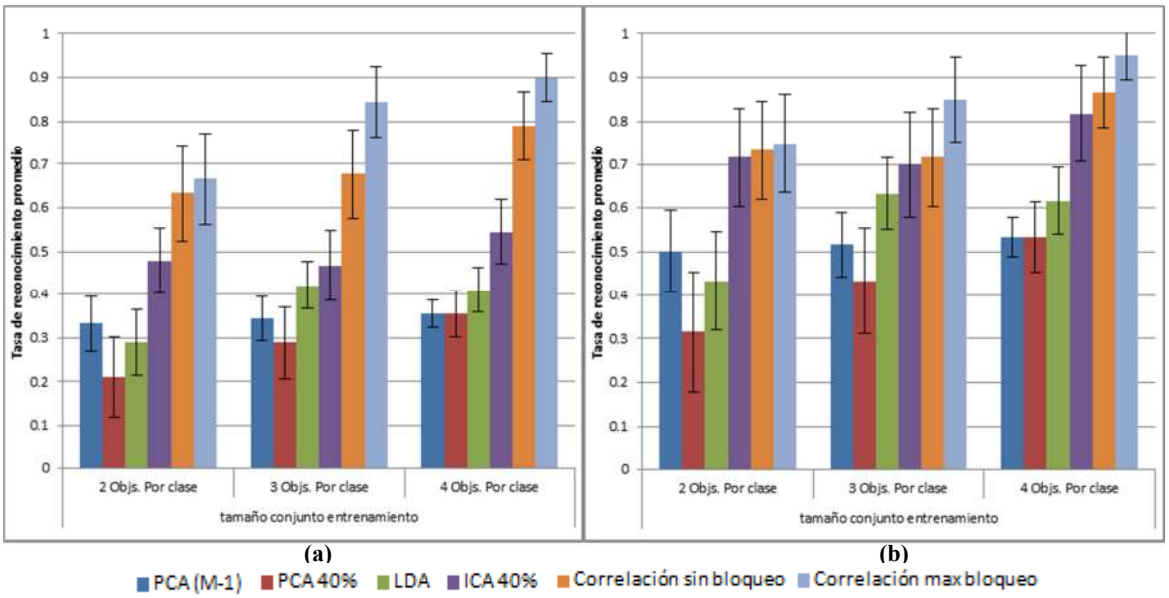


Figura 42. Tasa de reconocimiento (TR) promedio, al 95% de confianza, de todas las técnicas implementadas (escenas sintéticas). (a) TR de las 3 escenas sintéticas de prueba. (b) TR de 2 de las 3 escenas sintéticas de prueba que fueron segmentadas por el detector.

Dado que una de las escenas de prueba del conjunto A, la del individuo de la clase 2, no pudo ser segmentada por el detector utilizado, provocando que no pasara a la entrada del sistema y no se clasificara, en la figura se muestran tanto los resultados tomando en cuenta todas las escenas del conjunto de prueba A (inciso (a)), como aquellos considerando sólo las escenas de prueba segmentadas (inciso (b)). Quiere decir que se pueden presentar los resultados como aquellos que ilustran el desempeño por la detección y aquellos que lo ilustran por la clasificación puramente.

Sobre los resultados de tasa de reconocimiento (TR) promedio, obtenidos con LDA, los cuales muestran una oscilación al incrementar el tamaño de conjunto de entrenamiento (TR de tamaño 4 más bajo que TR de tamaño 3 en el inciso (a) como en el (b) de la Figura 42), se puede mencionar que el objeto de prueba de la clase 1 se confundió con regularidad con la clase 3, después de haber hecho una inspección de las imágenes de prueba y los resultados individuales de clasificación. Esto podría deberse a que los objetos de entrenamiento seleccionados, en la mayoría de las 30 selecciones aleatorias, no eran los más representativos. Además, el método de LDA (Fisherfaces) generalmente tiene mejor desempeño cuando se consideran más imágenes de entrenamiento por clase ya que se dispone de más información para poder discriminar y lograr la maximización de diferencias

entre clases y minimización de diferencias intraclase; mínimo se pueden tener dos imágenes por clase. Aunque se incrementa el número de imágenes de entrenamiento, puede ser que no sean suficientes para alcanzar un comportamiento típico. En cuanto al número de clases consideradas, no hay restricción excepto que por lo menos sean dos. Sin embargo, por lo general se utilizan 10 o más clases al utilizar este método para reconocimiento de rostros.

Por otro lado, es de notar que la TR promedio de ICA con 40% de componentes principales retenidos oscila, como en el caso de LDA pero de manera contraria, presentándose por igual para escenas sintéticas, ya sea tomando en cuenta todas las escenas segmentadas o no. Respecto a estas oscilaciones para entrenamiento de tamaño tres por clase, mediante inspección de la clasificación más a fondo, se puede comentar que se presentaron confusiones a la hora de clasificar el objeto de la clase 3 como perteneciente a la clase 2. Observando que en las selecciones de entrenamiento de la clase 3 se tuvieron en su mayoría objetos ligeramente mal alineados (un poco más alto el nivel de los ojos con respecto al resto de los objetos), podría argumentarse que esto influyó en que no se tuviera la mejor similitud del objeto de prueba de la clase 3 con los objetos de entrenamiento seleccionados de su clase, creando así la confusión con la clase 2. Consecuentemente, la técnica de ICA se presume más sensible, que las otras técnicas, a conjuntos de entrenamiento donde los objetos no estén alineados homogéneamente.

Los resultados de TR promedio con filtros de correlación sin aplicar la técnica de bloqueo de frecuencias, presentan mejores niveles de TR que los métodos clásicos (PCA, LDA, ICA) en escenas sintéticas, debido, en parte por la inhabilidad del detector utilizado de segmentar una de las escenas de prueba como se mencionó y en parte porque los filtros utilizados están precisamente diseñados con la información casi completa del escenario de estas escenas; es decir, se usa información del fondo esperado en la escena de prueba, como su media, para entrenar los filtros. Sin embargo, el desempeño en términos de TR del método ICA, queda muy cercano al obtenido con correlación (ver Figura 42(b)).

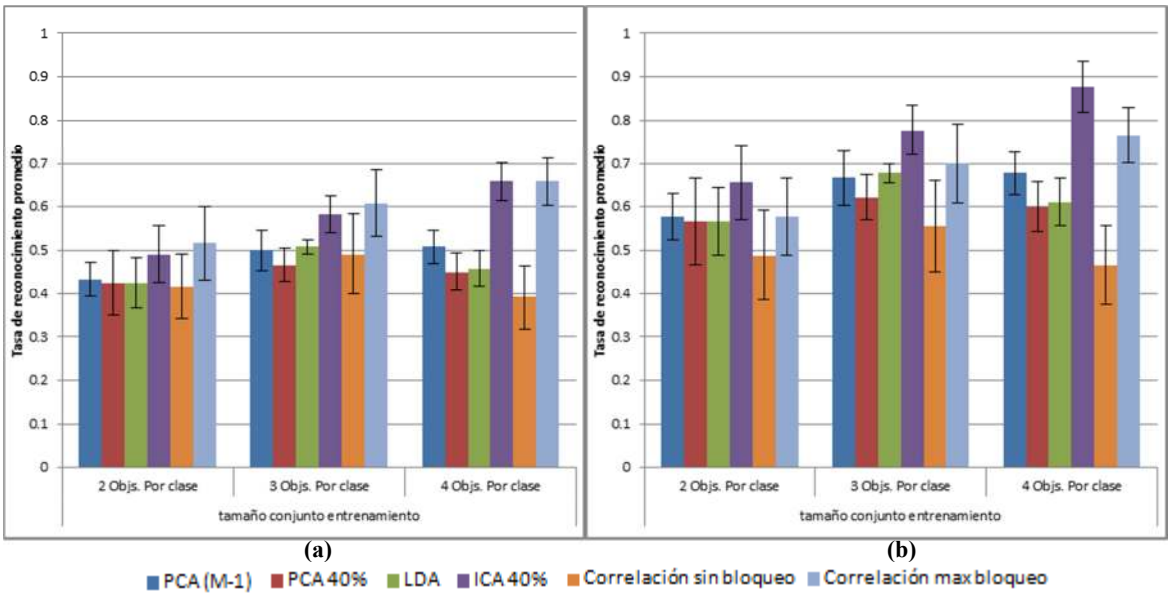


Figura 43. Tasa de reconocimiento (TR) promedio, al 95% de confianza, de todas las técnicas implementadas (escenas reales). (a) TR de las 4 escenas reales de prueba. (b) TR de 3 de las 4 escenas reales de prueba.

El conjunto B de prueba incluye dos imágenes de la clase 1 y una imagen de las clases 2 y 3, como se muestra en la Figura 41 (c). También se dividen en dos gráficas los resultados (ver Figura 43); cuando se toman en cuenta todas las escenas de prueba (inciso (a)) y cuando sólo se toman en cuenta las tres primeras escenas, las cuales fueron segmentadas satisfactoriamente (inciso (b)).

Los resultados del experimento con este conjunto B (ver Figura 43) muestran que la TR de la técnica de ICA no presenta oscilación, es decir, incrementa al mismo tiempo que incrementa el tamaño del conjunto de entrenamiento, como se hubiera esperado en un principio. Esto se atribuye a que los rostros segmentados de las escenas de prueba reales que entrega el detector, tienen mejor alineación que aquellos rostros segmentados de las escenas sintéticas, mejorando así la similitud con los rostros de los conjuntos de entrenamiento.

Con escenas reales de prueba, se puede observar que el desempeño de la técnica de correlación sin aplicar bloqueo de frecuencias es en general inferior al desempeño de las técnicas estadísticas. Dada una revisión visual de los elementos de entrenamiento y de prueba en cuestión, podría argumentarse, que la oscilación en la TR de correlación sin bloqueo, para tamaño 4 de entrenamiento por clase (Figura 43(b)), se puede deber a que hay una confusión del objeto de la clase 1 como perteneciente a la clase 3 por la forma o

contorno del rostro de prueba que resulta ser más similar a la forma de los objetos de entrenamiento de la clase 3. También se tiene que, para escenas reales y tamaño 4 objetos de entrenamiento por clase, el filtro de la clase 3 falla al localizar al objeto de prueba de su clase, por lo que se puede decir que dicho filtro confunde a estructuras que aparecen en la escena como un objeto de su clase. Aquí se observa la importancia de los objetos de entrenamiento para los filtros de correlación de tipo GOF, los cuales le dan mayor relevancia a la forma y contorno de los objetos con los que se diseña.

Respecto a las TR obtenidas de la técnica de correlación con y sin bloqueo de frecuencias (ver Figura 43(b)), se aprecia que, si bien la TR al aplicar bloqueo no se incrementa de manera constante al incrementar el tamaño de entrenamiento, sí lo hace monótonamente. Es decir, mientras que para la técnica de correlación sin bloqueo se observa un decremento de TR para 4 objetos de entrenamiento, al aplicar bloqueo se convierte en un incremento. De aquí que la aplicación de la técnica de bloqueo separa satisfactoriamente las clases, resultando en una mejora en la clasificación de los objetos de prueba en escenas reales.

Analizando el desempeño de las técnicas PCA e ICA con escenas reales (ver Figura 43), se observa cómo la TR es mayor que con escenas de pruebas sintéticas (ver Figura 42), aun cuando no se toma en cuenta la imagen que no es segmentada y que no pasa al sistema de reconocimiento (incisos (b) de ambas figuras). Aunque el desempeño de estos métodos es mejor si solamente se consideran aquellas escenas de prueba sintéticas que fueron segmentadas satisfactoriamente, cuando se trata de escenas reales, el desempeño siempre es más alto.

Por el contrario, el método LDA obtiene resultados similares cuando se trata de escenas sintéticas o reales, indistintamente. Es claro que el desempeño aumenta si sólo se consideran escenas de prueba segmentadas satisfactoriamente, pero comparando los niveles de TR para esta técnica en escenas sintéticas y en escenas reales, se encuentra que son, en general, muy similares dentro de un rango; entre 0.3 y 0.5 de TR, aproximadamente, cuando se consideran todas las escenas de prueba incluidas aquellas no segmentadas y entre 0.4 y 0.6 de TR, aproximadamente, cuando sólo se consideran las segmentadas.

En resumen, las técnicas que obtienen mejor TR promedio en las 30 selecciones son las de ICA, correlación sin bloqueo y correlación con bloqueo, para escenas de prueba sintéticas; y la de ICA y la de correlación con bloqueo para escenas de prueba reales. Sin embargo, si solamente se compara ICA con el método de correlación con bloqueo, al considerar las 4 escenas de prueba (Figura 43(a)), la técnica de correlación está por arriba de ICA en los primeros dos tamaños de entrenamiento. Por el contrario, al considerar sólo las escenas segmentadas, tres escenas en este caso (ver Figura 43(b)), el desempeño de ICA es mejor que con correlación.

En general, se observa una tendencia de aumento de la TR al incrementar el tamaño del conjunto de entrenamiento para todas las técnicas, aunque más marcada en el caso del experimento con el conjunto A de prueba.

Ahora bien, al tomar los resultados de TR máxima que se puede alcanzar con el método de correlación aplicando el bloqueo, se observa un incremento en la distancia de niveles DC promedio entre el filtro correcto y los filtros de otras clases, para todos los tamaños de entrenamiento (ver Figura 44). Las distancias de los DC promedios se muestran en la Tabla 1 y Tabla 2. También aquí se observa que los niveles DC pueden bajar o subir, sin embargo, independientemente de esto la distancia sólo incrementa monótonamente, como ya se mencionó. Lo mismo ocurre en escenas reales, en las que se incrementa la distancia entre el nivel DC promedio del filtro correcto y el de los otros filtros, para todos los tamaños de entrenamiento, como se muestra en la Figura 45 y la Tabla 3 y Tabla 4.

Tabla 1. Diferencia de niveles DC promedio en escenas de prueba sintéticas (correlación sin bloqueo).

Tamaño entrenamiento	Escena clase 1	Escena clase 2	Escena clase 3
2 objs. por clase	0.113	0.028	0.112
3 objs. por clase	0.175	0.073	0.079
4 objs. por clase	0.195	0.034	0.093

Tabla 2. Diferencia de niveles DC promedio en escenas de prueba sintéticas (máximo desempeño de correlación con bloqueo).

Tamaño entrenamiento	Escena clase 1	Escena clase 2	Escena clase 3
2 objs. por clase	0.113	0.036	0.112
3 objs. por clase	0.188	0.097	0.125
4 objs. por clase	0.195	0.096	0.171

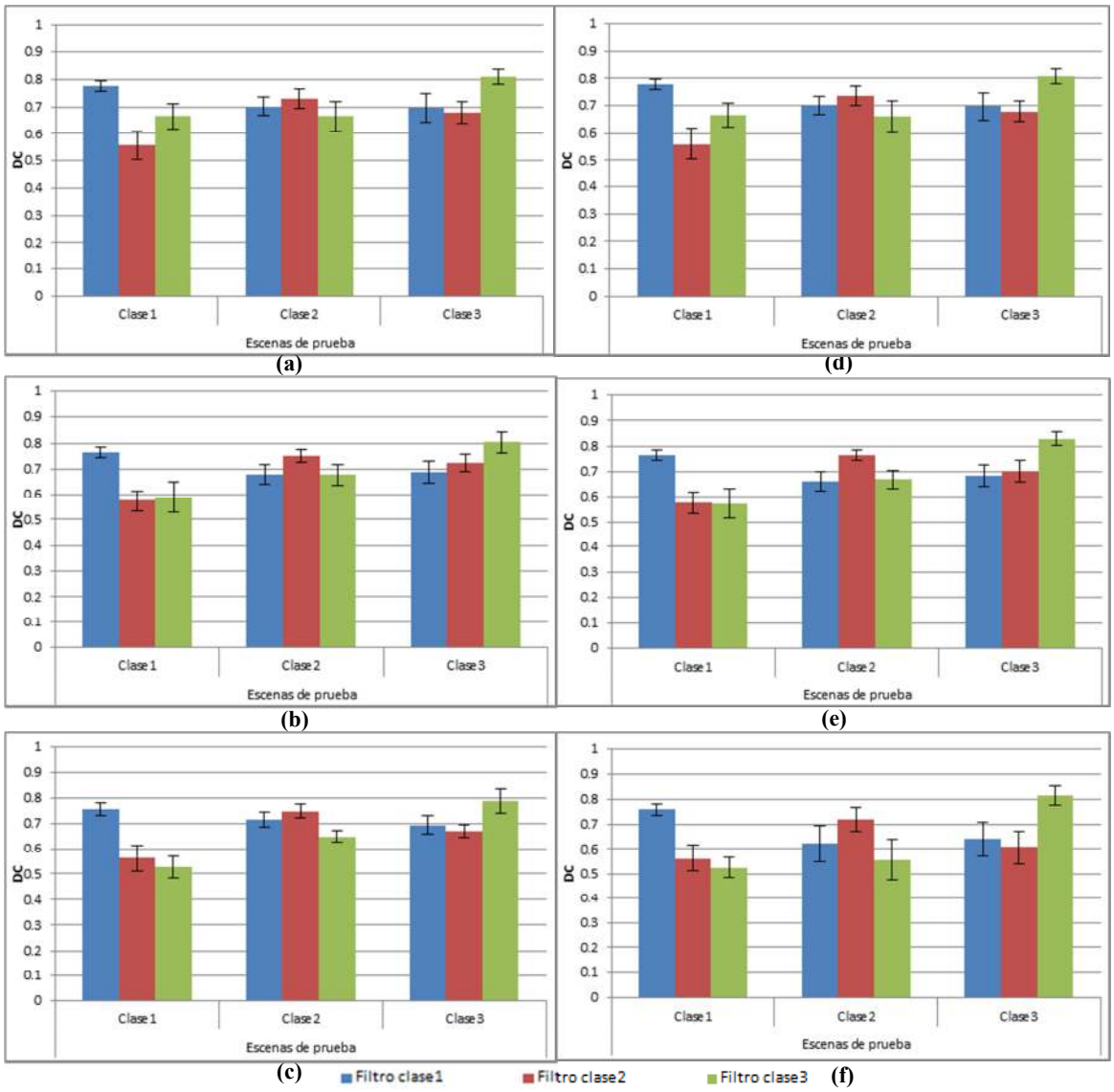


Figura 44. Desempeño de filtros SDF blanqueados en escenas sintéticas del experimento 1, con confianza del 95%. (a)-(c) Desempeño sin aplicar bloqueo. (d)-(f) Desempeño de los filtros al aplicar la técnica de bloqueo.

Observando los detalles del desempeño de los filtros en las cuatro escenas reales de prueba (ver Figura 45), se tiene que la cuarta escena no se puede clasificar correctamente. El filtro de la clase 3 confunde al objeto de la clase 1. Es por esto que el método de ICA obtiene mejor desempeño que el método de correlación con bloqueo aun cuando una de las escenas de prueba no haya podido entrar al sistema para clasificarla.

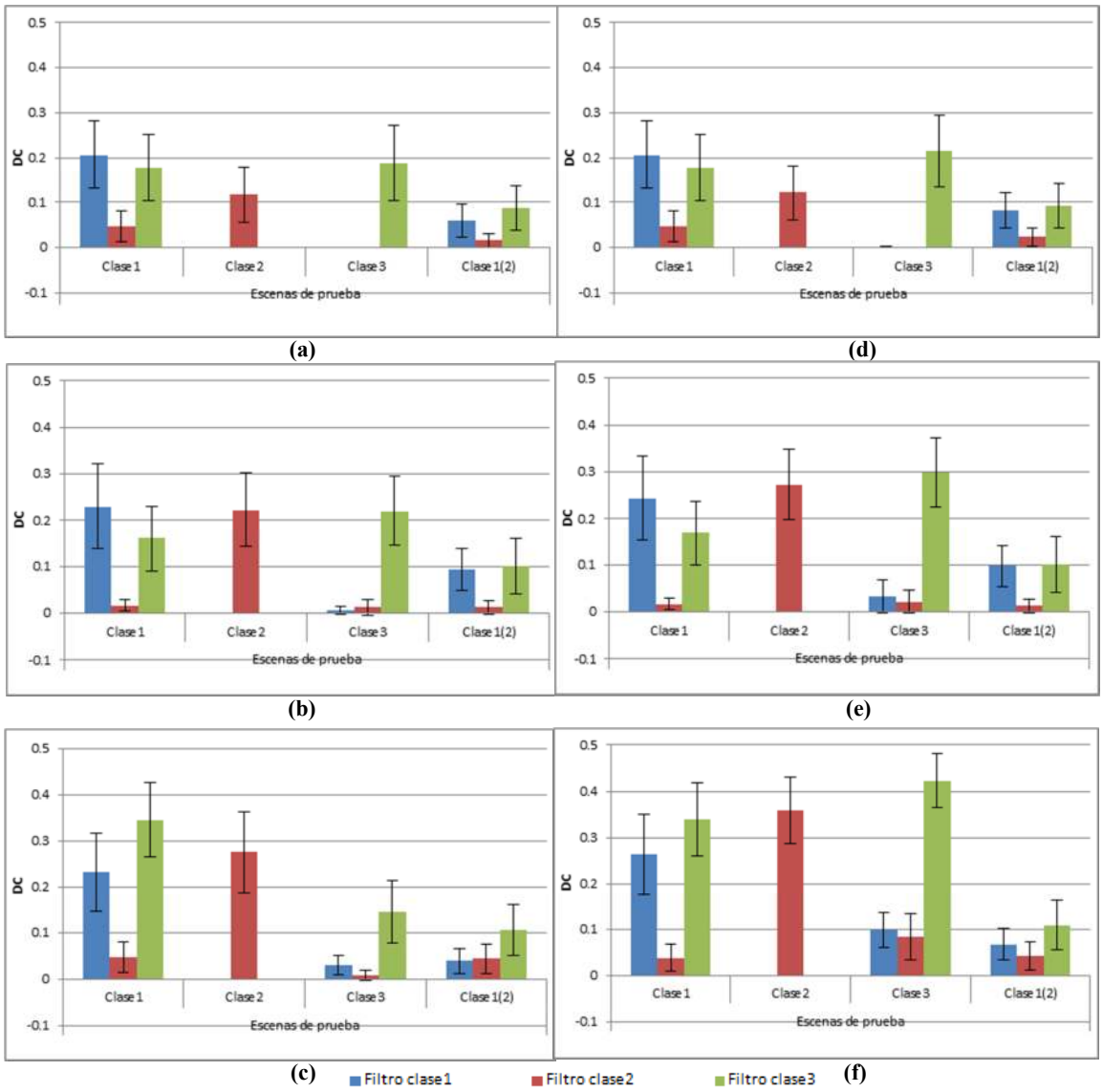


Figura 45. Desempeño de filtros SDF blanqueados con escenas reales, con confianza del 95%. (a)-(c) Desempeño sin aplicar bloqueo. (d)-(f) Desempeño de los filtros al aplicar la técnica de bloqueo.

Tabla 3. Diferencia de niveles DC promedio en escenas de prueba reales (correlación sin bloqueo).

Tamaño entrenamiento	Escena clase 1	Escena clase 2	Escena clase 3	Escena clase 1 (2da)
2 objs. por clase	0.028	0.117	0.188	-0.027
3 objs. por clase	0.070	0.223	0.207	-0.007
4 objs. por clase	-0.114	0.276	0.115	-0.066

Tabla 4. Diferencia de niveles DC promedio en escenas de prueba reales (máximo desempeño de correlación con bloqueo).

Tamaño entrenamiento	Escena clase 1	Escena clase 2	Escena clase 3	Escena clase 1 (2da)
2 objs. por clase	0.028	0.121	0.213	-0.011
3 objs. por clase	0.074	0.273	0.265	-0.003
4 objs. por clase	-0.077	0.360	0.325	-0.041

6.2. Experimento 2: Reconocimiento invariante a desplazamiento en escenas sintéticas.

Este experimento tiene como objetivo reconocer (detectar y clasificar) un rostro en cualquier ubicación de la escena de prueba, dado un conjunto de entrenamiento. Las escenas sintéticas se componen por una imagen de fondo y un rostro de entrenamiento superpuesto, como los mostrados en la Figura 41(a).

Se tomaron 30 desplazamientos aleatorios de los objetos de prueba en la escena de entrada. Se utilizaron conjuntos de entrenamiento con 2, 3 y 4 imágenes de entrenamiento por clase, previamente seleccionados.

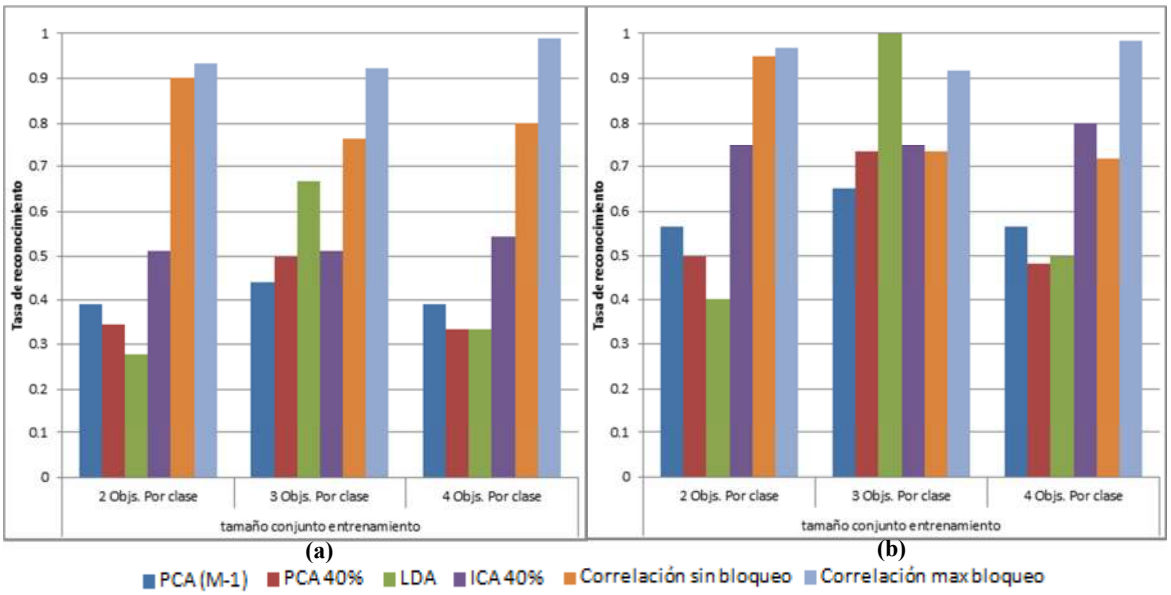


Figura 46. Tasa de reconocimiento del experimento 2 con 30 desplazamientos de los objetos en la escena de prueba, de todas las técnicas.

En la Figura 46 se ilustra el desempeño respecto a la tasa de reconocimiento de todas las técnicas considerando los desplazamientos para las tres clases. En el inciso (a) se aprecia que las técnicas de correlación, tanto sin aplicar la técnica de bloqueo como al aplicarla son las que tienen tasas de reconocimiento mayores, inclusive mayor que la

técnica de ICA, la cual obtuvo muy buenos resultados en el experimento 1. La mejoría que puede alcanzar el desempeño de la técnica de correlación aplicando el bloqueo de frecuencias se hace más notoria al utilizar 3 o 4 objetos de entrenamiento por clase. Sin embargo, al hablar de invarianza a desplazamiento, uno esperaría tener la misma TR en cualquier ubicación que se encuentren los objetos de prueba, respecto a la TR obtenida cuando los objetos de prueba están centrados, si esa es la referencia, indistintamente de qué tal alta sea. Para analizar este aspecto se establece un criterio definiendo la estabilidad del desempeño de una técnica de reconocimiento dada. Este criterio, llamado coeficiente de invarianza, corresponde a la suma de las varianzas de las TR por clase para todos los 30 desplazamientos.

Es importante mencionar que en este experimento, de las 30 escenas de prueba de la clase 2, no todas fueron detectadas y segmentadas en la etapa de preprocesamiento de los métodos estadísticos, similar a lo que sucedió con la escena de la misma clase (con el objeto centrado) en el experimento 1. Solamente una de las escenas de las 30 posiciones de la clase 2 fue segmentada. Por esta razón, de la misma manera como se presentaron los resultados del experimento 1, se analizan también tomando en cuenta sólo las escenas de las clases 1 y 3 que siempre son segmentadas. Los resultados de las TR sin tomar en cuenta las 30 escenas de prueba de la clase 2, se muestran en la Figura 46(b).

En la Figura 47 se muestra el desempeño en TR por clase de cada técnica. Se puede apreciar que para la clase 2, la TR de los métodos estadísticos es demasiado baja; esto ilustra lo mencionado en el párrafo anterior. Los coeficientes de invarianza a desplazamiento de las técnicas de reconocimiento implementadas para el caso en el que se toman en cuenta todas las escenas de las tres clases, sean segmentadas o no y que se ilustran en la Figura 47, se muestran en la Tabla 5. Para el caso cuando sólo se toman en cuenta las clases cuyas escenas fueron todas segmentadas, los coeficientes de invarianza a desplazamiento se muestran en la Tabla 6. Mientras más tienda a cero el coeficiente de invarianza, mayor estabilidad tendrá la técnica de reconocimiento en su desempeño. El coeficiente máximo que se puede obtener para una clase es de 0.259, mientras que para cuando se toman en cuenta todas las escenas de las 3 clases el coeficiente máximo es 0.776. Cuando se toman en cuenta sólo dos clases, el coeficiente máximo es 0.517.

De estas dos tablas se puede concluir que las técnicas que presentan mayor estabilidad de TR son las de correlación, aplicando el bloqueo de frecuencias y LDA; sus respectivos coeficientes mínimos están en negritas en las tablas mencionadas. Asimismo, se puede apreciar que la técnica de correlación sin bloqueo presenta un coeficiente un poco más grande, resultando ser menos invariante a desplazamiento, lo que se le atribuye al fondo no-homogéneo al que son sensibles los filtros diseñados, ya que queda fuera de la teoría de los mismos. Mientras tanto, la técnica de PCA en general presenta coeficientes más grandes para los tres tamaños de entrenamiento. Esto podría deberse a la variación en el fondo presente en las imágenes segmentadas por el desempeño del detector utilizado. Éste puede detectar un rostro, pero el área segmentada pude no ser la adecuada; es decir, que el tamaño del cuadro segmentado contenga más área del fondo de lo que se debería tener en el caso ideal. De aquí la relevancia del posprocesamiento que se le aplique a la información que arroja el detector.

También se aprecia, en la Figura 47(f), cómo el método de correlación con la técnica de bloqueo es el único que mantiene niveles de TR similares para las tres clases en los tres tamaños de entrenamiento. Respecto a la TR baja para escenas de la clase 1 y de la clase 3, con 3 y 4 objetos de entrenamiento, respectivamente, (ver Figura 47(e)), se debe a que los filtros de la clase correcta no pudieron obtener un DC lo suficientemente grande para asegurar la clasificación. Esto se puede observar detalladamente en la Figura 48(a) y (c) (escenas de la clase 1 y de la clase 3).

Tabla 5. Coeficientes de invarianza a desplazamiento de todas las técnicas, a partir de los conjuntos de entrenamiento seleccionados de cada tamaño, tomando en cuenta todas las escenas de las tres clases.

Coeficiente invarianza	PCA (M-1)	PCA (40%)	LDA	ICA (40%)	CORR SB*	CORR CB**
2 objs por clase	0.459	0.468	0.344	0.352	*0.263*	**0.186**
3 objs por clase	0.457	0.291	**0**	0.393	0.401	*0.208*
4 objs por clase	0.355	0.191	*0.067*	0.287	0.425	**0.033**

Nota. *CORR SB: Correlación sin bloqueo, **CORR CB: Correlación con bloqueo.

Tabla 6. Coeficientes de invarianza a desplazamiento de todas las técnicas, a partir de los conjuntos de entrenamiento seleccionados de cada tamaño, tomando en cuenta sólo las escenas de la clase 1 y 3.

Coeficiente invarianza	PCA (M-1)	PCA (40%)	LDA	ICA (40%)	CORR SB*	CORR CB**
2 objs por clase	0.425	0.434	0.310	0.318	*0.098*	**0.067**
3 objs por clase	0.424	0.257	**0**	0.360	0.257	*0.144*
4 objs por clase	0.322	0.157	*0.067*	0.248	0.392	**0.033**

Nota. *CORR SB: Correlación sin bloqueo, **CORR CB: Correlación con bloqueo.

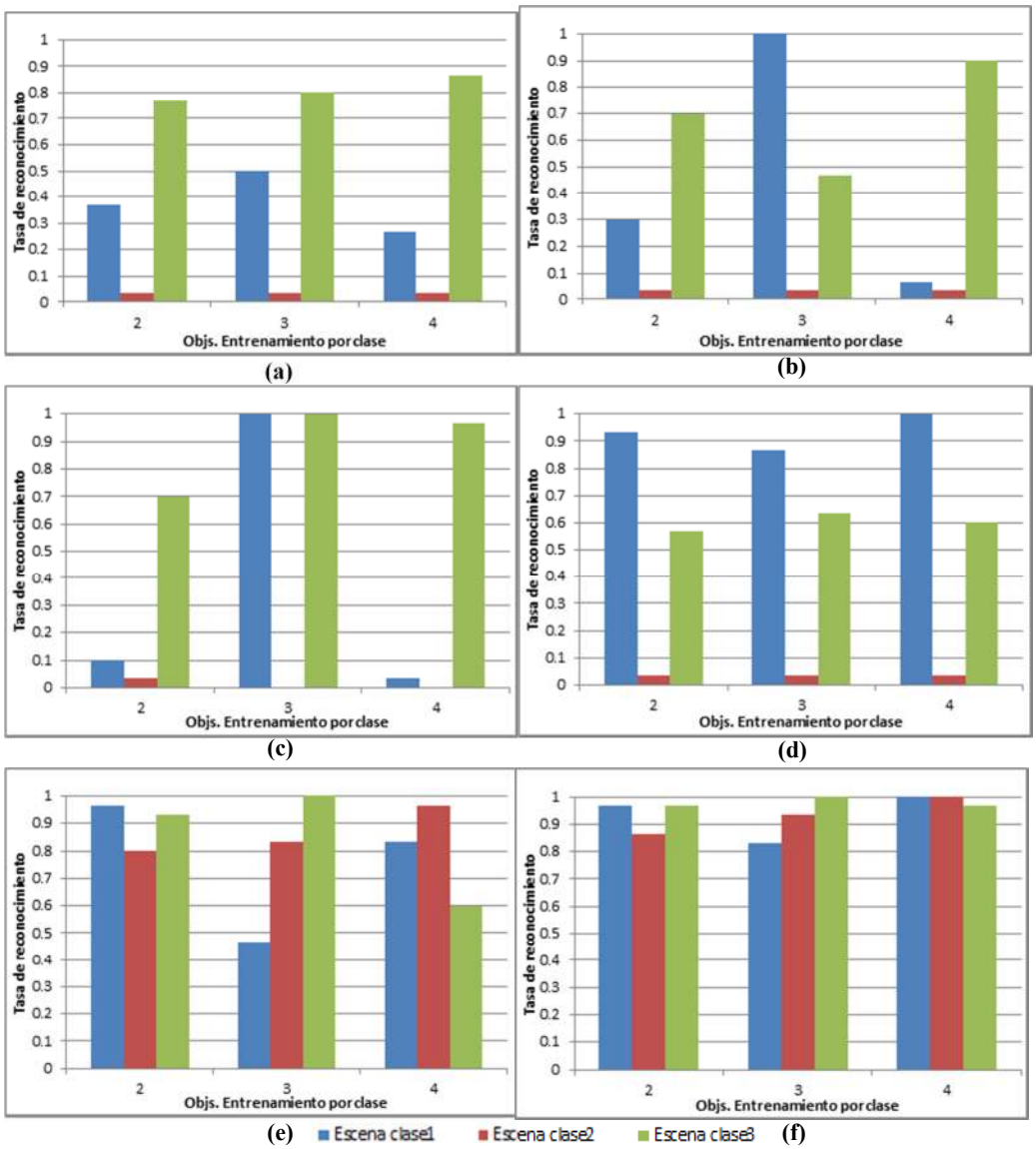


Figura 47. Tasas de reconocimiento obtenidas de 30 escenas con desplazamiento aleatorio de cada una de las 3 clases con las diferentes técnicas. (a) PCA (M-1); (b) PCA (40%); (c) LDA; (d) ICA (40%); (e) Correlación sin bloqueo; (f) Correlación con bloqueo.

Sin embargo, aplicando el bloqueo se logra incrementar la distancia entre clases, de manera que los niveles DC que obtienen los filtros son suficientes para clasificar correctamente los objetos de su clase. Los intervalos de confianza obtenidos para la DC promedio no se traslapan, mejorando el resultado que se tenía de los filtros sin bloqueo (ver Figura 48(d) y (f)) para escenas de la clase 1 y de la clase 3, con tamaños 3 y 4, respectivamente.

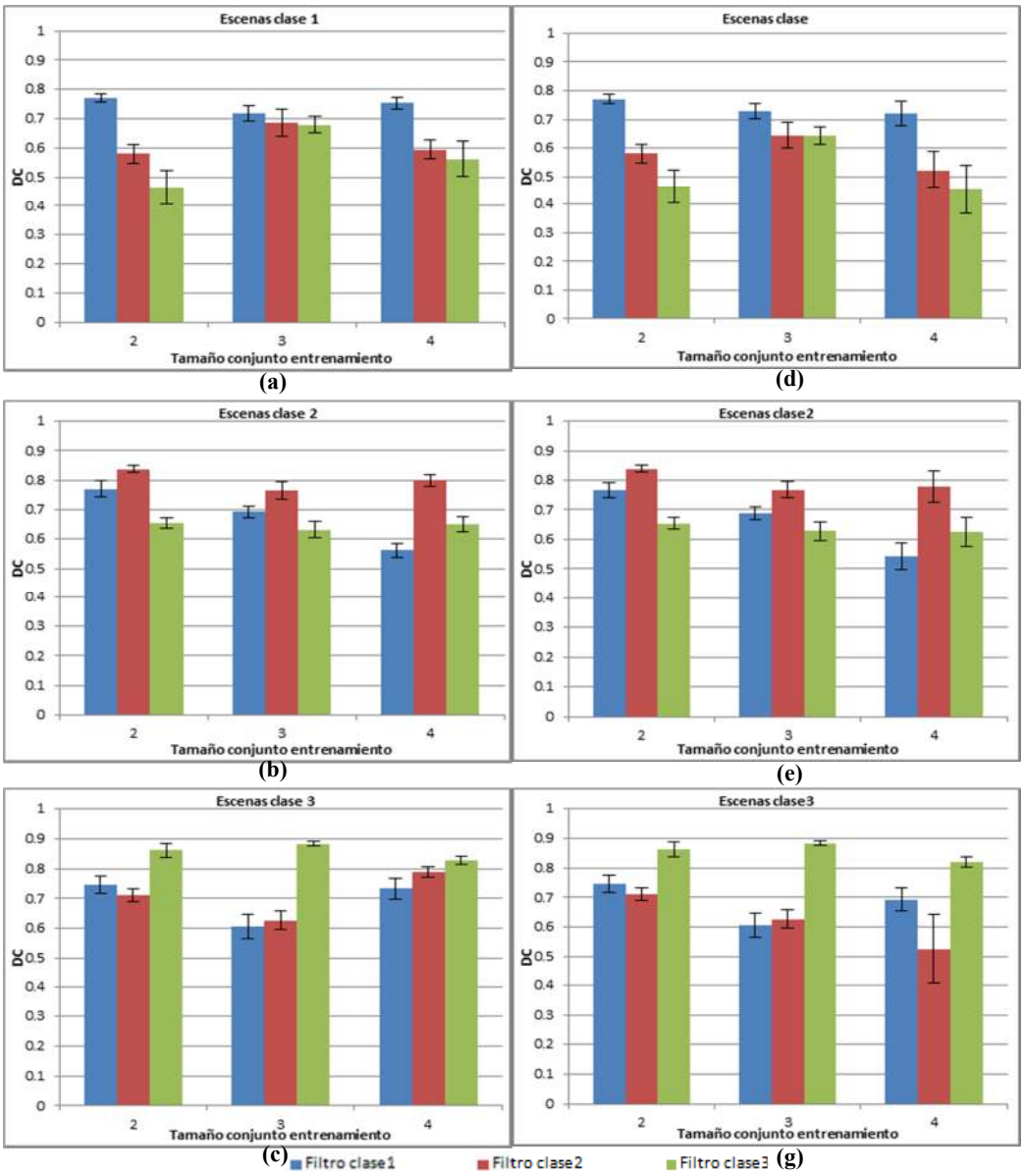


Figura 48. Desempeño al 95% de confianza, en términos del criterio DC, de cada filtro entrenado con 30 escenas sintéticas de las 3 clases, variando la ubicación del objeto de prueba en la escena de entrada. (a)-(c) Desempeño de los filtros sin bloqueo. (d)-(f) Desempeño de los filtros aplicando máscaras de bloqueo.

Como se mostró en el experimento 1, también se incrementa la separación entre las clases resultando en un incremento monótono al comparar la distancia entre el criterio DC promedio del filtro de la clase correcta y el DC promedio máximo de las otras clases. La Tabla 7 y la Tabla 8 muestran dichas distancias.

Tabla 7. Diferencia de niveles DC promedio de 30 desplazamientos de los rostros en escenas de prueba sintéticas (correlación sin bloqueo).

Tamaño entrenamiento	Escena clase 1	Escena clase 2	Escena clase 3
2 objs. por clase	0.191	0.067	0.115
3 objs. por clase	0.032	0.072	0.256
4 objs. por clase	0.158	0.149	0.0388

Tabla 8. Diferencia de niveles DC promedio de 30 desplazamientos de los rostros en escenas de prueba sintéticas (máximo desempeño de correlación con bloqueo).

Tamaño entrenamiento	Escena clase 1	Escena clase 2	Escena clase 3
2 objs. por clase	0.191	0.071	0.116
3 objs. por clase	0.085	0.080	0.256
4 objs. por clase	0.197	0.152	0.125

6.3. Resumen

En este capítulo se mostraron dos experimentos realizados para evaluar y comparar el desempeño de las técnicas tradicionales de reconocimiento de rostros, basadas en métodos estadísticos, presentadas en el capítulo 3, así como de la técnica propuesta basada en filtros de correlación, presentada en el capítulo 5. En el primer experimento se observa la desventaja de las técnicas tradicionales al depender del resultado de la segmentación automática que influye negativamente en su desempeño al aplicarlas en escenas capturadas en condiciones no controladas. Asimismo, se observa que la técnica propuesta puede resolver el problema de reconocimiento tanto en escenas sintéticas como en escenas reales. Además, el segundo experimento muestra la estabilidad en el desempeño de las técnicas respecto al desplazamiento que presentan los rostros en la escena de prueba, siendo la de Fisherfaces y la técnica de correlación propuesta aplicando el bloqueo de frecuencias, las más estables.

Conclusiones

En este trabajo se abordó el problema de reconocimiento de rostros en escenas del mundo real, las cuales están capturadas en condiciones no controladas; no se sabe la ubicación del rostro en la escena y existen estructuras complejas en el fondo. Se aplicó un enfoque de reconocimiento de patrones, mediante el diseño de filtros lineales de correlación, que ha demostrado eficiencia en la detección y reconocimiento de objetos bajo el modelo disjunto de imagen que caracteriza las escenas del mundo real. Asimismo se comparó el desempeño de técnicas tradicionales de reconocimiento de rostros que utilizan métodos estadísticos, adaptados al problema, mediante un detector de rostros ampliamente conocido: el de Viola y Jones. Los resultados muestran la eficiencia de los filtros de correlación para resolver el problema propuesto y resalta la relevancia del trabajo, pues no se encuentran muchos trabajos similares actualmente, con este enfoque en escenas reales.

La evaluación de las técnicas estadísticas de reconocimiento de rostros Eigenfaces, Fisherfaces y la Arquitectura I con ICA y de la técnica propuesta en este trabajo con filtros de correlación, destaca la desventaja de las primeras al ser aplicadas a escenas reales, bajo condiciones no controladas, al depender del desempeño de la etapa previa de segmentación. De esto, se vio la importancia del posprocesamiento que debe aplicarse al resultado de la detección. Los resultados arrojados del estudio comparativo, considerando una segmentación exitosa para las técnicas estadísticas, indican que el desempeño de la técnica de reconocimiento por correlación con filtros SDF blanqueados compite con las técnicas estadísticas. La técnica de mejoramiento de bloqueo de frecuencias con máscaras binarias logra maximizar el desempeño de la técnica propuesta de manera que obtiene los mejores resultados, junto con la técnica de la Arquitectura I con ICA, la cual se considera una del estado del arte en el área.

Por otro lado, la técnica de correlación propuesta de filtros SDF blanqueados se considera invariante a desplazamiento, así como también al aplicar la técnica de bloqueo de frecuencias propuesta. La técnica de Fisherfaces también reveló ser tolerante a desplazamientos del rostro en la escena de prueba. Por el contrario, el resto de las técnicas estadísticas no demuestran tanta invarianza debido, en parte al desempeño del detector

utilizado en la etapa previa y la calidad de la imagen segmentada con que alimenta a estas técnicas.

Ahora bien, aunque los resultados obtenidos con la técnica propuesta sean satisfactorios, se concluye que la formación del conjunto de entrenamiento del sistema es una etapa crítica para el desempeño de los filtros diseñados. Dado que la forma de los objetos es muy importante para éstos, la segmentación de las imágenes de entrenamiento puede introducir un poco de ruido e influir negativamente en la representatividad de los objetos de cada clase. Además, la base de datos utilizada presenta muchos factores de variación simultáneos que incrementan la complejidad del problema de reconocimiento; los rostros de cada clase presentan grandes diferencias entre ellas.

Por último, estos resultados son difíciles de comparar directamente con los que se encuentran en la literatura, sobre todo aquellos de los métodos estadísticos, pues el estudio realizado en esta tesis contempla conjuntos de entrenamiento y prueba pequeños. Además, la gran mayoría de las publicaciones abordan el problema con los rostros ya segmentados.

Trabajo a futuro

A partir del trabajo realizado y resultados obtenidos, se presentan a continuación algunas ideas que habría que considerar en el futuro, así como oportunidades de trabajo, siguiendo la misma línea de investigación.

Dado que se considera la etapa de selección para la formación del conjunto de entrenamiento como crítica, se recomendaría utilizar otro modelo para las imágenes en el enfoque de correlación. Por ejemplo, tomar toda la forma original de los rostros sin realizar una segmentación de los mismos para entrenar. Anteriormente se ha trabajado con este modelo en la detección de objetos y podría realizarse un estudio aplicándolo al reconocimiento de rostros.

Para la técnica de bloqueo de frecuencias, se recomendaría su implementación mediante un algoritmo de optimización, por ejemplo con algoritmos genéticos, ya que en el presente trabajo se abordó de manera iterativa y probablemente se pueda reducir la cantidad de cálculos y consecuentemente, el tiempo de ejecución del mismo.

Dado que en la literatura se han presentado resultados en el área considerando conjuntos de entrenamiento más grandes, se recomendaría incluir más rostros de entrenamiento y de prueba. Sin embargo como es difícil encontrar bases de datos que se apliquen al problema en situaciones reales y que proporcionen suficientes imágenes por cada individuo, se recomienda construir una base de datos propia que contenga las características adecuadas (escenarios reales, con varias perspectivas representativas de cada individuo), teniendo así más control de las variaciones que presenten las imágenes.

De este último punto se podría considerar la implementación de bancos de filtros de correlación para poder diseñar varios filtros por clase, tomando en cuenta múltiples perspectivas, evitando que baje su desempeño al incluir un gran número de imágenes de entrenamiento.

Referencias bibliográficas

Abate, A. F., Nappi, M., Riccio, D., & Sabatino, G. (2007). 2D and 3D face recognition: A survey. *Pattern Recognition Letters*, *28*(14), 1885–1906. doi:10.1016/j.patrec.2006.12.018

Aguilar González, P. M. (2011). *Diseño de filtros adaptativos para el reconocimiento de objetos mediante imágenes ruidosas de referencia*. (Tesis de Doctorado). Centro de Investigación Científica y de Educación Superior de Ensenada, Baja California.

AT&T Laboratories Cambridge, Cambridge University Computer Laboratory. (2002). The AT&T database of faces. http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

Bartlett, M. S. (2001). *Face Image Analysis by Unsupervised Learning*. Kluwer.

Bartlett, M. S. (2002). Face recognition by independent component analysis. http://mplab.ucsd.edu/~marni/code.html

Bartlett, M. S., Lades, H. M., & Sejnowski, T. J. (1998). Independent Component Representations for Face Recognition. *Proceedings of the SPIE Symposium on Electronic Imaging: Science and Technology; Conference on Human Vision and Electronic Imaging III* (pp. 528–539). San Jose, CA.

Bartlett, M. S., Movellan, J. R., & Sejnowski, T. J. (2002). Face Recognition by Independent Component Analysis. *IEEE transactions on neural networks*, *13*(6), 1450–1464.

Belhumeur, P. N., Hespanha, J. P., & Kriegman, D. J. (1997). Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, *19*(7), 711–720.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, *7*(6), 1129–59.

Bovik, A. C. (2005). *Handbook of Image and Video Processing* (2nd ed.). New York: Academic Press.

Brunelli, R. & Poggio, T. (1993). Face recognition: features versus templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 15*(10), 1042-1052.

Casasent, D. (1984). Unified synthetic discriminant function computational formulation. *Applied optics*, *23*(10), 1620.

Chen, T. & Liu, X. (2001). Face Expression database. Advanced Multimedia Processing Lab, Electrical and Computer Engineering Department, Carnegie Mellon University. http://chenlab.ece.cornell.edu

Comon, P. (1994). Independent component analysis-A newconcept? *Signal Processing, 36*(3), 287-314.

Cox, I. J., Ghosn, J., & Yianilos, P. N. (1996). Feature-based face recognition using mixture-distance. *Proceedings of IEEE conference on Computer Vision and Pattern Recognition*, 209-216.

Delac, K. (2007). Principal Component Analysis on the FERET database. http://www.face-rec.org/source-codes/

Freund,Y. & Schapire, R.E. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. In Pau Vitányi (Ed.), *Lecture Notes in Computer Science: Vol. 904. Computational Learning Theory, Eurocolt 95* (pp. 23–37). Springer.

Gonzalez, R. C., & Woods, R. E. (2002). *Digital Image Processing* (2nd ed.). New Jersey: Prentice Hall.

Gonzalez, R. C., & Woods, E. R. (2008). *Digital Image Processing* (3rd ed.). New Jersey: Pearson Education.

González-Fraga, J. A., Kober, V., & Álvarez-Borrego, J. (2006). Adaptive synthetic discriminant function filters for pattern recognition. *Optical Engineering*, *45*(5), 057005–1–057005–10. doi:10.1117/1.2205232

Hjelmas, E., & Low, B. K. (2001). Face Detection : A Survey. *Computer Vision and Image Understanding*, *83*(3), 236–274. doi:10.1006/cviu.2001.0921

Horner, J. L. & Gianino, P. D. (1984). Phase-only matched filtering. *Appl. Opt*. *23*(6), 812-816.

Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis*. NewYork: Wiley.

Jafri, R., & Arabnia, H. R. (2009). A Survey of Face Recognition Techniques. *Journal of Information Processing Systems*, *5*(2), 41–68.

Javidi, B. (Ed.). (2002). *Image Recognition and Classification: Algorithms, Systems, and Applications*. New York: Marcel Dekker.

Javidi, B., & Wang, J. (1994). Design of filters to detect a noisy target in nonoverlapping background noise. *Journal of the Optical Society of America A*, *11*(10), 2604–2612.

Kanade, T. (1973). *Picture Processing System by Computer Complex and Recognition of Human Faces*. (Tesis de doctorado) Kyoto University, Japan.

Kirby, M. & Sirovich, L. (1990). Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 12*(1), 103-108, 0162-8828.

Kober, V. I., & Ovseyevich, I. A. (2000). Phase-Only Filter with Improved Filter Efficiency and Correlation Discrimination. *Pattern Recognition and Image Analysis*, *10*(4), 514–519.

Kroon, D. J. (2010). Viola & Jones object detection. http://www.mathworks.com/matlabcentral/fileexchange/29437-viola-jones-object-detection

Kumar, B. V. K. V. (1992). Tutorial survey of composite filter designs for optical correlators. *Applied Optics*, *31*(23), 4773–4801.

Kumar, B. V. K. V., Mahalanobis, A., & Juday, R. (2005). *Correlation Pattern Recognition*. New York: Cambridge University Press.

Kumar, B. V. K. V., Savvides, M., Venkataramani, K., & Chunyan, X. (2002). Spatial Frequency Domain Image Processing for Biometric Recognition. *Proceedings 2002 International Conference on Image Processing, Vol. 1* (p. I–53–I–56).

Kumar, B. V. K. V., Savvides, M., & Xie, C. (2006). Correlation Pattern Recognition for Face Recognition. *Proceedings of the IEEE*, *94*(11), 1963–1976.

Li, S. Z., & Jain, A. K. (Eds.). (2005). *Handbook of Face Recognition*. New York: Springer. doi:10.1007/b138828

Lienhart, R., & Maydt, J. (2002). An Extended Set of Haar-like Features for Rapid Object Detection. *Proceedings in 2002 International Conference onImage Processing.* (p. I–900– I–903). Santa Clara, CA: IEEE.

Lin, S. H., Kung, S. Y. & Lin, L. J. (1997). Face recognition/detection by probabilistic decision-based neural network. *IEEE Transactions on Neural Networks, 8*(1)*,* 114-132.

Mahalanobis, A., Kumar, B. V. K. V., & Casasent, D. (1987). Minimum average correlation energy filters. *Applied Optics, 26*(17), 3633–3640.

Phillips, P. J., Moon, H., Rizvi, S. A., & Rauss, P. J. (2000). The FERET Evaluation Methodology for Face-Recognition Algorithms. *IEEE transactions on pattern analysis and machine intelligence*, *22*(10), 1090–1104.

Pratt, W. K. (2001). *Digital Image Processing: PIKS Inside* (3rd ed.). New York: Wiley.

Ramos Michel, E. M. (2008). *Reconocimiento Confiable de Objetos Degradados*. (Tesis de Doctorado). Centro de Investigación Científica y de Educación Superior de Ensenada, Baja California. Ensenada, México.

Ramos-Michel, E. M., & Kober, V. (2008). Adaptive composite filters for pattern recognition in linearly degraded and noisy scenes. *Optical Engineering*, *47*(4), 047204. doi:10.1117/1.2911020

Savvides, M., Kumar, B. V. K., & Pradeep, K. (2002). Face Verification using Correlation Filters. *Proceedings of the Third IEEE Automatic Identification Advanced Technologies* (pp. 56–61). Tarrytown, New York.

Sirovich, L. & Kirby, M. (1987). Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A, 4*(3), 519-524.

Tieu, K. & Viola, P. (2000). Boosting image retrieval. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1,* 228-235. doi:10.1109/CVPR.2000.855824

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

VanderLugt, A. B. (1964). Signal detection by complex filtering. *IEEE Transactions on Information Theory, 10*(6), 139-145.

Viola, P., & Jones, M. J. (2001). Robust Real-time Object Detection. *Proceedings of IEEE Workshop on Statistical and Computational Theories of Vision,* 1–25.

Viola, P., & Jones, M. J. (2004). Robust Real-Time Face Detection. *International Journal of Computer Vision*, *57*(2), 137–154.

Wagner, P. (2012). Face recognition with OpenCV2.http://www.face-rec.org/source-codes/

Wechsler, H. (2007). *Reliable Face Recognition Methods, System Design, Implementation and Evaluation*. New York: Springer.

Webber, M. Faces 1999 (Front). California Institute of Technology http://www.vision.caltech.edu/html-files/archive.html

Wiskott, L., Fellous, J.-M., Krüger, N., & von der Malsburg, C. (1997). Face Recognition by Elastic Bunch Graph Matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19* (7), 775-779

Witten, I. H., Frank, E. & Hall, M. (2011). *Data Mining, Practical Machine Learning Tools and Techniques* (3rd ed.). New York: Morgan Kaufman.

Yaroslavsky, L. P. (1993). The theory of optimal methods for localization of objects in pictures. In E. Wolf (Ed.), *Progress In Optics XXXII* (pp. 145–201). Amsterdam: Elsevier.

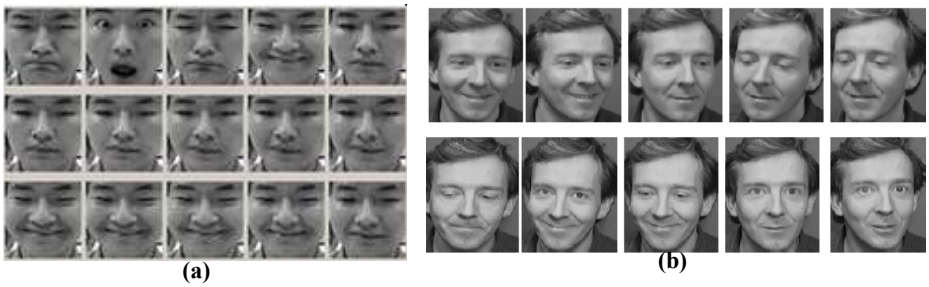
Zhao, W., Chellappa, R., Phillips, P. J., & Rosenfeld, A. (2003). Face Recognition : A Literature Survey. *ACM Computing Suveys*, *35*(4), 399–458.

.

Apéndice

Base de datos de rostros.

Dado que el objetivo de la investigación era de aplicar técnicas de reconocimiento de rostros en imágenes reales, es decir, en situaciones no controladas conteniendo fondo complejo, muchas de las bases de datos utilizadas para el reconocimiento de rostros que se encuentran en internet no eran adecuadas. Algunas contienen los rostros completamente segmentados, de manera que sólo se aprecia la expresión del individuo, como se muestra en la Figura 49(a); otras, aunque muestran el contorno del rostro completo, están capturadas en ambientes controlados, por ejemplo en un estudio fotográfico o por lo menos contienen un fondo homogéneo, como se aprecia en la Figura 49(b). Sólo se encontró una base de datos que contenía las características consideradas adecuadas para usarse como escenas completamente reales.



(a) (b)

Figura 49. Bases de datos de rostros segmentados para reconocimiento en ambiente controlado. (a) Ejemplo de rostros de la base de datos Face Expressions (recuperado de http://chenlab.ece.cornell.edu/projects/FaceAuthentication/#Download). (b) Ejemplo de la base de datos de rostros AT&T, antes base de datos ORL.

Las imágenes de rostro utilizadas en los experimentos de este proyecto corresponden a la base de datos Faces 1999 (Front), que pertenece al banco de imágenes Caltech-101 (imágenes de objetos de 101 categorías), del grupo de Visión Computacional de Caltech. Esta base de datos de rostros, recopilada por Markus Weber, contiene 450 imágenes de aproximadamente 28 individuos y 3 caricaturas de rostros; por lo menos 18 individuos cuentan con 20 imágenes. Los otros 10 individuos tienen diferente número de imágenes. Cada imagen tiene un tamaño de 896×592 pixeles y están en formato JPG. Todas

las imágenes presentan distintos tipos de iluminación, expresiones, posturas, escalas y fondos. En la Figura 50 se muestra un ejemplo de las imágenes de dicha base de datos.

Para tener la misma cantidad de imágenes por individuo, se consideró un subconjunto de 360 imágenes pertenecientes a 18 individuos (20 imágenes de cada uno), al cual se refiere como base de datos Caltech en la tesis.

Figura 50. Ejemplo de imágenes de la base de datos de rostros Face 1999 de Caltech.