UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIA E INGENIERÍA DE LA COMPUTACIÓN

RECONOCIMIENTO DE ROSTROS HUMANOS CON REDES NEURONALES Y TRANSFORMACIONES DE TRANSVECCIÓN

TESIS

QUE PARA OPTAR POR EL GRADO DE:

MAESTRO EN CIENCIAS (COMPUTACIÓN)

PRESENTA:

ALEJANDRO JOAQUÍN IBARRA GALLARDO

TUTOR
DR. ERNST KUSSUL
CENTRO DE CIENCIAS APLICADAS Y DESARROLLO TECNOLÓGICO

MÉXICO, D.F. JUNIO 2015.

Resumen

La tarea de reconocimiento de rostros humanos de manera automatizada es un área de investigación de gran relevancia en diversos ámbitos académicos y de desarrollo tecnológico, puesto que constituye un punto clave en diversos sistemas de interacción humano-computadora o humano, vigilancia o verificación de identidad.

No obstante su gran importancia, en la actualidad no existe un sistema capaz de lograr el grado de precisión del cerebro humano debido a una serie de dificultades relacionadas con la variabilidad de las imágenes capturadas en entornos no controlados.

La presente tesis tiene por objetivo realizar un estudio detallado del estado del arte en esta área de investigación, revisando los diferentes desafíos técnicos existentes, los enfoques propuestos para superarlos (con especial énfasis en el uso de redes neuronales), y finalmente, los recursos disponibles para evaluar el desempeño de los diferentes algoritmos.

Así mismo, esta tesis propone una estrategia para mejorar el rendimiento de un sistema reconocedor basado en redes neuronales en situaciones donde se presenta un cambio de pose del individuo. Dicho método consiste en alimentar el sistema con distorsiones adicionales de transvección.

Abstract

The task of the automatic facial recognition is a research field of great significance in several academic and technology areas, since it constitutes a key point in various systems of human-computer interaction, surveillance and authentication tasks.

In spite of its great relevance, currently there isn't a system able to achieve the degree of precision of the human recognition system due to some difficulties related to the variability of the captured images in uncontrolled environments.

This thesis is aiming to make a detailed study of the state of the art in this research field; reviewing several technical challenges currently existing, proposed approaches to overcome them (with special emphasis on the use of neural networks), and finally, the resources available to help us to evaluate the performance of different algorithms.

This thesis also proposes a strategy to improve the performance of a face recognition system based on neural networks in conditions where there is a change of position of the subject. This method consists in feeding the system with additional distortions of shear mapping.

Agradecimientos

La presente tesis representa la culminación de una etapa de mi vida en la que ha habido momentos de trabajo, presión y esfuerzo; pero también momentos de aventura, curiosidad y de ávido interés por descubrir cosas nuevas cada día. Momentos malos y momentos buenos, pero sobre todo, historias dignas de contar, siempre.

En esta pequeña pero importante sección de mi tesis deseo mostrar mi más sincero agradecimiento a todas las personas que hicieron todo esto posible.

En primer lugar deseo agradecer a mis padres, por darme siempre su amor, cariño y comprensión, por brindarme siempre su apoyo incondicional y por inculcarme el amor hacia el estudio.

También deseo agradecer a mis hermanas Paula y Josie, así como el resto de mi familia por estar a mi lado a pesar de la distancia y por encontrar siempre una manera de distraerme y hacerme sonreír.

Agradezco también de modo especial a mis asesores de tesis, el Dr. Ernst Kussul y la Dra. Tetyana Baydyk; por su enorme paciencia y ayuda para completar este trabajo, por sus buenos consejos y su confianza, y sobre todo, por su amistad.

A todo el resto del comité de revisión que conforma mi jurado, los doctores Francisco Javier García, Ernesto Bribiesca y Lucía Medina; por dedicar el tiempo necesario para revisar mi trabajo, darme sus comentarios y aportar sus conocimientos.

Gracias a cada uno de mis profesores, amigos y compañeros de generación, por sus conocimientos, apoyo y amistad.

A la Universidad Nacional Autónoma de México y su programa de Posgrado de Ciencia e Ingeniería en Computación, así como a todo el personal que ahí labora y hace posible toda esa gran área de estudios e investigación, muy especialmente a las secretarias Lourdes, Diana, Amalia y Cecilia y a los directores Dr. Fernando Arámbula y Dr. Jorge Luis Ortega por su confianza, solidaridad y apoyo.

Finalmente, al Consejo Nacional de Ciencia y Tecnología (CONACyT) y los proyectos PAPIIT IN102014 e IT102814 por proporcionar los fondos para realizar mis estudios, así como las presentaciones llevadas a cabo en el extranjero.

Índice general

Capítulo 1. Introducción **1**
1.1 Introducción y motivación de estudio 1
1.2 Objetivos de la tesis. 2
1.3 Justificación del estudio 3
1.4 Planteamiento del problema y solución propuesta 4
1.5 Alcances, limitaciones y dificultades esperadas 6
1.6 Estructura de la tesis 7
Capítulo 2. Antecedentes **8**
2.1 Conceptos generales 9
2.1.1 Visión artificial y áreas afines de estudio 9
2.1.2 Biometría y aplicaciones 11
2.1.3 Detección y reconocimiento de rostros 11
2.1.4 Madurez de las tecnologías 13
2.2 Fundamentos biológicos de la percepción facial 14
2.3 Desafíos técnicos 19
2.3.1 Resolución 21
2.3.2 Pose 21
2.3.3 Iluminación 22
2.3.4 Expresiones faciales 23
2.3.5 Oclusiones 24
2.3.6 Componentes variables del rostro 24
2.3.7 Edad 25
2.4 Técnicas de reconocimiento generales existentes 26
2.4.1 Clasificación 26
2.4.2 Métodos holísticos 26
2.4.2.1 Eigenfaces 26
2.4.2.2 Fisherfaces 28
2.4.2.3 Evolutionary Pursuit 30
2.4.2.4 Máquinas de Soporte Vectorial 32
2.4.3 Métodos estructurales 30
2.4.2.1 EBGM 33
2.4.2.2 LBP 34
2.4.2.3 LEM 36
2.4.2.4 Modelos de Apariencia Activa 37

Capítulo 3. Bases de datos para experimentación 40
3.1 Bases de Datos Estáticas Bidimensionales. 41
3.1.1 Base de datos ORL 41
3.1.2 Base de datos Yale 42
3.1.3 Base de datos Yale Extendida 43
3.1.4 Base de datos CAS-PEAL 44
3.1.5 Base de datos FERET 45
3.1.6 Base de datos FEI 47
3.1.7 Base de datos GTAV 48
3.1.8 Base de datos FRAV2D 50
3.1.9 Base de datos Achermann 52
3.2 Análisis comparativo 53
Capítulo 4. Redes Neuronales 54
4.1 Introducción 54
4.2 El perceptrón de Rosenblatt 57
4.3 El perceptrón multicapa (MLP) 58
4.4 Red Neuronal Convolucional 60
4.5 Clasificador neuronal RTC 64
4.6 Clasificador neuronal RSC 67
4.7 Clasificador neuronal LIRA 68
4.7.1 Selección de conexiones y activación de neuronas 63
4.7.2 Proceso de entrenamiento 64
4.7.3 Mejora de aprendizaje 66
Capítulo 5. Sistema de Reconocimiento de Rostros 73
5.1 Estructura general del clasificador 73
5.2 Extracción de propiedades 74
5.3 Descriptores locales aleatorios (RLD) 76
5.4 Codificación de características 80
5.5 Entrenamiento 85
5.6 Transformaciones de transvección 86
5.7 Implementación y ejecución 89
5.8 Experimentación y resultados 91
5.9 Discusión 95
Capítulo 6. Conclusiones y trabajo futuro 98
6.1 Conclusiones 98

6.2 Aportaciones 98
6.3 Trabajo futuro 99
Referencias 101

Índice de figuras

1.1. Diagrama de bloques de un sistema de reconocimiento facial representativo 1
2.1. Medidas faciales de uno de los estudios de Bertillon..... 8
2.2. Ejemplo de detección de rostros en una imagen estática 11
2.3. Etapas de un sistema de reconocimiento de rostros 12
2.4. Ubicación del giro fusiforme en el cerebro humano 15
2.5. Experimento de Young sobre el procesamiento de la información..... 16
2.6. Experimento de Sadr sobre la influencia de las diferentes características faciales 17
2.7. Ejemplos de separabilidad de clases en los problemas de detección y reconocimiento..... 19
2.8. Ejemplos del problema de separación de rostros con alto grado de similitud entre sí 20
2.9. Ejemplos de cambios de resolución y perspectiva de la base de datos SCFaceSurveillance 21
2.10. Ejemplos de variaciones de pose simples 21
2.11. Ejemplos de variaciones de pose compuestos..... 22
2.12. Ejemplos de rostros con distinta iluminación de la base de datos FEI..... 22
2.13. Ejemplos de cambios de expresión facial 23
2.14. Ejemplos de algunas oclusiones frecuentes 24
2.15. Ejemplos de algunas variaciones de apariencia frecuentes..... 24
2.16. Ejemplos de variaciones en la apariencia producidas por la edad..... 25
2.17. Ejemplos de Eigenfaces 27
2.18. Cuatro primeras Fisherfaces generdas a partir de un conjunto de 100 sujetos..... 29
2.19. Representación gráfica de cromosomas utlizandos en el algoritmo evolutivo..... 30
2.20. Cuatro representaciónes gráficas de vectores óptimos de EP 31
2.21. Separación de clases utilizando hiperplanos SVM 32
2.22. Redes adaptadas al rostro en diferentes poses..... 33
2.23. Representación gráfica del funcionamiento del algoritmo EBGM 33
2.24. Ejemplo de un operador binario básico..... 34
2.25. Descripción del rostro mediante patrones binarios locales 35
2.26. Bordes de un rostro obtenidos mediante el algoritmo LEM 36
2.27. Ejemplo de un rostro marcado con 122 puntos de manera manual..... 37
2.28. Modelos lineales de forma en un AAM independiente..... 38
2.29. Modelos lineales de apariencia en un AAM independiente..... 38
3.1. Ejemplo de un set completo de tomas de un individuo de la base de imágenes ORL 41
3.2. Ejemplo de imágenes para un individuo de la base de datos Yale..... 42
3.3. Imágenes recortadas de ejemplo pertenecientes a la base de datos Yale extendida..... 43
3.4. Ejemplo de un set completo de tomas de un individuo de la base de imágenes CAS-PEAL 45

3.5. Ejemplo de un set parcial de tomas de una participante de la base de imágenes FERET 46
3.6. Ejemplo de un set parcial de tomas de una participante de la base de imágenes FEI 47
3.7. Ejemplo de un set completo de tomas de un individuo de la base de imágenes GTAV 48
3.8. Ejemplo de un set parcial de tomas adicionales de la base de imágenes GTAV 49
3.9. Ejemplo de un set completo de tomas de un individuo de la base de imágenes FRAV 2D 51
3.10. Ejemplo de un set completo de tomas de un individuo de la base de imágenes Achermann ... 52
4.1. Clasificación de las redes neuronales en base a su tipo de aprendizaje y arquitectura 56
4.2. Diagrama simplificado del perceptrón de Rosenblatt 58
4.3. Estructura general de un Perceptrón Multicapa (MLP) 59
4.4. Operación básica de un clasificador basado en CNN 60
4.5. Lectura de la imagen a través de ventanas y transformación a vectores 61
4.6. Reducción de la dimensionalidad de la imagen original utilizando SOM y ventanas de 8x8 61
4.7. Capas y mapas de la CNN 62
4.8. Capas y mapas de convolución y sub-muestreo de la CNN 62
4.9. Estructura general de un clasificador neuronal RTC 64
4.10. Interpretación geométrica de la clasificación con RTC utilizando dos características 65
4.11. Región acotada por el clasificador RTC utilizando múltiples neuronas en la capa B 66
4.12. Zona de clasificación creada a partir del proceso de entrenamiento 66
4.13. Estructura de un clasificador neuronal LIRA binario 68
4.14. Ejemplo de un clasificador neuronal LIRA Grayscale 69
4.15. Ejemplo de distorsiones para mejorar el rendimiento de LIRA 72
5.1. Estructura general del clasificador PCNC 74
5.2. Puntos de interés seleccionados mediante el extractor de características 75
5.3. Estructura del sistema de reconocimiento de propósito general 77
5.4. Esquema detallado RLD 77
5.5. Esquema de permutaciones horizontales y verticales 81
5.6. Ejemplos de permutaciones horizontales y verticales realizadas 83
5.7. Resultados de las permutaciones X y Y 84
5.8. Programa utilizado para crear imágenes de prueba 88
5.9. Diferentes factores de inclinación de una imagen de muestra 89
5.10. Primera etapa de la ejecución del programa: generación de máscaras 89
5.11. Segunda etapa de la ejecución del programa: codificación de características 90
5.12. Tercera etapa de la ejecución del programa: proceso de entrenamiento 90
5.13. Etapa final: Reconocimiento de individuos 91

Índice de tablas

3.1. Distribución de las imágenes de la base de datos CAS-PEAL 45

3.2. Niveles de resolución del mallado proporcionado por el escáner 51

3.3. Tabla comparativa de las bases de datos consideradas para la experimentación 54

4.1. Características y diferencias destacables entre los enfoques de la IA 56

5.1. Definición detallada de las pruebas realizadas con la base de imágenes FRAV2D 92

5.2. Comparaciones de resultados de SVM y PCNC 93

5.3. Resultados de los experimentos 6 y 9 considerando transformaciones skewing 94

5.4. Resultados de los 15 experimentos en la base de datos FRAV2D 94

5.5. Comparación de PCNC con otros algoritmos de uso general usando ORL 95

5.6. Comapración de PCNC con otros algoritmos utilizando la base de datos Achermann 96

Capítulo 1

Introducción.

1.1 Introducción y motivación de estudio

El reconocimiento artificial de rostros humanos es una de las ramas de investigación y desarrollo tecnológico que más ha cobrado auge en los últimos años dentro del campo de estudio de las ciencias computacionales denominado visión por computadora, así como también en campos afines e interdisciplinarios como la inteligencia artificial o el reconocimiento de patrones.

Existen diversos motivos para esta creciente tendencia, pero destacan dos de manera específica: el incremento de aplicaciones del ámbito académico, gubernamental y recreativo que se han hecho evidentes recientemente, y la disponibilidad en las últimas décadas de tecnologías de costo bajo y moderado para su desarrollo [Zhao03].

Si bien es cierto que en la actualidad existen gran cantidad de sistemas públicos y comerciales que han alcanzado cierto grado de madurez y éxito, es necesario señalar que hasta la fecha solo se ha logrado la fiabilidad completa bajo ciertas condiciones que rara vez se dan en el mundo real, lo cual es el objetivo al que apuntan la mayoría de las aplicaciones mencionadas.

Por ejemplo, el reconocimiento de individuos en imágenes adquiridas mediante videos de vigilancia, los cuales generalmente presentan ángulos de captura pronunciados, baja resolución, y en muchas ocasiones también una vista parcial del sujeto a reconocer; permanece como una cuestión abierta.

En pocas palabras, la creación de un sistema automatizado de reconocimiento de rostros que iguale de manera razonable al sistema de percepción humano está lejos de ser una realidad.

Es debido a lo anteriormente expuesto que la presente tesis tiene por objetivo el estudio y evaluación de rendimiento de los métodos y herramientas existentes en la actualidad para esta tarea, así como la presentación de una propuesta de mejora en los problemas que permanecen sin resolver.

1.2 Objetivos de la tesis.

4.7.1 Objetivos generales

El objetivo principal del presente trabajo es la realización de un estudio detallado de los algoritmos, técnicas y herramientas existentes en la actualidad para el reconocimiento de rostros humanos en diferentes condiciones de pose, iluminación y expresiones faciales, entre otras.

Así mismo, se busca evaluar y comparar de manera objetiva el rendimiento de diversos algoritmos bajo las diferentes condiciones planteadas, para posteriormente realizar una propuesta enfocada a mejorar el rendimiento en situaciones de variación de pose a partir de una red neuronal.

4.7.2 Objetivos particulares

Para alcanzar los objetivos principales, se plantean y describen los siguientes objetivos parciales o específicos:

1. Realización de un estudio bibliográfico sobre el campo del reconocimiento facial
    - Se deberán considerar tanto las técnicas clásicas o referentes en el campo, como de los avances más recientes (estado del arte) a fin de tener un panorama general de la tecnología existente.
    - El estudio debe incluir varias técnicas basadas en redes neuronales.
2. Obtener varios conjuntos de prueba con rostros humanos (bases de datos)
    - Las bases de datos consideradas deben ser públicas y disponibles para su utilización gratuita y legal en publicaciones académicas.
    - Deben considerar un número suficiente de individuos para que los resultados de la evaluación tengan significación estadística.
    - Se busca que la base de datos considere gran número de variaciones controladas para cada individuo, específicamente de pose, para incluir rotaciones suaves y pronunciadas en los 2 o más ejes coordenados.

3. Evaluación del rendimiento de los algoritmos investigados sobre las bases de datos adquiridas.
    - Se debe obtener el código fuente de los algoritmos considerados, o en su defecto, recolectar los resultados reportados en sus correspondientes artículos.
    - Las imágenes de la(s) base(s) de datos a considerar se deberán dividir en conjuntos de entrenamiento y de pruebas. Este último debe considerar diferentes variaciones, especialmente de pose.
    - La evaluación deberá considerar tanto el desempeño general (sobre el total de imágenes de prueba) como el desempeño específico, es decir sobre cada imagen del conjunto de pruebas, a fin de obtener una evaluación detallada bajo las diferentes condiciones presentadas.
4. Proponer una técnica o metodología que ayude a mejorar el desempeño frente a variaciones de pose del individuo.
    - La técnica o metodología propuesta se hará a partir de una red neuronal preexistente.
    - Se pretende, aunque no es necesario, que está técnica pueda aplicarse a sistemas de reconocimiento diferentes al de rostros humanos.
5. Implementar dicha propuesta y analizar los resultados obtenidos en comparación con los algoritmos previos considerados.

1.3 Justificación del estudio

La elección del reconocimiento de rostros humanos como el tema de la presente tesis se realizó en consideración a los siguientes aspectos:

1. Es un tema de investigación que conlleva numerosas aplicaciones prácticas y que se encuentra en creciente desarrollo. Entre sus muchas aplicaciones podemos citar:
    - Seguridad de la información y control de acceso mediante biometría.
    - Reforzamiento de la ley en tareas de vigilancia en aeropuertos, oficinas gubernamentales, bancos, etc.
    - Interacción humano-máquina en tareas de robótica.
    - Reforzamiento de control parental de contenidos.
    - Identificación automatizada de sujetos en documentos oficiales como pasaportes, licencias de conducir, credenciales de elector, etc.
    - Programas de entrenamiento, realidad virtual, realidad aumentada y

videojuegos.

- Etiquetado automático de fotografías en álbumes públicos y personales.

2. Es un tema de investigación que permanece abierto a pesar de la gran cantidad de avances que se han realizado en la materia, pues aún se encuentra alejado de una comparación razonable con el sistema que poseemos los seres humanos. Tomando esta referencia como punto de comparación, aún queda pendiente cómo mejorar los algoritmos para hacerlos robustos en los siguientes aspectos:
   - Variaciones de pose pronunciados (translación y rotaciones en los 3 ejes coordenados).
   - Variaciones de iluminación.
   - Variaciones severas de escala y resolución.
   - Variaciones de aspecto del individuo (cambios de peinado, color de cabello, presencia de gafas o anteojos, sombreros o gorras, maquillaje, etc.).
   - Variaciones de tiempo/edad.
   - Variaciones de expresión facial.
   - Variaciones de medios de representación (caricaturas o dibujos semi-realistas).
3. Es un tema muy interesante que requiere la aplicación de conocimientos de diversas disciplinas como:
   - Ciencias computacionales
     - ✓ Procesamiento de imágenes
     - ✓ Inteligencia Artificial
     - ✓ Visión por computadora
   - Matemáticas
     - ✓ Estadística
     - ✓ Reconocimiento de patrones
   - Neurología
   - Psicología

Estos puntos se tratarán de manera más profunda en el siguiente capítulo.

1.4 Planteamiento del problema y solución propuesta

Cualquier sistema de reconocimiento de rostros (incluyendo probablemente el humano) puede ser dividido conceptual y funcionalmente en tres grandes bloques [Zhao03]. En la primera etapa se lleva a cabo la detección del rostro en la imagen, la cual consiste en determinar si en

ella se encuentran una o más caras humanas. Una vez que se ha confirmado la presencia de una o varias de ellas en la imagen, se procede a obtener su posición o posiciones para llevar a cabo un análisis individual más detallado. A dicho análisis se le conoce como extracción de características y su objetivo es recolectar los aspectos más descriptivos de la cara para su posterior identificación. Este proceso se encuentra en la segunda etapa. Finalmente, la última etapa consiste en hacer uso de las características extraídas para clasificar al individuo en base a la comparación con sujetos conocidos. Esta etapa se conoce como identificación o reconocimiento. En la figura 1.1 se muestra un diagrama que condensa estas ideas.

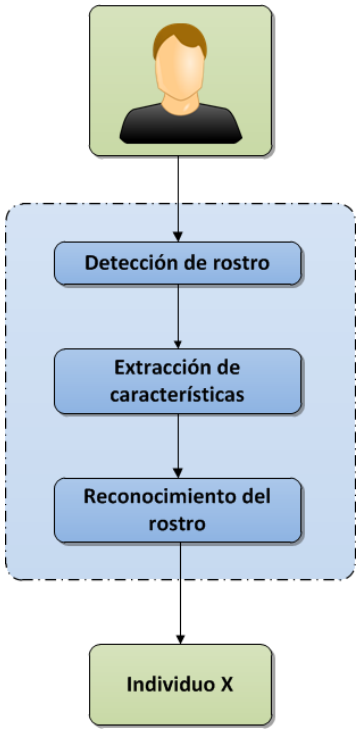


Figura 1.1. Diagrama de bloques representativo de un sistema reconocedor de rostros humanos genérico. El sistema debe indicar como salida la clase a la que pertenece el individuo a partir de una imagen de entrada que contenga un rostro humano.

Cabe señalar que el proceso de detección de rostros ha sido resuelto con un éxito bastante aceptable, por lo que la mayor parte de la investigación actual se centra en las dos etapas restantes. Con esta finalidad se han creado diversas bases de datos que contienen exclusivamente los rostros a identificar considerando específicamente las variaciones que quedan pendientes por resolver, principalmente de pose e iluminación.

El presente estudio tiene como propósito mejorar el reconocimiento de imágenes con variaciones de pose pronunciadas. El problema que esto encierra, es la gran pérdida de información que se producto del hecho de que el rostro es una entidad tridimensional plasmada en un medio bidimensional, por lo que una variación severa de pose del individuo representa un

cambio muy grande de información entre las imágenes de entrenamiento y las imágenes a reconocer.

Para hacer frente a este problema, se propone la utilización de una red neuronal conocida como PCNC (Permutation Coding Neural Classifier, Clasificador Neuronal de Códigos Permutados) entrenada con imágenes transformadas para compensar dicha pérdida de información. En el capítulo 4, este planteamiento será tratado con profundidad.

1.5 Alcances, limitaciones y dificultades esperadas

El primer problema que se debe resolver es la búsqueda y selección de una (o más) base de datos adecuada para la realización de experimentos y posterior evaluación de resultados. En este aspecto se requieren considerar varios puntos, entre los que cabe resaltar los siguientes:

- Tipo. Existen varios tipos de bases de datos en la red. Estas se clasifican en función del objetivo que persiguen. De esta manera existen las bidimensionales (imágenes comunes), tridimensionales (incluyen información de profundidad) y multimodales (pueden incluir, además de imágenes 2D, otro medio de identificación como imágenes 3D, audio de voz, etc.). Para la presente tesis requerimos únicamente las imágenes bidimensionales.
- Disponibilidad. De las múltiples bases de datos existentes, algunas están disponibles completamente gratuita, otras exclusivamente para el ámbito académico, algunas más solo permiten legalmente el uso para experimentación y comparación, pero no la difusión de las imágenes en medios impresos (artículos, presentaciones, etc.) y finalmente existen algunas más de pago.
- Extensión. Se requiere una base de datos con un número considerable de clases (número de individuos) y un número también extenso de variaciones para cada clase (se busca principalmente un gran número de rotaciones controladas).
- Popularidad. Por último en orden, pero no en importancia está la aceptación que la base de datos tenga entre la comunidad especializada, pues esto facilita la comparación y presentación de resultados.

La segunda dificultad que se espera es como se describió con anterioridad, la pérdida de información en la imagen al presentarse una rotación en el rostro del individuo. Para compensar esta falta de información, es necesario generar imágenes nuevas, lo cual conlleva mayor tiempo de procesamiento y consumo de recursos computacionales. Por lo que se debe aclarar que, aunque se pretende hacer un sistema lo más eficiente posible, el objetivo primordial de la tesis

es investigar la forma de mejorar en el reconocimiento de imágenes con variación de pose, y no la optimización de los algoritmos requeridos para llevarla a cabo.

1.6 Estructura de la tesis

Esta tesis está dividida en siete capítulos. El primero de ellos es la presente introducción. En el segundo capítulo se desarrollaran a profundidad los conceptos generales y se hará una revisión del estado del arte, para mostrar las técnicas que se siguen actualmente, algunos conceptos históricos, descubrimientos biológicos recientes y teorías psicológicas que pueden ser de utilidad en el futuro. Posteriormente, en el tercer capítulo se trata en profundidad una de las herramientas más necesarias para la investigación: las bases de datos de rostros humanos. En el capítulo cuarto se describen detalladamente otros algoritmos de reconocimiento basados en redes neuronales, algunos de los cuales sirven como base del algoritmo empleado. En el quinto capítulo se describe en profundidad el clasificador neuronal utilizado, la descripción de la solución propuesta, la implementación y ejecución del algoritmo, y finalmente la presentación de las pruebas y los resultados obtenidos. Finalmente, en el capítulo sexto se muestran las conclusiones finales del trabajo y se enlistan algunas propuestas de investigación futura.

Capítulo 2

Antecedentes.

El ser humano siempre ha utilizado de manera natural los rasgos de su apariencia y otros elementos únicos e individuales, tales como la voz, para reconocerse entre sí. Sin embargo, fue hasta mediados del siglo XIX cuando Alphonse Bertillon, jefe de investigación criminal del departamento de policía en París, se dio a la tarea de sistematizar la labor del reconocimiento de personas haciendo uso de la antropometría [Jain04], es decir la medición del cuerpo humano, haciendo particular énfasis en la medición de rasgos faciales (fig 2.1). Dicho sistema recibió el nombre de Bertillonage, y fue ampliamente adoptado en occidente, constituyendo los cimientos no únicamente de la criminalística forense, sino también de otras ramas de investigación posteriores, tales como la biometría.

Figura 2.1. Medidas faciales de uno de los estudios de Bertillon.

Esta breve introducción histórica nos sirve para poner bajo relieve la importancia que ha tenido, y sigue teniendo, el reconocimiento sistemático de rostros humanos. Pero el reforzamiento de la seguridad mediante la biometría es solo una de las posibles aplicaciones que este campo de estudio tiene.

En el presente capítulo se proporcionará una breve descripción de los campos de estudio que engloban o son afines al reconocimiento facial, así como algunos de los conceptos generales más relevantes en la materia, tales como reconocimiento de patrones, procesamiento de imágenes, biometría, etc.

Además se establecerá la diferencia entre la detección y el reconocimiento de rostros, se describirán algunas de las bases biológicas y psicológicas del sistema de reconocimiento humano de rostros, se detallará la problemática y los retos técnicos existentes, y finalmente se hará una revisión de los métodos y enfoques que existen actualmente para realizar esta tarea.

2.1 Conceptos generales

2.1.1 Visión artificial y áreas afines al estudio

La visión es uno de los mecanismos de percepción sensorial más importantes, no únicamente en el ser humano sino también en la gran mayoría de las especies animales, puesto que dependemos de este para obtener información sobre nuestro entorno, y de esta manera, tomar decisiones sobra la manera de interactuar con él.

En el caso del hombre se estima que aproximadamente el 70% de la información percibida por él es captada a través de su sistema visual [Rodr12]. Así pues, las personas con discapacidad visual, tienen, por lo general mayores problemas para desenvolverse satisfactoriamente dentro de su entorno en comparación con una persona que goce plenamente de sus capacidades visuales.

Establecida la importancia que tiene en el sistema de visión en los organismos biológicos, no es difícil comprender la trascendencia científica y tecnológica que tiene la posibilidad de dotar de un sistema similar a una máquina; con la finalidad de que esta sea capaz de tomar decisiones y relacionarse con su entorno de manera adecuada. El campo de las ciencias computacionales dedicado a este cometido se denomina Visión Computacional o Visión por Computador.

La Visión Computacional formalmente es definida como el campo de estudio de las ciencias computacionales que incluye métodos de adquisición, procesamiento, análisis y comprensión de imágenes, y en general datos de alta dimensionalidad provenientes del mundo real con la finalidad de producir información simbólica o numérica que sirvan para la toma de decisiones [Klet14, Shap01, Morr04, Jahn00]. Para lograr este objetivo, se sirve de modelos construidos con la ayuda de la geometría, la física, la probabilidad y la estadística [Fors02]. Además de esto, la Visión Computacional involucra fuertemente tópicos de la Inteligencia Artificial (IA) tales como el reconocimiento de patrones, el aprendizaje automático y la

psicología cognitiva. Debido a esto, la Visión por Computadora es frecuentemente clasificada como un subcampo de la IA.

La IA, por su parte, es definida como el "estudio y diseño de agentes inteligentes", donde un agente inteligente es un sistema capaz de percibir su entorno y tomar acciones que maximicen su posibilidades de éxito [Russ03]. Si tomamos en cuenta esta definición, resulta evidente que la visión artificial representa un aspecto primordial en la IA, al menos en lo referente a la interpretación de estímulos visuales.

Un campo de estudio que tiene una importancia fundamental tanto en la IA como con la visión computacional es el aprendizaje automático. El aprendizaje automático (o su equivalente en inglés, machine learning) es definido como el campo de estudio que proporciona a las computadoras la habilidad de aprender cosas para las que no ha sido explícitamente programadas [Simo13].

El aprendizaje automático se divide en dos ramas. En el aprendizaje supervisado, se le proporcionan a la máquina cierto número de entradas de ejemplo, así como las salidas deseadas, y con la ayuda de un "maestro" se persigue el objetivo de crear una regla general que mapee todas las entradas y las salidas de manera correcta. Los sistemas anti-spam, los antivirus y los sistemas de reconocimiento (rostros, caracteres, etc.) son un ejemplo de este tipo de aprendizaje.

En el aprendizaje no supervisado en cambio, no existe un "maestro", ni se le proporcionan a la máquina entradas de muestra previamente clasificadas. En su lugar, se le asigna a la máquina la tarea de agrupar entradas similares (clustering) por sí misma [Bish06]. Esta tarea puede constituir un objetivo por sí mismo (descubrimiento de patrones) o puede servir como base para otras tareas. Algunas aplicaciones de minería de datos, como el modelado de tópicos relacionados, por ejemplo, representan esta clase de aprendizaje.

Un término muy común en las tareas de visión artificial es el reconocimiento de patrones. Bishop considera que este término es casi un sinónimo del aprendizaje automático [Bish06], debido a que ambas tareas están estrechamente emparentadas, especialmente en el aprendizaje no supervisado. Duda y Hart, definen el reconocimiento de patrones como la tarea de tomar datos "crudos" y realizar una acción basados en la "categoría" de dicho patrón [Duda99].

Finalmente, un campo de estudio con el que la visión artificial guarda una relación muy estrecha es el procesamiento de imágenes digitales. El procesamiento de imágenes puede definirse como el conjunto de procedimientos que se realizan sobre una imagen para su almacenamiento, transmisión o tratamiento [Rodr12]. El interés de los métodos de procesamiento de imágenes digitales se fundamenta en dos áreas principales de aplicación: a) mejora de la calidad para la interpretación humana; b) procesamiento de datos de la escena para

la percepción de las máquinas de manera autónoma [Gonz96]. Como ejemplo de esto último, se puede señalar el mejoramiento en la calidad de la imagen (pre-procesamiento) y la extracción de características.

2.1.2 Biometría y aplicaciones

La visión computacional en general, y el reconocimiento de rostros en particular están estrechamente relacionados con una de sus aplicaciones más comunes: la biometría.

La biometría es el conjunto de estudios, técnicas y herramientas que buscan identificar a los seres humanos de manera automática, y surge con la llegada de las tecnologías de la información. La biometría emplea gran variedad de rasgos conductuales y físicos para lograr este propósito. Entre los primeros podemos mencionar la firma, el tecleo y el modo de andar y expresarse, mientras que entre los segundos, los más populares, podemos mencionar la lectura de las huellas dactilares, la retina, el iris, la configuración de las venas, la geometría de la palma de la mano, la voz (también considerada conductual), y finalmente la apariencia del rostro.

El reconocimiento de rostros tiene la ventaja de ser un método pasivo de identificación, esto es, un método que no requiere la cooperación del individuo para llevarse a cabo. Entre las desventajas principales se encuentra el hecho de que al ser un método no intrusivo, las lecturas obtenidas presentan gran cantidad de irregularidades y alteraciones que dificultan la identificación, tales como cambios de iluminación, pose, expresiones faciales, etc. (estos y otros factores serán descritos con profundidad en la sección 2.4).

2.1.3 Detección y reconocimiento de rostros

En el procesamiento digital de imágenes existe un conjunto de conceptos y términos relacionados con la identificación de rostros que es importante diferenciar de manera correcta, dado que cada uno designa tareas distintas. Entre dichos conceptos, los más comunes son reconocimiento, detección, verificación y seguimiento de rostros.

Figura 2.2. Ejemplo de detección de rostros en una imagen estática.

La tarea de detección puede ser definida como un problema consistente en localizar un número desconocido de rostros en un ambiente no controlado a partir de un conjunto de imágenes estáticas o video [Hjel01] (Véase la figura 2.2).

La tarea del reconocimiento de rostros por su parte puede ser planteada de la siguiente manera: dado un conjunto de imágenes o una secuencia de video, el objetivo es identificar o verificar a una o más personas en la escena usando una base de datos que contiene una muestra de rostros pertenecientes a individuos conocidos [Zhao03].

La identificación y la verificación de rostros por otro lado, son dos formas distintas de utilizar un sistema de reconocimiento, donde la primera consiste en introducir un rostro desconocido al sistema para que este devuelva como respuesta la identidad del sujeto en cuestión a parir de la información existente en su base de datos; mientras que la segunda consiste en confirmar o rechazar a alguien en base a su semejanza con alguien que el individuo afirma ser [Zhao03]. El primer planteamiento constituye el reconocimiento de rostros propiamente dicho, mientras que el segundo constituye la premisa en la que se basa la biometría.

Finalmente, el seguimiento de rostros es una tarea específicamente enfocada en las secuencias de video, cuyo objetivo es mantener la ubicación real de uno o más rostros dentro de la secuencia, a pesar de su movimiento dentro de esta.

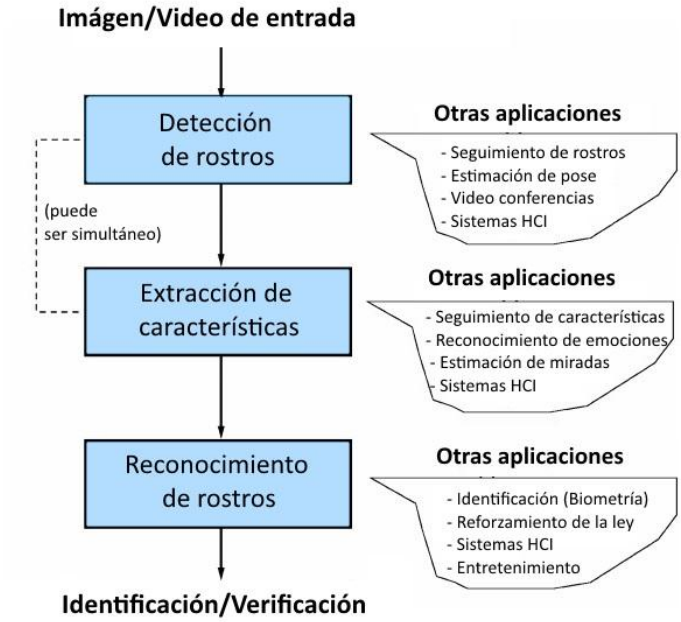


Figura 2.3. Etapas de un sistema de reconocimiento de rostros. Se puede apreciar que la detección constituye un módulo dentro de él, así como el hecho de que cada módulo puede tener aplicaciones de manera independiente.

La detección y el reconocimiento de caras son dos procesos estrechamente vinculados, tanto entre sí como con el resto de las tareas antes descritas (fig. 2.3), pues ambos son procesos claramente definidos dentro de cualquier sistema que tenga por objetivo el mundo real. Esto

debido a que la detección proporciona la localización inicial en la tarea de seguimiento, y también las coordenadas para delimitar el rostro para la función de reconocimiento. Y como se señaló con anterioridad, la verificación es una tarea específica del reconocimiento.

Cabe señalar que la detección de rostros tiene aplicaciones adicionales que pueden incluir o no el reconocimiento/seguimiento de rostros. Entre dichas aplicaciones se puede mencionar una fase que podemos considerar intermedia, que consiste en identificar elementos individuales del rostro, tales como ojos o boca con la finalidad de rastrear la mirada en sistemas de interacción humano computadora (HCI, por sus siglas en inglés) o identificar emociones a partir de los gestos faciales.

2.1.4 Madurez de las tecnologías

Si se pretendiera hacer una descripción a grandes rasgos del estado del arte en el reconocimiento de rostros, la mejor forma de hacerlo es a través de cada uno de los diferentes módulos que lo integran, puesto que se ha alcanzado distinto grado de éxito en ellos. También es necesario asumir cierta clase de restricciones en las posibles variables que se pueden presentar, tales como pose, iluminación, expresión facial, etc.

Así pues, tomando en cuenta un sistema constituido por varios procesos como detección de rostros, seguimiento, alineación, extracción de características y clasificación, podemos decir lo siguiente: la detección y el seguimiento en tiempo real en lugares interiores normales es un problema relativamente resuelto, mientras que se requiere algo más de trabajo en escenarios exteriores. Una vez hecha la detección y el seguimiento de manera satisfactoria, contando con una buena alineación (por lo general con la cooperación del individuo) y asumiendo que se cuenta con una resolución suficientemente buena para localizar los componentes faciales, que no se presentan expresiones faciales exageradas y que la iluminación no produce grandes sombras en la cara, se puede decir que el reconocimiento de rostros es una tarea suficientemente bien resuelta.

Sin embargo, si asumimos un entorno diario sin restricciones y sin la cooperación del sujeto, tal como lo que se presenta en un aeropuerto o un banco la tarea es aún una cuestión abierta. Es probable que aún se requieran años de esfuerzo para producir soluciones prácticas a tal tipo de problemas [Stan05].

2.2 Fundamentos biológicos de la percepción facial

El concepto de percepción facial hace referencia al proceso de comprensión e interpretación de un rostro individual, particularmente (pero no exclusivamente) el rostro humano, especialmente en relación al modo en que la información asociada es procesada por el cerebro.

La percepción facial es un proceso complejo, puesto que involucra áreas del cerebro muy extensivas y diversas. Las tareas que el sistema visual humano desempeña en la percepción de rostros es así mismo, bastante diversa. Algunas de las tareas que podemos mencionar son: identificación de rostros, evaluación de emociones, estimación de la orientación en la mirada, clasificación de género, raza, edad, atractivo físico, etc. [Gold09].

Existe evidencia de que además de desempeñar estas tareas de manera inconsciente, el cerebro también utiliza conocimiento contextual para llevar a cabo la identificación de rostros. Por ejemplo, para reconocer a alguien en un lugar donde se supone que debería de estar [Chel95].

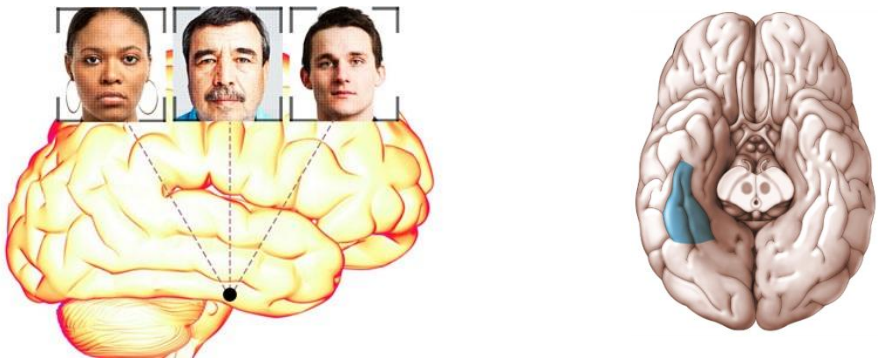
Considerando todos estos factores, resulta casi imposible tratar de emular la sorprendente capacidad del cerebro humano para realizar esta tarea con la tecnología actual. Sin embargo, comprender algunos de los aspectos fundamentales de su funcionamiento podrían ser de utilidad para igualarlo en futuro no muy lejano, e incluso superarlo. Considérese por ejemplo el hecho de que cerebro tiene limitantes en cuanto al número de rostros que puede recordar de manera correcta, mientras que las computadoras poseen una capacidad, al menos en el aspecto de almacenamiento, infinitamente mayor.

Es debido a lo anterior que en el presente estudio se ha decidido incluir algunos de los descubrimientos neurológicos y psicológicos más relevantes en la materia, que pueden servir como pistas potenciales en el desarrollo de sistemas de reconocimiento automáticos. En algunos casos, más que descubrimientos, se trata de cuestiones de investigación parcialmente resueltas o tema de debate, pero se ha decidido incluirlos de todos modos por considerar potencialmente útiles todas las perspectivas. Tales cuestiones son las siguientes:

- **¿Es el reconocimiento facial un proceso dedicado?**

  La teoría de la existencia de un sistema de procesamiento dedicado proviene de 3 evidencias fundamentales. A) Los rostros son más fácilmente reconocibles por los humanos que otros objetos cuando son presentado en orientación vertical. B) Los pacientes con un padecimiento denominado prosopagnosia son incapaces de reconocer rostros previamente conocidos, pero usualmente no presentan otro tipo de agnosia, por lo que pueden reconocer a la gente por sus voces, color de pelo, ropa, etc. Aunque los

pacientes son capaces de reconocer elementos individuales, tales como ojos, nariz o boca, así como reconocer si una imagen es una cara o no, son incapaces de reunir estos elementos con el propósito de llevar a cabo una identificación exitosa [Rivo13]. Los pacientes con prosopagnosia generalmente presentan daño en una región cerebral conocida como córtex temporal inferior, o giro fusiforme (fig. 2.4). C) Se argumenta que los infantes tienen la predisposición innata de sentir atracción por los rostros. Los neonatos parecen sentir preferencia por los patrones que describen un rostro sobre aquellos que no lo hacen.



(a) (b)

Figura 2.4. Ubicación del giro fusiforme en el cerebro humano. El reconocimiento de rostros se lleva a cabo en su mayor parte en un área del cerebro conocida como lóbulo temporal inferior, o giro fusiforme. (a) Vista lateral del cerebro (b) Vista inferior.

Una de las teorías más ampliamente aceptadas en este sentido indica que la comprensión de rostros es un proceso que involucra varias etapas [Bruc86]. Haxby, Hoffman y Gobbini proponen un modelo de funciones distribuido, donde el giro fusiforme tiene la tarea de tratar la información asociada a los aspectos invariantes del rostro, mientras que los aspectos variables (expresión, mirada) son procesados en zonas periféricas del cerebro como el surco temporal superior [Haxb00].

- **¿El reconocimiento facial se hace de manera holística o analiza características de manera separada?**

Los rostros humanos con mucha frecuencia pueden ser reconocidos por el cerebro humano a partir de información muy pequeña, tal como elementos individuales del rostro, como ojos, labios, etc. Esta capacidad se ve incrementada con la familiaridad que tengamos con el rostro, siendo especial el caso de los personajes famosos.

Sin embargo, cuando las características en la mitad superior de una cara son combinadas con la mitad inferior de otra cara, las identidades de los individuos son muy

difíciles de reconocer (fig. 2.5).

Esto parece indicar que el contexto holístico (es decir, el procesamiento global) afecta la manera en la que las características individuales son procesadas [Sinh06]. Cuando ambas mitades están desalineadas, presumiblemente interrumpiendo el procesamiento holístico, ambas identidades son más fáciles de reconocer.

(a) (b)

Figura 2.5. Experimento de Young sobre el procesamiento de la información: trate de reconocer la identidad de los personajes en (a) y en (b). En la mayoría de los experimentos fue más fácil la identificación con los rasgos separados que con ambas mitades combinadas en una sola imagen, mostrando que el cerebro lleva a cabo un procesamiento holístico con los rasgos observados de manera individual para ayudar (y en este caso para interferir) en la identificación del rostro.

Estos resultados sugieren que al ser consideradas de manera individual, las características personales en ocasiones son suficientes para llevar a cabo el reconocimiento del individuo. En el contexto de la cara, sin embargo, la relación geométrica entre cada característica y el resto del rostro puede "reemplazar" el diagnóstico de tal característica.

Como resultado se puede decir que, aunque el procesamiento de las características individuales desempeña un papel muy relevante en el reconocimiento de rostros, el procesamiento holístico que involucra la configuración e interdependencia entre ellas es al menos igualmente relevante.

- **¿Cuáles son las características más importantes en el reconocimiento facial?**

  En lo que se refiere a la identificación de rostros, no todas las características tienen la misma importancia. Los resultados experimentales típicamente subrayan la importancia de los ojos, seguidos de la boca y finalmente la nariz. Sin embargo, una característica que había recibido poco importancia hasta hace poco son las cejas. En recientes estudios [Sinh06] se ha demostrado que las cejas no solo son importantes, sino

que pueden ser una de las características más relevantes, junto con los ojos.

Para comprobar esta teoría, los investigadores realizaron un experimento consistente en tomar las imágenes de 50 personajes famosos y borrar digitalmente sus cejas. Posteriormente pidieron a los participantes identificar a cada uno de los personajes y como resultado, se obtuvo un porcentaje de identificación significativamente peor en las fotografías de los personajes sin cejas que en las fotografías normales, e incluso peor que en las que carecían de ojos (fig. 2.6).

Existen varias posibilidades para explicar este resultado. Primero que nada, las cejas son un elemento significativo para expresar emociones, por lo que el cerebro humano puede estar inconscientemente inclinado a observar las cejas para realizar la tarea de identificación. Segundo, las cejas son elementos faciales "estables" en relación a otros, tales como el cabello o los labios. Además, puesto que se encuentran en la parte superior de la cara son menos susceptibles a cambios de iluminación o sombras. Finalmente, hablando del rostro en términos de procesamiento de imágenes, las cejas son elementos de alto contraste, por lo que tienen alta tolerancia a las degradaciones producidas por distancia o iluminación.

Figura 2.6. Experimento de Sadr et al. [Sinh06] sobre la influencia de las diferentes características faciales. Imágenes modificadas de los rostros del presidente Nixon y la actriz Winona Ryder. Estas imágenes fueron utilizadas para investigar la influencia de las características individuales (específicamente las cejas) en la identificación de rostros.

Por otra parte, es necesario señalar que existen otros elementos individuales que son igualmente importantes a la hora de caracterizar un rostro, como el pelo, o la forma de la cara; si bien no son elementos muy fiables para el reconocimiento automático, puesto que en son altamente susceptibles a cambios debido a estética en el primer caso, o variación de pose, en el segundo.

También es importante mencionar que los elementos tienen diferente importancia en función de la pose que el rostro tenga. Así, mientras la nariz no juega un papel importante en las imágenes de frente, comienza a ganar importancia conforme se va acercando a la vista de perfil [Stan05, Sin06].

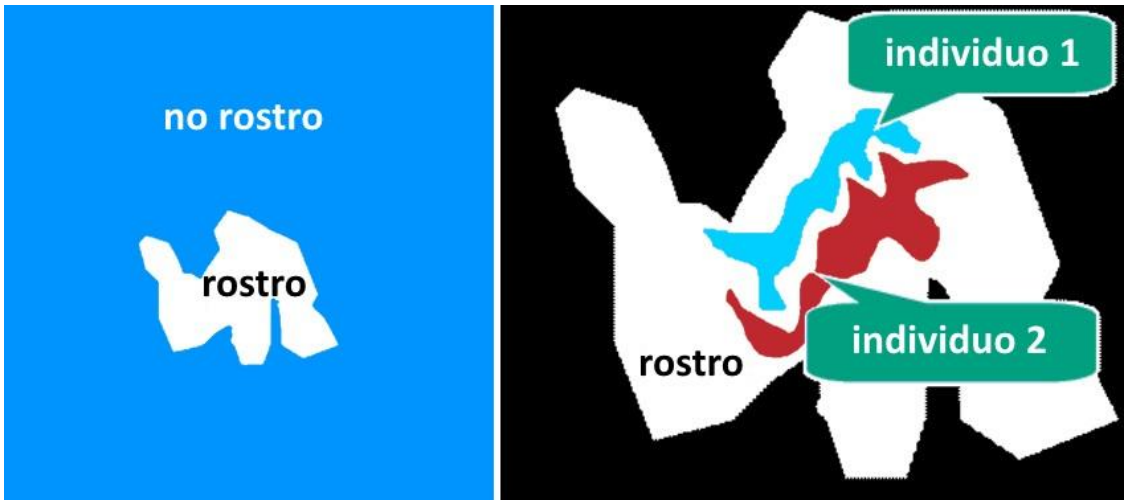
- **El rol del análisis de la frecuencia espacial.**

  Algunos estudios [Chel95] demuestran que la información en las bandas de baja frecuencia espacial juega un rol dominante en el reconocimiento de rostros. También muestran que, dependiendo de la tarea a realizar, la frecuencia desempeña roles diferentes. Por ejemplo, la tarea de identificación del sexo es realizada de manera exitosa usando componentes de baja frecuencia, mientras que las tareas de identificación requieren el uso de componente de alta frecuencia.

  La baja frecuencia contribuye a la descripción global, mientras que los componentes de alta frecuencia contribuyen a la descripción detallada que se requiere para llevar a cabo la tarea de identificación.

2.3 Desafíos técnicos

En la tarea de reconocimiento de rostros humanos existe una serie de dificultades técnicas que impiden llevar a cabo una clasificación eficaz con métodos relativamente simples de visión e inteligencia artificial. Observando la distribución de algunas clases de muestra en un espacio de características podemos notar con facilidad que este es un problema no-lineal y no-convexo. En la figura 2.7 por ejemplo, podemos apreciar la distribución de clases creada a partir de un ejemplo real, donde se intenta separar en un primer momento las imágenes que contienen rostros de las que carecen de ellos (tarea de detección), y posteriormente llevar a cabo la clasificación de dos individuos distintos dentro del espacio de rostros (tarea de reconocimiento) [Stan05].



(a) (b)

Figura 2.7. Ejemplos de la separabilidad de clases en los problemas de detección y reconocimiento de rostros, respectivamente. (a) Rostros contra universo de "no rostros". (b) Rostros de dos individuos dentro de un espacio previamente separado de los "no rostros".

Las imágenes (a) y (b) de la figura 2.7 permiten observar la dificultad para modelar funciones de clasificación que resulten completamente fiables y efectivas en las tareas de detección y reconocimiento respectivamente. En el caso del reconocimiento, esto se hace especialmente notorio al observar la irregularidad en las manchas que representan a los 2 individuos. En términos generales podemos decir que las figuras roja y azul tienen poca homogeneidad intrínseca y heterogeneidad extrínseca. Esto es el resultado de dos problemas que se presentan comúnmente en las tareas de clasificación:

- **Variaciones inter-clase pequeñas**. Es decir, cuando la diferencia entre dos o más clases distintas es muy pequeña o insuficiente, lo que ocasiona que al colocar las clases en un espacio de características, éstas se distribuyan de manera muy próxima la una de

la otra, por lo que el clasificador las considerará como parte de una misma clase de manera errónea. En el caso de los rostros humanos, este problema es relativamente común, por ejemplo en el caso de parientes cercanos, y más notablemente, de gemelos. (Figura 2.8)

- **Variaciones intra-clase grandes**. Este problema surge cuando existen grandes diferencias entre dos muestras de una misma clase, lo que da como resultado que el clasificador las considere dos clases distintas. En el caso del reconocimiento de rostros humanos, este es un problema muy común, comprendiendo diversas variaciones que se presentan habitualmente en imágenes no controladas, tales como cambios de pose, iluminación, expresiones faciales, etc.

  Un estudio muy extendido [Adin97] señala que en lo que respecta al reconocimiento de rostros, algunas de las variaciones existentes en una misma clase son mayores que las existentes entre la variación de clases; es decir que los cambios que puede presentar una persona en distintas imágenes pueden hacerla más diferente con respecto a sí misma que con respecto a otras personas. Esto pone de relieve la dificultad que implica la tarea de reconocimiento de rostros, así como la necesidad de conocer en profundidad las posibles variaciones que pueden existir, con la finalidad de elaborar métodos encaminados a compensar dichos cambios. Debido a esto, las variaciones más comunes se describen con mayor detalle a continuación.

(a) (b)

Figura 2.8. Ejemplos del problema de separación de rostros distintos con alto grado de similitud entre sí. (a) Gemelos (b) Parientes.

2.3.1 Resolución

Definimos la resolución como la cantidad de detalle que puede observarse en una imagen, mismo que es frecuentemente descrito mediante la cantidad de pixeles por unidad de área (pulgadas cuadradas, centímetros cuadrados, etc.), hablándose de alta resolución si se puede apreciar gran cantidad de detalle y de baja resolución en caso contrario.

La baja resolución afecta notablemente a todas las aplicaciones de la visión computacional, por lo que existe gran cantidad de investigación al respecto en el área. Sin embargo, es en el reconocimiento de rostros humanos donde el problema tiene particular interés debido a la gran cantidad de situaciones del mundo real que presentan esta desventaja.

Figura 2.9. Ejemplos de cambios de resolución (y perspectiva) en imágenes de la base de datos SCFace-surveillance [Grgi11], enfocada a evaluar el reconocimiento en imágenes de cámaras de vigilancia.

Considérese por ejemplo el hecho de que la gran mayoría de cámaras de seguridad son de baja resolución (fig 2.9), debido tanto al costo del equipo como a la cantidad de horas de grabación que deben ser almacenadas.

2.3.2 Pose

Definimos la variación de pose como el cambio producido por la rotación del objetivo (en nuestro caso, el rostro) con respecto a la cámara. Equivalentemente, también se podemos entender esta variación como un cambio en el ángulo de captura de la cámara.

Figura 2.10. Ejemplos de variaciones de pose en un solo eje: cabeceo (eje X), guiñada (eje Y) y ladeo (eje Z). Las poses simples son relativamente menos frecuentes que las compuestas en aplicaciones reales, pero son la forma más habitual de encontrar en las bases de datos de prueba, para permitir la normalización de resultados.

El cambio de pose es uno de los problemas más frecuentes tanto en la detección como en el reconocimiento de rostros, debido a que la mayoría de las aplicaciones del mundo real carecen de la posibilidad de obtener tomas frontales debido a la ubicación de la cámara o la falta de cooperación del individuo (ej. cámaras de seguridad). Es por ello que este problema representa el principal motivo de estudio de la presente tesis.

Se debe considerar el hecho de que las rotaciones pueden producirse en cualquiera de los tres ejes coordenados (fig. 2.10), o bien, en una combinación de ellos (fig 2.11). En la bibliografía especializada en inglés, con frecuencia se define con un nombre específico a la rotación en cada eje, siendo sus traducciones más aproximadas: cabeceo (rotación sobre eje X), guiñada (rotación sobre el eje Y) e inclinación/ladeo (rotación sobre el eje Z). Con frecuencia nos referiremos a las rotaciones con estos términos, o también empleando el término "perfil" en términos fraccionarios (ej. perfil completo, medio perfil, cuarto de perfil, etc.) para denominar a las rotaciones sobre el eje Y.

Figura 2.11. Ejemplos de variaciones de pose en uno, dos y tres ejes coordenados. Esta forma es la forma en la que se encuentran la mayoría de capturas del mundo real.

2.3.3 Iluminación

La iluminación es otro de los retos fundamentales que enfrenta el reconocimiento de rostros, debido la frecuencia con la que se presenta. En este aspecto se debe considerar que los cambios pueden ser producidos por diversas condiciones lumínicas, tales como el tipo de fuente (natural, artificial), la dirección de proveniencia de la luz, su intensidad y su color (fig. 2.12).

(a) (b) (c)

Figura 2.12. Ejemplos de rostros con distinta iluminación de la base de datos FEI [Oliv06]: (a) iluminación natural, (b) iluminación artificial de tono azul proveniente del lado izquierdo, (c) iluminación de baja intensidad proveniente de la luz ambiental residual.

Así mismo, las variaciones pueden producirse por las condiciones de adquisición de la imagen, tales como la luminosidad/reflectancia del entorno del rostro o la sensibilidad a la luz y ajustes de la cámara (cuando se adquiere con medios distintos) con la que se capturan las imagénes.

Como se señala en un estudio [Adin97], "las variaciones en las imágenes de una misma cara debidas a la iluminación y punto de vista son mayores que las debidas al cambio de identidad".

2.3.4 Expresiones faciales

Dada su utilidad como elemento de comunicación no verbal, el rostro humano ha desarrollado un rostro conformado por gran cantidad de músculos capaces de mostrar distintas gesticulaciones faciales. Esta posibilidad afecta la capacidad de identificación de los sistemas de reconocimiento facial, pues modifican de manera significativa la geometría del rostro (fig. 2.13), además de que en ocasiones muestran elementos no presentes en la(s) imagen(es) de muestra (como lengua o dientes) y ocultan otros (como los ojos).

Figura 2.13. Ejemplos de expresiones faciales comunes y exageradas. Nótese como la geometría de varios de los elementos que lo conforman (cejas, boca, ojos) se ve alterada de manera significativa, pese a lo cual, sigue siendo reconocible por un ser humano.

Como nota cabe señalar que en la actualidad existen algoritmos que buscan identificar, no únicamente la identidad del individuo, sino también su estado anímico en función de su expresión facial; mientras que algunos otros se enfoca exclusivamente en la segunda cuestión.

2.3.5 Oclusiones

Dentro de la tarea de reconocimiento de rostros, definimos oclusión como la obstrucción parcial de la cara del individuo por cualquier elemento ajeno a ella, tales como prendas de vestir o accesorios (lentes, bufandas, sombreros, gorras, etc.), partes del cuerpo del propio individuo (manos, brazos, etc.), otras personas (en las imágenes no controladas se presenta con frecuencia) u objetos. Ver figura 2.14.

Figura 2.14. Ejemplos de algunas de las oclusiones más frecuentes: gafas, manos y cabello. También pueden presentarse debido a gorros, bufandas, objetos y otras personas.

2.3.6 Componentes variables del rostro

Mención aparte merecen las variaciones en el rostro de los individuos producidas de manera más o menos consiente por ellos mismo para cambiar su apariencia. Entre tales cambios podemos enumerar el cambio de peinado, el cambio de color de cabello, el maquillaje en las mujeres, la presencia o ausencia de barba y bigote en los hombres y la presencia de adornos tales como los aretes (fig. 2.15).

Todos ellos producen distintos grados de alteración en el aspecto del individuo, que van desde lo imperceptible hasta lo altamente significativo. Es debido a ello que estas variaciones merecen ser consideradas de manera separada a las oclusiones, puesto que carecen de regularidad y por lo tanto su impacto es altamente variable.

Figura 2.15. Ejemplos de variaciones en la apariencia de los individuos producidas por la presencia de componentes variables, tales como la presencia de pelo en el rostro en el caso de los hombre, o bien, el maquillaje en el caso de las mujeres.

2.3.7 Edad

Finalmente, entre las variaciones intra-clase debemos mencionar los cambios producidos en el rostro del individuo en el transcurso del tiempo. Entre estos cambios podemos mencionar la proporción de la cara (se alarga en la edad adulta), el cambio de color y la cantidad del cabello y la aparición de marcas faciales, tales como arrugas, manchas o cicatrices (fig. 2.16).

Aunque este tipo de variaciones reciben una consideración menor en los sistemas de reconocimiento de rostros debido tanto a su dificultad como a la menor cantidad de aplicaciones en la vida cotidiana; es necesario considerar su existencia en etapas más avanzadas de investigación. Así mismo cabe señalar que las principales líneas de investigación hoy en día centran sus esfuerzos en determinar la edad del individuo, más que su identidad.

Figura 2.16. Ejemplos de variaciones en el individuo producidas por la edad. Se pueden apreciar el cambio de color en el cabello y la aparición de marcas en la piel.

2.4 Técnicas de reconocimiento generales existentes

2.4.1 Clasificación

Se han propuesto gran cantidad de métodos para el reconocimiento de rostros en los últimos 40 años. Muchos de ellos siguen enfoques similares, y como consecuencia, han surgido diferentes clasificaciones para facilitar su comprensión y estudio. A pesar de esta variedad de clasificaciones, la más común es la que se basa en el modo de extraer las características de la imagen. En esta clasificación existen los siguientes grupos:

- ***Métodos holísticos***. Este tipo de métodos utilizan la imagen completa como entrada para el sistema. La mayoría de las técnicas empleadas se basan en métodos estadísticos. A esta clase pertenece las técnicas de Eigenfaces y Fisherfaces; detallados más adelante.
- ***Métodos estructurales***. Este tipo de métodos, al contrario de los métodos holísticos, se basa en las características individuales del rostro tales como los ojos, nariz, boca, etc. y no en la imagen completa. Sus técnicas se basan principalmente en el análisis local de texturas, detección de contornos, etc. A esta clase pertenecen los algoritmos EBGM, LBP y las redes CNN, por ejemplo.
- ***Métodos híbridos***. Estos métodos utilizan una combinación de los dos métodos anteriores, de la misma manera en la que se supone que trabaja el sistema de reconocimiento de rostros humano. En teoría se supone que este enfoque brinda lo mejor de los dos anteriores. A este enfoque pertenece el algoritmo de PCA-Modular o Modular Eigenfaces.

A continuación se brinda una breve descripción de los enfoques más importantes hasta la fecha.

2.4.2 Métodos holísticos

2.4.2.1 Eigenfaces

El algoritmo de Eigenfaces se basa en el concepto de reducción de la dimensionalidad de datos utilizado en el análisis estadístico; más concretamente, en el Análisis de Componentes Principales, más conocido como PCA, por sus siglas en inglés.

Fue desarrollado inicialmente por Kirby y Sirovich [Siro87, Kirb90] como una manera de representar eficientemente las imágenes de rostros humanos mediante un número reducido de coeficientes correspondientes a los valores más significativos (eigenvalores o valores propios). Posteriormente, esta técnica fue utilizada por Turk y Pentland [Turk91a, Turk91b] en la tarea de reconocimiento de rostros.

El planteamiento original de Kirby y Sirovich se basó en la idea de que todas las imágenes pertenecientes a una misma categoría (por ejemplo, las de rostros humanos) comparten gran cantidad de similitudes; de manera que cada imagen individual podría ser expresada como la combinación lineal de K componentes separados. Los autores utilizaron análisis de componentes principales para encontrar dichos K elementos; recordando que PCA permite encontrar las variables que presentan mayor varianza, lo que equivale a mayor información según el concepto de entropía de Shannon.

Kirby y Sirovich apuntaron inicialmente esta idea a la tarea de compresión de imágenes relacionadas entre sí, de tal modo que cada imagen pudiera ser reconstruida a partir de un conjunto de datos menor al original. Así por ejemplo, para reconstruir un rostro humano, se podría partir de un "rostro promedio" $R0$ que aportaría el 50% de la información; posteriormente se le añadirían características de otro rostro individual $R1$ para alcanzar el 75%, más un rostro $R2$ para alcanzar el 85%, y así sucesivamente hasta completar o acercarse de manera significativa al 100%. Los autores denominaron Eigenpictures (o imágenes propias, en español) a todas las imágenes que se basaran en este principio.

Figura 2.17. Ejemplos de Eigenfaces. Imágenes obtenidas al descomponer los rostros originales en sus componentes principales.

Posteriormente, en 1991, Turk y Pentland se basaron en estas ideas para desarrollar un algoritmo de reconocimiento automatizado de rostros humanos. Al ser un algoritmo de propósito específico enfocado en las imágenes de caras, denominaron Eigenfaces tanto al algoritmo como al nuevo conjunto de imágenes de rostros con dimensión reducida (fig 2.17).

El algoritmo de Turk y Pentland inicia convirtiendo cada imagen en un vector individual conformado por la concatenación de todas las filas de pixeles de dicha imagen. Posteriormente dicho vector es colocado dentro de una matriz T, donde cada columna de la misma es una imagen. El algoritmo prosigue extrayendo la "cara promedio", definida como $\Psi = \frac{1}{M}\sum_{n=1}^{M}\Gamma_n$, siendo $\Gamma_1, \Gamma_2, \Gamma_3 \ldots \Gamma_M$ el conjunto de rostros de entrenamiento. Luego, se obtiene la diferencia de cada cara con respecto a la cara promedio ($\Phi_i = \Gamma_i - \Psi$) con la finalidad de conocer los elementos característicos de cada rostro.

El tercer paso consiste en calcular los eigenvectores y eigenvalores de la matriz de covarianza a partir de las imágenes obtenidas Φ_i. Cada eigenvector obtenido tiene las mismas

dimensiones que la imagen original, por lo que puede ser representada como la imagen de una cara, dando lugar a las imágenes características observadas en la figura 2.17.

Como cuarto y último paso en el proceso de creación de las eigenfaces, se selecciona el número de componentes principales que se utilizarán en la caracterización de rostros, siendo conservados N primeros elementos que presentes los eigenvalores más altos.

Para realizar la clasificación de rostros, los autores proyectan estos eigenvectores en un "espacio facial" con tantas dimensiones como número de eigenfaces seleccionadas. A partir de este espacio, se clasifica la imagen desconocida en base a su proximidad a una clase k de rostro. El método utilizado por Turk y Pentland fue la distancia euclidiana, definida como $\epsilon_k = ||(\Omega - \Omega_k||$. Utilizando un par de umbrales fijados de manera empírica es posible utilizar esta distancia para determinar si la cara pertenece a una clase existente (identificación), si es un rostro desconocido, o si la imagen contiene un rostro humano o no (detección).

La importancia de las Eigenfaces radica en que debido a su rapidez y simplicidad se puede considerar como el primer algoritmo de reconocimiento de rostros técnicamente viable, así como el primero capaz de generar resultados prácticos convincentes bajo condiciones aceptables. Debido a ello, se convirtió en la técnica más ampliamente utilizada en reconocimiento de imágenes controladas [Zhan09].

No obstante sus ventajas, este algoritmo, como casi todos los métodos holísticos, presentan una disminución de rendimiento muy pronunciada en rostros que presentan cambios de pose mayores a los 15°. Además, también es vulnerable a cambios grandes de escala e iluminación.

2.4.2.2 Fisherfaces

El algoritmo de reconocimiento de Fisherfaces fue propuesto por Belhumeur et al. en 1997 [Belh97] como una evolución del método de Eigenfaces. Al igual que la técnica presentada por Turk y Pentland 6 años antes, Fisherfaces es un método holístico basado en el análisis estadístico de los componentes de la imagen. Pero a diferencia de Eigenfaces, que se basa únicamente en el Análisis de Componentes Principales (PCA), Fisherfaces utiliza Análisis Discriminante Lineal (LDA) o Análisis Discriminante de Fisher (de donde deriva el término).

LDA es una técnica de análisis estadístico que permite mejorar la capacidad discriminante entre diferentes clases cuando se dispone de múltiples muestras (imágenes) por clase (individuo). Los autores utilizaron esta propiedad con el objetivo específico de mejorar los resultados obtenidos por Eigenfaces en situaciones donde existieran variaciones de iluminación (pero se mantuviera la misma pose). Al igual que con la técnica de Turk y Pentland, es posible visualizar los resultados mediante imágenes, como se aprecia en la figura 2.18.

Figura 2.18. Cuatro primeras Fisherfaces generadas a partir de un conjunto de 100 sujetos.

Durante el proceso de entrenamiento de esta técnica, se busca aumentar la razón de diferencia entre las variaciones extra-clase en relación con las intra-clase. La diferencia extra-clase es caracterizada utilizando una matriz de dispersión extra-clase S_B, la cual calcula las diferencias sumarias entre la media de la clase μ_i y la media global μ. La diferencia intra-clase es representada en una matriz de dispersión intra-clase S_W que calcula las diferencias sumarias entre las imágenes individuales x_k con relación a la media de la clase μ_i. Los vectores y valores propios son entonces calculados para maximizar la razón entre S_B y S_w, expresada como $S_B w_i = \lambda_i S_w w_i$, con $i = 1,2,\dots,m$ y donde w_i son las m mayores vectores propios generalizados y λ_i son los correspondientes valores propios generalizados.

Utilizando este método, se reduce la dimensión del espacio de proyección de vectores en relación con PCA, pero se puede decir que mientras PCA logra una *mayor* dispersión, LDA logra una *mejor* dispersión, lo que consecuentemente proporciona una mejor clasificación.

Uno de los problemas que es necesario superar con esta técnica es la singularidad en la matriz de dispersión intra-clase. Para lograrlo las imágenes son pre-procesadas mediante PCA para reducir su dimensión y proporcionar un nivel manejable con la técnica LDA. Este método representa el algoritmo de Fisherfaces, propiamente dicho y formalmente está dado por:

$$W_{opt}^T = W_{fld}^T W_{pca}^T \tag{2.1}$$

Donde

$$W_{pca} = \arg\max |W^T S_T W| \tag{2.2}$$

$$W_{fld} = \arg\max \left| \frac{W^T W_{pca}^T S_B W_{pca} W}{W^T W_{pca}^T S_W W_{pca} W} \right| \tag{2.3}$$

siendo W_{opt}, W_{pca}, W_{fld} las matrices óptima, de análisis de componente principales y de análisis discriminante de Fisher, respectivamente.

Cabe señalar que para que esta técnica sea realmente efectiva, se requiere tener varias imágenes por individuo o su eficacia será exactamente igual a la de Eigenfaces. Los autores lograron mejorar significativamente los resultados de Eigenfaces en galerías de imágenes con variaciones de iluminación; pero no así de pose, donde ambos algoritmos resultan altamente sensibles a rotaciones pronunciadas [Tan06].

2.4.2.3 Evolutionary Pursuit (EP)

El algoritmo de búsqueda evolutiva (EP, por su siglas en inglés) fue propuesto por Liu et al. en el año 2000 [Liu00] y utiliza un algoritmo genético para intentar encontrar una base de rostros óptima tanto para propósitos de compresión de datos como para la identificación de patrones. Para lograr este fin, EP utiliza un conjunto de rotaciones sobre ejes definidos en un espacio PCA "blanqueado". El algoritmo evolutivo de este método es guiado mediante una función de ajuste que intenta balancear la precisión en la clasificación (riesgo empírico) y la habilidad para generalizar (intervalo de confianza).

La base de rostros óptima que el algoritmo trata de encontrar, es por lo tanto expresado matemáticamente como el subconjunto óptimo de vectores buscado a través de sucesivas transformaciones de rotación realizadas sobre los vectores base de la(s) imagen(es) de entrenamiento, que tienen el ajuste óptimo entre precisión en la clasificación y la habilidad para generalizar. Dicho subconjunto óptimo representa por lo tanto la evolución de los vectores transformados a partir del conjunto de vectores iniciales que representan cada una de las imágenes.

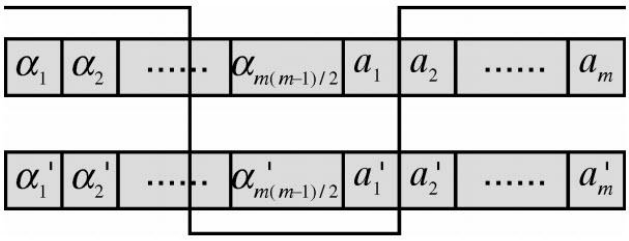


Figura 2.19. Representación gráfica de los cromosomas utilizados en el algoritmo evolutivo. En este algoritmo cada vector de 10 bits se utiliza para representar un ángulo. La figura también indica el segmento de intercambio entre vectores que se lleva a cabo al realizar cruzas.

El algoritmo de búsqueda evolutiva (EP) consta de 5 etapas primordiales. Dichas etapas son descritas detalladamente a continuación:

1. **Reducir la dimensión de las imágenes originales mediante PCA**. Este paso consiste en calcular las matrices de eigenvectores Φ y eigenvalores Λ, de la matriz de covarianza, Σ_X, usando descomposión de valores singulares (SVD) o método de Jacobi. Se escogen los primeros m vectores iniciales de Φ como vectores base y se proyectan el conjunto original de imágenes sobre dichos vectores para formar el conjunto de características Z en este espacio reducido PCA.
2. **"Blanquear" el espacio Z**. Este paso consiste en transformar el espacio Z y derivar el nuevo espacio característico V en el espacio PCA blanqueado.
3. **Establecer la matriz base**. Consiste en establecer los vectores unitarios obtenidos

$[\varepsilon_1\varepsilon_2 \dots \varepsilon_m]$ en una matriz unitaria $m \times m$ tal que: $[\varepsilon_1\varepsilon_2 \dots \varepsilon_m] = I_m$.

4. **Llevar a cabo el bucle evolutivo**. Este bucle se repetirá hasta que se encuentre la solución buscada o hasta que se alcance un número de iteraciones prefijado.
    a. Realizar diversas rotaciones entre parejas de vectores base de acuerdo a un orden definido para obtener el conjunto de ángulos de rotación $\alpha_1^{(k)}, \alpha_2^{(k)}, \dots, \alpha_{2m(m-1)/2}^{(k)}$, los cuales serán utilizados en la representación de cromosomas individuales (fig. 2.19). Cada rotación de vector será derivada de acuerdo a la siguiente ecuación:

    $$[\xi_1\xi_2 \dots \xi_m] = [\varepsilon_1\varepsilon_2 \dots \varepsilon_m]Q_k \quad (2.4)$$

    donde $Q_k \in \mathbb{R}^{m\times m}$ es una matriz de rotación, ε_m representan los vectores base y ξ_m los nuevos vectores transformados.
    b. Calcular la función de ajuste para evaluar la precisión y la generalización. La función de ajuste utilizada es la siguiente:

    $$\zeta(F) = \zeta_a(F) + \lambda\zeta_s(F) \quad (2.5)$$

    donde ζ_a representa el término de precisión en la clasificación, ζ_s es el término de generalización o separación de clases y λ es una constante positiva que determina la importancia del segundo término en relación al primero. Esta función se calcula sobre los ι ejes de proyección $\eta_1, \eta_2, \dots, \eta_\iota$, elegidos entre el conjunto de vectores base $\{\xi_1^{(k)}, \xi_2^{(k)}, \dots, \xi_m^{(k)}\}$ sobre los que se realizó la rotación.
    c. Encontrar el conjunto de ángulos y los subconjuntos de ejes de proyección que maximizan la función de ajuste y conservar dichos cromosomas como la mejor solución temporal.
    d. Establecer un nuevo conjunto de ángulos de rotación y subconjuntos de ejes de proyección e iterar sobre ellos.
5. **Llevar a cabo el reconocimiento**. Utilizando la base óptima encontrada, $T = [\Theta_{i1}\Theta_{i2} \dots \Theta_{il}]$ (fig. 2.20), realizar el reconocimiento de rostros con la ayuda de alguna medida de similitud.

Figura 2.20. Cuatro representaciones gráficas de vectores óptimos, obtenidos al derivar 26 vectores con el algoritmo EP.

2.4.2.4 Máquinas de Soporte Vectorial (SVM)

Las máquinas de soporte vectorial (SVM por sus siglas en inglés) fueron propuestas por Cortes y Vapnik en 1995 [Cort95] como un algoritmo clasificador de aprendizaje supervisado y propósito general. Desde entonces ha sido utilizado extensivamente en una gran cantidad de problemas de clasificación incluyendo el reconocimiento de rostros [Phil98, Guo00].

El algoritmo consiste en, dado un conjunto de puntos o elementos en un espacio determinado (y donde tales puntos pertenecen a 2 clases distintas), encontrar un hiperplano tal que separe la mayor cantidad de elementos de la misma clase en cada lado. Lo anterior se logra maximizando la distancia de cada clase al hiperplano de decisión, denominado hiperplano de separación óptima (OSH por sus siglas en inglés). Los puntos más cercanos al plano constituyen los así llamados vectores soporte, el término que da nombre al algoritmo (fig 2.21).

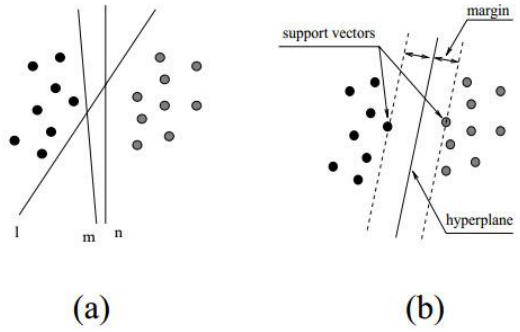


Figura 2.21. Separación de clases usando hiperplanos: (a) arbitrarios (b) separación óptima de clases.

El problema de reconocimiento de rostros es complicado, dado que una limitación importante de SVM es que está diseñado para permitir la separación óptima únicamente de 2 clases. En [Guo00] se utilizan Eigenfaces para la extracción de características y posteriormente las clases se separan con la ayuda de SVM. Con este método, se obtiene mayor precisión que con el método de aproximaciones de Eigenfaces normales.

2.4.3 Métodos estructurales

2.4.3.1 EBGM

Uno de los algoritmos más exitosos de reconocimiento basado en características locales es el *Elastic Bunch Graph Matching*; más comúnmente conocido por sus siglas EBGM y propuesto por Wiskott et al. en 1997 [Wisk97].

En este algoritmo, cada rostro es descrito mediante un conjunto de *wavelets* de Gabor situados en los componentes faciales (ojos, boca, nariz, etc.) y mediante redes elásticas de grafos etiquetados que se basan en la Arquitectura de Enlace Dinámico (DLA por sus siglas en inglés) propuesta por Lades et al. en 1993 [Lade93].

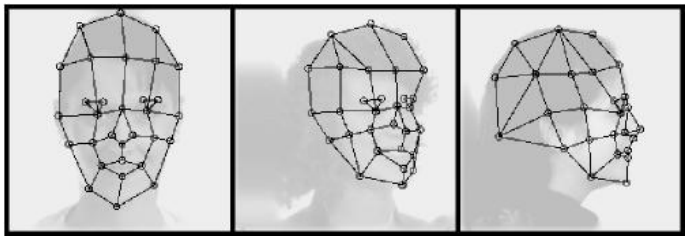


Figura 2.22. Redes adaptadas al rostro en diferentes poses. Los nodos son fijados de manera automática por el algoritmo. Nótese la correspondencia entre estos y los diferentes marcadores fiduciarios del rostro.

Los wavelets de Gabor actúan como descriptores locales de textura alrededor de cada marcador fiduciario de la red. Los marcadores de referencia o fiduciarios son puntos de la imagen que describen regiones importantes en el rostro, tales como bordes o esquinas, ya sea de su contorno o estructura interna (fig 2.22). Una vez obtenidos estos puntos, los autores aplican 40 convoluciones del kernel de Gabor en la zona que los rodea para obtener un conjunto de 40 coeficientes de Gabor. Este conjunto es denominado por ellos como un jet J (fig. 2.23).

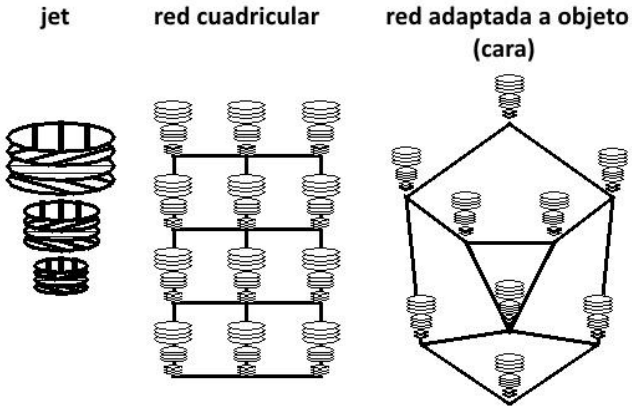


Figura 2.23. Representación gráfica del funcionamiento del algoritmo EBGM. El rostro es descrito mediante un grafo que actúa como una malla elástica capaz de ajustarse a la cara del individuo. Cada nodo es descrito mediante un jet obtenido mediante sucesivas convoluciones del kernel de Gabor sobre la zona.

La medida de similaridad entre rostros es obtenida a través de la comparación de jets y las distancias entre ellos. La comparación de jets es definida como la multiplicación de las magnitudes de sus coeficientes. Estas propiedades de los wavelets de Gabor fueron utilizadas tanto para la localización de componentes como para el reconocimiento.

A pesar de ser una de las técnicas más complejas y computacionalmente costosas, el algoritmo EBGM se desempeña de manera notablemente mejor que los métodos holísticos, sobre todo en los cambios de pose, gracias a la flexibilidad de la plantilla deformable en la que se basa.

2.4.3.2 Patrones Binarios Locales (LBP)

Ojala et al. propusieron, en el año 2002, un método de clasificación de texturas con invariancia a rotación mediante el uso de patrones binarios locales [Ojal02]. A partir de este trabajo, Ahonen et al. [Ahon04, Ahon06] aplicaron exitosamente estos descriptores en el reconocimiento de rostros en el año 2004.

El patrón binario local (LBP por sus siglas en inglés) es un operador de textura muy simple pero eficiente que se obtiene evaluando un vecindario de pixeles en función su valor central (fig. 2.24). El resultado es considerado como una cadena binaria y esta es utilizada como una característica con propósitos de clasificación.

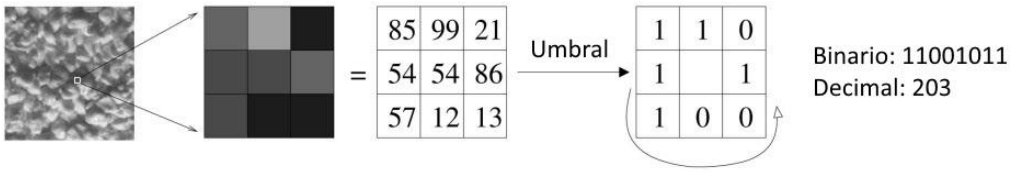


Figura 2.24. Operador binario básico. Se elige como umbral el valor del pixel central y si el valor es mayor o igual a él, el valor se transforma al valor binario 1. En caso contrario, su valor binario será 0.

Dos importantes propiedades de este operador son su simplicidad computacional y su alto poder discriminante. La primera propiedad se traduce en una baja demanda de recursos y una velocidad de procesamiento muy elevada; mientras que la segunda le proporciona robustez frente a cambios monótonos en escala de grises causados, por ejemplo, por variaciones de iluminación [Li07].

Cuando se utiliza el operador LBP para reconocimiento de texturas, la información del número de ocurrencias de códigos es condensada en un histograma. De este modo, la clasificación puede llevarse a cabo mediante cálculos convencionales de similaridades entre histogramas. Sin embargo, cuando se requiere describir una imagen no homogénea, como un rostro, la utilización de un histograma global produce una pérdida de información espacial.

Para soslayar esta dificultad, Ahonen et al. dividieron la imagen del rostro en varias

regiones locales y utilizaron los descriptores de textura para extraer información de cada región de manera independiente. Luego, los descriptores fueron concatenados para formar una descripción global, como se muestra en la figura 2.25.

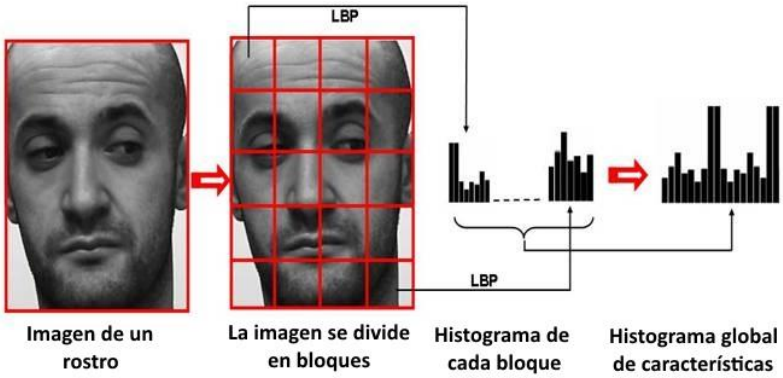


Figura 2.25. Descripción del rostro mediante patrones binarios locales.

De manera práctica, el histograma final de Ahonen contiene la descripción de 3 niveles diferentes de localidad: la evaluación LBP a nivel de pixeles, la suma de estas evaluaciones dentro de la pequeña región delimitada por el bloque, y finalmente la descripción global obtenida al concatenar los histogramas regionales. Es necesario destacar que las regiones no necesariamente deben ser rectangulares, tener el mismo tamaño y forma, o cubrir la imagen por completo. También es posible que las regiones se solapen parcialmente entre ellas, así como extender el método al dominio espacio-temporal [Zhao07].

Para llevar a cabo la clasificación, los histogramas de dos imágenes son comparados calculando la distancia Ji-cuadrada χ ponderada, la cual es definida como:

$$\chi_w^2(x,\xi) = \sum_{j,i} w_j \frac{(x_{i,j} - \xi_{i,j})^2}{x_{i,j} + \xi_{i,j}} \tag{2.6}$$

donde x y ξ son los histogramas normalizados a ser comparados, los índices i e j se refieren al i-esimo binario del histograma correspondiente a la j-esima región local y w_j es el peso de la región j.

En comparación con los métodos holísticos, el algoritmo de Ahonen es más robusto a cambios de pose debido a que no requiere la localización de los componentes del rostro de manera exacta, sino información regional mucho más flexible. Se ha demostrado [Zhan09] que el algoritmo de Ahonen es capaz de alcanzar tasas de reconocimiento perfecto en rostros con rotación inferior a los 15°. Sin embargo, cuando la rotación se incrementa, la división de la imagen el bloques resulta problemática, debido a la desalineación de las regiones (debido a que, por ejemplo, la región de una imagen frontal podría convertirse en fondo en una donde el rostro gire 45°).

2.4.3.3 Mapas de líneas de contorno (LEM)

La información de los contornos del rostro también puede ser utilizada para reconocimiento. El enfoque de mapas de líneas de contorno (Line Edge Map, LEM) fue propuesto inicialmente por Gao en 2002 [Gao02], y la idea consiste en proporcionar una medida de distancia entre dos mapas de contornos y realizar la comparación de rostros basándose en estas mediciones.

El mapa de bordes LEM de una imagen facial se obtiene de acuerdo a la siguiente secuencia (1) se extraen los bordes del rostro, (2) adelgazamiento de contornos (operación de morfología matemática conocida como 'esqueleto'), (3) ajuste poligonal de líneas. Un ejemplo de los resultados que se generan hasta este punto se muestra en la figura 2.26.

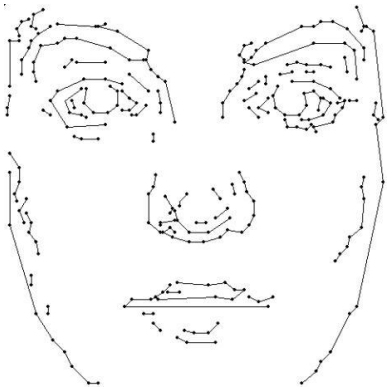


Figura 2.26. Bordes de un rostro obtenidos mediante el algoritmo LEM.

La medida de similitud entre dos rostros LEM, los autores definieron una nueva distancia de Haussdorf, la cual calcula la distancia entre dos segmentos como la raíz de suma de cuadrados (RSS) de tres componentes de distancia: distancia paralela, distancia de orientación y distancia perpendicular.

Posteriormente la distancia típica Hausdorff sobre conjuntos de puntos fue extendida a LEMs basados en la distancia definida individual de segmentos de líneas. Para llevar a cabo el reconocimiento, cada imagen de la galería de entrenamiento fue convertida a una imagen LEM para posteriormente ser comparada contra las LEMs de la galería de prueba utilizando la distancia de segmentos de línea de Haussdorff.

El mismo autor propuso una mejora a su algoritmo en 2005 [Gao05], sustituyendo los descriptores de características por Puntos de Esquinas Direccionales (abreviado DCP en inglés), los cuales detectan esquinas en la imagen (que no necesariamente corresponden a componentes faciales). Un DCP es representado mediante sus coordenadas cartesianas y dos atributos direccionales que apuntan a los puntos-esquina anterior y posterior. Aquí la distancia de dos DCP es medida calculando el costo de deformación mediante translación, rotación y operaciones

de apertura/cierre, así como promediando los costos mínimos como el resultado de disimilitud.

El algoritmo de DCP demostró ser robusto a variaciones de iluminación; una característica heredada de los mapas de contorno, puesto que una esquina en general puede ser considera como un borde de bordes. Sin embargo, tanto DCP como LEM mostraron ser sumamente sensibles a cambios de pose, tales como rotaciones de profundidad, puesto que estas generaron distorsiones significativas en los mapas de contornos.

2.4.3.4 Modelos de Apariencia Activa (AAM)

Los Modelos de Apariencia Activa (Active Appearance Models: AAM) son modelos estadísticos de forma y apariencia propuestos inicialmente por Cootes et al. en 1998 [Coot98] [Coot01].

Los AAM se construyen durante la fase de entrenamiento utilizando las imágenes base y un conjunto de coordenadas que indican los puntos de referencia en cada imagen (fig 2.27).

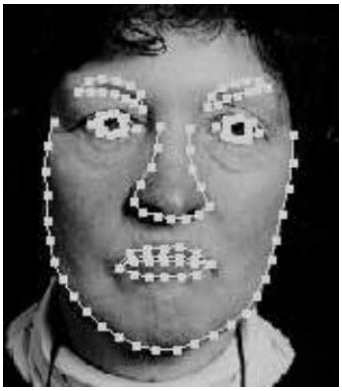


Figura 2.27. Ejemplo de un rostro marcado con 122 puntos de marcaje manual.

Está técnica está estrechamente relacionada con los Modelos de Forma Activa (Active Shape Model, ASM), propuestos por el mismo autor en 1995 [Coot95]; pero a diferencia de estos últimos, los AAM si toman en cuenta la textura de la imagen (apariencia), y no únicamente la forma.

Existen 2 tipos de modelos AAM; los que utilizan los modelos de apariencia y forma de manera separada (llamados independientes) y los que combinan ambos utilizando un único conjunto de parámetros lineales (llamados AAM combinados).

Los modelos AAM independientes definen matemáticamente la forma s como el conjunto de coordenadas de los v vértices que construyen la malla (fig. 2.28) de la siguiente manera:

$$s = (x_1, y_1, x_2, y_2, \dots, x_v, y_v)^T \tag{2.7}$$

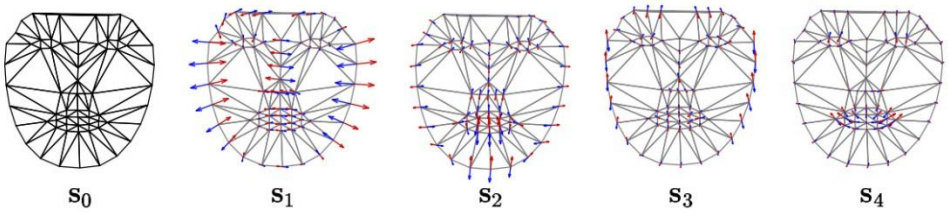


Figura 2.28. Modelos lineales de forma en un AAM independiente. Cada modelo consiste de una malla base triangular S_0 más una combinación lineal de n vectores de forma s_i.

Para permitir variaciones lineales de forma, se consideran combinaciones lineales s_i que son añadidas a la forma base expresada en la ecuación 2.7, por lo que el modelo de forma queda definido como:

$$s = s_0 + \sum_{i=1}^{n} p_i s_i \tag{2.8}$$

Donde p_i representa los parámetros de la forma. Al igual que la forma, el modelo de apariencia $A(x)$ también permite variaciones utilizando una combinación de una apariencia base $A_0(x)$ más un conjunto de m imágenes de apariencia $A_i(x)$:

$$A(x) = A_0(x) + \sum_{i=1}^{m} \lambda_i A_i(x) \tag{2.9}$$

Dónde los coeficientes λ_i representan los parámetros de apariencia y donde toda x denota un conjunto de pixeles $x = (x, y)^T$ que yacen dentro de la malla base s_0. Un ejemplo de un modelo de apariencia se puede apreciar en la figura 2.29.

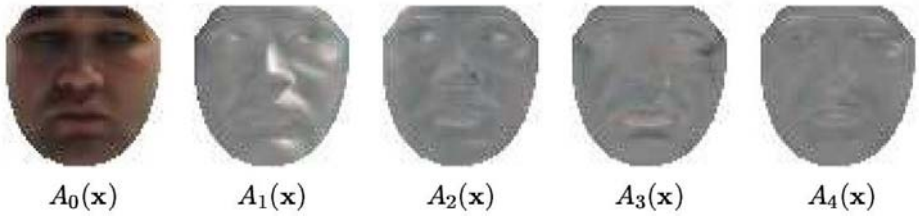


Figura 2.29. Modelos lineales de apariencia en un AAM independiente. Los modelos consisten de una imagen base A_0 más una combinación lineal de m imágenes de apariencia A_i.

Tanto el modelo de apariencia como el de forma hacen uso del Análisis de Componentes Principales para calcular los modelos base y sus combinaciones lineales.

Los AAM combinados utilizan un mismo conjunto de parámetros $c = (c_1, c_2, \dots, c_l)^T$ para ambos modelos; a diferencia de los independientes, que usan parámetros distintos (p y λ). De esta forma ambos modelos queda acoplados y se expresan de la siguiente manera:

$$s = s_0 + \sum_{i=1}^{l} c_i s_i \tag{2.10}$$

$$A(x) = A_0(x) + \sum_{i=1}^{l} c_i A_i(x) \tag{2.11}$$

Lo anterior permite a los AAM combinados representar la misma imagen con menos parámetros, haciendo la representación más eficiente y estricta, aunque con la desventaja de dificultar más el ajuste.

Una vez calculados ambos modelos, el algoritmo lleva a cabo el reconocimiento calculando las diferencias entre el rostro a identificar y los modelos obtenidos haciendo uso del método de mínimos cuadrados.

Capítulo 3

Bases de datos para experimentación.

Desde comienzos de la historia de la visión computacional se ha hecho notable la necesidad de disponer de conjuntos de imágenes que posean la mayor amplitud, control y estandarización posible con el propósito de llevar a cabo experimentos y comparaciones de manera fiable y precisa.

En el caso de las bases de datos utilizadas para la detección y reconocimiento de rostros humanos es necesario señalar que esta necesidad es especialmente evidente debido a la gran cantidad de posibles variaciones que pueden afectar la apariencia de un mismo individuo, tales como cambios de pose, iluminación, expresiones faciales, oclusiones, cabello/peinado, accesorios o prendas, maquillaje y, por supuesto, la edad.

Debido a esto, se han creado una gran cantidad de bases de datos tanto públicas como privadas que se han puesto a disposición del sector académico y científico por un lado, y del sector industrial y comercial por el otro.

Dichas bases de datos han sido creadas bajo diferentes condiciones, metodologías de adquisición y distribución, y, más importante que todo, con distinto grado de precisión y control. Las bases de datos pueden contener únicamente las imágenes 2D, o pueden proporcionar adicionalmente (o exclusivamente) información tridimensional. También existen bases que proporcionan un medio adicional para la identificación, como la voz. Además existen otras dedicadas al reconocimiento de rostros en imágenes dinámicas, es decir, video. Esto en parte porque las bases de datos pueden diferir en el o los propósito que persiguen, ya que algunas

fueron creadas únicamente la detección de rostros humanos, otras únicamente para el reconocimiento de rostros humanos, otras para la identificación de emociones mediante las expresiones faciales, y por supuesto, otras que persiguen varios o todos estos fines.

Debido a todos estos factores, es necesario hacer un análisis y comparación de diversas bases de datos de imágenes de rostros para elegir la más adecuada a la intención que perseguimos.

En nuestro caso, estamos interesados exclusivamente en una base de datos de imágenes estáticas en 2D que posean una amplia variedad de poses controladas, de preferencia con giros en los 3 ejes coordenados y que sean públicas y disponibles para la publicación de resultados.

3.1 Bases de Datos Estáticas Bidimensionales.

3.1.1 Base de datos ORL

La base de datos ORL (Olivetti Research Laboratory) fue creada entre 1992 y 1994 [Sama94] en los laboratorios de AT&T de la Universidad de Cambridge, Inglaterra. Está formada por 10 imágenes distintas de 40 individuos (35 hombres y 5 mujeres) que fueron fotografiados sobre un fondo negro. Las tomas fueron realizadas en distintos periodos de tiempo y contienen pequeñas variaciones en iluminación, expresiones (neutrales, con o sin lentes, sonriendo o no sonriendo) y detalles faciales (con o sin lentes). Sin embargo, cabe mencionar que estas variaciones no fueron hechas de manera sistemática y controlada, por lo que están presentes de manera desordenada en algunos conjuntos solamente.

Figura 3.1. Ejemplo de un set completo de tomas de un individuo de la base de imágenes ORL. Contiene ligeras variaciones en postura.

Cada imagen fue cuidadosamente recortada para dejar exclusivamente el rostro del

individuo. Debido a esto y a que las variaciones incluidas son muy ligeras, esta base de datos ha sido ampliamente superada por diversos algoritmos, por lo que se considera saturada para propósitos de significación estadística [Phil02] y por lo tanto, es prácticamente obsoleta, aunque sigue siendo considerada como referente histórica en el área.

Actualmente se conoce como "The Database of Faces". Un ejemplo de un conjunto completo de tomas de un individuo se muestra en la figura 3.1.

3.1.2 Base de datos Yale

Esta base de datos fue creada en la Universidad de Yale por el investigador Peter N. Belhumeur en 1997 [Belh97]. Consta de 15 individuos (14 hombres y una mujer) con 11 tomas por cada uno. Las 165 imágenes totales están en escala de gris con formato GIF y cuentan con una resolución de 320 x 243 pixeles.

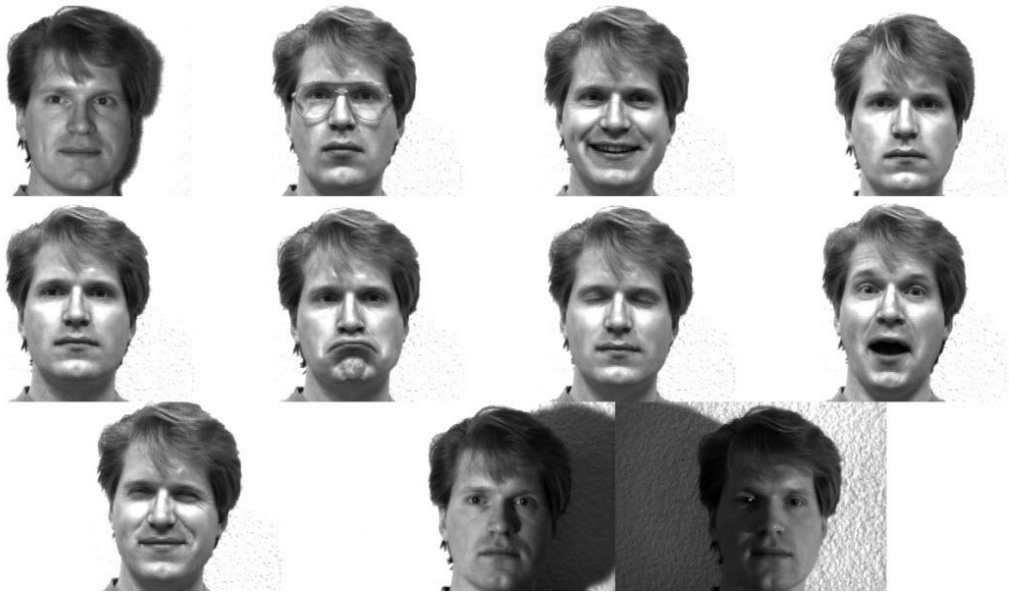


Figura 3.2. Ejemplo de imágenes para un individuo de la base de datos Yale. Podemos apreciar variaciones en expresión facial, iluminación y cambio de detalles (gafas).

Cada uno de los conjuntos se compone de una toma normal, 6 cambios de expresiones faciales (neutral, alegre, triste, somnoliento, sorprendido y haciendo un guiño), 3 cambios de iluminación (luz central, luz desde la derecha, luz desde la izquierda) y finalmente una toma donde los individuos utilizan anteojos. Todas las tomas se realizaron de frente hacia la cámara, por lo que esta base no incluye variaciones de pose.

Al igual que la base de datos ORL es una base de datos básica, pero también es una referencia histórica en el área de reconocimiento. En la figura 3.2 puede apreciarse un conjunto completo de tomas como muestra.

3.1.3 Base de datos Yale Extendida

La Universidad de Yale creó una segunda base de datos bajo la dirección de Athinodoros S. Georghiades y Peter N. Belhumeur en 2001 [Geor01]. Consta de 16,128 imágenes de 28 sujetos con 9 poses y 64 condiciones de iluminación para cada una. Cada imagen está en escala de grises y tiene una resolución de 640 x 480 pixeles codificados en formato PGM. Adicionalmente, se incluye un conjunto de imágenes recortadas que muestran exclusivamente la parte central del rostro del individuo, para tareas específicas de reconocimiento.

Figura 3.3. Imágenes recortadas de ejemplo pertenecientes a la base de datos Yale extendida. Se aprecian notablemente los cambios en la fuente luminosa.

Cada una de las 64 tomas con iluminación distinta para cada pose fue capturada en un tiempo de alrededor de 2 segundos de diferencia [Gros05], por lo que los movimientos involuntarios del rostro y las expresiones faciales distintas son prácticamente inexistentes.

Las 9 poses consideradas se distribuyen de la siguiente manera: una frontal, 5 poses a 12° y 3 poses a 24° con respecto al eje de la cámara. Finalmente, la base de datos está dividida en cuatro subconjuntos de acuerdo al ángulo existente entre la fuente de luz y el eje de la cámara (12°, 25°, 50° y 77°). En la figura 3.3 podemos apreciar algunas imágenes de muestra del conjunto de rostros recortados.

3.1.4 Base de datos CAS-PEAL

Esta base de datos fue creada por la Academia China de Ciencias (Chinese Academy of Sciences, CAS) con la finalidad de proporcionar la mayor cantidad de condiciones posible. Se consideraron variaciones de pose, expresión, accesorios e iluminación (Pose, Expression, Accesory, Lighting). Las capturas se realizaron entre Agosto de 2002 y Abril del 2003.

La base de datos contiene imágenes de entre 66 y 1040 individuos (595 hombres y 445 mujeres) distribuidos en siete categorías: pose, expresión, accesorios, iluminación, fondo, distancia y tiempo [Gao04]. La distribución de estos individuos y la cantidad de tomas puede apreciarse en la tabla 3.1.

Número de sujetos	Condiciones		Resolución	No. De Imágenes
1040	Pose	21		
377	Expresiones Faciales	6		
438	Accesorios	6		
233	Iluminación	9-15	360x480	30,900*
297	Fondo	2-4		
296	Distancia	1-2		
66	Tiempo	2		
*Originalmente se capturaron un total de 99,594 imágenes, de las cuales están disponibles bajo demanda 30,900.				

Tabla 3.1. Distribución de las imágenes de la base de datos CAS-PEAL

Las 6 expresiones faciales que se consideraron son neutral, sonrisa, disgusto (ceño fruncido), sorpresa, ojos cerrados y boca abierta. Las variaciones de accesorios o detalles faciales que se consideraron fueron 3 tipos de lentes y 3 tipos de sombreros o gorros. Las variaciones de fondo fueron entre 2 y 4 cortinas distintas de color uniforme. Las variaciones de tiempo consideradas fueron 2 tomas distintas con un periodo de separación de 6 meses entre una y otra.

Las variaciones de iluminación consideraron iluminación natural constante junto con 15 lámparas fluorescentes colocadas en rangos de ±90°, ±45° y 0° en azimuth con ±45° y 0° en elevación. Las 15 diferentes tomas con cada una de las fuentes lumínicas tardaron alrededor de 2 minutos en total [Gros05], por lo que se observan pequeñas variaciones entre ellas.

Finalmente, para el subconjunto de poses, se utilizaron 9 cámaras dispuestas en semicírculo con una separación de 22.5° entre sí, por lo que los ángulos considerados fueron ±90°, ±67.5, ±45°, ±22.5 y 0°. Se tomaron 9 imágenes en posición frontal de cada individuo y adicionalmente se le pidió a cada individuo voltear la cabeza hacia arriba y hacia abajo en ángulos de 30°, por lo que en total, se tomaron 27 fotografías en distinta pose, de las cuales están disponibles públicamente 21 de ellas (faltan las tomas de ±90°). El conjunto completo disponible de un individuo puede observarse en la figura 3.4.

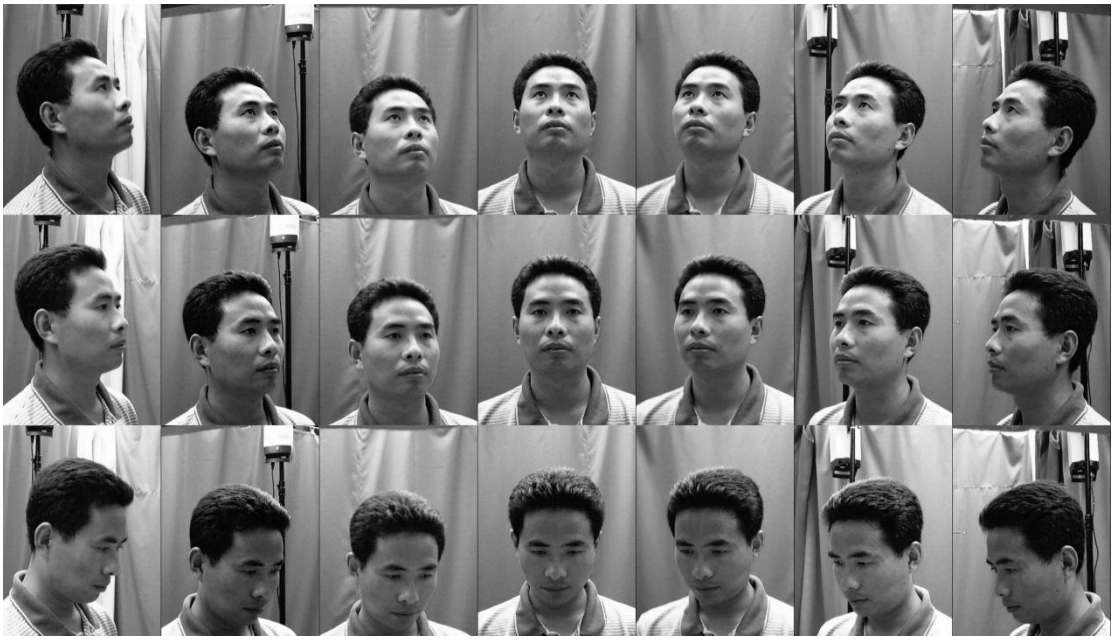


Figura 3.4. Conjunto disponible de las variaciones de pose para un individuo de la base de datos CAS-PEAL. Se pueden apreciar claramente las rotaciones sobre los ejes X e Y, así como los cambios de iluminación en el rostro por efectos de las sombras. Nótese también la ausencia de un fondo homogéneo.

3.1.5 Base de datos FERET

La base de datos FERET (Facial Recognition Technology) fue reunida en la Universidad George Mason y las instalaciones del US Army Research Laboratory como parte del programa FERET, el cuál fue patrocinado por el US Department of Defense Counterdrug Technology Development Program [Phil98, Phil00]. Dado que esta base de datos ha sido extensamente utilizada por diversos grupos de investigación y evaluadores independientes, y a que los subconjuntos de muestra se distribuyen junto con la base de datos, la comparación directa entre el desempeño de un reconocedor determinado y los resultados previos publicados es posible. Hasta el 2005, la base de datos había sido distribuida a más de 460 grupos de investigación [Gros05].

Las imágenes fueron tomadas en 15 sesiones entre Agosto de 1993 y Julio de 1996. El conjunto proporcionado originalmente constaba de 14,051 imágenes que tenían una resolución de 256 x 384 pixeles, pero posteriormente el NIST (National Institute of Standards and Technology) creó imágenes a color con resoluciones más altas (512 x 768) de la mayoría de las imágenes en escala de gris originales (no proporcionadas con la base de datos original para distribución).

La documentación de la base de datos lista 24 categorías de imágenes. Las categorías fa y *fb* fueron tomadas en sucesión inmediata y tienen variaciones de expresión facial típicamente

sutiles, tales como una sonrisa. Las imágenes de la categoría *fc* fueron capturadas con diferente cámara y con cambios de iluminación.

Figura 3.5. Ejemplo de un set parcial de tomas de una participante de la base de imágenes FERET. En las 10 primeras tomas podemos ver variaciones de pose y expresión facial, mientras que en las 10 tomas siguientes apreciamos una nueva sesión completa tras cierto lapso de tiempo, en la que también se observan variaciones de pose y expresión.

Posteriormente se crearon varios conjuntos duplicados para registrar variación en el tiempo. El subconjunto *dupI* contiene duplicados de conjuntos con entre 0 y 1031 días transcurridos entre las dos sesiones (promedio 251 días y mediana 72). Un segundo subconjunto, *dupII*, tiene al menos 18 meses de diferencia entre las dos sesiones (promedio 627 días, mediana 569).

Las categorías restantes cubren un amplio rango de variaciones de pose. Las categorías comprendidas entre *ba* y *bi* constan de 200 individuos y fueron capturadas pidiéndole al individuo rotar la cabeza y el cuerpo en ángulos con un rango comprendido entre -60° y +60°.

Adicionalmente, un subconjunto de poses distintas fue etiquetado con las categorías *pr*, *pl*, *qr*, *ql*, *hr* y *hl* para denominar a las poses con perfil derecho, perfil izquierdo, cuarto de perfil derecho, cuarto de perfil izquierdo, medio perfil derecho y medio perfil izquierdo, respectivamente.

La información proporcionada con la base de datos incluye la fecha de captura de la imagen, si el individuo usaba o no gafas (para todos los conjuntos tomados) así como la ubicación, manualmente determinada, de los ojos y el centro de la boca (para 3816 imágenes únicamente). En la figura 3.5 podemos apreciar algunas imágenes de muestra de la base de datos.

3.1.6 Base de datos FEI

Esta base de datos brasileña contiene un conjunto de imágenes tomadas entre Junio de 2005 y Marzo de 2006 en el Artificial Intelligence Laboratory de la Faculta de Ingeniería Industrial (Faculdade de Engenharia Industrial, FEI) de São Bernardo do Campo, São Paulo, Brasil [Oliv06].

Figura 3.6. Ejemplo de un set parcial de tomas de una participante de la base de imágenes FEI. En el vemos 10 variaciones de pose (rotaciones en el eje Y). También observamos 2 variaciones de expresión facial: neutra y sonriente.

Contiene 14 imágenes distintas de 200 individuos diferentes (100 hombres y 100 mujeres con edades de entre 19 y 40 años), para un total de 2800 imágenes. Cada imagen tiene una resolución de 640 x 480 pixeles y se encuentran en formato JPG. Todas las imágenes están

completamente a color y fueron tomadas sobre un fondo blanco homogéneo.

Las variaciones incluyen 10 cambios de pose, 2 de expresión facial y 2 de iluminación. Los cambios de expresión facial considerados fueron una toma neutra y una con el individuo sonriendo. Los cambios de iluminación presentes en cada conjunto son una disminución tenue y una pronunciada en la iluminación.

Finalmente, los cambios de pose considerados comprenden una variación progresiva de aproximadamente 18° entre cada toma. Podemos observar un set parcial de una de las voluntarias fotografiadas en la figura 3.6.

3.1.7 Base de datos GTAV

La base de datos GTAV (Audio Visual Technologies Group) fue creada con el propósito específico de poner a prueba la solidez de los algoritmos de reconocimiento de rostro frente a cambios pronunciados de pose e iluminación. El grupo que lo creó pertenece a la Universitat Politècnica de Catalunya, en Barcelona, España. Consta de 1,767 fotografías de 44 individuos (27 hombres y 17 mujeres) en formato BMP a todo color y con una resolución de 320 x 240 pixeles.

Figura 3.7. Ejemplo de un set completo de tomas de un individuo de la base de imágenes GTAV con iluminación ambiente. Apreciamos rotaciones en un rango de 180º sobre el eje coordenado Y. De izquierda a derecha y de arriba hacia abajo, rotaciones en: 0º, -30º, -45º, -60º, -90º, 30º, 45º, 60º y 90º

Las variaciones de pose consideradas son 9 rotaciones en el eje coordenado Y en los siguientes grados: 0º, -30º, -45º, -60º, -90º, 30º, 45º, 60º y 90º. Cada una de las tomas fue numerada en este orden específico. Para cada una de las poses se realizaron 3 tomas con distinta iluminación, siendo la primera de ellas la luz natural o ambiente; la segunda, "luz fuerte" procedente de un ángulo de 45º y la tercera, luz "medianamente fuerte" procedente de una fuente casi frontal al individuo.

Adicionalmente, los creadores incluyeron al menos 10 fotografías extra con distintas variaciones en expresión facial (sonriendo, mostrando la lengua, boca abierta, haciendo guiños, etc.), vestimenta (anteojos, gafas de sol), oclusiones (rostro cubierto por el cabello y/o una o dos manos) y/o poses extras (cabeza inclinada hacia abajo, a lo lados o levantada).

Figura 3.8. Ejemplo de un set parcial de las tomas adicionales de la base de imágenes GTAV. Apreciamos variaciones de vestimenta, oclusiones, expresiones faciales, poses e iluminación, de manera desordenada.

Además, se alternó la iluminación para añadir variedad en las tomas. Sin embargo, cabe hacer notar que estas tomas adicionales carecen de un orden regular y una metodología fija. En las figuras 3.7 y 3.8 podemos apreciar varias muestras de esta base de datos.

3.1.8 Base de datos FRAV2D

Esta base de datos fue creada por el laboratorio FRAV (Face Recognition and Artificial Vision Group) de la Universidad Rey Juan Carlos de Madrid, España. Este grupo de investigación creó una base de datos multimodal que consta de 3 tipos de imágenes. El primer tipo está conformado por imágenes tridimensionales representadas mediante un mallado triangular capturado mediante un escáner láser Minolta. El segundo tipo está conformado por imágenes de rango o mapas de profundidad, mientras el tercero contiene imágenes con información bidimensional o de textura. Los creadores denominan a estos 3 tipos de imágenes 3D, 2.5D y 2D, respectivamente [Cond06].

Las imágenes fueron adquiridas bajo condiciones controladas de pose e iluminación. Para el control de la pose se colocaron marcas en el lugar donde se tomaron las fotografías para posteriormente indicar a cada individuo la dirección en la que tenía que girar la cabeza. Para la iluminación se contemplaron dos tipos: controlada y no controlada. En la primera se situó la cara del individuo entre un par de haces de luz provenientes de focos de halógeno, mientras que para la segunda se colocó adicionalmente una fuente lumínica cenital fluorescente, la cual es variable.

La base de datos consta de imágenes de 105 individuos (81 hombres y 24 mujeres) de raza caucásica. El período de las tomas abarcó 10 meses, desde septiembre del 2004 hasta junio del 2005, y cada individuo tiene el mismo número de imágenes (16 de cada tipo), adquiridas en idénticas condiciones. Todas las imágenes fueron adquiridas teniendo el sujeto los ojos cerrados, para garantizar una absoluta seguridad frente al láser. Ningún individuo usó sombreros, gafas u otra prenda o detalle que alterara el aspecto facial.

De las 16 imágenes, se tomaron 4 imágenes frontales, ocho con giros en diferente sentido y grado, dos con gestos y dos con iluminaciones diferentes. En cada captura se consideró una sola variación para permitir el análisis de la influencia de cada factor por separado. Cabe señalar también que los giros se realizaron sobre cada uno de los 3 ejes coordenados de manera separada.

Resolución del mallado	Tasa de reducción	Número de puntos
r1	1/1	18.535
r2	1/4	4.657
r3	1/9	2.060
r4	1/16	1.161

Tabla 3.2. Niveles de resolución del mallado proporcionados por el escáner

La resolución de las imágenes bidimensionales o de textura es de 400x400 pixeles, mientras que para las imágenes de rango es de 200x200 pixeles. Para las imágenes tridimensionales, se capturaron 4 distintos niveles de resolución, detallados en la tabla 3.2. El formato de salida utilizado es mapa de bits (BMP).

Finalmente, señalaremos que debido a la disponibilidad, afinidad y regularidad en las condiciones de captura, esta base de datos será considerada como la principal para la presente tesis. En la figura 3.9 podemos apreciar las imágenes de un set bidimensional completo de esta base de datos.

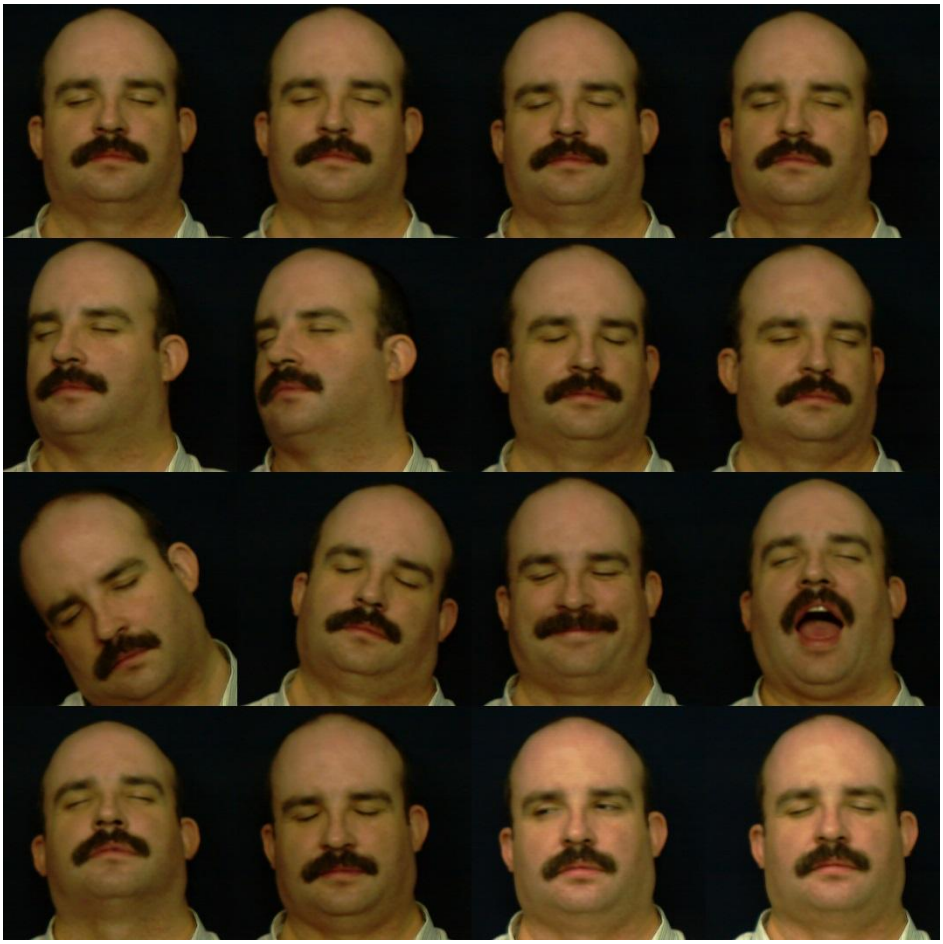


Figura 3.9. Ejemplo de un set completo de tomas de un individuo de la base de imágenes FRAV 2D. Apreciamos variaciones ligeras y pronunciadas en la pose del individuo (giros en los 3 ejes coordenados), en las expresiones faciales (sonrisa y boca abierta) y en las condiciones de iluminación.

3.1.9 Base de datos Achermann

La base de datos Achermann (también conocida como base de datos Berna), fue creada por el laboratorio de Visión Computacional e Inteligencia Artificial (FKI por sus siglas en alemán) de la Universidad de Berna, Suiza.

La base consta de 312 imágenes de 30 individuos en 10 poses distintas (las 12 fotografías adicionales incluyen gestos y expresiones adicionales, pero se encuentran distribuidas de manera irregular). Las imágenes se encuentran en formato RAS (graficos raster propietario de Sun Microsystems) en escala de grises con 8 bits de profundidad y una resolución de 512 x 342 pixeles.

Figura 3.10. Ejemplo de un set completo de tomas de un individuo de la base de imágenes Acherman. En este conjunto podemos apreciar, vistas de arriba hacia abajo y de izquierda a derecha; variaciones de pose en cabeceo (últimas 4 imágenes) y guiñada/rotación en profundidad (imágenes 3-6).

Todas las tomas se realizaron sobre un fondo blanco y con iluminación controlada. El conjunto completo de tomas de un individuo está constituido de la siguiente manera:

- 2 tomas frontales (imágenes 1 y 2).
- 2 tomas con rotación de ≈ 20° hacia la izquierda (imágenes 3 y 4).
- 2 tomas con rotación de ≈ 20° hacia la derecha (imágenes 5 y 6).
- 2 tomas con rotación de ≈ 20° hacia arriba (imágenes 7 y 8).
- 2 tomas con rotación de ≈ 20° hacia abajo (imágenes 9 y 10).

En varias tomas los individuos tienen además ligeras variaciones de expresiones faciales, pero estás no están controladas y se presentan aleatoriamente en todo el conjunto. Se puede observar un set completo en la figura 3.10.

Debido a la disponibilidad de esta base de datos y su regularidad en los ángulos de captura, esta base de datos también será utilizada a fin de ampliar la comparativa entre los algoritmos generales de Eigenfaces [Gao02], LEM [Gao02] y DCP [Gao05], así como el algoritmo de recuperación de pose cilíndrica tridimensional [Gao01] , el cual cuenta con una técnica de compensación de pose.

3.2 Análisis comparativo

A manera de resumen se ofrece una tabla comparativa donde se pueden apreciar la cantidad de individuos e imágenes totales que se capturaron, la resolución de las imágenes y las variaciones consideradas.

Base de Datos	Individuos	Imágenes	Resolución	Variaciones	
ORL	40	400	92 x 110	Pose Expresiones Faciales Accesorios	IC IC IC
Yale	15	165	320 x 243	Iluminación Expresiones Faciales Accesorios	3 6 1
Yale Extendida B	28	5,850	640 x 480	Pose Iluminación	9 64
CAS-PEAL	66-1040	30,900	360 x 480	Pose Iluminación Expresiones Faciales Accesorios Fondo Distancia Tiempo	21 9-15 6 6 2-4 1-2 2
FERET	1199	14,051	256 x 384	Pose Iluminación Expresiones Faciales Tiempo	9-20 2 2 2
FEI	200	2,800	640 x 480	Pose Iluminación Expresiones Faciales	10 2 2
GTAV	44	1,767	320 x 240	Pose Iluminación Expresiones Faciales Accesorios Oclusiones	9 3 IC IC IC
FRAV	105	1,680	400 x 400 200 x 200	Pose Iluminación Expresiones Faciales	8 2 2
Achermann	30	312	512 x 342	Pose	8

Tabla 3.3. Tabla comparativa de las bases de datos consideradas para la experimentación. Las siglas IC (Irregularmente Capturada) indican que si bien las variaciones están presentes, no fueron capturadas y ordenadas de manera regular y sistemática.

Capítulo 4

Redes Neuronales.

4.1 Introducción

Dentro de la Inteligencia Artificial (IA) existen dos enfoques principales: el enfoque descendente y el enfoque ascendente. El enfoque descendente (top-down, en la bibliografía especializada) pretende emular funciones específicas de la mente humana mediante la utilización de extensas cantidades de conocimiento representado y manipulado mediante un conjunto exhaustivo de reglas lógicas pre-programadas. Un ejemplo clásico de este enfoque son los Sistemas Expertos. El enfoque ascendente (bottom-up en inglés), por el contrario, busca imitar el funcionamiento del cerebro humano desde su base mecánica, reproduciendo las propiedades observables de los organismos biológicos. Su estrategia fundamental es permitir que el sistema se desarrolle a partir de su propia experiencia – al igual que lo hacen los mecanismos biológicos - haciendo hincapié en los aspectos perceptivos del ser humano, como el aprendizaje, la generalización, las capacidades asociativas y sensoriales, la autonomía y el reconocimiento de patrones. Los ejemplos más comunes y representativos de este enfoque son los Algoritmos Evolutivos y las Redes Neuronales.

Ambos enfoques se desempeñan mejor que el otro frente a distintos problemas. Así, mientras el enfoque descendente (propio de la IA clásica y de las computadoras convencionales) se desempeña mejor en tareas claramente definidas, precisas, y altamente repetibles (tales como el razonamiento formal o el cálculo), el enfoque ascendente lo supera en tareas donde la información a procesar es masiva, redundante e imprecisa, como el reconocimiento de patrones,

la percepción, el control, etcétera. El primer grupo de tareas es denominado comúnmente "de alto nivel" y generalmente pertenece al entorno intelectual humano, mientras el segundo se denomina "de bajo nivel" y es más propio del mundo real donde se desenvuelven todos los organismos biológicos [Mart07]. En la tabla 4.1 se detallan las diferencias entre ambos enfoques.

	Enfoque descendente	Enfoque ascendente
Ejemplos	IA Clásica, computadoras	Redes neuronales, Algoritmos evolutivos
Bases	Psicología, Lógica	Biología
Filosofía	Que hace el cerebro	Como lo hace el cerebro
Funcionamiento	Reglas lógicas definidas	Generalización a partir de ejemplos
Desarrollo	Programación	Entrenamiento
Campos de dominio	Lógica, conceptos, reglas	Reconocimiento de patrones, gestalt
Arquitecturas	Von Neumann, separación hardware/software	Paralelas, distribuidas, adaptativas, autoorganizantes

Tabla 4.1. Características y diferencias destacables entre los enfoques de la IA.

A modo de ejemplo ilustrativo de las diferentes áreas de competencia de estos enfoques, consideremos el hecho de que en la actualidad se han desarrollado sistemas expertos capaces de derrotar al ser humano en ajedrez (Deep Blue) y concursos de conocimiento (Watson), además de realizar otras tareas, como los diagnósticos médicos (MYCIN, Caduceus) de manera aceptable, en tanto que no se ha conseguido crear un sistema que sea capaz de identificar una mosca y atraparla en vuelo, tal como es capaz de hacerlo sin mayor esfuerzo una simple rana.

Este ejemplo demuestra de manera muy clara la ventaja que tienen los mecanismos naturales, por más simples que sean, sobre las más avanzadas herramientas de cómputo, cuando hablamos de procesar información tan irregular como la que captamos de nuestro entorno cotidiano.

Es debido a lo anterior que en áreas como el procesamiento de imágenes y la visión artificial las redes neuronales son usadas de manera extensiva, existiendo una gran cantidad de

algoritmos y metodologías de implementación; muchos de ellos orientados a resolver un problema específico de estas áreas.

No obstante la abundancia de tipos de redes neuronales existentes, podemos decir que gran mayoría pueden ser clasificados en modelos más reducidos y comunes; en base al tipo de neurona que utilizan, su arquitectura o topología de conexión y su algoritmo de aprendizaje. En la figura 4.1 se muestran varios modelos de clasificación, así como algunas de las redes más representativas [Mart07].

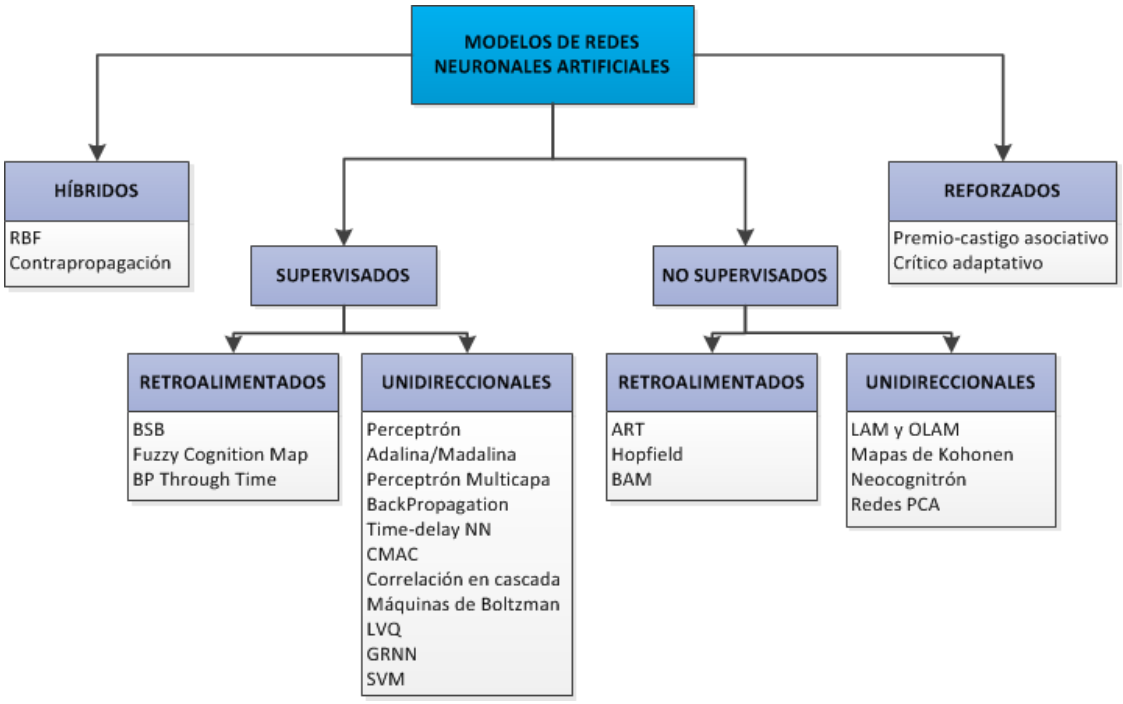


Figura 4.1. Clasificación de las redes neuronales en base a su tipo de aprendizaje y arquitectura

Como se señaló con anterioridad, muchas de estas redes neuronales fueron diseñadas para resolver problemas específico, tales como el reconocimiento de caracteres impresos y manuscritos, el reconocimiento de retina y de huellas dactilares para biometría, la identificación de objetos y escenas para navegación robótica, etc.

En el objetivo que nos atañe, la identificación de rostros humanos, existen también diversas redes neuronales diseñadas específicamente para cada uno de los procesos que lo conforman. Así por ejemplo, existen redes encargadas de optimizar el brillo, el contraste, etc. en la etapa de pre-procesamiento, redes para llevar a cabo únicamente la detección del rostro y finalmente, redes para llevar a cabo la clasificación de los mismos en el proceso de análisis de la imagen.

Muchas de las redes neuronales presentadas en la figura 4.1 han sido aplicadas con distinto grado de éxito al problema de reconocimiento de rostros. Presentar un estudio pormenorizado de cada uno de ellas excede el alcance del presente trabajo. No obstante, el presente capítulo presenta una perspectiva general del funcionamiento de algunas de las más destacadas, así como algunos trabajos relacionados que influyeron en el desarrollo de la red neuronal en la que se centra esta tesis.

4.2 El perceptrón de Rosenblatt

Este modelo neuronal fue introducido por Frank Rossenblatt a finales de la década de los cincuenta [Rose62]. Su estructura se inspira en las primeras etapas de procesamiento de los sistemas de percepción animal (por ejemplo, la visión), en los cuales la información va atravesando sucesivas capas neuronales que realizan un procesamiento de más alto nivel de manera progresiva.

El perceptrón es un modelo unidireccional y consta de una capa de entrada o sensores (capa S), una intermedia o asociativa (capa A) y finalmente, una capa de salida o reacción (capa R). En términos generales, cada una de estas capas funciona de la siguiente manera:

- **Capa S**. Si hiciéramos una analogía con el sistema de visión biológico, las neuronas de esta capa se corresponderían con las células fotorreceptoras de la retina. En términos técnicos, la función de la capa S es leer la imagen de entrada, es decir, convertir la información de la imagen en datos que puedan ser procesados por las capas superiores.
- **Capa A**. Esta segunda capa es el subsistema encargado de realizar la extracción de características de la imagen. Las neuronas de la capa A están conectadas a la capa de entrada (S) mediante conexiones no entrenables seleccionadas aleatoriamente y cuyos pesos pueden ser positivos (con valor de 1) o negativos (con valor de -1). El conjunto de conexiones entre estas dos capas es lo que se considera el extractor de características. En lo referente a los estados de las neuronas de esta capa, estos pueden ser activos (con valor de 1) o inactivos (con valor igual a 0).
- **Capa R**. Finalmente, la función de la tercera capa (capa R) es realizar la clasificación propiamente dicha de la imagen y ofrecerla como salida del sistema. Para lograrlo, la capa R se conecta con la capa asociativa mediante conexiones que se modifican durante el proceso de entrenamiento hasta alcanzar un óptimo ajuste. En esta capa, cada neurona corresponde a cada una de las clases consideradas.

Podemos apreciar un diagrama simplificado del perceptrón de Rosenblatt en la fig. 4.2

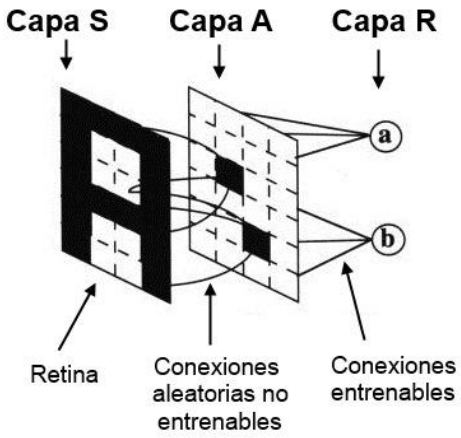


Figura 4.2. Diagrama simplificado del perceptrón de Rosenblatt.

La función de activación de las neuronas de la capa de salida es de tipo escalón, también conocida como función de Heaviside. De esta manera, su operación matemática puede ser descrita de acuerdo a la siguiente fórmula:

$$y_i = H(\sum_{j=1}^{n} w_{ij}x_j - \theta), \qquad \forall i, 1 \leq i \leq m \tag{4.1}$$

Dónde w_{ij} representa los pesos de las conexiones entre las neuronas x_j hacia las neuronas i de la capa de salida, menos el umbral (con frecuencia opcional) θ, encargado de restar potencial a la función de activación.

No obstante su éxito inicial, el perceptrón original de Rosenblatt, constituido por solo dos capas, pronto mostró grandes carencias; como el hecho de que únicamente permite discriminar clases que sean linealmente separables. Para superar esta limitante, en la década de los 80s se creó una nueva versión en la que se añadían capas de adicionales de perceptrónes, conocidas como capas ocultas, dando origen al perceptrón multicapa (MLP, por sus siglas en inglés).

4.3 El perceptrón multicapa (MLP)

El perceptrón multicapa es una evolución del perceptrón simple de Rosenblatt. Su origen tuvo lugar debido a una publicación de Papert y Minksy, en 1969, en la que demuestran matemáticamente que tanto el perceptrón simple como las redes ADALINE son incapaces de resolver problemas no lineales, tales como la operación booleana OR exclusivo o XOR. La combinación de varios perceptrones en capas superpuestas permitía resolver operaciones no separables linealmente, pero no existía un mecanismo definido para adaptar los pesos de la capa

oculta. Debido a este hecho, la comunidad científica abandonó casi completamente la investigación no únicamente sobre el perceptrón, sino de las redes neuronales en general durante casi 2 décadas.

Un descubrimiento fundamental para renovar el interés de la comunidad científica en las RNAs fue el algoritmo de retropropagación de errores o backpropagation (BP), el cual es comúnmente utilizado para realizar el entrenamiento del MLP y una gran cantidad de sus variantes.

Dicho algoritmo fue introducido inicialmente por Paul Werbos en sus tesis doctoral [Werb74], pero no tuvo un gran impacto de manera inmediata. Fue hasta 1984 cuando el algoritmo fue redescubierto por David Parker y de manera separada por Rumelhart, Hinton y MacClelland [Widr90], quienes lo presentaron y popularizaron ante la comunidad internacional.

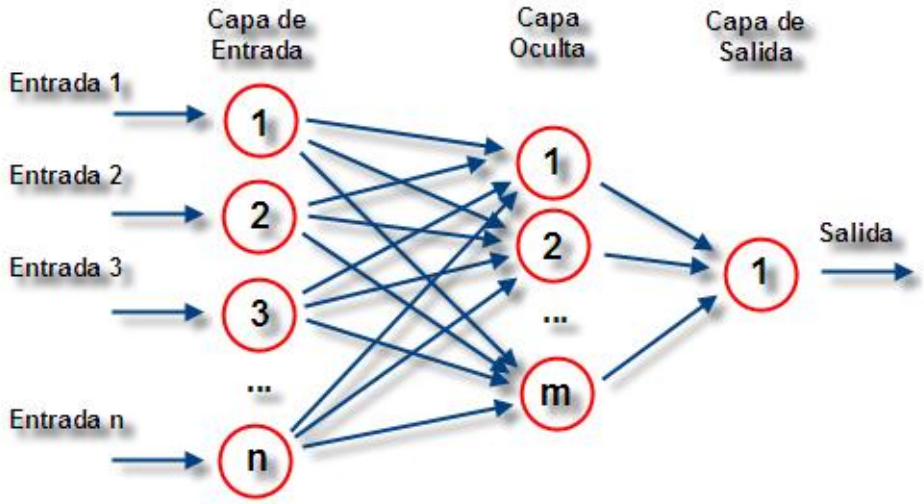


Figura 4.3. Estructura general del Perceptrón Multicapa (MLP).

En la figura 4.3 se muestra el esquema básico de su arquitectura. Esta se constituye por 3 tipos de capas fundamentales:

- Capa de entrada. Constituye la capa inicial y al igual que en el perceptrón simple, su función fundamental es transformar la imagen de entrada en información asimilable por las capas superiores. Esta capa no es entrenable.
- Capas ocultas. Esta capa constituye la principal diferencia con respecto al perceptrón simple. Su principal función es la de ofrecer una conexión asociativa entre las capas de entrada y de salida. Las neuronas de estas capas permiten añadir flexibilidad a la red al permitir la separación no lineal, de tal manera que el MLP puede ser considerado un aproximador universal de funciones. El número de capas ocultas varía en función de la complejidad de las clases que se desean separar.
- Capa de salida. Finalmente, esta capa será la encargada de discriminar las clases, pues

cada neurona corresponderá a una clase.

Esta red se ha convertido en la tendencia más popular dentro de las redes neuronales artificiales hasta la fecha, pues existen múltiples variantes basadas en su funcionamiento.

4.4 Red Neuronal Convolucional

Esta red neuronal pertenece al tipo de redes unidireccionales retroalimentadas de aprendizaje supervisado. En términos generales es una variante del perceptrón multicapa (MLP), y se basa en procesos biológicos de visión. Su idea fundamental consiste en disponer neuronas en un mosaico de manera tal que estas respondan a regiones superpuestas en el campo visual. Entre sus ventajas más destacables está el hecho de que se requiere una mínima cantidad de pre-procesamiento de la imagen. Su importancia en el presente estudio radica en que ha sido bastamente utilizado en múltiples aplicaciones comerciales de reconocimiento de imágenes.

En la figura 4.4 se ilustra el funcionamiento básico de un clasificador basado en una red neuronal convolucional (CNN, por sus siglas en inglés). En ocasiones se utilizan una red neuronal adicional, conocida como mapa auto-organizado de Kohonen (Self Organized Map, SOM) para reducir la dimensionalidad de la imagen de entrada y facilitar y agilizar su procesamiento. El diagrama muestra que esta reducción de dimensionalidad puede ser llevada a cabo también mediante Análisis de Componentes Principales (PCA, por sus siglas en inglés) o bien por un tercer método como la transformación de Karhunen-Loeve.

Posteriormente, la imagen comprimida es mostrada a la CNN, la cual se encargará de determinar a qué clase pertenece. Cabe hacer notar que el MLP puede ser reemplazado por otro clasificador, como el de vecinos más cercanos, aunque lo más usual es utilizar este, o alguna de sus múltiples variantes.

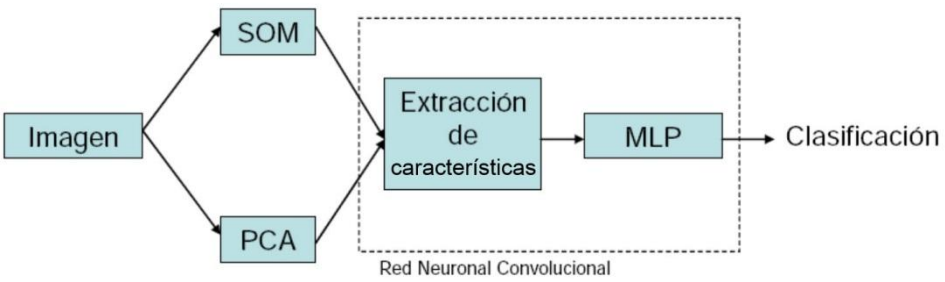


Figura 4.4. Operación básica de un clasificador basado en CNN

Tanto si se utilizan mapas auto-organizados como PCA el funcionamiento el algoritmo es similar: se comienza extrayendo sucesivas ventanas de n x n pixeles de la imagen de entrenamiento, con m pixeles de solapamiento, habitualmente la mitad de n.

Cada una de estas ventanas se muestra como vector (figura 4.5) al reductor de dimensionalidad (SOM o PCA), y en el caso del SOM (algoritmo que se utilizará como ejemplo) se elegirá la neurona ganadora y se actualizarán sus pesos y los de la vecindad.

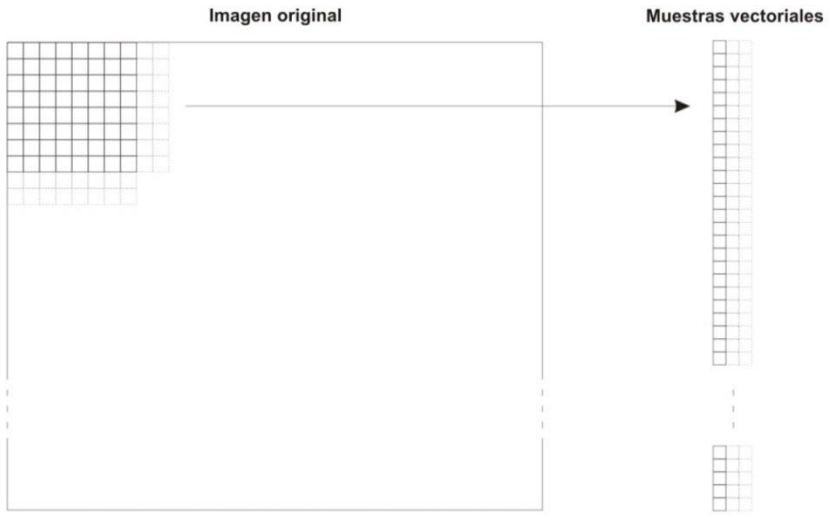


Figura 4.5. Lectura de la imagen a través de ventanas y transformación a vectores

Si al SOM se le presentan ventanas de 8 x 8 pixeles, por ejemplo, se creará un vector de 64 puntos. Al elegir a la neurona ganadora este quedará reducido a uno de 3 puntos, con las coordenadas de la neurona ganadora. Al combinar los puntos resultantes, se generarán 3 imágenes de 1/16 del tamaño original cada una (figura 4.6).

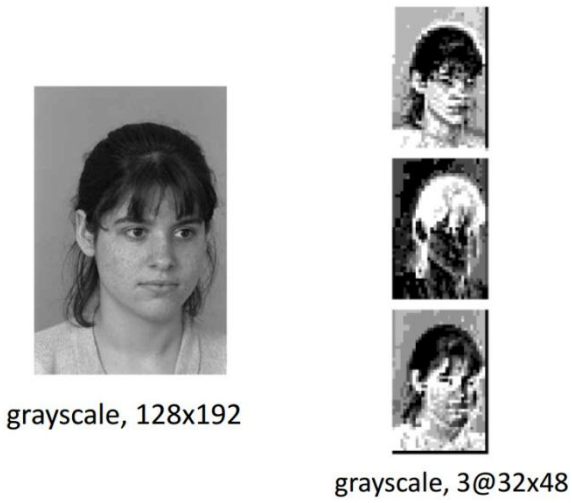


Figura 4.6. Reducción de la dimensionalidad de la imagen original utilizando SOM y ventanas de 8x8

De esta forma, se tiene una compresión con un radio de aproximadamente 1:5. En este caso, el SOM actúa como algoritmo de clustering o compresión para definir los vectores más representativos de una imagen.

Posteriormente, en la red neuronal convolucional propiamente dicha, se suceden capas y mapas (figura 4.7) en los que se alternan la convolución y el sub-muestreo (fig. 4.8).

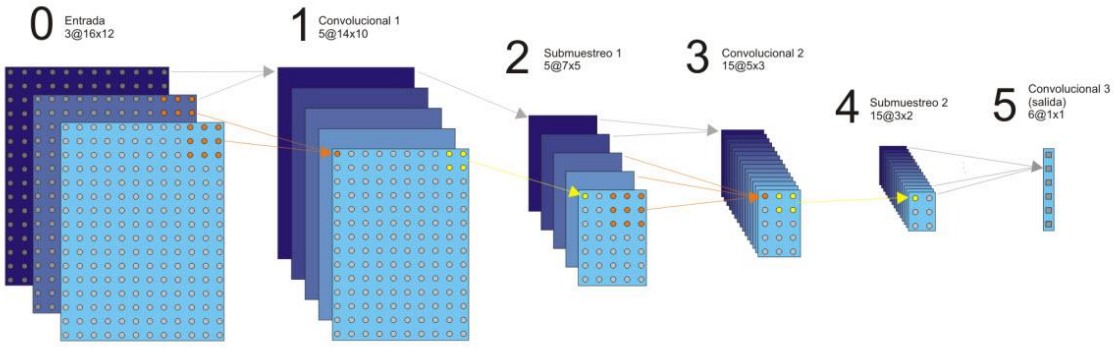


Figura 4.7. Capas y mapas de la CNN

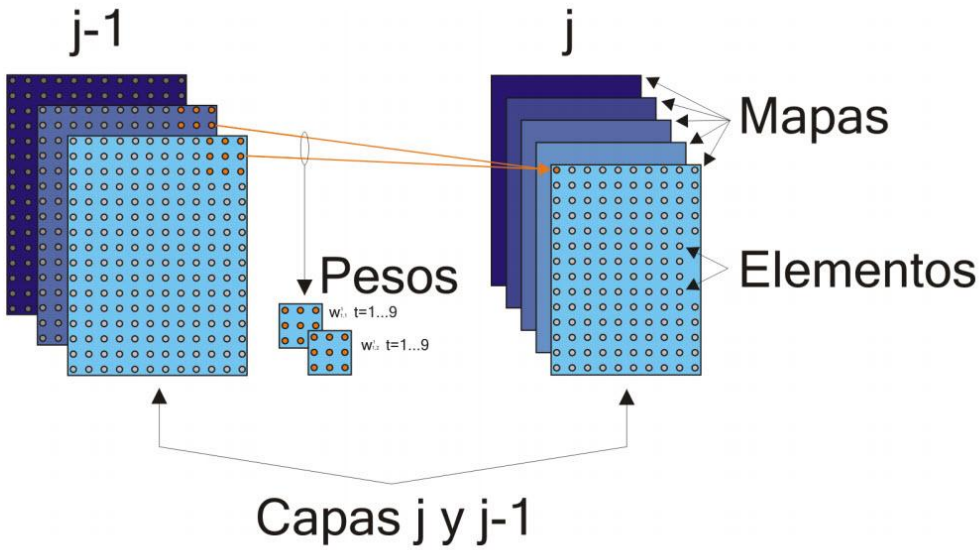


Figura 4.8. Capas y mapas de convolución y sub-muestreo de la CNN

Cada punto en las imágenes representa una neurona tipo perceptrón y todos los perceptrónes de un mismo mapa comparten el mismo peso. La entrada de cada neurona es una sección fija de 1 o varios mapas de tamaño variable de la capa anterior y su función de activación es sigmoidea.

En la CNN se destacan 3 ideas fundamentales:

- Se utilizan campos recpetivos locales donde la distribución espacial de la entrada es importante.
- Los pesos de las conexiones son compartidos
- Se utiliza sub-muestreo espacial.

Estos dos últimos facilitan la generalización [Lecu98] al reducir la resolución espacial al tiempo que se incrementa el número de mapas.

Una vez que las imágenes de entrenamiento han sido suficientemente reducidas, son utilizadas para entrenar la CNN. A cada imagen se le resta la media y se le divide por la desviación estándar para evitar la saturación de la red.

La CNN es una red de entrenamiento supervisado que utiliza el algoritmo de *bacpropagation* para la etapa de aprendizaje, es decir, el error obtenido en cada clasificación es difundido hacia las conexiones de la capa anterior, y esta a su vez, los difunde a la capa anterior, y así sucesivamente hasta llegar a un punto donde la clasificación obtiene resultados satisfactorios.

Las redes neuronales convolucionales fueron introducidas en 1980 por Kunihiko Fukushima [Fuku80] y posteriormente mejoradas en 1998 por Yann LeCun, León Bottou, Yoshua Bengio y Patrick Haffner [Lecu98]. En el año 2003 fueron generalizadas por Sven Behnke [Behn03] y simplificadas por Patrick Simard, David Steinkraus y John C. Platt en el mismo año [Sima03]. Finalmente, en 2011 fueron refinadas por Dan Ciresan et al. [Cire11] e implementadas en un GPU con notables resultados de desempeño.

Han sido empleadas también para clasificar rostros humanos utilizando la base de imágenes ORL [Lawr97], con resultados superiores a los obtenidos mediante Eigenfaces.

4.5 Clasificador neuronal RTC

El clasificador neuronal de umbrales aleatorios (Random Threshold Classifier, por sus siglas en inglés) fue desarrollado y probado en 1994 por Kussul et al. [Kuss94]. Se trata de un clasificador de 3 capas similar al perceptrón multicapa. Su primera capa corresponde a las características de entrada, la segunda a las neuronas binarias de decisión, y la última a las neuronas de clasificación. Su arquitectura general es mostrada en la figura 4.9.

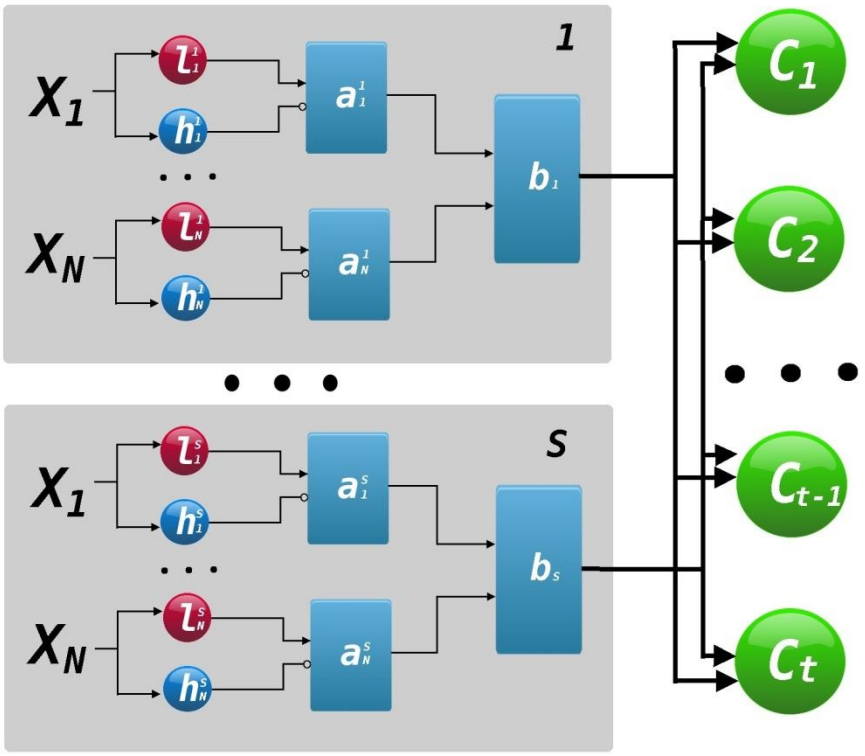


Figura 4.9. Estructura general de un clasificador neuronal RTC.

El clasificador se compone de *S* bloques neuronales, cada uno compuesto por una neurona de salida $(b_1, \dots, b_s)$ y un conjunto *N* de neuronas de entrada $(a_1, \dots, a_N)$. El número de neuronas a corresponde al número de características de entrada X_j. Para cada una de estas características hay dos neuronas umbrales, h_j^i y l_j^i, dónde *i (i=1,...,s)* representa el número del bloque al que pertenece, mientras que *j (j=1,...n)* representa el número de característica evaluada. El umbral l_j^i representa el umbral inferior y h_j^i el superior en un rango de valores que oscila en función de la característica utilizada (histograma de brillo, histograma de contraste, histograma de microcontornos). Cada uno de estos umbrales es seleccionado de forma aleatoria, pero el umbral l_j^i siempre debe cumplir la condición de ser menor que el umbral superior h_j^i. La

salida de la neurona l_j^i se conecta a la entrada excitatoria de la neurona a_j^i, mientras la neurona h_j^i se conecta a la entrada inhibitoria. La salida de la neurona a_j^i aparecerá, sí y solo sí, la salida de la neurona umbral inferior l_j^i es 1, y la salida de su neurona umbral superior h_j^i es 0. Todas las neuronas a de un bloque *S* están conectadas a la neurona de salida b_j, la cual ofrece el resultado del bloque neuronal. La neurona b_j actúa como un operador AND lógico, pues únicamente tendrá salida 1 cuando todas sus entradas, alimentadas por las neuronas a_j^i, estén excitadas.

Cada una de las neuronas b_j (y por lo tanto, cada uno de los bloques), se conectará por medio de conexiones entrenables a cada una de las neuronas de la capa de clasificación *C* $(c_1, \dots, c_t,)$, dónde *t* representa el número clases en las que se desea separar el conjunto de imágenes. Las neuronas de la capa *C* tienen un peso asociado que es modificado por las neuronas b_j en cada etapa del entrenamiento. En cada una de estas etapas se elige como neurona ganadora aquella que tiene el valor de excitación más alto.

Para llevar a cabo el entrenamiento, los autores utilizaron la regla de aprendizaje de Hebb para modificar las conexiones entrenables. Esta consiste en que una vez obtenida la respuesta del clasificador, si esta es incorrecta, el sistema disminuye todos los pesos asociados a la neurona equivocada y al mismo tiempo aumenta los de la neurona correcta. Este proceso de aprendizaje, denominado supervisado o con maestro tiene dos condiciones de terminación:

- Número de errores menor que el definido
- Número de iteraciones superior al definido.

De esta manera, las características utilizadas para describir a *n* objetos crean un espacio de convergencia multidimensional en el cual es posible clasificar un nuevo objeto. La interpretación geométrica de este principio puede ayudar a comprenderlo mejor. Sean X_1 y X_2 dos características utilizadas para describir una clase (fig. 4.10):

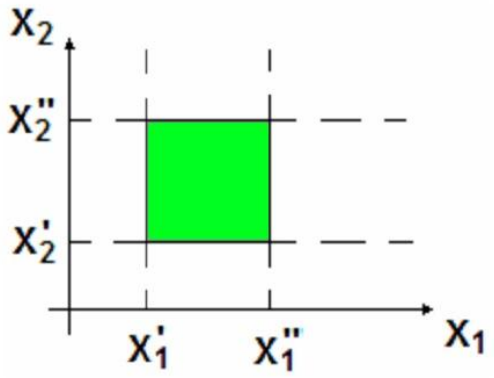


Figura 4.10. Interpretación geométrica de la clasificación con RTC utilizando dos características

En la figura 4.10, representamos en el eje de las abscisas los valores asociados a la característica X_1, mientras en el eje de las ordenas representamos los valores posibles de la característica X_2. En ambos ejes, los valores X_1' y X_2' denotan el umbral inferior l_j^i de cada característica, mientras que los valores X_1'' y X_2'' denotan los umbrales superiores h_j^i de ambas. La zona verde representa una región de intersección creada por a partir de los dos pares de umbrales aleatorios. Resulta evidente que todo punto que se encuentre dentro de dicha zona será clasificado como perteneciente a la clase que tiene dicho rango de características. Cabe señalar que como los umbrales son aleatorios, la región acotada tendrá también un tamaño aleatorio.

Al incrementarse el número de neuronas $b_1, \ldots, b_s$, el número de rectángulos en diferentes planos espaciales acotan aún más la región de clasificación como podemos apreciar en la figura 4.11, donde X_1^* y X_2^*, señalan un punto perteneciente a la clase acotada por múltiples rectángulos.

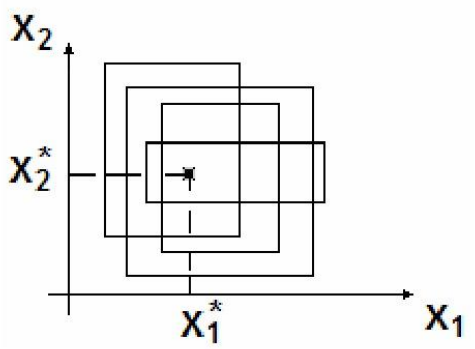


Figura 4.11. Región acotada por el clasificador RTC utilizando múltiples neuronas en la capa *B*

Posteriormente, durante el tiempo de entrenamiento, la adaptación de pesos mediante el proceso de aprendizaje permitirá delimitar una región más regular a partir de las hiper-regiones creadas por las neuronas de la capa *B* en el hiper-plano de características (fig. 4.12).

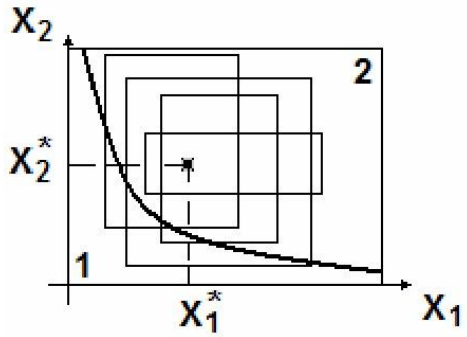


Figura 4.12. Zona de clasificación creada a partir del proceso de entrenamiento.

Los autores del algoritmo señalan haber creado diferentes versiones del mismo para aplicaciones de reconocimiento de texturas naturales, reconocimiento de escritura, reconocimiento de palabras, reconocimiento de voz, sonidos de micromáquinas y de series de señales en general, entre otras. En todas ellas, los autores reportan buenos resultados.

Como ventaja adicional, los autores también señalan que este clasificador tiene un tiempo rápido de entrenamiento y reconocimiento, en comparación con otros clasificadores neuronales como *backpropagation* que requieren mayor tiempo de entrenamiento o ART que requiere mayor tiempo para el reconocimiento.

4.6 Clasificador neuronal RSC

El clasificador neuronal de subespacios aleatorios (Random Subspace Classifier, por sus siglas en inglés) se desarrolló a partir del clasificador neuronal RTC. Su característica principal es que no todos los parámetros $(X_1, \dots, X_n)$ participan en los procesos de entrenamiento y reconocimiento de imágenes, sino que se elige un subconjunto aleatorio del universo disponible. Esto se debe a que cuando la dimensión del espacio de entrada *N* (fig. 4.10) se incrementa, es necesario aumentar el margen entre el umbral superior h_j^i y el umbral inferior l_j^i de las neuronas de entrada. En algunos casos, ambos umbrales se separarán tanto que alcanzarán los límites superior e inferior de la variable X_i, por lo que la correspondiente neurona a_j^i siempre tendrá como salida un 1 y no aportará información alguna sobre los datos de entrada. De esta forma, solo una pequeña fracción de las neuronas a_j^i modificará el resultado.

Debido a esto, en [Baid05] los autores modificaron el clasificador RTC para incluir en cada bloque neuronal b_j (fig. 4.9) únicamente una pequeña fracción de neuronas a_j^i seleccionadas aleatoriamente a partir del vector de entrada, las cuales son denominadas *subespacios aleatorios*. Cabe notar que para cada bloque neuronal se seleccionan distintas neuronas a_j^i, por lo que el espacio total es representado por múltiples subespacios de entrada. De esta manera se disminuye el espacio paramétrico y con ello se reduce el tiempo de procesamiento y los recursos de cómputo necesarios.

4.7 Clasificador neuronal LIRA

El clasificador neuronal de área receptiva limitada (LImited Receptive Area, por sus siglas en inglés) se basa en los principios del perceptrón de tres capas de Rosenblatt.

Baidyk et al. [Baid04] proponen 4 grandes modificaciones al esquema de Rosenblatt. Estas modificaciones son un cambio en el esquema de conexiones aleatorias de la capa S, la adaptación del clasificador para reconocer imágenes en escala de grises, una mejora en el proceso de entrenamiento, y finalmente, un cambio en la regla de selección de ganador.

Los autores proponen dos variantes del clasificador LIRA [Make08]: LIRA binario y LIRA Grayscale. El primero está diseñado para reconocer imágenes binarias, mientras que el segundo, una extensión del anterior, puede reconocer imágenes en escala de grises. La figura 4.13 muestra la estructura general de un reconocedor LIRA binario, mientras que la figura 4.14 muestra la estructura de un LIRA Grayscale.

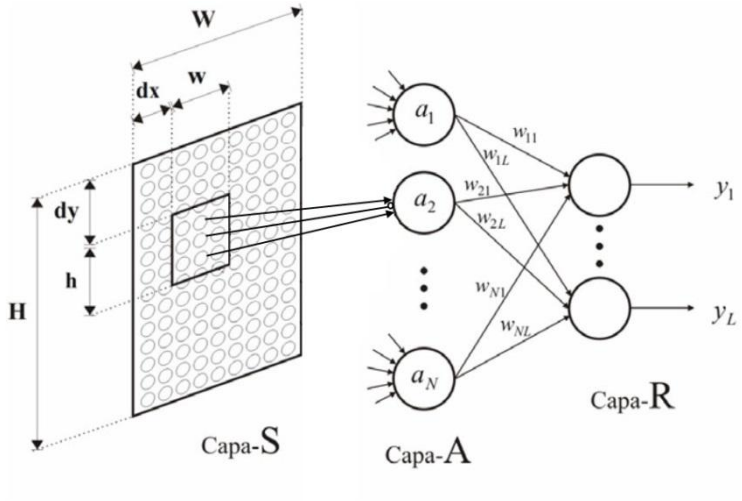


Figura 4.13. Estructura de un clasificador neuronal LIRA binario.

LIRA Grayscale al igual que el perceptrón de Rosenblatt y de LIRA binario, contiene una capa de entrada (*S*), una capa asociativa (*A*) y una capa de salida (*R*), pero a diferencia de estos, incluye una capa adicional, denominada capa intermedia (*I*), la cual se ubica entre la capa *S* y la capa *A* y su propósito es ampliar la funcionalidad de la extracción de características en imágenes en escala de grises.

Las neuronas de la capa *S* representan el nivel de gris de un pixel de la imagen. Dado que la representación estándar de la escala de grises es de 256 tonalidades, el valor de salida de la capa *S* oscila entre 0 y 255, correspondiendo el 0 al negro absoluto y el 255 al blanco.

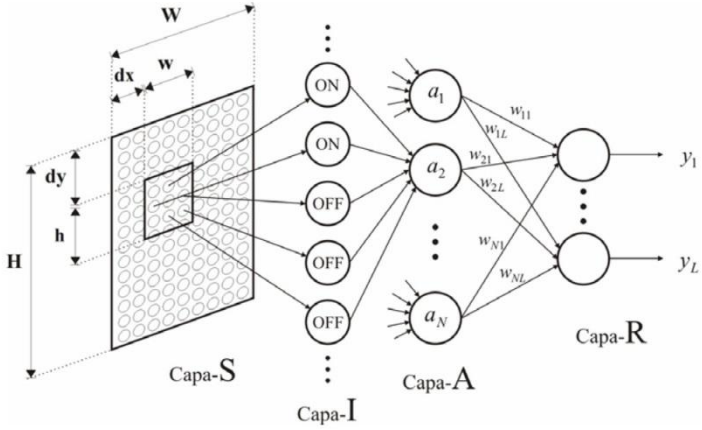


Figura 4.14. Estructura de un clasificador neuronal LIRA Grayscale.

La función de la capa *I* es describir una pequeña área de la capa *S*, llamada ventana, mediante *m* descriptores de dos tipos: neuronas ON y neuronas OFF. Ambas neuronas tienen dos estados como valores de salida, 0 y 1.

La capa *S* se conecta a la capa *I*, y esta a su vez a la capa *A*. Las neuronas de la capa A ofrecen una síntesis de las características extraídas por la capa *I*. Con base a los resultados ofrecidos por las neuronas de la capa inferior conectadas a ella, cada neurona de la *A*, tendrá una salida con valores {0,1} para indicar si determinada característica se encuentra presente en la ventana seleccionada (valor 1) o no (valor 0).

Finalmente, cada neurona de la capa *A* se conectará con cada neurona de la capa *R* (capa de salida), mediante conexiones de pesos modificables durante el proceso de entrenamiento. El número de neuronas de la capa *R* corresponde al número de clases que se desean reconocer. La función de activación de estas neuronas es de tipo lineal [Kuss10].

4.7.1 Selección de conexiones y activación de neuronas

El procedimiento de conexión entre las diferentes capas de la red es el siguiente. Sea *N* el número total de neuronas de la capa asociativa. Para cada neurona asociativa a_k, donde k =1,…,*N* se selecciona aleatoriamente un área rectangular en la capa *S*, la cual denominaremos ventana, con *h* x *w* neuronas como se muestra en la figura 4.14. La posición de esta ventana en la imagen se elige de acuerdo a las siguientes fórmulas:

$$dx_i = random_i(W_s - w), \tag{4.2}$$

$$dy_i = random_i(H_s - h), \tag{4.3}$$

donde d_x y d_y representan las distancias en los ejes x y y respectivamente, i representa la

posición de una neurona en la capa asociativa A, W_S y H_s denotan el ancho y el alto de la capa S respectivamente y finalmente $random_i(z)$ es un número aleatorio que se encuentra uniformemente distribuido en el rango $[0, z]$.

De esta ventana se eligen aleatoriamente m puntos (neuronas) que serán divididos, también de forma aleatoria en puntos "positivos" p y puntos "negativos" n. Cada punto positivo y negativo será conectado a una neurona ON y OFF respectivamente, de la capa intermedia I, y tendrá asociado un umbral aleatorio T_{mk} seleccionado el rango [0, 255]. Este grupo de m neuronas en la capa intermedia se conectará a la neurona a_k.

Sea x_{ij} una neurona de entrada de la capa S. La salida de la neurona ON será activo (igual a 1) si su valor de entrada x_{ij} es mayor o igual al umbral T_{pk}; y será 0 en otro caso, es decir:

$$\varphi_{on}(x_{ij}) = \begin{cases} 1, & x_{ij} \geq T_{pk} \\ 0, & x_{ij} < T_{pk} \end{cases} \tag{4.4}$$

La salida de la neurona OFF será activo (igual a 1) si su valor de entrada x_{ij} es menor o igual al umbral T_{nk}; y será 0 en otro caso, es decir:

$$\varphi_{off}(x_{ij}) = \begin{cases} 1, & x_{ij} \leq T_{nk} \\ 0, & x_{ij} > T_{nk} \end{cases} \tag{4.5}$$

Las neuronas de la capa asociativa son similares a las compuertas AND de lógica digital. Su salida será 1 (estado activo) si todas las neuronas ON y OFF conectadas a ella a través de su entrada se encuentran también en estado activo (valor igual a 1), en caso contrario su salida será igual a 0 o inactiva. Cada neurona de la capa A actúa como descriptor local al indicar si determinada característica se encuentra presente o ausente en la imagen observada.

4.7.2 Proceso de entrenamiento

Tanto para LIRA binario como para LIRA Grayscale se utiliza un proceso de entrenamiento idéntico [Make08]. Antes de iniciar el entrenamiento, los pesos de las conexiones entre las neuronas de la capa A y R son inicializados a cero. Posteriormente se ejecuta el siguiente algoritmo:

- ***Paso 1.*** El proceso de entrenamiento inicia cuando se presenta una imagen a la capa S, la cual se encarga de transformarla en datos interpretables por el clasificador, para que posteriormente la capa I y la capa A extraigan y codifiquen sus características, respectivamente. Cuando la imagen es codificada, las excitaciones de la capa R son calculadas de la siguiente manera:

$$E_i = \sum_{j=1}^{N} a_j \cdot w_{ji} \tag{4.6}$$

donde E_i es la salida (excitación) de la *i*-ésima neurona de la capa *R*, a_j es la salida (0 o 1) de la *j*-ésima neurona de la capa *A* y w_{ji} es el peso de la conexión entre la j-ésima neurona de la capa *A* y la i-ésima neurona de la capa *R*.

- ***Paso 2.*** Para garantizar la robustez en el reconocimiento del clasificador, tras realizar el cálculo de las salidas de todas las neuronas de la capa R, se lee la clase correcta (denotada E_r) a la que pertenece la imagen bajo reconocimiento, y su excitación es recalculada de acuerdo a la siguiente fórmula:

$$E_r^* = E_r \cdot (1 - T_E) \tag{4.7}$$

donde $0 \leq T_E \leq 1$ determina la excitación adicional que la neurona correcta debe tener. A continuación la neurona con la excitación más grande es seleccionada. La neurona ganadora representa la clase reconocida.

- ***Paso 3***. Una vez que se ha obtenido la neurona ganadora (la neurona con mayor excitación en la red neuronal), si la neurona correspondiente a la clase real (denotada por *r*) es igual a la neurona ganadora (denotada como *g*), es decir *r* = *g*, entonces no se modifica ningún peso en las conexiones, pero si *r* ≠ *g*, entonces:

$$\forall k, \;\; w_{kr}(t+1) = w_{kr}(t) + a_k \tag{4.8}$$

$$\forall k, \;\; w_{kg}(t+1) = w_{kg}(t) - a_k \tag{4.9}$$

$$si\big(w_{kg}(t+1) < 0\big) \rightarrow w_{kg}(t+1) = 0 \tag{4.10}$$

donde $w_{ki}(t)$ es el peso de la conexión entre la k-ésima neurona de la capa *A* y la i-ésima neurona de la capa *R* antes del refuerzo, $w_{ki}(t+1)$ es el peso de la misma conexión después del refuerzo, y finalmente, a_k es el valor de salida (0 o 1) de la k-esima neurona de la capa *A*.

El anterior algoritmo se repite iterativamente. Cuando todas las imágenes del conjunto de entrenamiento son presentadas, se calcula el número total de errores. Si este número es mayor que un porcentaje preestablecido, se ejecuta un nuevo ciclo de entrenamiento. En caso contrario el entrenamiento es finalizado. El proceso también finaliza si se supera un número máximo preestablecido de ciclos de entrenamiento.

Dado que las características extraídas y codificadas en las capas inferiores (*S*, *I* y *A*) son las mismas para cada ciclo de entrenamiento, para ahorrar tiempo y poder de cómputo se optó por realizar el proceso de codificación de características únicamente una vez para todas las imágenes y luego almacenar en disco el conjunto de neuronas activas para cada imagen. Posteriormente durante la etapa de entrenamiento se utilizan dichos conjuntos y no las imágenes

directamente, con lo que se consigue un decremento sustancial en el tiempo y poder de cómputo requerido.

4.7.3 Mejora del aprendizaje

LeCun et al. [Lecu98] muestran que es posible mejorar el desempeño de los sistemas de reconocimiento añadiendo distorsiones durante el proceso de aprendizaje. Estas distorsiones pueden comprenden desplazamientos horizontales y verticales, rotaciones a izquierda y derecha, así como combinaciones de una o más de ellas. En la figura 4.15 se muestran algunos ejemplos de estas distorsiones en un carácter escrito (el número 1). El clasificador neuronal LIRA ha sido utilizado exitosamente en el reconocimiento de piezas de micromecánica [Baid04, Baid08, Make08], de insectos en agricultura [Baid08] y de caracteres escritos [Kuss94].

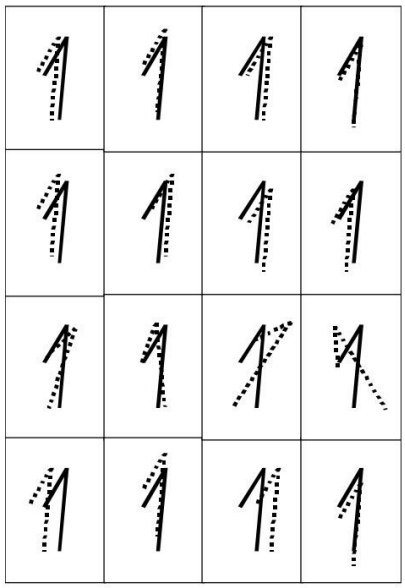


Figura 4.15. Ejemplos de distorsiones de una imagen para mejorar el rendimiento del clasificador neuronal LIRA.

Capítulo 5

Sistema de Reconocimiento de Rostros.

En el presente capítulo se describe el método desarrollado para mejorar el reconocimiento de rostros que presentan variación de pose. La primera parte del capítulo está dedicada a describir el algoritmo de la red neuronal utilizada, para posteriormente continuar con la descripción del método de transformación usado para mejorar los resultados, y finalmente, se detallan los experimentos realizados y los resultados obtenidos.

5.1 Estructura general del clasificador

El Clasificador Neuronal de Permutación de Códigos (Permutation Coding Neural Classifier, PCNC) fue propuesto por Kussul et al. en 2003 [Kuss03, Kuss06b] como un método de reconocimiento de imágenes en general. Ha sido probado para las tareas de reconocimiento de caracteres manuscritos de la base de datos MNIST (obteniendo una tasa de errores de 0.37%) [Kuss06a]; en el reconocimiento de objetos de micromecánica (obteniendo tasa de reconocimiento de 92.5%) [Kuss05]; y en el reconocimiento de rostros humanos de la base de datos ORL (obteniendo una tasa de error de 0.1% en 10 experimentos) [Kuss10].

PCNC se basa en la estructura genérica del paradigma de redes neuronales asociativas-proyectivas (APNN) descritas en [Kuss91a, Kuss91b]. A este paradigma también pertenecen las redes neuronales RTC, RSC y LIRA, descritas en el capítulo anterior. La estructura de PCNC

consta de 3 partes que trabajan de forma serial: extractor de características, codificador y clasificador. Dichas partes se pueden apreciar en el esquema general mostrado en la imagen 5.1.

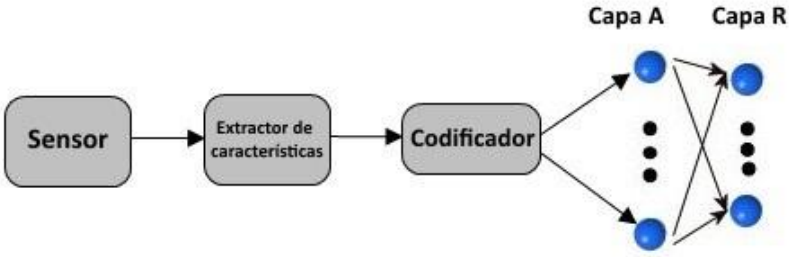


Figura 5.1. Estructura general del clasificador PCNC.

A grandes rasgos, el clasificador PCNC funciona de la siguiente manera:

- Al clasificador le es presentada una imagen en escala de grises para ser leída. Esta etapa corresponde a la capa sensor S.
- Posteriormente la capa descriptora D se encarga de extraer un conjunto de características o propiedades.
- Dichas propiedades son presentadas al codificador, el cual se encarga de transformarlas en un vector de gran dimensión a fin de aumentar su separabilidad lineal.
- Finalmente, el conjunto de vectores de características es entregado al clasificador neuronal (capas A y R) para ser procesado con propósitos de entrenamiento, prueba o reconocimiento dentro de una clase previamente entrenada.

A continuación se describen en detalle cada una de las partes que conforman el sistema, así como los conceptos más relevantes del algoritmo.

5.2 Extracción de propiedades

La extracción de propiedades comienza con la presentación de una imagen en escala de grises, a partir de la cual se seleccionan una serie de puntos específicos. Existen diferentes métodos de selección de dichos puntos, pero la idea fundamental es elegir aquellos que representen las propiedades más discriminantes y significativas para su clasificación. Una selección descuidada de puntos específicos en imágenes relativamente grandes resulta en una gran cantidad de estos, con el consecuente incremento masivo en la demanda de recursos de procesamiento. Hasta ahora, el algoritmo ha utilizado dos métodos para elegir estos puntos específicos: por umbral de brillo y por extracción de contornos. El primero de ellos consiste en seleccionar un umbral de brillo B y todos aquellos puntos cuyo brillo b_{ij} sea mayor a dicho

umbral serán considerados puntos específicos. El segundo método requiere la utilización de un algoritmo de extracción de contornos para ser aplicado sobre la imagen. El presente trabajo utiliza extracción de contornos mediante el operador Sobel [Sobe68].

Una vez que se ha obtenido un punto específico (o punto de interés) en la imagen se define un rectángulo de dimensión HxW dentro de esta, en cuyo centro se encuentra precisamente el punto de interés (fig. 5.2).

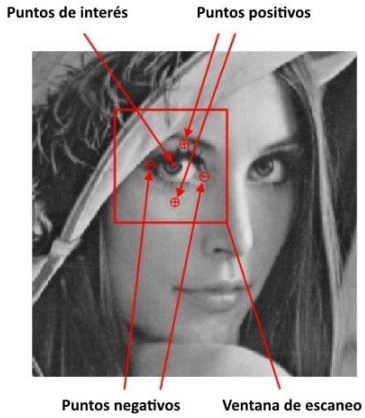


Figura 5.2. Puntos de interés seleccionados mediante el extractor de características

A partir de este rectángulo, denominado ventana, se extraen múltiples propiedades mediante la utilización de descriptores locales conocidos como RLD (tratados en detalle en el siguiente apartado). La idea fundamental de los RLD es detectar una característica específica dentro de la ventana mediante la localización de p puntos positivos y n puntos negativos. Se dice que la característica buscada existe si todos los puntos, tanto positivos como negativos están activos.

Se procura que ninguna de las ventanas de búsqueda salga del área de la imagen. Esto se puede lograr mediante imágenes normalizadas, donde se trata de evitar que los puntos de interés se encuentren en las orillas. No obstante, si la ventana es relativamente grande, o si se selecciona gran cantidad de puntos de interés, existen mayores probabilidades de que las ventanas sobrepasen los límites de la imagen. Para solucionar lo anterior el tamaño de la imagen se expande artificialmente $w/2$ pixeles hacia cada lado y $h/2$ pixeles hacia arriba y hacia abajo, siendo h y w la altura y anchura de la imagen en pixeles, respectivamente. Se utiliza el color blanco como relleno en las áreas expandidas, de manera que esto represente la ausencia de todo punto específico posible.

El extractor de propiedades hace uso de muchas características $F_i(i = 1, \dots, S)$ donde S por lo general se encuentra en el orden de los miles. El extractor examina las S propiedades para cada uno de los puntos específicos definidos; para posteriormente ser entregadas al codificador.

5.3 Descriptores locales aleatorios (RLD)

Después de analizar los distintos paradigmas de clasificadores de imágenes basados en redes neuronales, es posible generalizar la idea de que la extracción de características constituye al mismo tiempo un componente de la red y una etapa del proceso de clasificación.

En la red neuronal PCNC la extracción de características es llevada a cabo mediante Descriptores Aleatorios Locales (Random Local Descriptors, RLD) los cuales surgieron a partir de una evolución de ideas presentadas en trabajos sucesivos [Rose62, Huan88, Kuss05, Kuss06b].

Así, por principio, los RLD parten de la idea de los descriptores aleatorios de la imagen presentados por Frank Rosenblatt en 1962 [Rose62]. En su trabajo, Rosenblatt presenta el perceptrón, una red neuronal constituida por 3 capas: una capa sensitiva, una asociativa y una de reconocimiento o reacción. La capa asociativa está constituida por un conjunto de neuronas y cada una de ellas desempeña el rol de descriptor aleatorio de la imagen. Dicha neurona se conecta a algunos puntos de la retina (pixeles de la imagen) seleccionados de manera aleatoria y calcula su función a partir del brillo de dichos puntos.

Posteriormente, también en los años 60, Hubel y Wiesel presentan la idea de descriptores locales a partir de descubrimientos realizados en el sistema de visión animal [Hube62, Hube68]. En sus estudios, prueban que en la corteza visual de los animales existen descriptores locales que corresponden a la orientación local de elementos del contorno, movimientos, etcétera. El conjunto de descriptores locales investigados es probablemente incompleto, puesto que en los experimentos únicamente se detectaron los descriptores de aquellas imágenes que habían sido preparadas de antemano para presentar a los animales, por lo que es probable que no todos los descriptores de la corteza visual de los animales hayan sido descubiertos. Dicha desventaja quizás puede ser resuelta mediante los descriptores aleatorios de Rosenblatt, dado que pueden ofrecer una gama más amplia, pero que son menos eficaces debido a que utilizan descriptores del mismo tamaño de la imagen de entrada

Así, pues, a partir de estas dos ideas, los descriptores aleatorios de Rosenblatt y los descriptores locales de Hubel y Wiesel; es como surge la idea de descriptores capaces de describir áreas locales de una imagen. Baidyk et al. [Baid04] aplicaron estos descriptores en la red neuronal denominada LIRA (descrita en el capítulo 4) y obtuvieron buenos resultados en el reconocimiento de caracteres manuscritos de la base de datos MNIST. LIRA sin embargo tiene la desventaja de ser sensible a desplazamientos de imagen. Fue posible solventar dicho problema para los desplazamientos pequeños aumentando la galería de entrenamiento de manera artificial, pero para los desplazamientos grandes fue necesario recurrir a un método más robusto.

En la figura 5.3 se muestra una estructura general más detallada del sistema PCNC, donde se puede apreciar claramente una red neuronal constituida por múltiples capas. La primera capa S, o capa sensorial, corresponde a la imagen de entrada. Posteriormente se encuentra la capa D, o capa de descripción, la cual está constituida a su vez por dos subcapas. La primera de ellas, denominada D_1 contiene los RLD de nivel más bajo y sirve para proporcionarle información a la capa D_2, que contiene los RLD de nivel más alto. Luego se encuentra la capa A, la cual contiene las neuronas asociativas, y finalmente tenemos la capa R o de reconocimiento, donde cada una de sus neuronas pertenece a cada clase a reconocer.

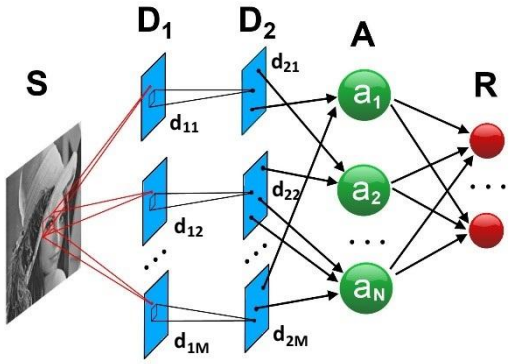


Figura 5.3. Estructura del sistema de reconocimiento de propósito general.

En la figura 5.4 podemos apreciar con mayor detalle la manera en la que se encuentran conformados los RLD dentro del esquema planteado en la figura anterior. En la figura podemos observar que cada RLD contiene varias neuronas (denotadas con los números 1-5) que se encuentran conectadas a diversos puntos de un fragmento de la imagen original; la cual corresponde a un área local de descripción, a saber, la ventana de tamaño HxW. Las neuronas 2-5 son denominadas neuronas simples y sirven para probar pixeles de la capa S seleccionados de manera aleatoria dentro de la ventana. Existen dos clases de neuronas simples, las neuronas *ON* y las neuronas *OFF* (similares a las neuronas homónimas en las redes neuronales naturales) y tienen un tipo de salida binario, es decir, 0 o 1.

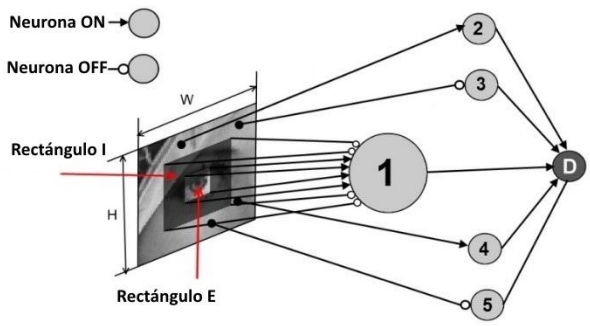


Figura 5.4. Esquema detallado RLD.

Las neuronas *ON* se activan si el gradiente de brillo b_i del pixel correspondiente es mayor o igual que el umbral T_i de la neurona, y permanecen inactivas en caso contrario, es decir:

$$\varphi_{ON}(b_i) = \begin{cases} 1, & b_i \geq T_i \\ 0, & b_i < T_i \end{cases} \tag{5.1}$$

Las neuronas *OFF* por su parte, se activan si el gradiente de brillo b_i del pixel correspondiente es menor que el umbral T_i de la neurona, y permanecen inactivas en caso contrario, es decir:

$$\varphi_{OFF}(b_i) = \begin{cases} 1, & b_i < T_i \\ 0, & b_i \geq T_i \end{cases} \tag{5.2}$$

Los umbrales de ambos tipos de neuronas son seleccionados de manera aleatoria en el intervalo dinámico de valores de la imagen de entrada, es decir $T_{min} \leq T_i \leq T_{max}$. En la figura 5.4 las neuronas con números 2 y 4 corresponden a las neuronas ON, mientras que las neuronas 3 y 5 corresponden a las neuronas OFF.

La neurona marcada con el número 1 es denominada neurona compleja pues consta de un mayor número de conexiones. Tiene conexiones de excitatorias con todos lo pixeles pertenecientes al rectángulo E de la figura 5.4; así como conexiones inhibitorias con todos los pixeles localizados en el rectángulo I, excluyendo los que ya pertenecen al rectángulo E. Ambos rectángulos se encuentran ubicados en el centro de la ventana HxW. Para la tarea de reconocimiento de rostros las conexiones excitatorias deben ser inversamente proporcionales al área del rectángulo *E*, mientras que las inhibitorias deben ser inversamente proporcionales al área del rectángulo *I*. Así, las neuronas complejas detectan los puntos más informativos de la imagen [Ibar14].

La neurona compleja realiza un filtrado en base a los pesos de todas sus conexiones. Se utilizan dos tipos de filtros. El primero de ellos se utilizó para el reconocimiento de caracteres manuscritos y consiste en seleccionar los puntos donde el brillo es mayor que el umbral predefinido. El segundo se utiliza para aplicaciones micromecánicas y reconocimiento de rostros humanos y consiste en seleccionar los puntos donde el gradiente de brillo es mayor que el umbral predefinido. Los puntos seleccionados por el filtro son denominados puntos de interés.

La última neurona de la figura 5.4 corresponde al descriptor D de la imagen. Esta neurona se activará, si y solo sí, todas las neuronas conectadas a ella (1-5 en el diagrama) están activadas. En otras palabras, la neurona D actúa como una compuerta AND lógica, por lo que es llamada AND-neurona.

En la figura 5.3 podemos observar que la capa D_1 consta de un número de planos $d_1, d_2, \ldots d_M$. Cada uno de estos planos está conformado por un número de AND-neuronas igual al número de pixeles de la imagen. Además del número de neuronas, cada plano preserva la

topología de la imagen original, lo que significa que cada neurona del plano d_{ij} se encuentra ubicada en la misma posición que el pixel central de la ventana HxW. La topología de conexiones entre la capa sensorial y las neuronas simples (2-5 de la figura 5.4) es la misma en el intervalo de cada plano d_{1j} (figura 5.3). La topología de conexiones entre la capa sensorial y la neurona compleja (1 de la figura 5.4) es igual para todas las neuronas en todos los planos de la capa D_1. Cada plano de esta capa es un descriptor que tiene por objetivo detectar la presencia de una característica específica en cualquier lugar de la imagen, por lo que el número de planos corresponde al número de características extraídas. Para el reconocimiento de caracteres manuscritos de la base de datos MNIST se utilizaron 12,800 características y para el reconocimiento de rostros de la base de datos ORL se utilizaron 200. En esta tesis, para el reconocimiento de rostros de la base de datos FRAV2D se utilizaron 400. Este parámetro fue fijado de manera experimental.

La capa D_2 (figura 5.3) al igual que la capa D_1 también está conformada por un conjunto de M planos descriptores que a su vez están conformados por un conjunto de neuronas. Cada neurona del plano d_{2j} se encuentra conectada con todas las neuronas del plano d_{1j} localizadas dentro de la ventana rectangular correspondiente, pero a diferencia de las neuronas de la capa anterior, las neuronas de este plano se activarán siempre que exista al menos una neurona activa en la ventana del plano d_{1j} correspondiente. Este mecanismo equivale a la función OR lógica, por lo que la neurona es llamada OR-neurona. En cada plano de esta capa, al igual que en la anterior, se preserva la topología de todas las neuronas.

La capa asociativa A contiene N elementos encargados de recolectar la actividad de las neuronas en la capa subyacente. Dichas neuronas son seleccionadas de manera aleatoria. Esta capa también es la encargada de realizar las interconexiones obligatorias que serán detalladas posteriormente.

La implementación directa de algoritmo descrito demanda gran consumo de tiempo y recursos de cómputo, por lo que se utilizó la siguiente estrategia para reducir ambos a una cantidad factible. Por cada pixel de la imagen de entrada se calculó la actividad de la neurona compleja 1 (fig 5.3). Si dicha neurona se activa, se realizan los cálculos secuenciales únicamente para los elementos neuronales conectados a esta neurona. Siguiendo este principio hasta los cálculos de la capa R, es posible disminuir drásticamente la cantidad de tiempo y recursos requeridos, dado que el número de neuronas activas es notablemente inferior al número total de neuronas requeridas en el planteamiento original. De esta forma, es posible simular el sistema propuesto del reconocimiento en tiempo real.

5.4 Codificación de características

El codificador de características es uno de los elementos más importantes del clasificador PCNC y sirve para permitir la separación lineal entre las clases. Es por eso que es necesario describirlo en detalle.

Una vez extraídas las características de la capa D_1, la capa D_2 se encarga de codificarlas en un vector binario:

$$V = \{v_i \mid v_i = \{0,1\}, i = (1, \dots, N)\} \tag{5.3}$$

Para cada característica F_k se crea un vector binario adicional para cumplir funciones auxiliares:

$$U = \{u_i \mid u_i = \{0,1\}, i = (1, \dots, N)\} \tag{5.4}$$

Este vector contiene K número de 1s donde $K \ll N$. Al menos 100 veces menor. En los experimentos, el número de unos era 16, mientras que N oscilaba entre 32,000 y 512,000. Las posiciones de los 1s en el vector U_k son determinadas mediante un procedimiento aleatorio para cada propiedad F_k. Dicho procedimiento genera todos los vectores U_k para cada característica y los guarda en memoria no volátil. Tales vectores binarios son denominados *máscaras* de las propiedades F_k.

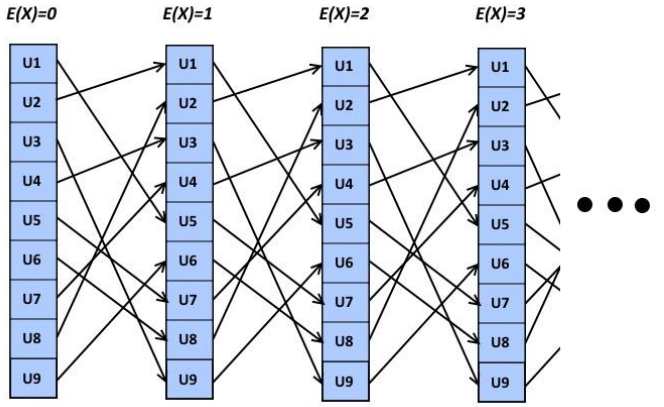
El propósito de las máscaras es ofrecer un método para comparar la ubicación de una característica F_k que ha sido detectada en dos imágenes distintas. Para lograr este objetivo, se procede de la siguiente manera:

- Una vez que se han encontrado y definido las S características a partir del conjunto de entrenamiento, se detectan los puntos de interés para cada imagen de prueba o reconocimiento.
- Para cada uno de estos puntos de interés encontrados, se prueba la existencia de las S propiedades mediante los RLD. Cada RLD escanea la imagen entera y si detecta la característica correspondiente, envía su vector auxiliar U al codificador.
- El codificador utiliza un esquema de permutaciones para generar un nuevo vector U^*. La cantidad de permutaciones depende de la localización de la característica en la imagen. Se consideran dos permutaciones diferentes, una correspondiente a las direcciones horizontales (X) y otra a las verticales (Y).

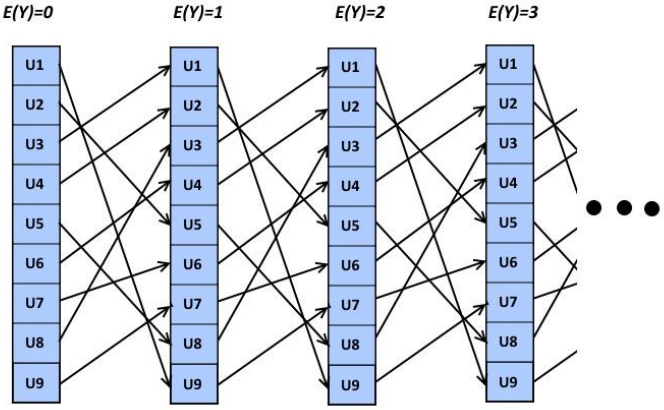
A través de las permutaciones se busca que los vectores U^* satisfagan la condición de estar fuertemente correlacionados si la distancia entre la ubicación de las características es pequeña; y débilmente correlacionados si la distancia es grande.

En la figura 5.5 se muestra un ejemplo de las permutaciones realizadas sobre el vector binario U para transformarlo en el vector U^*. Como se mencionó anteriormente, estas pueden ser

horizontales (fig. 5.5a) o verticales (fig. 5.5b) dependiendo de la ubicación de la característica. Es muy importante hacer notar que ambos son esquemas de permutación distintos.



(a)



(b)

Figura 5.5. Esquema de permutaciones (a) horizontales y (b) verticales.

El esquema de permutación presentado en la figura 5.5 se crea de la siguiente manera para el esquema de permutaciones horizontales:

1. Para el primer componente U_1 de la primera columna, se selecciona aleatoriamente un componente de la siguiente columna (U_5 en este caso) y se realiza la conexión.
2. Para el siguiente componente de la misma columna 0, se selecciona un componente de la columna 1 que no hay sido seleccionado previamente y se conecta.
3. Se repite el paso 2 hasta que todos los componentes estén conectados.
4. Se repite la estructura de conexiones entre las siguientes columnas.

Tras completar las permutaciones horizontales, se aplican las permutaciones verticales

siguiendo el mismo procedimiento. Para determinar la cantidad de permutaciones que se deben realizar, primero se define una distancia de correlación D_c, la cual es tomada como punto de referencia para comparar la distancia euclidiana entre dos características detectadas en una imagen. Por ejemplo, sea F_k una propiedad detectada en dos puntos distintos $P_1(x_1, y_1)$, y $P_2(x_2, y_2)$. La distancia euclidiana está dada por:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{5.5}$$

Utilizando la distancia de correlación, decimos que existe correlación si $d < D_c$ y que no están correlacionado sí $d \geq D_c$. Para tratar de cumplir con esta propiedad, se calculan los siguientes valores:

$$X = \frac{i}{D_c}, \qquad Y = \frac{j}{D_c}, \tag{5.6}$$

$$E(X) = int(X), \qquad E(Y) = int(Y), \tag{5.7}$$

$$R(X) = i - E(X) \cdot D_c \qquad R(Y) = j - E(Y) \cdot D_c \tag{5.8}$$

$$P_x = int(R(X) \cdot \frac{N}{D_c}) \qquad P_y = int(R(R) \cdot \frac{N}{D_c}) \tag{5.9}$$

donde $E(X)$ y $E(Y)$ representan la parte entera de X e Y respectivamente; $R(X)$ y $R(Y)$ representan las partes fraccionarias; i y j representan las coordenadas de la característica detectada y N representa el número de neuronas (componentes del vector). Las partes enteras $E(X)$ y $E(Y)$ indican el número de permutaciones completas que se requieren realizar horizontal y verticalmente, respectivamente; mientras que las partes fraccionarias indican el número de permutaciones adicionales que se efectuaran sobre los P_x y P_y primeros componentes no ceros del vector. Los valores de P_x y P_y oscilan en el rango [*0, N*).

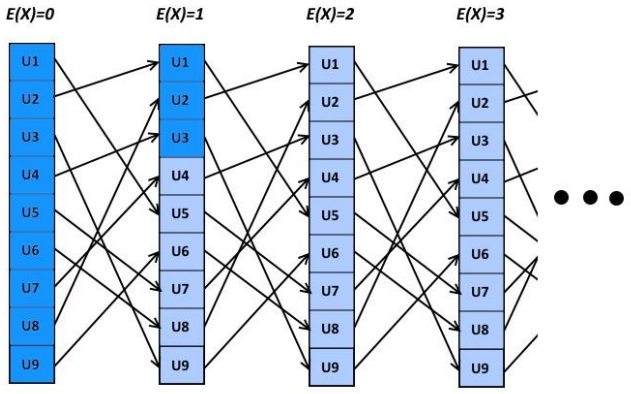
Como ejemplo considérese la característica F_k detectada en el punto (11,22), teniendo una distancia de correlación de 8 y un número de neuronas igual a 9. En este caso, tenemos que *j*=11, *i*=22, D_c=8, *N*=9, y tras aplicar los cálculos definidos anteriormente, obtenemos *E(X)*=1, *E(Y)*=2, P_x=3 y P_y=6. A partir de estos valores, se procede de la siguiente manera:

1. Primero se realizan las permutaciones horizontales. Para lo cual, en nuestro ejemplo,

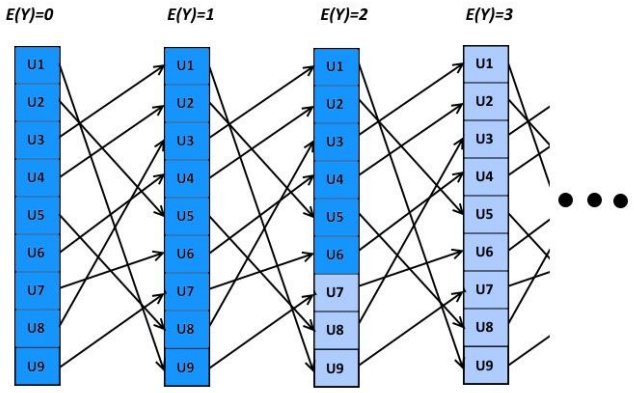
se aplica una permutación completa de todos los elementos, para después aplicar la misma una vez más, pero únicamente a los primeros 3 componentes no ceros del vector U_k. Los componentes que no hayan sido definidos mediante esta última permutación parcial, se copian del vector anterior, para conformar el vector U'.

2. Una vez terminadas las permutaciones horizontales, se aplican las verticales. En nuestro ejemplo se aplican 2 permutaciones completas, más una adicional sobre los primeros 6 elementos no ceros del vector. Es importante señalar que estos desplazamientos se efectúan sobre el vector U' obtenido en la permutación anterior, y no sobre el original. Al igual que en el paso anterior, los elementos no afectados por la permutación parcial se copian de la columna anterior.

Este ejemplo se muestra en la figura 5.6, donde aparecen en tono más oscuro los elementos permutados.



(a)



(b)

Figura 5.6. Ejemplo de permutaciones (a) horizontales y (b) verticales. Las permutaciones realizadas tienen un color más oscuro.

Así pues, el vector final U^*se obtiene al aplicar sucesivamente estos dos esquemas de permutación sobre la máscara vector U de la propiedad F_k (fig. 5.7).

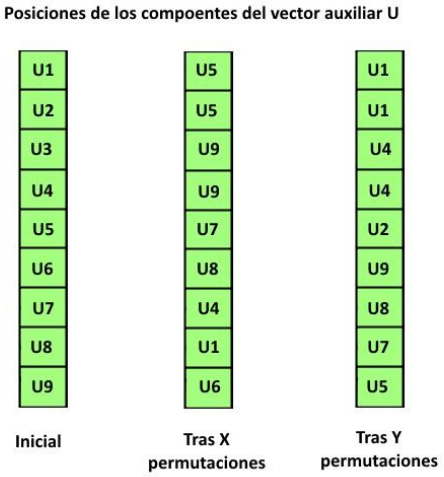


Figura 5.7. Resultados de las permutaciones X y Y.

Una vez calculados todos los vectores U_r^* se crea un vector código a partir de todas las propiedades detectadas en la imagen. El código vector se define como:

$$v_i = \bigcup_r u_{ri}^* \tag{5.10}$$

donde v_i es el i-ésimo componente del vector código V, u_{ri}^* es el i-ésimo componente del vector U_r^*, el cual corresponde a la característica detectada F_k y finalmente ∪ es el signo de disyunción. Después de obtener el vector código, se realiza un procedimiento de conexión (*binding*) [Kuss03, Plat95, Rach01] sobre el vector obtenido V. El *binding* utilizado, se define de la siguiente manera:

$$C = V \tag{5.11}$$

$$C = (C \downarrow)\&(\neg V) \qquad \text{(se repite R veces)} \tag{5.12}$$

$$V = C \tag{5.13}$$

donde C representa un vector auxiliar, $C \downarrow$ es una operación de corrimiento cíclico bit por bit, $\neg V$ es una operación de negación binaria y & representa la operación de conjunción entre los dos valores. Esta operación se repite R veces y al terminar se asigna el valor del vector resultante C al vector V.

El propósito principal del proceso de *binding* es reducir el número de neuronas en la capa asociativa incrementando de esta manera la velocidad de cómputo. Un segundo propósito es el de introducir dependencia en la actividad de la neurona asociativa en el resto de las actividades de otras neuronas asociativas. Esta propiedad es útil para el reconocimiento dado que el clasificador de una capa utilizado en el sistema no permite tomar en cuenta tales dependencias.

Existen diferentes algoritmos para el proceso de *binding*. Una descripción detallada de estos algoritmos se puede ver en [Plat95].

5.5 Entrenamiento

Para comenzar el proceso de entrenamiento, los pesos de todas las conexiones entre las neuronas de la capa asociativa A y la capa de reconocimiento R se establecen en 0. El proceso se repite iterativamente y consta de las siguientes 3 etapas secuenciales:

- ***Primera etapa***. La primera imagen a reconocer es presentada al clasificador PCNC. La imagen es codificada y la excitación E_i en las neuronas de la capa R son calculadas de la siguiente manera:

$$E_i = \sum_{j=1}^{N} a_j \cdot w_{ji} \tag{5.14}$$

  donde E_i es la excitación en la i-ésima neurona de la capa R; a_j es la excitación en la j-ésima neurona de la capa A; y w_{ji} es el peso de la conexión entre la j-ésima neurona de la capa asociativa y la i-ésima neurona de la capa de reconocimiento.

- ***Segunda etapa***. Para garantizar la robustez del reconocimiento, tras calcular la excitación de las neuronas de la capa R, se lee la clase correcta a la que pertenece la imagen con la que se está entrenando. La excitación E_c de la neurona correspondiente de la capa R es recalculada de acuerdo a la siguiente ecuación:

$$E_c^* = E_c \cdot (1 - T_E) \tag{5.15}$$

  donde $0 \leq T_E \leq 1$ determina la excitación adicional que la clase correspondiente debe tener. Este valor oscila experimentalmente entre 0.1 y 0.5.

- ***Tercera etapa***. Se recalculan los pesos de las conexiones en base a los resultados. Denotemos mediante el subíndice j a la neurona ganadora, y con el subíndice c a la neurona de la clase correcta. Si $c = j$, entonces no se realiza ninguna acción. En cambio, si $j \neq c$, entonces se realizan las siguientes modificaciones en los pesos de las conexiones:

$$w_{ic}(t+1) = w_{ic}(t) + a_i \tag{5.16}$$

$$w_{ij}(t+1) = w_{ij}(t) - a_i \tag{5.17}$$

$$si\ \left(w_{ij}(t+1) < 0\right) \Rightarrow w_{ij}(t+1) = 0 \tag{5.18}$$

donde $w_{ij}(t)$ y $w_{ij}(t+1)$ son los peso de las conexiones entre la i-ésima neurona de la capa A

yla j-ésima neurona de la capa R antes y después de la modificación, respectivamente; mientras que a_i es la señal de salida (0 o 1) de la i-ésima neurona de la capa A.

El proceso continúa presentando la siguiente imagen al clasificador hasta agotar el conjunto de imágenes de entrenamiento. Una vez completadas todas las imágenes, se calcula el número total de errores de entrenamiento. Si este número es mayor que un porcentaje predefinido del total de imágenes, entonces se ejecuta un nuevo ciclo de entrenamiento. El proceso se detiene cuando se alcanza un porcentaje menor o igual al establecido, o cuando el número de ciclos de entrenamiento alcanza una cantidad límite preestablecida. Para el reconocimiento de caracteres esta cantidad es de 30 ciclos de entrenamiento, mientras que para el reconocimiento de rostros se utilizaron 200.

Para acelerar el proceso de entrenamiento, el proceso de codificación se ejecuta una sola vez y se almacena en memoria no volátil, dado que este proceso genera los mismos resultados en cada ejecución. De esta manera, en los ciclos subsecuentes de entrenamiento no se utilizan las imágenes en sí, sino los códigos almacenados en disco. De esta manera se ahorran sustancialmente tanto tiempo como recursos computacionales.

5.6 Transformaciones de transvección

Se sabe que el desempeño de un sistema de reconocimiento puede ser mejorado incrementando el conjunto de imágenes de entrenamiento [Lecu98]. Dado que no siempre es posible contar con imágenes reales para tal propósito, se debe recurrir a otras técnicas para incrementar artificialmente la galería. Se puede decir que existen dos técnicas principales para hacer esto; una de ellas es la generación de vistas virtuales y otra es el uso de sencillas transformaciones geométricas [Zhan09]. Existen varias técnicas para generar vistas virtuales, pero en general todas se basan en la generación de una imagen artificial haciendo uso de técnicas de reconstrucción tridimensional y de algunos conocimientos previos de la imagen a generar; en este caso, del rostro. Las transformaciones geométricas, por su parte, hacen uso de herramientas más sencillas y no requieren conocimiento previo del objeto a identificar. Aunque en apariencia las vistas virtuales pudieran considerarse el mejor enfoque, su renderización demanda gran cantidad de tiempo y recursos de procesamiento, además de limitar la técnica al reconocimiento de ciertas imágenes en específico. Es debido a lo anterior que se eligieron las transformaciones geométricas como técnica para mejorar el desempeño del clasificador.

En [Kuss13] se utilizaron 12 desplazamientos de δ pixeles hacia los lados, y hacia arriba y abajo. Como resultado se obtuvo una reducción notable en la tasa de errores, la cual será detallada más adelante. Se compararon estos resultados con los obtenidos utilizando máquinas

de soporte vectorial (SVM) y estos favorecieron a PCNC en todos los experimentos, salvo en 2. Dichos experimentos consistían en identificar rostros con rotaciones leves y pronunciadas a partir de imágenes en posición normal. Debido a esto, se decidió añadir una nueva transformación que reforzara la capacidad de reconocimiento del clasificador. El criterio para elegir la transformación a utilizar es que esta debe estar en concordancia con la base de datos a utilizar y los recursos computacionales de los que se disponga. De esta manera, la técnica seleccionada fue la transvección, también conocida en la bibliografía técnica como *shearing* o *skewing*.

La transvección es una transformación afín que consiste en desplazar cada punto (pixel, si se trata de una imagen) en una dirección fija, por una cantidad proporcional a su distancia signada a partir de una línea que es paralela a dicha dirección [Knud99]. A diferencia de la rotación, que mantiene la forma original de la figura y que solamente cambia de ángulo, la transvección es una transformación que desplaza un eje de manera tal que este deja de ser perpendicular con respecto a otro, causando una deformación en el objeto. La transvección en el plano bidimensional requiere 2 parámetros (factores de transvección), ya que puede ser aplicada de manera vertical u horizontal [Klaw12]. En el plano $\mathbb{R}^2$, la transvección horizontal se define de la siguiente manera:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} x + my \\ y \end{bmatrix} = \begin{bmatrix} 1 & m \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{5.19}$$

Mientras que la transvección vertical se define de la siguiente manera:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} x \\ y + mx \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ m & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{5.20}$$

En ambos casos, m representa el factor de transvección. También es posible aplicar esta transformación de manera simultánea, siendo posible alcanzar un efecto similar a la rotación con determinados factores de transvección [Van14]. De manera condensada, ambas transvecciones se pueden expresar de la siguiente manera:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} x + m_x y \\ y + m_y x \end{bmatrix} = \begin{bmatrix} 1 & m_x \\ m_y & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{5.21}$$

donde m_x y m_y dos factores de transvección distintos. Para realizar el procesamiento de las imágenes de la galería de referencia, utilizaremos un modo de procesamiento de pixeles crudos, esto es, se recalculará la nueva posición de cada uno de ellos, en lugar de hacer uso de coordenadas homogéneas. Para esto se utilizarán las siguientes ecuaciones para realizar las transvecciones horizontal y vertical:

$$Sh(x) = x + \sigma y - \frac{W\sigma}{2} \tag{5.22}$$

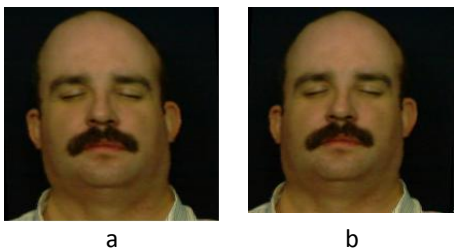
$$Sh(y) = y + \sigma x - \frac{H\sigma}{2} \tag{5.23}$$

donde x e y representan las coordenadas originales del pixel; $Sh(x)$ y $Sh(y)$ representan las nuevas coordenadas de x e y, tras las transvecciones horizontal y vertical, respectivamente; σ representa el factor de transvección; H representa el alto de la imagen y W la anchura de la misma. Las dimensiones de la imagen son representadas en cantidad de pixeles, mientras que el factor σ de transvección es un valor que oscila entre [0, 1].

Con propósitos de prueba, se creó un programa en Visual C++ 2010 para aplicar la transformación en las imágenes de rostros frontales de la base de datos FRAV2D. Su algoritmo consistió en calcular la nueva posición de cada pixel de la imagen, para posteriormente renderizarlos todos en una nueva imagen (fig. 5.8).

Figura 5.8. Programa utilizado para crear imágenes de prueba.

El algoritmo implementado permite tanto la transformación horizontal como la vertical por varios factores de inclinación. Para los espacios vacíos producidos por la transformación se añadieron pixeles negros. En la figura 5.9 se muestran algunos ejemplos de la transformación de rostros usando diferentes ángulos de inclinación (descritos en grados). Posteriormente, estas nuevas imágenes fueron añadidas al conjunto de imágenes de entrenamiento para el clasificador.



a b

c d

Figura 5.9. Diferentes factores de inclinación: a) 5° a la izquierda, b) 5° a la derecha, c) 10° a la izquierda, d) 10° a la derecha

5.7 Implementación y ejecución

El clasificador PCNC se implementó con el entorno de desarrollo de Microsoft Visual C++ 2010. En su desarrollo no se añadió o hizo uso de ninguna librería que no perteneciera al entorno, tal como la biblioteca OpenCV, la cual incluye métodos y rutinas de visión computacional. La ejecución y las pruebas se llevaron a cabo en un equipo de cómputo con Microsoft Windows 7. El equipo utilizado consta de un procesador Intel i7-3610QM de 2.30 GHz y 8 GB de memoria RAM DDR3.

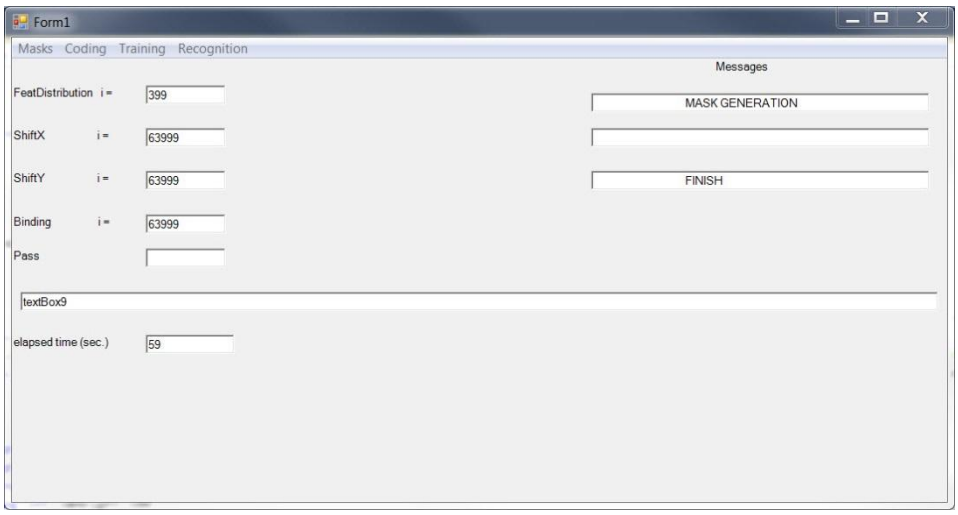


Figura 5.10. Primera etapa de la ejecución del programa: Generación de máscaras.

En las pruebas ejecutadas con este equipo, se requirió un total 61 segundos en promedio para el proceso de creación de máscaras binarias (fig. 5.10) y 1275 segundos (21 minutos, 25 segundos) en promedio para la codificación de características (fig. 5.11).

Figura 5.11. Segunda etapa de la ejecución del programa: Codificación de características.

La etapa con mayor demanda de tiempo y recursos de cómputo es generalmente la etapa de entrenamiento. Con las imágenes añadidas de la transformación de transvección, se requirieron 6928 segundos en promedio (1 hora, 55 minutos aprox.) para completar esta etapa (fig. 5.12).

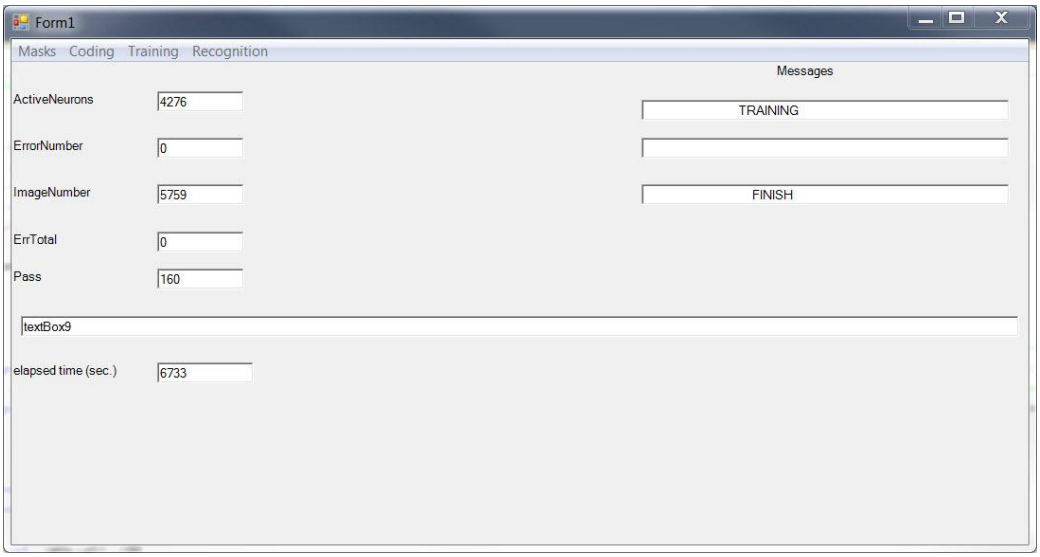


Figura 5.12. Tercera etapa de la ejecución del programa: Proceso de entrenamiento.

Finalmente, la etapa de reconocimiento es la más rápida de todas, lo que hace viable el sistema, una vez entrenado. En todas las pruebas realizadas se requirieron en promedio 6.7 segundos para identificar 1 imagen de prueba en cada ejecución.

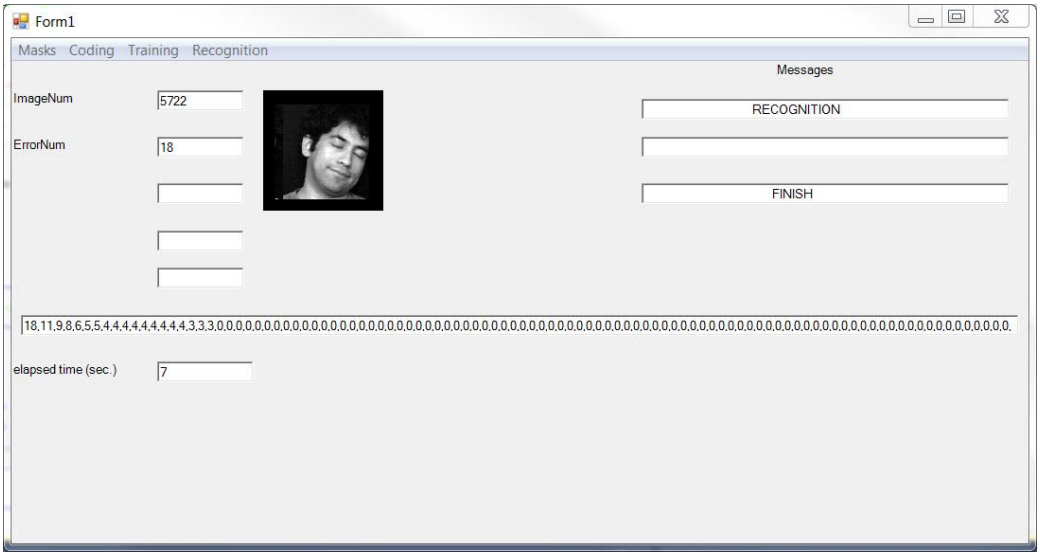


Figura 5.13. Etapa final: Reconocimiento de individuos.

5.8 Experimentación y resultados

En la tabla 5.1 se muestra la definición de pruebas realizadas en experimentos previos [Kuss13]. La tabla muestra cuales imágenes de la base de rostros FRAV2D son utilizadas en el conjunto de entrenamiento y cuales en el conjunto de pruebas. En el capítulo 3 se mencionó que la base de datos FRAV2D contiene 16 tomas por persona en diferentes poses y condiciones de iluminación y expresión. Los comentarios en la primera y segunda columna describen estas condiciones en cada una de la(s) imagen(es).

Número de prueba	Conjunto de entrenamiento	Conjunto de pruebas
1	3 frontal 1,2,3	1 frontal 4
2	4 frontal 1,2,3,4	1 expresión (sonrisa) 11
3	4 frontal 1,2,3,4	1 expresión (boca abierta) 12
4	4 frontal 1,2,3,4	2 iluminación 15,16
5	4 frontal 1,2,3,4	2 rotaciones 5° φ 7,8
6	4 frontal 1,2,3,4	1 rotación pequeña en Z 10
7	4 frontal 1,2,3,4	2 rotaciones X 13,14
8	4 frontal 1,2,3,4	2 rotaciones 25° Y 5,6
9	4 frontal 1,2,3,4	1 rotación pronunciada en Z 9

10	2 frontal, 2 iluminaciones 1,2,15,16	2 frontal 3,4
11	3 frontal, 1 rotación 5° ϕ, 1,2,3,7	1 frontal, 1 rotación 5° ϕ, 4,8
12	3 frontal, 1 rotación 5° ϕ, 1 iluminación 1,2,3,7,15	1 frontal, 1 rotación 5° ϕ, 1 iluminación 4,8,16
13	1,2,3,7,15	4,8,11,16
14	1,2,3,4	11,12
15	1,2	3,4

Tabla 5.1. Definición detallada de las pruebas realizadas con la base de imágenes FRAV2D.

En la tabla 5.2 presentamos la comparación de resultados previos entre el clasificador PCNC y el clasificador SVM [Kuss13]. En este caso fueron utilizadas 12 distorsiones verticales y horizontales, esto es, corrimientos de varios pixeles hacia la izquierda, la derecha, hacia arriba y hacia abajo de la imagen original. Los peores resultados fueron obtenidos para los experimentos 6 y 9 (tabla 5.2), por lo que se decidió mejorar estos resultados mediante distorsiones adicionales añadidas a la galería de entrenamiento.

Número de prueba	Error porcentual con PCNC (12 distorsiones)	Error porcentual con SVM
1	0.78	1.94
2	1.94	5.13
3	1.94	10.68
4	2.45	8.70
5	3.5	14.60
6	33.7	33.93
7	16.4	12.62
8	19.0	27.02
9	54.3	41.14
10	0.59	1.94
11	0.39	4.85
12	0.33	4.17
13	0.44	4.09
14	1.4	8.90
15	0.98	1.46

Tabla 5.2. Comparación de resultados de SVM y PCNC utilizando la base de datos FRAV2D.

En la tabla 5.4 presentamos los nuevos resultados con las distorsiones de transvección adicionales. Estas transformaciones geométricas se realizaron sobre las 4 imágenes frontales de entrenamiento base en diferentes ángulos de distorsión y luego se añadieron a la galería. Posteriormente se seleccionó una combinación de ellas para llevar a cabo el entrenamiento. En

la tabla podemos ver en la segunda columna el número de imágenes adicionales que fueron incluidas por cada ángulo de transvección, mientras que en la tercera y cuarta columna podemos ver el error porcentual promedio considerando 5 ejecuciones del programa.

Número de experimento	Número de imágenes por ángulo	Error promedio (Experimento 6)	Error promedio (Experimento 9)
1	4(-15°) 2(-10°) 2(-5°) 2(+5°) 2(+10°) 4(+15°)	25.5%	42.2%
2	4(-15°) 4(-10°) 4(+10°) 4(+15°)	23.5%	41.2%
3	4(-15°) 4(-10°) 4(-5°) 4(+5°) 4(+10°) 4(+15°)	23.5%	43.1%
4	4 (-15°) 4(-10°) 4(+10°) 4(+15°)	22.5%	45.1%

Tabla 5.3. Resultados de los experimentos 6 y 9 considerando transformaciones de transvección sobre 4 imágenes frontales de la base de datos FRAV2D.

El porcentaje global de error obtenido es de 23.75% para el experimento 6, mientras que para el experimento 9 se obtuvo un valor medio porcentual de 42.9%.

Con base a los resultados de la tabla 5.3, se eligieron las distorsiones que obtuvieron el mejor rendimiento global (experimento número 2, es decir, 4 transformaciones de ±10° y ±15°) para probar el desempeño en el resto de las pruebas especificadas en la tabla 5.1. Los resultados se muestran en la tabla 5.4. Se consideraron 5 ejecuciones para obtener el promedio de errores del algoritmo PCNC considerando imágenes adicionales.

Número de prueba	Error porcentual con PCNC (12 distorsiones)	Error porcentual con PCNC (12 distorsiones + 16 imágenes)	Error porcentual con SVM
1	0.78	0.0	1.94
2	1.94	0.5	5.13
3	1.94	1.75	10.68
4	2.45	2.75	8.70
5	3.5	4.25	14.60
6	33.7	23.5	33.93

7	16.4	11.75	12.62
8	19.0	13.75	27.02
9	54.3	41.2	41.14
10	0.59	0.25	1.94
11	0.39	0.0	4.85
12	0.33	0.0	4.17
13	0.44	0.25	4.09
14	1.4	0.5	8.90
15	0.98	0.0	1.46

Tabla 5.4. Comparación de resultados de SVM y PCNC con y sin imágenes adicionales utilizando la base de datos FRAV2D.

A fin de realizar una comparación del rendimiento base del algoritmo PCNC (sin distorsiones ni imágenes adicionales) en comparación con otros algoritmos, se utilizó la base de datos ORL para ejecutar pruebas [Kuss04b]. Estas pruebas consistieron en seleccionar 5 imágenes de manera aleatoria para entrenamiento, dejando las restantes 5 para pruebas. Durante la investigación bibliográfica se encontraron 3 algoritmos que realizaron el mismo experimento, Eigenfaces [Lawr97, Zhan09], Redes Neuronales Convolucionales con Mapas Autoorganizantes (SOM+CN) [Lawr97, Zhan09] y Redes Neuronales basadas en Decisiones Probabilísticas (PDBNN) [Lin97, Zhan09]. La tabla de errores porcentuales se muestra en la tabla 5.5.

Experimento	Base de datos	% Errores
5 aleatorias / 5 restantes	Eigenfaces	11.5
	SOM+CN	3.8
	PDBNN	4.0
	PCNC	0.1

Tabla 5.5. Comparación del algoritmo PCNC con otros algoritmos de uso general utilizando la base de datos ORL.

En la tabla 5.5 se puede apreciar que el algoritmo PCNC sin distorsiones ni galería adicional tiene el mejor desempeño de las pruebas. Sin embargo, dado que la base de datos ORL tiene imágenes normalizadas y con ángulos de pose muy limitados, se considera una base de datos prácticamente obsoleta. Debido a ello, se procedió a utilizar una base de datos más compleja con la que se pudieran realizar comparaciones con otros algoritmos. La base de datos Achermann descrita en el capítulo 3 se eligió debido a razones de disponibilidad y utilidad. En estas pruebas se utilizó una imagen frontal para entrenamiento y se dejaron 4 imágenes para pruebas. Estas 4 imágenes contienen 2 rotaciones en profundidad o guiñadas hacia la izquierda, y dos hacia la derecha de aproximadamente 20° con respecto a la posición frontal.

A fin de evaluar la influencia de las distorsiones adicionales sobre el algoritmo PCNC, se consideraron 4 casos: el algoritmo PCNC sin distorsiones y con 4, 8 y 12 distorsiones. Se consideraron 5 ejecuciones para obtener el promedio y los resultados se muestran en la tabla 5.6:

Experimento	Base de datos	% Errores
1 frontal / 4 guiñadas (±20°)	Eigenfaces	34.88
	LEM	27.91
	DCP	31.39
	Recuperación de pose cilíndrica 3D	20.0
	PCNC (sin distorsiones)	39.2
	PCNC (4 distorsiones)	21.25
	PCNC (8 distorsiones)	15.0
	PCNC (12 distorsiones)	11.04
	PCNC (4 distorsiones + 2 imágenes con transvección)	20.42
	PCNC (8 distorsiones + 2 imágenes con transvección)	13.13
	PCNC (12 distorsiones + 2 imágenes con transvección)	12.42

Tabla 5.6. Comparación del algoritmo PCNC con y sin distorsiones contra otros algoritmos usando la base de datos Achermann.

Los algoritmos de comparación considerados fueron Eigenfaces, Mapa de Líneas de Contorno (LEM), Puntos de Esquinas Direccionales (DCP) y Recuperación cilíndrica de pose 3D [Gao01]. Los resultados de la tabla son los reportados en [Gao02, Gao05, Gao01 y Zhan09]. Cabe señalar que las 3 primeras técnicas son algoritmos generales sin métodos de compensación de pose, mientras que la cuarta si considera un método de compensación.

Por último, es necesario señalar que adicionalmente a las distorsiones consideradas, se consideraron pruebas para evaluar el desempeño de las distorsiones más 2 imágenes con transformación de transvección de ±15°, obteniéndose ligeramente mejores resultados que con las respectivas pruebas sin imágenes adicionales en los casos de 4 y 8 distorsiones.

5.9 Discusión

En el presente capítulo se realizó la descripción de la red neuronal utilizada, así como las técnicas utilizadas para mejorar su desempeño. Posteriormente se llevaron a cabo experimentos para evaluar el impacto de estas técnicas en el algoritmo y como resultado se obtuvo un valor medio porcentual de 23.5% como tasa de errores para el experimento 6, y 42.9% como tasa de errores para el experimento 9 (tabla 5.3). En relación a la investigación anterior [Kuss13], se obtuvo una mejoría de 10.2% y 11.4% respectivamente.

Adicionalmente, las imágenes transformadas permitieron mejorar el rendimiento del clasificador en casi todas las pruebas planteadas restantes (tabla 5.4), notablemente en los experimentos 7 y 8, lo que permite decir que las imágenes añadidas contribuyen a mejorar la capacidad de generalización del algoritmo.

En la tabla 5.5 se incluyó una tabla comparativa del algoritmo PCNC de propósito general contra otros algoritmos de clasificación específica de rostros a fin de separar el impacto que tiene el algoritmo por si solo y el que corresponde a las transformaciones añadidas.

En la comparativa se utilizó la base de datos ORL y los algoritmos comparados fueron Eigenfaces, Redes Neuronales Convolucionales (SOM+CN) y Redes de Decisión Probabilística (PDBNN). En dicha tabla se puede apreciar con claridad que el peor resultado fue obtenido por el método de Eigenfaces, mientras que los 3 método restantes, basados en redes neuronales obtuvieron resultados notablemente mejores. Esto se debe a que el método de Eigenfaces utiliza un método holístico de extracción de características, lo cual lo vuelve muy vulnerable incluso a los pequeños cambios de pose de la base de datos ORL. Los restantes 3 métodos, en cambio, comparten las características de estar basados en redes neuronales y de realizar la extracción de características de manera local, lo cual les permite una mayor tolerancia a cambios de pose.

Esta tolerancia, sin embargo, está limitada a ángulos de rotación relativamente pequeños (aprox. 30° como máximo), especialmente en rotaciones de profundidad (sobre el eje Y) o cabeceos (sobre el eje X). Para compensar estos cambios de pose es necesario utilizar métodos de extracción de características invariantes a cambios de pose, o bien, utilizar métodos para compensar estas variaciones [Zhan09].

Entre los métodos de compensación de pose más importantes cabe mencionar el enfoque basado en vistas reales, los métodos de transformación de imágenes, los métodos de transformación en el espacio característico y finalmente la modelación tridimensional del rostro. El enfoque basado en vistas reales amplia el conjunto de entrenamiento base mediante imágenes del sujeto adquiridas en diferentes ángulos de toma. Los métodos de transformación de la imagen utilizan transformaciones geométricas para crear imágenes sintetizadas artificialmente que simulan estos cambios de pose. Las transformaciones en el espacio característico buscan mejorar la separabilidad lineal mediante el mapeo lineal o bien, modelar explícitamente el cambio de pose usando aproximaciones lineales. Finalmente, los métodos de modelación tridimensional hacen una reconstrucción virtual en 3D del rostro para simular las poses, haciendo una estimación del rostro y realizando un mapeo de las texturas de la cara base.

En el presente trabajo, se utilizó el método de transformación de imágenes por ser el más directo y barato en términos de costo computacional, obteniéndose una significativa mejoría superior al diez por ciento en las pruebas de reconocimiento de rostros con rotación sobre el eje Z, además de una mejoría considerable en el resto de las pruebas.

Finalmente, en la tabla 5.6 se muestra la comparación entre PCNC y otros 4 algoritmos utilizando la base de datos Achermann, la cual considera variaciones de pose mayores que ORL. Como se puede apreciar, el algoritmo PCNC de propósito general mostró mejor rendimiento al

añadir ligeras distorsiones e imágenes transformadas. Esto es especialmente notable considerando que el algoritmo de recuperación de pose cilíndrica de [Gao01] utiliza un método de compensación de pose para mejorar su rendimiento.

Lo anterior permite concluir que el algoritmo general PCNC tiene un mejor rendimiento en relación a otros algoritmos en imágenes dónde no existen cambios pronunciados (superiores a 30°) de pose. Adicionalmente, se puede inferir que es posible mejorar su desempeño (en el reconocimiento de rostros u otras imágenes) de manera notable incorporando métodos más avanzados de compensación de la pose (tales como modelación 3D), o bien haciendo más flexibles y generales los métodos de extracción de características. Estas propuestas se incluyen en mayor profundidad en las líneas de investigación futura del siguiente capítulo.

Capítulo 6

Conclusiones y trabajo futuro.

En el presente capítulo se presentan las conclusiones finales del proyecto desarrollado, así como de los resultados obtenidos. De igual forma se plantean algunas propuestas de trabajo futuro, como sugerencias para profundizar la investigación o mejorar el desempeño del algoritmo PCNC.

6.1 Conclusiones

El reconocimiento artificial de rostros humanos se ha convertido en una de las ramas de investigación y desarrollo con más auge en la actualidad debido a su amplia gama de aplicaciones en sistemas de vigilancia, en seguridad de datos y en ocio o entretenimiento.

La enorme variabilidad que un rostro puede presentar por diferentes motivos dificulta enormemente las tareas de detección y de reconocimiento del mismo.

Debido a ello se han creado una gran cantidad de algoritmos enfocados a mejorar el rendimiento de manera general, o bien, bajo una o más condiciones de variación en específico. El estudio del estado del arte nos ofreció la posibilidad de conocer algunos de los principales enfoques existentes actualmente para tratar el problema. En este punto es necesario mencionar que si bien varios métodos son muy diferentes entre sí, la mayoría de ellos comparten algunos enfoques o aspectos que no solo pueden resultar útiles para investigaciones posteriores, sino que además son muy interesantes en sí mismos.

En esta tesis se presentó una propuesta para mejorar el rendimiento de un clasificador neuronal existente, PCNC, mediante la ampliación artificial de imágenes de la galería de

entrenamiento. Se alcanzó el objetivo de mejorar la tasa de reconocimiento en relación con los resultados obtenidos en estudios anteriores.

No obstante, las tasas de error obtenidas en las imágenes con variación de pose aún pueden ser consideradas como relativamente altas, sobre todo teniendo en mente aplicaciones del mundo real. Debido a ello, es necesario profundizar el estudio de técnicas y herramientas adicionales para hacer más robusto el clasificador frente a estas variaciones.

6.2 Aportaciones

La presente tesis ha presentado un trabajo de investigación sobre la tarea de reconocimiento facial automático, centrándose específicamente en el problema de reconocimiento bajo condiciones de cambio de pose del individuo. Como resultado de dicho estudio, las principales aportaciones fueron:

- Ofrecer un estudio profundo del estado del arte de los algoritmos generales de reconocimiento, así como de las técnicas de compensación más comúnmente aplicadas para subsanar los cambios pronunciados de posición.
- Ofrecer un estudio detallado de las principales bases de datos de rostros que existen actualmente en esta área de investigación, así como llevar a cabo una recopilación de varias de ellas para realizar una comparación posterior con otros algoritmos.
- Estudiar y realizar una descripción detallada del clasificador de imágenes de uso general PCNC, así como adaptarlo para las tareas de reconocimiento de rostros de las bases de datos FRAV y Achermann.
- Identificar e implementar un método de compensación de la imagen basado en transformaciones geométricas para mejorar el reconocimiento del clasificador neuronal previamente existente.
- Evaluar el rendimiento del clasificador PCNC llevando a cabo una comparación con otros algoritmos de reconocimiento actuales.
- Identificar puntos de mejoría en el clasificador para ser llevados a cabo en investigaciones futuras utilizando las bases de imágenes adquiridas.

6.3 Trabajo futuro

Las siguientes propuestas de trabajo futuro se presentan con dos objetivos: ofrecer ideas para mejorar el desempeño y eficiencia del presente trabajo o bien, sugerir formas de como apoyarse en el para lograr otros objetivos. Las propuestas son las siguientes:

- A fin de mejorar la velocidad del sistema, se debe considerar transformar el algoritmo secuencial utilizado actualmente, a un algoritmo de procesamiento en paralelo para ser implementado en una tarjeta gráfica o similar que cuente con múltiples núcleos de procesamiento. Esta idea se considera altamente factible debido a que se aplica una misma gama de operaciones básicas a un conjunto de imágenes de manera independiente.
- Realizar pruebas utilizando una base de datos de imágenes más amplia en cuanto condiciones de captura, tal como CAS-PEAL o FEI o GTAV. O bien, una base de datos más extendida (con propósitos de comparación) como FERET o CMU-PIE.
- Estudiar métodos adicionales de transformación geométrica, tales como escalamiento bidimensional, o bien transformaciones artificiales en profundidad (tercera dimensión), como rotación o transvección.
- Estudiar el rendimiento utilizando vistas virtuales simples o bien, la creación de imágenes de mosaico.
- Hacer uso del conocimiento previo del objeto que se desea reconocer, en este caso el rostro. El presente algoritmo asigna la misma importancia a todas las características detectadas. Sin embargo, a partir de las ideas del algoritmo de Eigenfaces y de los estudios biológicos se desprende que algunas características son ayudan a discriminar mejor las clases.
- Investigar un método para lograr la invariancia a rotaciones en el eje Z. Una manera posible de lograr esto es emplear el algoritmo utilizado por LBP.
- En la búsqueda de puntos de interés, se plantea la posibilidad de remplazar el detector de bordes Sobel empleado actualmente por un detector SIFT o SURF.

Referencias

[Adin97] ADINI, Yael; MOSES, Yael; ULLMAN, Shimon. *Face recognition: The problem of compensating for changes in illumination direction*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1997, vol. 19, no 7, p. 721-732.

[Ahon04] AHONEN, Timo; HADID, Abdenour; PIETIKÄINEN, Matti. *Face recognition with local binary patterns*. European Conference in Computer Vision 2004. Springer Berlin Heidelberg, 2004. p. 469-481.

[Ahon06] AHONEN, Timo; HADID, Abdenour; PIETIKAINEN, Matti. *Face description with local binary patterns: Application to face recognition*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2006, vol. 28, no 12, p. 2037-2041.

[Baid04] BAIDYK, Tatiana, et al. *Flat image recognition in the process of microdevice assembly*. Pattern Recognition Letters, 2004, vol. 25, no 1, p. 107-118.

[Baid05] BAIDYK, Tatiana; KUSSUL, Ernst; MAKEYEV, Oleksandr. *Texture recognition with random subspace neural classifier*. WSEAS Transactions on circuits and systems, 2005, vol. 4, no 4, p. 319-324.

[Baid08] BAIDYK, Tatiana, et al. *Limited receptive area neural classifier based image recognition in micromechanics and agriculture*. International Journal of Applied Mathematics and Informatics, 2008, vol. 2, no 3, p. 96-103.

[Behn03] BEHNKE, Sven, ed. *Hierarchical neural networks for image interpretation*. Vol. 2766. Springer, 2003.

[Belh97] BELHUMEUR, Peter N..; HESPANHA, João P.; KRIEGMAN, David. *Eigenfaces vs. fisherfaces: Recognition using class specific linear projection*.Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1997, vol. 19, no 7, p. 711-720.

[Bish06] BISHOP, Christopher M., et al. *Pattern recognition and machine learning*. Springer, 2006.

[Bruc86] BRUCE, Vicki; YOUNG, Andy. *Understanding face recognition*. British journal of psychology, 1986, vol. 77, no 3, p. 305-327.

[Cond06] CONDE, C. *Verificación facial multimodal: 2D y 3D*. Tesis Doctoral. Universidad Rey Juan Carlos, Madrid, España, 2006.

[Cort95] CORTES, Corinna; VAPNIK, Vladimir. *Support-vector networks*. Machine learning, 1995, vol. 20, no 3, p. 273-297.

[Chel95] CHELLAPPA, Rama; WILSON, Charles L.; SIROHEY, Saad. *Human and machine recognition of faces: A survey*. Proceedings of the IEEE, 1995, vol. 83, no 5, p. 705-741.

[Cire11] CIREŞAN, Dan C., et al. *Flexible, high performance convolutional neural*

networks for image classification. Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Two. AAAI Press, 2011.

[Coot95] COOTES, Timothy F., et al. *Active shape models-their training and application*. Computer vision and image understanding, 1995, vol. 61, no 1, p. 38-59.

[Coot98] COOTES, Timothy F.; EDWARDS, Gareth J.; TAYLOR, Christopher J. *Active appearance models*. En Computer Vision—ECCV'98. Springer Berlin Heidelberg, 1998. p. 484-498.

[Coot01] COOTES, Timothy F.; EDWARDS, Gareth J.; TAYLOR, Christopher J. *Active appearance models*. IEEE Transactions on pattern analysis and machine intelligence, 2001, vol. 23, no 6, p. 681-685.

[Coot02] COOTES, Timothy F., et al. *View-based active appearance models*. Image and vision computing, 2002, vol. 20, no 9, p. 657-664.

[Duda99] DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern classification*. John Wiley & Sons, 1999.

[Fors02] FORSYTH, David A.; PONCE, Jean. *Computer vision: A modern approach*. Prentice Hall Professional Technical Reference, 2002.

[Fuku80] FUKUSHIMA, Kunihiko. *Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position*. Biological cybernetics 36.4 (1980), p. 193-202.

[Gao01] GAO, Yongsheng, et al. *Fast face identification under varying pose from a single 2-D model view*. IEE Proceedings-Vision, Image and Signal Processing, 2001, vol. 148, no 4, p. 248-253.

[Gao02] GAO, Yongsheng; LEUNG, Maylor KH. *Face recognition using line edge map*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2002, vol. 24, no 6, p. 764-779.

[Gao04] GAO, Wen, et al. *The CAS-PEAL large-scale Chinese face database and evaluation protocols*. Technique Report No. JDL-TR_04_FR_001, Joint Research & Development Laboratory, CAS, 2004.

[Gao05] GAO, Yongsheng; QI, Yutao. *Robust visual similarity retrieval in single model face databases*. Pattern Recognition, 2005, vol. 38, no 7, p. 1009-1020.

[Geor01] GEORGHIADES, Athinodoros S.. ; BELHUMEUR, Peter N.. ; KRIEGMAN, David. *From few to many: Illumination cone models for face recognition under variable lighting and pose*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2001, vol. 23, no 6, p. 643-660.

[Gold09] GOLDSTEIN, E. Bruce (ed.). *Encyclopedia of perception*. Sage Publications, 2009.

[Gonz96] GONZÁLEZ, Rafael C.; WOODS, Richard E. *Tratamiento digital de imágenes*.

Addison-Wesley, 1996.

[Grgi11] GRGIC, Mislav; DELAC, Kresimir; GRGIC, Sonja. *SCface–surveillance cameras face database*. Multimedia tools and applications, 2011, vol. 51, no 3, p. 863-879

[Gros05] GROSS, Ralph. *Face databases. Handbook of Face Recognition.* Springer New York, 2005. p. 301-327.

[Guo00] GUO, Guodong; LI, Stan Z.; CHAN, Kapluk. *Face recognition by support vector machines*. En Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on. IEEE, 2000. p. 196-201.

[Haxb00] HAXBY, James V.; HOFFMAN, Elizabeth A.; GOBBINI, M. Ida. *The distributed human neural system for face perception*. Trends in cognitive sciences, 2000, vol. 4, no 6, p. 223-233.

[He05] HE, Xiaofei, et al. *Face recognition using Laplacianfaces*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2005, vol. 27, no 3, p. 328-340.

[Hjel01] HJELMÅS, Erik; LOW, Boon Kee. *Face detection: A survey*. Computer vision and image understanding, 2001, vol. 83, no 3, p. 236-274.

[Hopf82] HOPFIELD, John J. *Neural networks and physical systems with emergent collective computational abilities*. Proceedings of the national academy of sciences, 1982, vol. 79, no 8, p. 2554-2558.

[Hopf84] HOPFIELD, John J. *Neurons with graded response have collective computational properties like those of two-state neurons*. Proceedings of the national academy of sciences, 1984, vol. 81, no 10, p. 3088-3092.

[Hube62] HUBEL, David H.; WIESEL, Torsten N. *Receptive fields, binocular interaction and functional architecture in the cat's visual cortex*. The Journal of physiology, 1962, vol. 160, no 1, p. 106.

[Hube68] HUBEL, David H.; WIESEL, Torsten N. *Receptive fields and functional architecture of monkey striate cortex*. The Journal of physiology, 1968, vol. 195, no 1, p. 215-243.

[Huan88] HUANG, William Y.; LIPPMANN, Richard P. *Neural net and traditional classifiers*. Neural information processing systems. 1988. p. 387-396.

[Ibar14] IBARRA, Joaquín; KUSSUL, Ernst; BAIDYK, Tatiana; CRUZ, Zamira. *Face Recognition with Permutation Coding Neural Classifier improved with skewing transformations*. Proceedings of the International Conference on Machine Vision and Machine Learning, 2014.

[Jahn00] JAHNE, Bernd (ed.). *Computer vision and applications: a guide for students and practitioners*. Academic Press, 2000.

[Jain04] JAIN, Anil K., et al. *Biometrics: a grand challenge*. Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. IEEE, 2004. p.

935-942.

[Kirb90] KIRBY, Michael; SIROVICH, Lawrence. *Application of the Karhunen-Loeve procedure for the characterization of human faces*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1990, vol. 12, no 1, p. 103-108.

[Klet14] KLETTE, Reinhard. *Concise Computer Vision*. Springer, 2014.

[Knud99] KNUDSEN, Jonathan. *Java 2D graphics*. O'Reilly Media, Inc., 1999.

[Kuss91a] KUSSUL, E. M.; RACHKOVSKIJ, D. A.; BAIDYK, T. N. *Associative-projective neural networks: Architecture, implementation, applications*. Proceedings of the Fourth International Conference on Neural Networks & their Applications", Nimes, France. 1991. p. 463-476.

[Kuss91b] KUSSUL, E. M.; RACHKOVSKIJ, D. A.; BAIDYK, T. N. *On image texture recognition by associative-projective neurocomputer*. Proceedings of the ANNIE'91 Conference on Intelligent engineering systems through artificial neural networks. 1991. p. 453-458.

[Kuss94] KUSSUL, Ernst M., et al. *Adaptive high performance classifier based on random threshold neurons*. Cybernetics and Systems, 1994, vol. 94, p. 1687-1695.

[Kuss98] KUSSUL, Ernst M., et al. *Application of random threshold neural networks for diagnostics of micro machine tool condition*. Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on. IEEE, 1998. p. 241-244.

[Kuss03] KUSSUL, Ernst; BAIDYK, Tatiana; et al. *Permutative coding technique for handwritten digit recognition system*. Proceedings of the International Joint Conference on Neural Network, 2003. p. 2163-2168.

[Kuss04a] KUSSUL, Ernst; BAIDYK, Tatiana. *Improved method of handwritten digit recognition tested on MNIST database*. Image and Vision Computing, 2004, vol. 22, no 12, p. 971-981.

[Kuss04b] KUSSUL, E.; BAIDYK, Tatiana; KUSSUL, Maksym. *Neural network system for face recognition*. Circuits and Systems, 2004. ISCAS'04. Proceedings of the 2004 International Symposium on. IEEE, 2004, vol. 5, p. V-768-V-771.

[Kuss05] KUSSUL, E.; BAIDYK, Tatiana; WUNSCH, D. C. *Image recognition systems with permutative coding*. Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on. IEEE, 2005. p. 1788-1793.

[Kuss06a] KUSSUL, Ernst, et al. *Image recognition systems based on random local descriptors*. Neural Networks, 2006. IJCNN'06. International Joint Conference on. IEEE, 2006. p. 2415-2420.

[Kuss06b] KUSSUL, Ernst M., et al. *Permutation coding technique for image recognition systems*. Neural Networks, IEEE Transactions on, 2006, vol. 17, no 6, p. 1566-1579.

[Kuss07] KUSSUL, E.; BAIDYK, Tatiana; MAKEYEV, Oleksandr. *Pairwise permutation coding neural Classifier*. Neural Networks, 2007. IJCNN 2007. International Joint Conference on. IEEE, 2007. p. 1847-1852.

[Kuss10] KUSSUL, Ernst; BAIDYK, Tatiana; WUNSCH, Donald C. *Neural networks and micromechanics*. Germany. Springer, 2010.

[Kuss13] KUSSUL, E.; BAIDYK, Tatiana; CONDE, Cristina; MARTÍN, Isaac, CABELLO, Enrique. *Face Recognition improvement with distortions of images in training set*. Proceedings of International Joint Conference on Neural Networks, 2013.

[Lade93] LADES, Martin, et al. *Distortion invariant object recognition in the dynamic link architecture*. Computers, IEEE Transactions on, 1993, vol. 42, no 3, p. 300-311.

[Lawr97] LAWRENCE, Steve, et al. *Face recognition: A convolutional neural-network approach.* Neural Networks, IEEE Transactions on 8.1 (1997): 98-113.

[Lecu98] LECUN, Yann, et al. *Gradient-based learning applied to document recognition*. Proceedings of the IEEE, 1998, vol. 86, no 11, p. 2278-2324.

[Li07] LI, Stan Z., et al. *Illumination invariant face recognition using near-infrared images*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2007, vol. 29, no 4, p. 627-639.

[Lin97] LIN, Shang-Hung; KUNG, Sun-Yuan; LIN, Long-Ji. *Face recognition/detection by probabilistic decision-based neural network*. Neural Networks, IEEE Transactions on, 1997, vol. 8, no 1, p. 114-132.

[Liu00] LIU, Chengjun; WECHSLER, Harry. *Evolutionary pursuit and its application to face recognition*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2000, vol. 22, no 6, p. 570-582.

[Make08] MAKEYEV, Oleksandr, et al. *Limited receptive area neural classifier for texture recognition of mechanically treated metal surfaces*. Neurocomputing, 2008, vol. 71, no 7, p. 1413-1421.

[Mart07] MARTIN, Bonifacio, & SANZ, Afredo. (2007). *Redes neuronales y sistemas borrosos*, 3a. ed., Editorial Alfaomega Ra-Ma. México, D.F.

[Morr04] MORRIS, Tim. *Computer Vision and Image Processing*. Palgrave Macmillan, 2004.

[Ojal02] OJALA, Timo; PIETIKAINEN, Matti; MAENPAA, Topi. *Multiresolution gray-scale and rotation invariant texture classification with local binary patterns*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2002, vol. 24, no 7, p. 971-987.

[Oliv06] OLIVEIRA JR, L. L.; THOMAZ, C. E. *Captura e alinhamento de imagens: Um banco de faces brasileiro*. Relatório de iniciação científico, Departamento de Engenharia Elétrica da FEI, São Bernardo do Campo, SP, 2006, vol. 10.

[Phil98] PHILLIPS, P. Jonathon, et al. *The FERET database and evaluation procedure for face-recognition algorithms*. Image and vision computing, 1998, vol. 16, no 5, p. 295-306.

[Phil00] PHILLIPS, P. Jonathon, et al. *The FERET evaluation methodology for face-recognition algorithms*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2000, vol. 22, no 10, p. 1090-1104.

[Phil02] PHILLIPS, P. J., NEWTON, E., *Meta-Analysis of Face Recognition Algorithms*, Proc. Fifth Int'l Conf. Automatic Face & Gesture Recognition, pp. 235-241, 2002.

[Plat95] PLATE, Tony A. *Holographic reduced representations*. Neural networks, IEEE transactions on, 1995, vol. 6, no 3, p. 623-641.

[Rach01] RACHKOVSKIJ, Dmitri A.; KUSSUL, Ernst M. *Binding and normalization of binary sparse distributed representations by context-dependent thinning*. Neural Computation, 2001, vol. 13, no 2, p. 411-452.

[Rivo13] RIVOLTA, D. *Prosopagnosia. When All Faces Look the Same*. Springer, 2013.

[Rodr12] RODRÍGUEZ, Roberto, & SOSSA, Juan Humberto. *Procesamiento y Análisis Digital de Imágenes*. 2012, Editorial Alfaomega Ra-Ma. México, D.F.

[Rose62] ROSENBLATT, Frank. *Principles of neurodynamics*. 1962.

[Russ03] RUSSELL, Stuart J.; NORVIG, Peter. *Artificial Intelligence: A Modern Approach*. 2nd ed. Prentice Hall, 2003.

[Sama94] SAMARIA, F.S., HARTER, A.C., *Parameterisation of a stochastic model for human face identification*. IEEE Workshop on Applications of Computer Vision. pp. 138–142, 1994.

[Shap01] SHAPIRO, Linda G.; STOKHAM, George C. *Computer Vision*. Prentice Hall, 2001.

[Sima03] SIMARD, Patrice, David STEINKRAUS, and John C. PLATT. *Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis*. ICDAR. Vol. 3. 2003.

[Simo13] SIMON, Phil. *Too Big to Ignore: The Business Case for Big Data*. John Wiley & Sons, 2013.

[Sinh06] SINHA, Pawan, et al. *Face recognition by humans: Nineteen results all computer vision researchers should know about*. Proceedings of the IEEE, 2006, vol. 94, no 11, p. 1948-1962.

[Siro87] SIROVICH, Lawrence; KIRBY, Michael. *Low-dimensional procedure for the characterization of human faces*. JOSA A, 1987, vol. 4, no 3, p. 519-524.

[Stan05] STAN, Z. Li; ANIL, K. Jain. *Handbook of face recognition*. 2005.

[Tan06] TAN, Xiaoyang, et al. *Face recognition from a single image per person: A survey*. Pattern Recognition, 2006, vol. 39, no 9, p. 1725-1745.

[Sobe68] SOBEL, Irwin; FELDMAN, Gary. *A 3x3 isotropic gradient operator for image processing*. Stanford Artificial Project, 1968. p. 271-272.

[Turk91a] TURK, Matthew A.; PENTLAND, Alex P. *Face recognition using eigenfaces*. Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on. IEEE, 1991. p. 586-591.

[Turk91b] TURK, Matthew; PENTLAND, Alex. *Eigenfaces for recognition*. Journal of cognitive neuroscience, 1991, vol. 3, no 1, p. 71-86.

[Werb74] WERBOS, Paul. *Beyond regression: New tools for prediction and analysis in the behavioral sciences*. 1974.

[Widr90] WIDROW, B., & LEHR, M. A. (1990). *30 years of adaptive neural networks: perceptron, madaline, and backpropagation*. Proceedings of the IEEE, 78(9), 1415-1442.

[Wisk97] WISKOTT, Laurenz, et al. *Face recognition by elastic bunch graph matching*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1997, vol. 19, no 7, p. 775-779.

[Zhan09] ZHANG, Xiaozheng; GAO, Yongsheng. *Face recognition across pose: A review*. Pattern Recognition, 2009, vol. 42, no 11, p. 2876-2896.

[Zhao03] ZHAO, W., CHELLAPPA, R., PHILLIPS, P. J., & ROSENFELD, A., *Face recognition: A literature survey*. ACM Computing Surveys (CSUR), 35(4), pp. 399-458, 2003.

[Zhao07] ZHAO, Guoying; PIETIKAINEN, Matti. *Dynamic texture recognition using local binary patterns with an application to facial expressions*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2007, vol. 29, no 6, p. 915-928.