



UNIVERSIDAD DE BUENOS AIRES
FACULTAD DE CIENCIAS EXACTAS Y NATURALES
DEPARTAMENTO DE COMPUTACIÓN

EIGENFUNGI

**Desarrollo de un método de Data Mining para la detección automática
de patrones en microscopía aplicada a micología médica**

**Tesis presentada para optar al título de
Doctor de la Universidad de Buenos Aires
en el área de Computación**

Marcela Leticia Riccillo

Directores de tesis: Dr. Oscar Bustos

Dr. Marcelo A. Soria

Consejera de Estudios: Dra. Ana Silvia Haedo

Lugar de trabajo: Departamento de Computación, FCEyN, UBA

Buenos Aires, 2007

*A mis padres
Ana María y Jorge*

*A mis hermanos
Lorena, Esteban, Nancy, Diego y Daniel*

y a Nicolás y Facundo (B-M-R)

Agradecimientos

Agradezco profundamente a mis directores Dr. Bustos y Dr. Soria por el apoyo y ayuda que me han brindado, en especial a mi consejera Dra. Haedo por acompañarme en la realización de este trabajo.

También a todos aquellos quienes de una u otra manera me ayudaron en este largo camino de estudio y formación, en especial a Natalia Debandi.

Agradezco al Instituto Malbrán por la posibilidad de contar con las muestras de hongos microscópicos para la validación del método desarrollado.

Agradezco a la Facultad de Ciencias Exactas y Naturales y a sus profesores, por la formación de excelencia brindada.

Agradezco enormemente a mi familia, por soportarme (en el sentido de soporte y de aguante) durante este tiempo y siempre.

Y muy especialmente, gracias a Dios por permitirme llegar a este momento.

Resumen

En este trabajo desarrollamos un método automático para el reconocimiento de especies de hongos microscópicos, que denominamos *eigenfungi*.

Está basado en la metodología para reconocimiento de rostros denominada *eigenfaces*, a la que se le introducen varias modificaciones que mejoran su exactitud en el análisis de imágenes microscópicas de hongos.

En los últimos años se registra un incremento en las infecciones causadas por hongos. Debido a la necesidad de entrenamiento específico que requiere el análisis microscópico, el diseño e implementación de herramientas informáticas que asistan al personal recibe creciente atención.

Este método transforma las imágenes y aplica técnicas propias de Data Mining, considerando al conjunto de imágenes como una base de datos. Se fundamenta en la aplicación en imágenes del Análisis de Componentes Principales (PCA) que descompone datos multidimensionales a un subespacio de menor dimensión pero preservando las características esenciales de los datos tratados.

No necesita de recortes manuales de los objetos por parte del experto humano y requiere de pocas imágenes para el entrenamiento.

Para la elaboración y validación de la metodología, se estudiaron imágenes de hongos microscópicos de las seis especies principales de dermatofitos, obtenidas de muestras provistas por el Departamento de Micología del Instituto Nacional de Enfermedades Infecciosas (INEI), ANLIS "Carlos G. Malbrán".

Se compararon los resultados obtenidos, con variantes del algoritmo PCA (generación de multiespacios, utilización de distancia Manhattan, combinación con preprocesamientos), y posteriormente con otro método de Data Mining aplicado al reconocimiento de rostros llamado *fisherfaces* que se basa en el Análisis Discriminante.

Palabras clave: *Eigenfungi, Eigenfaces, Fisherfaces, Hongos microscópicos, Reconocimiento de patrones, Análisis de Componentes Principales, Análisis Discriminante, Transformada de Hotelling*

Abstract

In this thesis we present an automatic method for the recognition of microscopic fungi, that we called *eigenfungi*.

This method is based on the methodology for face recognition called *eigenfaces*, with some modifications that improve their precision as recognizer of microscopical fungi images.

In the last years an increase in the infections caused by fungi is registered. Design and implementation of computer-aided tools for staff attendance is very important, due to the necessity for specific training that microscopic analysis require.

This method transforms images and applies *Data Mining* techniques, treating images like databases. It based on the application in images of the Principal Component Analysis (PCA). The PCA takes apart multidimensional data in a smaller dimensional subspace but preserving the principal characteristics of the data.

It doesn't need manual cuts of objects by expert humans and it requires few images for training.

We compared the results with variants of PCA algorithm –generation of multispaces, Manhattan distance, preprocessing- and then we compared the method with another Data Mining method for face recognition called fisherfaces based on Discriminant Analysis (LDA).

For the construction and validation of this methodology, we studied microscopic fungus images of the six principal species of dermatophytes, obtained from samples provided by the Micology Department of Instituto Nacional de Enfermedades Infecciosas (INEI), ANLIS "Carlos G. Malbrán".

Key words: *Eigenfungi, Eigenfaces, Fisherfaces, Microscopic fung, Pattern Recognition, Principal Component Analysis, Discriminant Analysis, Hotelling Transform*

Tabla de Contenidos

Parte I: Conceptos Generales	14
1. Introducción	14
1.1. Antecedentes	14
1.2. CMEIAS	16
1.3. Aplicaciones para la identificación de hongos	17
1.4. Objetivo del trabajo	18
1.5. Organización	18
2. El Reino de los Hongos	19
2.1. Características	19
2.2. Miosis	20
2.3. Epidemiología	20
2.4. Observaciones microscópicas	21
2.5. Dermatofitos	22
2.6. Descripción de las principales dermatofitosis	23
2.7. Descripción de las principales especies de dermatofitos	24
2.8. Recolección y manipulación de las muestras	28
2.9. Diagnóstico	31
2.10. Antifúngicos	32
Parte II: Métodos de Data Mining	33
3. Análisis univariado y multivariado	33
3.1. Análisis univariado – Estadísticos descriptivos	33
3.2. Análisis multivariado	35
3.3. Covarianza	36
3.4. Correlación de las variables	36
4. Análisis de Componentes Principales	38
4.1. Obtención de las nuevas componentes	38
4.2. Autovalores y Autovectores	38
4.3. Autovalores y autovectores generalizados	39
4.4. Proyecciones en los nuevos ejes	40
4.5. Propiedades de PCA	41
4.6. PCA para compresión de datos	43
5. Análisis Discriminante	45
5.1. Reglas de clasificación	45
5.2. Estimaciones de las probabilidades de una clasificación errónea	46
5.3. Funciones discriminantes canónicas	47
Parte III: Reconocimiento de Rostros	48
6. Tratamiento de Imágenes	48
6.1. Histograma de una imagen	51
6.2. Filtros y Convolución	52
6.3. Transformación de las imágenes en datos	53
6.4. Transformada de Hotelling	54

6.5.	Compresión de imágenes	55
7.	Técnicas de reconocimiento de rostros	57
7.1.	Técnicas por características	57
7.2.	Problemática	58
7.3.	Análisis de bajo nivel	59
7.4.	Análisis de rasgos	60
7.5.	Modelos de silueta activa	60
Parte IV: Método de reconocimiento Eigenfunghi		62
8.	Desarrollo del método	62
8.1.	Diferencias entre rostros y hongos	64
8.2.	Descripción del método	65
8.3.	Cálculo de los eigenfunghi	65
Parte V: Pruebas experimentales		69
9.	Características de las pruebas	69
9.1.	Software desarrollado para las pruebas	70
9.2.	Tipos de pruebas	70
10.	Primeras pruebas - eigenimages	71
10.1.	Aplicación de los eigenfunghi a los objetos	73
11.	Pruebas con dermatofitos	74
11.1.	Pruebas binarias: E. floccosum versus M. canis	74
11.2.	Pruebas con todas las especies: eigenfaces	75
11.3.	Pruebas con todas las especies: eigenfunghi	75
11.4.	Pruebas con preprocesamientos	77
11.5.	Resultados obtenidos	81
11.6.	Preprocesamiento seleccionado	83
11.7.	Otros preprocesamientos con porcentajes altos	84
12.	Pruebas de Robustez	86
12.1.	Ruido gaussiano	86
12.2.	Ruido "sal y pimienta"	87
13.	Pruebas con dos muestras	90
13.1.	Pruebas con todas las especies: eigenfaces	90
13.2.	Pruebas con todas las especies: eigenfunghi	91
13.3.	Pruebas con preprocesamientos	92
13.4.	Preprocesamiento seleccionado	95
13.5.	Otros procesamientos con valores altos	96
13.6.	Resumen de preprocesamientos	96
14.	Errores de origen	98
15.	Pruebas de predicción	100
15.1.	Pruebas con todas las especies: eigenfaces	100
15.2.	Pruebas con todas las especies: eigenfunghi	101
15.3.	Pruebas con preprocesamientos	102
15.4.	Preprocesamiento seleccionado	105
15.5.	Resumen de preprocesamientos	106
Parte VI: Comparaciones con otros métodos de Data Mining		107
16.	Variantes de PCA	107

16.1.	Multiespacios	107
16.2.	Distancia Manhattan	109
16.3.	Combinación de ambas variantes	110
16.4.	Combinación de ambas variantes con preprocesamientos	110
17.	Método de Fisherfaces	112
17.1.	Cálculo de las fisherfaces	112
17.2.	Comparación entre las eigenfaces y las fisherfaces	113
17.3.	Fisherfaces aplicado a los dermatofitos	114
17.4.	Fisherfaces con preprocesamientos	116
Parte VII: Conclusiones y trabajos futuros		117
17.5.	Características del método	119
Bibliografía		122
Anexo A – Software utilizado		129
	ImageJ	129
	Fisherfaces	130
Anexo B – Software desarrollado		132
1.1.	Características de las imágenes	132
1.2.	Armado de directorios	133
1.3.	Descripción del programa	133
1.4.	Variantes del programa	134
1.5.	Código Matlab Eigenfungi	135
Anexo C – Ejemplos de PCA y LDA		139
Análisis de Componentes Principales (PCA)		139
Análisis Discriminante Linear de Fisher (LDA)		144

Índice de Figuras

Fig. 1.1 – Imagen de una comunidad de bacterias y ejemplos de formas detectadas...	16
Fig. 1.2 – Ejemplo de pantalla del programa CMEIAS donde se observa una imagen analizada y la cantidad de morfotipos encontrados y clasificados	16
Fig. 2.1 – <i>Microsporum canis</i> – Se observan hifas y esporas	19
Fig. 2.2 - <i>Flammulina velutipes</i> – Se observa un conjunto de setas	20
Fig. 2.3 - Cultivos de hongos filamentosos.....	21
Fig. 2.4 – Imagen de un dermatofito <i>E. floccosum</i>	22
Fig. 2.5 – Ejemplos de microconidias.....	22
Fig. 2.6 – Ejemplos de macroconidias	22
Fig. 2.7 – Ejemplo de imagen microscópica de <i>E. floccosum</i> y colonia	25
Fig. 2.8 – Ejemplo de imagen microscópica de <i>M. canis</i> y colonias	25
Fig. 2.9 – Ejemplo de infección de <i>M. canis</i> en animales	26
Fig. 2.10 – Ejemplo de imagen microscópica de <i>M. gypseum</i> y colonia	26
Fig. 2.11 – Ejemplo de imagen microscópica de <i>T. mentagrophytes</i> y anverso y reverso de colonia	27
Fig. 2.12 – Lesión en un brazo (tinea corporis) causada por <i>T. mentagrophytes</i>	27
Fig. 2.13 – Ejemplo de imagen microscópica de <i>T. rubrum</i> y colonia.....	28
Fig. 2.14 – Ejemplo de imagen microscópica de <i>T. tonsurans</i> y cultivos.....	28
Fig. 2.15 – Recolección de muestras de especímenes en lesiones.....	29
Fig. 2.16 – Preparación de cultivo [DAV04]	31
Fig. 2.17 - Adhesión de conidios de <i>Trichophyton mentagrophytes</i> al estrato córneo	32
Fig. 3.1 – Se observan distintos tipos de relación entre dos variables en forma gráfica. En el primer caso, aparentemente no habría una relación entre las variables. En el segundo caso, la relación no es lineal. En los casos siguientes se ven correlaciones altas, positiva y negativa respectivamente.	37
Fig. 4.1 – PCA conforma un nuevo sistema de ejes. X1 y X2 conforman las variables originales y Y1 y Y2 las nuevas componentes.....	44
Fig. 5.1 – Gráfico de dispersión de las funciones discriminantes canónicas de las especies de Iris con los centroides de cada uno de los grupos (Anexo C – Ejemplos de PCA y LDA)	47
Fig. 6.1 - Se observa al ampliar la imagen que la misma está conformada de píxeles de diferentes colores.....	48
Fig. 6.0.1 – Ejemplo de imagen en tonos de grises y matriz de valores de píxeles.....	49
Fig. 6.0.2 – Ejemplo de imagen en color y matriz de valores de píxeles	49
Fig. 6.0.3 – Superficie de una imagen de un dermatofito <i>E. floccosum</i>	50
Fig. 6.0.4 – Superficie de una imagen de un dermatofito <i>M. canis</i>	50
Fig. 6.0.5 – Ejemplo de histograma de imagen	51
Fig. 6.0.6 – Ejemplo de histograma de imagen oscura [FOT05].....	52
Fig. 6.0.7 – Ejemplo de histograma de imagen clara [FOT05].....	52
Fig. 6.0.8 – Definición de una máscara de convolución	53
Fig. 6.0.9 – Aplicación de una convolución píxel a píxel	53
Fig. 6.0.10 – Se observa cómo se arma una variable con las filas de píxeles de una imagen.....	54
Fig. 7.1 - Ejemplo de rostro humano	57

Fig. 7.2 – La detección de una cara en una imagen es el paso previo al reconocimiento del individuo	58
Fig. 7.3 – Diferentes poses de una misma persona [GRO01]	59
Fig. 7.4 – Variaciones de iluminación [GRO01]	59
Fig. 7.5 – En este ejemplo se observan dos rostros con distintos rasgos faciales (imágenes extraídas de [FAC02]).....	59
Fig. 7.6 – Se dificulta el reconocimiento de la boca (también en el caso de los ojos) debido a que las diferentes posiciones de los labios hacen cambiar la luminosidad, colores y geometría de la misma (imágenes extraídas de [FAC02])	60
Fig. 7.7 – Puntos clave para el modelo frontal (extraído de [SEV06])	61
Fig. 7.8 – Puntos clave para el modelo lateral (extraído de [SEV06])	61
Fig. 8.1 - Ejemplos de imágenes de entrenamiento.....	62
Fig. 8.2 - Ejemplos de eigenfaces.....	63
Fig. 8.3 - Ejemplos de imágenes de persona, de M. canis y de T. tonsurans	64
Fig. 8.4 - Ejemplos de imágenes originales	66
Fig. 8.5 - Ejemplos de eigenfungi	67
Fig. 9.1 – Ejemplo de cambio de tamaño y paso de color a tonos de grises	69
Fig. 10.1- Ejemplos de imágenes de entrenamiento.....	71
Fig. 10.2 - Ejemplos de eigenimages	72
Fig. 11.1 - Ejemplos de imágenes de entrenamiento de la especie E. floccosum y de M. canis.....	74
Fig. 11.2 - Ejemplos de eigenfungi de la prueba E. floccosum versus M. canis	74
Fig. 11.1 – Comparación de porcentajes de acierto entre los métodos eigenfaces y eigenfungi.....	76
Fig. 11.2 – Ejemplo de detección de contornos en una imagen de M. canis.....	77
Fig. 11.3 – Ejemplo de imagen de E. floccosum luego de aplicarle la binarización....	78
Fig. 11.4 – Ejemplo de imagen de T. tonsurans con corrección de histograma	78
Fig. 11.5 – Ejemplo de imagen de M. canis con suavizado de bordes.....	79
Fig. 11.6 – Ejemplo de imagen de E. floccosum transformada por la Transformada de Fourier.....	80
Fig. 11.7 – Ejemplo de imagen de M. canis con desenfoque Gaussiano.....	81
Fig. 11.10 – Comparación de porcentajes de acierto entre los métodos eigenfaces, eigenfungi y eigenfungi con preprocesamiento de suavizado de bordes y ecualización de histograma.	84
Fig. 11.12– Comparación de porcentajes de acierto entre el preprocesamiento de suavizado de bordes y ecualización de histograma, con suavizado solo y desenfoque gaussiano.	85
Fig. 12.1 – Imagen de M. canis a la cual se le aplicó ruido gaussiano.....	86
Fig. 12.2 – Imagen de M. canis a la cual se le aplicó ruido gaussiano y luego suavizado de bordes con corrección de histograma	87
Fig. 12.3 – Imagen de T. tonsurans a la cual se le aplicó ruido “sal y pimienta”	88
Fig. 12.4 – Imagen de T. tonsurans a la cual se le aplicó ruido “sal y pimienta” y luego suavizado de bordes con corrección de histograma	88
Fig. 12.5 – Comparación de porcentajes de acierto entre imágenes sin degradar, con ruido y con ruido sal y pimienta.....	89
Fig. 13.1 – Comparación de porcentajes de acierto entre los métodos eigenfaces y eigenfungi.....	92

Fig. 13.2 – Comparación de porcentajes de acierto entre los métodos eigenfaces, eigenfunji y eigenfunji con preprocesamiento de suavizado de bordes y ecualización de histograma.	96
Fig. 15.1 – Comparación de porcentajes de acierto entre los métodos eigenfaces y eigenfunji	101
Fig. 15.2 – Comparación de porcentajes de acierto entre los métodos eigenfaces, eigenfunji y eigenfunji con preprocesamiento de suavizado de bordes y ecualización de histograma.	105
Fig. 16.1 – Ejemplos gráficos de las distancias entre dos puntos, Euclídea en el primer caso y Manhattan en el segundo.....	109
Fig. A.1 – Ejemplos de menú y ventanas de ImageJ.....	129
Fig. A.2 – Barra de menú de ImageJ	130
Fig. A.3 – Menú de opciones de Fisherfaces	130
Fig. A.4 – Ejemplo de información de composición de una base	131
Fig. B.1 – Distribución de directorios.....	133

Índice de Tablas

Tabla 10.0.1 – Porcentajes de acierto eigenfaces con objetos	72
Tabla 10.1.1 – Porcentajes de acierto eigenfungi con objetos	73
Tabla 11.2.1 – Porcentajes de acierto eigenfaces con dermatofitos	75
Tabla 11.3.1 – Porcentajes de acierto eigenfungi con dermatofitos	76
Tabla 11.5.1 – Porcentajes de acierto eigenfungi combinado con detección de contornos	81
Tabla 11.5.2 – Porcentajes de acierto eigenfungi combinado con transformada de Fourier	82
Tabla 11.5.3 – Porcentajes de acierto eigenfungi combinado con imágenes binarizadas	82
Tabla 11.6.1 – Porcentajes de acierto eigenfungi combinado con suavizado de bordes y corrección de histograma	83
Tabla 11.7.1 – Porcentajes de acierto eigenfungi combinado con suavizado de bordes	84
Tabla 11.7.2 – Porcentajes de acierto eigenfungi combinado con desenfoque gaussiano y corrección de histograma	85
Tabla 12.1.1 – Porcentajes de acierto eigenfungi combinado con suavizado de bordes y corrección de histograma, previa degradación con ruido.....	87
Tabla 12.2.1 – Porcentajes de acierto eigenfungi combinado con suavizado de bordes y corrección de histograma, previa degradación con ruido “sal y pimienta”	88
Tabla 13.1.1 – Porcentajes de acierto eigenfaces con dos muestras.....	90
Tabla 13.2.1 – Porcentajes de acierto eigenfungi con dos muestras.....	91
Tabla 13.3.1 – Porcentajes de acierto eigenfungi con dos muestras combinado con detección de contornos	92
Tabla 13.3.2 – Porcentajes de acierto eigenfungi con dos muestras combinado con transformada de Fourier	93
Tabla 13.3.3 – Porcentajes de acierto eigenfungi con dos muestras combinado con imágenes binarizadas	94
Tabla 13.3.4 – Porcentajes de acierto eigenfungi con dos muestras combinado con suavizado de bordes	94
Tabla 13.4.1 – Porcentajes de acierto eigenfungi con dos muestras combinado con suavizado de bordes y corrección histograma	95
Tabla 13.5.1 – Porcentajes de acierto eigenfungi con dos muestras combinado con desenfoque gaussiano y corrección histograma.....	96
Tabla 13.6.1 – Resumen de porcentajes de acierto eigenfungi con dos muestras combinado con varios preprocesamientos	97
Tabla 15.1.1 – Porcentajes de acierto eigenfaces para predicción	100
Tabla 15.2.1 – Porcentajes de acierto eigenfungi para predicción	101
Tabla 15.3.1 – Porcentajes de acierto eigenfungi para predicción combinado con detección de contornos	102
Tabla 15.3.2 – Porcentajes de acierto eigenfungi para predicción combinado con corrección de histograma	103
Tabla 15.3.3 – Porcentajes de acierto eigenfungi para predicción combinado con transformada de Fourier	103

Tabla 15.3.4 – Porcentajes de acierto eigenfungi para predicción combinado con imágenes binarizadas.....	104
Tabla 15.3.5 – Porcentajes de acierto eigenfungi para predicción combinado con suavizado de bordes y corrección de histograma.....	104
Tabla 15.3.6 – Porcentajes de acierto eigenfungi para predicción combinado con desenfoque gaussiano y corrección de histograma	105
Tabla 15.5.1 – Resumen de porcentajes de acierto eigenfungi para predicción combinado con varios preprocesamientos	106
Tabla 16.1.1 – Porcentajes de acierto variante eigenfungi multiespacios	108
Tabla 16.2.1 – Porcentajes de acierto variante eigenfungi distancia Manhattan	109
Tabla 16.3.1 - Porcentajes de acierto variante eigenfungi combinando multiespacio y distancia Manhattan	110
Tabla 16.4.1 - Porcentajes de acierto variante eigenfungi combinando multiespacio y distancia Manhattan con preprocesamiento de suavizado de bordes	111
Tabla 17.3.1 – Comparación porcentajes de acierto eigenfungi y fisherfaces, especie E. floccosum versus el resto	115
Tabla 17.3.2 – Comparación porcentajes de acierto eigenfungi y fisherfaces, pruebas totales	115
Tabla 17.4.1 – Comparación porcentajes de acierto fisherfaces y fisherfaces combinado con preprocesamiento suavizado de bordes y corrección de histograma, especie E. floccosum versus el resto	116

Parte I: Conceptos Generales

1. Introducción

La utilización de técnicas de microscopía es fundamental en numerosos procedimientos analíticos y de control en ingeniería, química, biología, medicina, entre otras áreas. En medicina el análisis microscópico se usa en investigación, diagnóstico de anomalías patológicas y detección de agentes infecciosos.

La micología médica, el estudio de los hongos que causan enfermedades, es una de las disciplinas donde el entrenamiento del personal requiere especial importancia. En muchas patologías la única forma de identificar el agente causante de la enfermedad es mediante análisis microscópico; y a su vez, sólo mediante la correcta identificación del hongo responsable de la infección, el médico es capaz de indicar un tratamiento adecuado, dado que las micosis podrían producir daños irreversibles o hasta llevar a la muerte del paciente.

Por otra parte, en los últimos años se registra un incremento en las infecciones causadas por hongos, principalmente debido a causas que comprometen el funcionamiento normal del sistema inmune de los pacientes, como la desnutrición, la epidemia de SIDA o la inmuno-supresión que sigue a los trasplantes de órganos.

Debido a la necesidad de entrenamiento específico que requiere el análisis microscópico, el diseño e implementación de herramientas informáticas que asistan al personal recibe creciente atención.

Existen varios ejemplos de aplicaciones del procesamiento de imágenes para microbiología general [LIU01], pero no tantos para micología médica. Uno de los motivos es que las imágenes micológicas tienen una complejidad mucho mayor que aquellas que contienen exclusivamente bacterias, que son morfológicamente menos complejas.

El reconocimiento de patrones en imágenes ha sido estudiado por años y muchos trabajos se han publicado en el área, pero todavía existe una gran distancia para alcanzar las características de la visión humana con respecto a velocidad y exactitud [XIA04].

En este trabajo desarrollamos un método de reconocimiento automático de hongos microscópicos, basado en un método específico utilizado para el reconocimiento de rostros denominado *eigenfaces* [TRK91].

La identificación de rostros es un campo donde ya se han probado técnicas de agrupamiento, registro y clasificación, y se encuentra en plena evolución, por lo que todavía no cuenta con un cuerpo maduro de técnicas y procedimientos.

1.1. Antecedentes

Para el reconocimiento de rostros se estudian técnicas desde hace décadas, pero hoy esta tarea crece en importancia debido a la búsqueda de nuevos y más eficientes sistemas de

seguridad, por ejemplo en aeropuertos, accesos a información confidencial, ingresos restringidos en empresas, etc.

Existen diversas técnicas para la identificación de rostros humanos, las cuales podrían clasificarse en dos grandes grupos:

- la identificación por características - toman en cuenta en las imágenes modelos colorímetros y proporciones geométricas de la disposición de los componentes del rostro, curvaturas de huesos, etc.
- las aproximaciones estadísticas – por ejemplo las *eigenfaces*, representan las imágenes como bases numéricas y detectan patrones mediante la aplicación de métodos multivariados

Las aplicaciones estadísticas revolucionaron el área de identificación de caras. En 1991, M. Turk y A. Pentland [TRK91] presentan un método de reconocimiento basado en el *Análisis de Componentes Principales* al que denominaron *Eigenfaces*. Este término proviene del prefijo alemán *eigen* que significa propio y proviene de la característica del método de encontrar un nuevo sistema de coordenadas con los *autovectores* de la matriz de covarianzas del conjunto original, en inglés *eigenvectors*. Se caracterizan las imágenes de entrenamiento según sus distancias a las nuevas imágenes obtenidas y luego, por cada nueva imagen, se analiza a cuál de los individuos se acerca más para su reconocimiento.

En 1997, Belhumeur y otros [BEL97] presentan las *Fisherfaces* que se basan en el método estadístico Análisis Discriminante Lineal de Fisher. Este análisis encuentra funciones que caracterizan los grupos en los que se clasifican una serie de datos. Según los autores, las *fisherfaces* no se ven afectadas por diferencias significativas de iluminación y expresiones faciales en las imágenes utilizadas.

Posteriormente, fueron desarrolladas otras técnicas basadas en variaciones de estos métodos, como por ejemplo *Independent Component Analysis* o *ICA* de Bartlett y otros [BAR98], que proyectan los datos sobre vectores básicos estadísticamente independientes. *Mixture of Principal Component* de Deepak y otros [DEE02], que usa una mezcla de eigen-espacios para capturar variaciones en los datos.

También encontramos otras variantes como *PCA 2-dimensional* de Yang y otros [YAN04] que, en vez de usar vectores como PCA, utiliza matrices 2D así la matriz de imágenes no debe ser transformada en un vector para la extracción de características. Otro ejemplo, *PCA Diagonal* de Zhang y otros [ZHA06], que busca los vectores de proyección óptimos desde imágenes diagonalizadas, sin transformar la imagen a un vector.

También existen algunas aplicaciones de redes neuronales, tanto para la identificación de individuos, como para la detección de rostros en imágenes como el trabajo de Rowley y otros [ROW98].

En el caso de reconocimiento de microorganismos, vemos algunas implementaciones de redes neuronales como el trabajo de Widmer y otros [WID02] que entrenan un perceptrón para el reconocimiento del *Cryptosporidium parvum*. Y por ejemplo Verpoulos y otros [VER98] utilizan una red neuronal para la identificación de bacilos de tuberculosis.

1.2. CMEIAS

CMEIAS [LIU01] es un sistema de análisis de imágenes que permite identificar bacterias y comunidades microbianas. Fue desarrollado por la Universidad de Michigan, USA, como un *plugin* para el software UTHSCSA ImageTool.

El programa usa varias funcionalidades de medición y dos clasificadores de objetos para extraer tamaños y formas de imágenes digitales de microorganismos y clasificarlos según sus morfotipos.

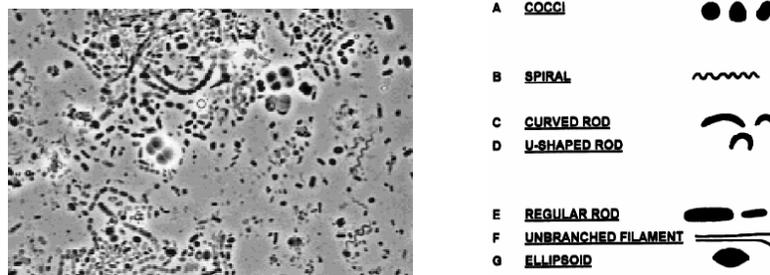


Fig. 1.1 – Imagen de una comunidad de bacterias y ejemplos de formas detectadas

El primer clasificador usa una función de medición para analizar comunidades simples conteniendo unos pocos morfotipos, como cocos, filamentos, etc.

Un segundo clasificador, que es un árbol jerárquico, utiliza un subconjunto optimizado de múltiples funciones de medición para analizar comunidades significativamente más complejas que contienen gran diversidad morfológica, como espirales, varas curvas, varas en "U", varas rectas, *cocci*, filamentos, elipsoides, etc.

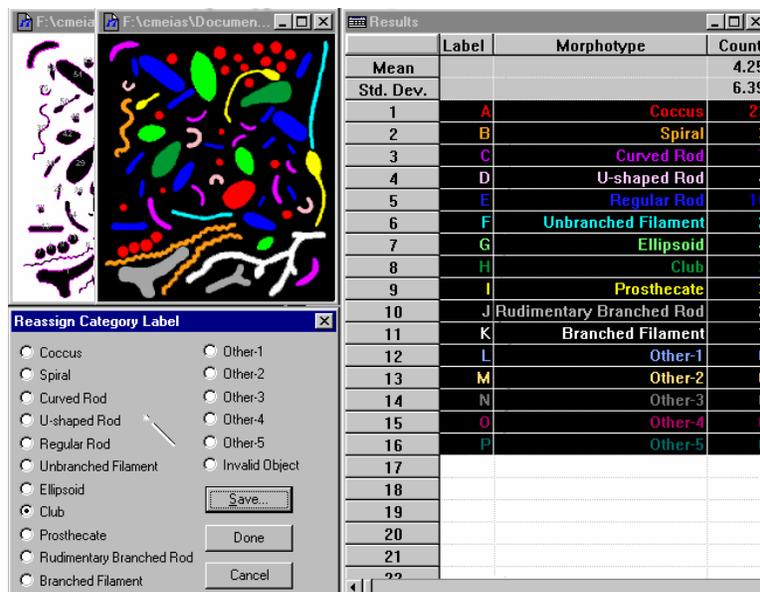


Fig. 1.2 – Ejemplo de pantalla del programa CMEIAS donde se observa una imagen analizada y la cantidad de morfotipos encontrados y clasificados

1.3. Aplicaciones para la identificación de hongos

En el campo de la micología, los avances en desarrollos automáticos para la identificación automática son casi inexistentes.

En el trabajo de Dorge y otros [DOR00] se muestra un método para el reconocimiento del *Penicillium* pero a nivel de imágenes macroscópicas de colonias. Inglis y otros [ING01] muestran un método semiautomático para identificar hifas en muestras microscópicas, que requiere de un preprocesamiento por parte del experto humano para una demarcación previa de las imágenes.

En mayo de 2007, Debandi, Haedo y Soria [DEB07] aplican Máquinas de Soporte Vectorial (SVM) para el reconocimiento de dermatofitos. El objetivo de las SVM es lograr la separación lineal de los datos de aprendizaje elevando la dimensión del espacio vectorial y así eliminar las no linealidades originales del problema. Son similares a las redes neuronales, salvo que utilizan el concepto de *Kernel* (núcleo o función) que se define según el tipo de problema. Más aún, una SVM con un *kernel sigmoideo* es equivalente a una red neuronal de dos capas con backpropagation. El método incorpora el principio de minimización del riesgo estructural, que a diferencia del principio de *Minimización del Riesgo Empírico* usado por las redes neuronales, minimiza el riesgo esperado en lugar de minimizar el error del conjunto de entrenamiento o hipótesis.

A continuación, se describen los *kernels* más utilizados

Función Kernel	Tipo de clasificador
$K(x, y) = x \cdot y$	Lineal
$K(x, y) = \exp(-1 x - y ^2)$	RBF Gaussiano
$K(x, y) = (1 + x \cdot y)^d$	Polinomio de grado d
$K(x, y) = \tanh(x \cdot y - \bullet)$ (sólo para algunos \bullet)	Función sigmoidea

Para el reconocimiento de los dermatofitos, utilizan un *kernel* polinómico de grado 2 combinado con una serie de preprocesamientos para incrementar la exactitud de clasificación del método.

El método de las SVM está planteado para clasificaciones binarias, es decir, solamente entre 2 clases. Por lo que esto dificulta la clasificación de más de 2 especies (como por ejemplo las 6 especies de dermatofitos estudiadas).

Para la clasificación multiclase se debe recurrir a dos técnicas principales: uno-vs-uno y uno-vs-todos. La clasificación uno-vs-uno consiste en un entrenamiento de a pares obteniendo $k(k-1)/2$ funciones de clasificación. Para seleccionar el valor final de las clasificación se suele utilizar un sistema que asigna a la clase con mayor cantidad de *votos*.

La segunda técnica, uno-vs-todos, consiste en obtener una función f_i por cada clase i , colocando como grupo positivo los elementos de esta clase y como negativo los

elementos de las clases restantes. Finalmente, para asignar la clase final a la que corresponde un elemento, se tomará el máximo valor obtenido al aplicar todas las funciones.

1.4. Objetivo del trabajo

En este trabajo desarrollamos un método automático para el reconocimiento de especies de hongos microscópicos, que llamamos *eigenfungi*. Está basado en la metodología para reconocimiento de rostros denominado *eigenfaces*, al que se le introducen varias modificaciones que mejoran su exactitud en el análisis de imágenes microscópicas de hongos.

El método *eigenfaces* presentado por Turk & Pentland [TRK91] en 1991, utiliza un entrenamiento para el reconocimiento de las imágenes consistente en una adaptación de la transformada de Hotelling (o de Karhunen y Loève). Esta transformada está basada en las propiedades estadísticas de la imagen y sus principales aplicaciones son la compresión y la rotación de la misma y representa una aplicación en el área de imágenes del método estadístico del Análisis de Componentes Principales (PCA).

El PCA es un método estadístico de análisis que descompone datos multidimensionales a un subespacio de menos dimensiones pero preservando las características esenciales de los datos tratados. Las nuevas componentes principales o factores (independientes entre sí) representan una combinación lineal de las variables originales [TOR03].

Para la elaboración y validación de la metodología, se estudiaron imágenes de hongos microscópicos de las seis especies principales de dermatofitos, obtenidas de muestras provistas por el Departamento de Micología del Instituto Nacional de Enfermedades Infecciosas (INEI), ANLIS "Carlos G. Malbrán".

1.5. Organización

En los siguientes capítulos explicaremos el método y sus propiedades. Primero, mostraremos cómo se clasifican los hongos, sus características y las afecciones que producen en los seres humanos.

Luego veremos los fundamentos matemáticos y estadísticos del método: el Análisis de Componentes Principales y su aplicación en imágenes, la Transformada de Hotelling.

Posteriormente analizaremos el método de *Eigenfungi* y sus diferencias con las *eigenfaces*. Luego mostramos las pruebas realizadas, con una muestra de cada especie y después con dos muestras.

En el siguiente capítulo comparamos el método con algunas variantes en el algoritmo de PCA. Por último veremos las características de otro método de Data Mining aplicado al reconocimiento de rostros llamado *fisherfaces* y lo compararemos con los resultados obtenidos.

2. El Reino de los Hongos

Los seres naturales fueron clasificados por Aristóteles en tres reinos: animal, vegetal y mineral. Los hongos inicialmente fueron clasificados como plantas, pero tanto a éstos como a las bacterias se los consideró luego como un reino aparte.

En 1969 Whittaker determina cinco reinos:

- reino animal (Animalia)
- reino vegetal (Plantae)
- reino moneras (Monera)
- reino hongos (Fungi)
- reino protoctistas (Protoctista)

En 1990, Woese determina como nivel superior el Dominio, dividido en reinos y subreinos. Indica que los seres se clasifican en tres reinos:

Dominio Archaea - microorganismos unicelulares que carecen de núcleo por lo que su ADN está libre en el citoplasma. Se subdivide en los reinos Euryarchaeota, Crenarchaeota, Koryarchaeota y Nanoarchaeota.

Dominio Bacteria

Dominio Eukarya - organismos celulares con núcleo (eucariotas). Se subdivide en los reinos Animalia, Fungi, Plantae y Protista

2.1. Características

Los hongos son organismos celulares eucariotas (que tienen núcleo definido) que se alimentan absorbiendo nutrientes que obtienen de la degradación de componentes químicos de mayor complejidad y peso molecular. Para lograr esto, secretan enzimas. La mayoría está constituida por filamentos denominadas *hifas*. El conjunto de hifas constituye el *micelio*.



Fig. 2.1 – *Microsporum canis* – Se observan hifas y esporas

Generalmente se reproducen a través de partículas de protoplasma denominadas *esporas*. Éstas son diseminadas en algunos hongos por un cuerpo fructífero denominado *seta*.



Fig. 2.2 - *Flammulina velutipes* – Se observa un conjunto de setas

Las esporas sexuales surgen como resultado de la fusión nuclear seguida por la meiosis, proceso que reduce el número de cromosomas a la mitad. Cuando los dos núcleos provienen de la misma colonia (talo) se dice que el hongo es homotálico. En caso contrario se denomina heterotálico.

Los hongos son utilizados en distintas industrias, como alimento, levaduras, bebidas fermentadas, antibióticos, etc. Las enfermedades producidas por hongos se denominan *micosis*.

2.2. Micosis

Las micosis superficiales, ampliamente distribuidas en el mundo, son afecciones producidas por el parasitismo fúngico en las estructuras córneas de la piel y sus faneras (pelos y uñas). Se excluye de estas infecciones aquellas en donde haya compromiso de mucosas y de tejidos blandos, que involucran más allá de la dermis.

Las micosis pasaron a ser un importante problema sanitario, por lo cual se requiere que todos los centros de salud y los laboratorios de diagnóstico microbiológico de cierta importancia, cuenten con personal capaz de diagnosticar una patología de este tipo.

2.3. Epidemiología

Las micosis presentan un cuadro clínico más o menos definido, sin embargo muchas veces es necesario realizar un diagnóstico diferencial con otras enfermedades infecciosas o no, debido a la similitud de sus cuadros clínicos.

Es importante el estado inmunológico del paciente, ya sea por enfermedades de base inmunológica o por tratamientos debilitantes o inmunosupresores.

En el caso de las micosis superficiales es necesario conocer los siguientes datos del paciente:

- Trayectoria residencial y ocupacional, a fin de determinar si visitó el área endémica de alguna de las micosis.
- Hábitos de esparcimiento, para detectar contacto con el reservorio del agente causal.
- Medicación que recibe o recibió, tanto antifúngica como de otra naturaleza.
- Enfermedades que sufre, ya que hay algunas que están estrechamente asociadas a las micosis.

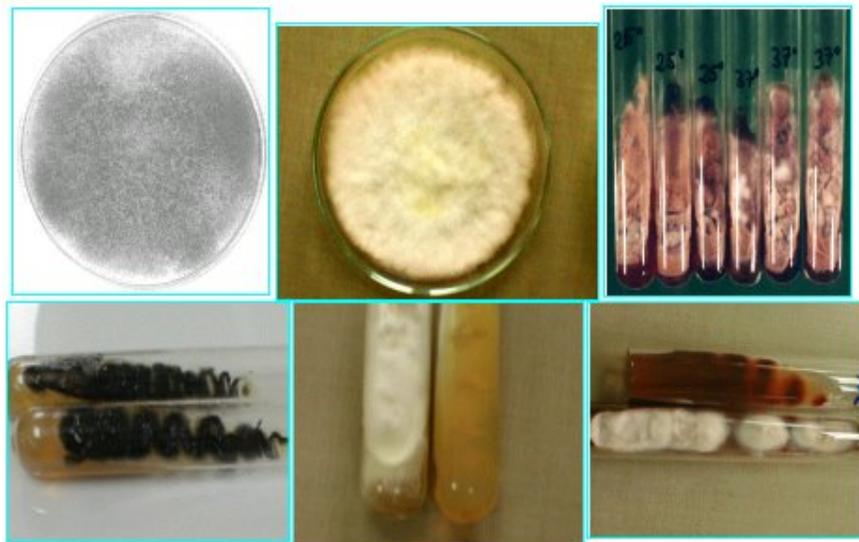


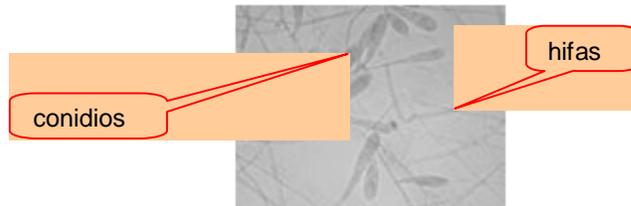
Fig. 2.3 - Cultivos de hongos filamentosos

Todo eso lleva a:

- 1) Elegir el método adecuado de examen directo para detectar el agente
- 2) Seleccionar los medios y temperaturas de cultivo más aptos para el desarrollo del hongo
- 3) Interpretar el rol que cumple el agente aislado cuando éste no es un patógeno primario
- 4) Sugerir los análisis complementarios que se crea conveniente

2.4. Observaciones microscópicas

Los hongos están conformados por filamentos denominados *hifas*. El conjunto de hifas, que en algunas especies puede llegar a kilómetros de longitud, se llama *micelio*. Los *conidios* o *esporas* son elementos de fructificación, que posibilitan la reproducción.

Fig. 2.4 – Imagen de un dermatofito *E. floccosum*

Microconidias – Se refiere a un conidio pequeño, generalmente unicelular. Es un elemento de reproducción asexual (espora asexual). Se debe observar la cantidad realtiva, forma, tamaño y disposición.

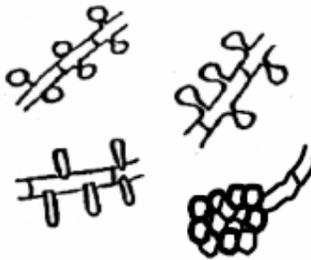


Fig. 2.5 – Ejemplos de microconidias

Macroconidias – Se refiere a un conidio o espora de reproducción asexual, se distingue del microconidio por su tamaño y por ser pluricelulares. Se debe observar la cantidad relativa, forma, tamaño, disposición y pared celular (lisa o equinulada, gruesa o fina).



Fig. 2.6 – Ejemplos de macroconidias

2.5. Dermatofitos

Los dermatofitos son hongos queratinofílicos que causan infecciones de los tejidos epidérmicos humanos y animales. Se encuentran distribuidos taxonómicamente en tres géneros: *Microsporum*, *Trichophyton* y *Epidermophyton*.

A diferencia de otros agentes de micosis superficiales, estos hongos penetran y parasitan todos los tejidos queratinizados del organismo (estructuras córneas de la piel, pelo y uñas), dando lugar a síntomas entre leves y severos.

Debido a las similitudes existentes entre las diferentes especies, es posible ver que un tipo clínico de infección puede ser causado por diferentes dermatofitos, o que una misma especie esté involucrada en varios tipos de enfermedades.

Las dermatofitosis o tiñas pueden ser desde asintomáticas a muy pruriginosas y dolorosas. Se diseminan por contacto directo o indirecto interhumano o animal-hombre.

Las 6 especies principales de dermatofitos son:

Epidermophyton floccosum

Trichophyton mentagrophytes

Microsporum canis

Trichophyton rubrum

Microsporum gypseum

Trichophyton tonsurans

2.6. Descripción de las principales dermatofitosis

Tiña	Características
<p data-bbox="423 1016 586 1045">Tinea capitis</p> 	<p data-bbox="753 1016 1334 1150">Infección fungosa del pelo y cuero cabelludo, caracterizada por lesiones eritematosas, escamosas, alopécicas y a veces con erupciones ulcerosas profundas</p>
<p data-bbox="418 1367 591 1396">Tinea barbae</p>	<p data-bbox="753 1367 1334 1501">Infección fúngica crónica en cara y cuello, con lesiones escamosas rodeadas de un borde vesículo pustuloso o de pústulas foliculares profundas, con pelos quebradizos</p>
<p data-bbox="410 1520 599 1549">Tinea corporis</p>	<p data-bbox="753 1520 1334 1717">Infección en la piel carente de pelo. Las lesiones van desde escamosas simples hasta eritema y granulomas profundos. Presentan un área central escamosa y periferia que avanza activamente y suele estar tachonada de vesículas y pústulas costrosas</p>

	
<p>Tinea imbricata</p>	<p>Lesiones anulares múltiples con aspecto de tatuaje</p>
<p>Tinea cruris</p>	<p>Infección aguda o crónica localizada en ingle, zona suprapúbica, puede extenderse a coxis y región perianal.</p> <p>Placas bilaterales de dermatitis escamosa, con vesículas o pústulas en la periferia. Porción central parda con escamas purpúreas</p>
<p>Tinea pedis</p> 	<p>Infección que invade especialmente membranas interdigitales y planta de los pies. Placas con aspecto de salvado, vesículas en las plantas, fisura o epidermis macerada con mal olor y húmeda en las membranas interdigitales, placas hiperqueratósicas en los talones</p>
<p>Tinea unguium</p>	<p>Invasión de la placa de las uñas. Uñas de color anormal, engrosadas y deformes, friables y quebradizas, rugosas, cubiertas con surcos.</p>

2.7. Descripción de las principales especies de dermatofitos

a) *Epidermophyton floccosum*

Patogenicidad

Agente de tinea corporis, cruris, pedis y onicomicosis. No invade el pelo.

Características microscópicas

Macroconidias abundantes, claviformes de paredes lisas, base ancha y roma, y extremo distal redondeado, con 1 a 4 células que pueden nacer aisladas o en racimos.

No forma microconidias.

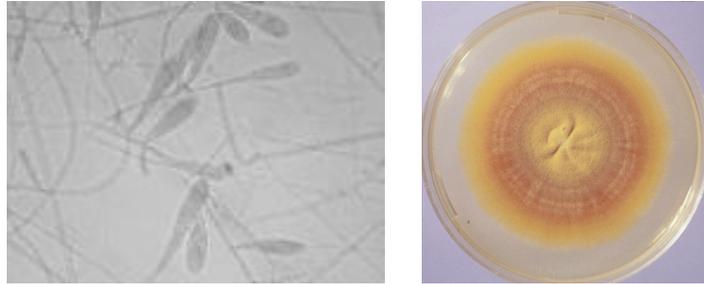


Fig. 2.7 – Ejemplo de imagen microscópica de *E. floccosum* y colonia

b) Microsporum canis

Patogenicidad

Causa en el hombre tinea capitis y tinea corporis. Se adquiere por contagio de animales.

Características microscópicas

Macroconidias abundantes, grandes (40-150 μm x 8-20 μm), fusiformes con un ápice prominente y curvado, de paredes gruesas y verrugosas.

Microconidias ausentes o escasas, piriformes, claviformes o truncadas, a lo largo de las hifas.

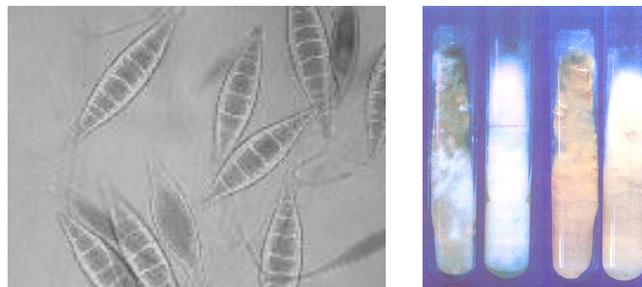


Fig. 2.8 – Ejemplo de imagen microscópica de *M. canis* y colonias



Fig. 2.9 – Ejemplo de infección de *M. canis* en animales

c) *Microsporum gypseum*

Patogenicidad

Agente de tiña capitis, barbae y corporis en niños y adultos. Causa tiña en animales como perros y caballos.

Características microscópicas

Macroconidias elipsoides a fusiformes con extremos redondeados de 25-69 μm x 8-15 μm de paredes moderadamente gruesas equinuladas.

Microconidias escasas, claviformes a los lados de las hifas.

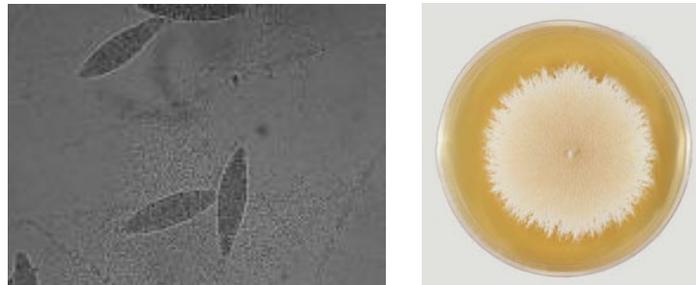


Fig. 2.10 – Ejemplo de imagen microscópica de *M. gypseum* y colonia

d) *Trichophyton mentagrophytes*

Patogenicidad

Todos los tipos de tiña en hombres y animales. El ataque del pelo es tipo ectotrix con esporas pequeñas.

Características microscópicas

Macroconidias muy escasas o ausentes, más abundantes en medios enriquecidos.

Microconidias abundantes, subesféricas y nacen en racimos sobre conidióforos o aisladas a los lados de las hifas.

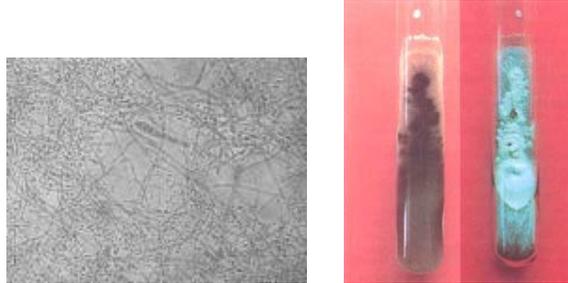


Fig. 2.11 – Ejemplo de imagen microscópica de *T. mentagrophytes* y anverso y reverso de colonia



Fig. 2.12 – Lesión en un brazo (tinea corporis) causada por *T. mentagrophytes*

e) *Trichophyton rubrum*

Patogenicidad

Causa tinea corporis, pedis, cruris, manum y onicomicosis. Parásito ocasional de animales.

Características microscópicas

En el tipo veloso, las macroconidias están ausentes y las microconidias son escasas, delgadas, claviformes de 3-5 μm x 2-3 μm y se disponen lateralmente a lo largo de las hifas.

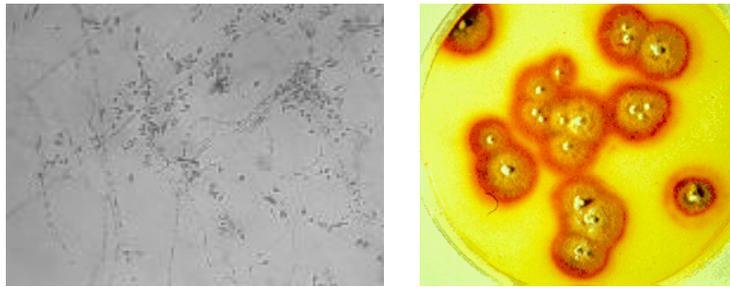


Fig. 2.13 – Ejemplo de imagen microscópica de *T. rubrum* y colonia

e) *Trichophyton tonsurans*

Patogenicidad

Agente causal de tinea capitis a punto negro, tinea corporis y tinea unguium. El ataque del pelo es endotrix.

Características microscópicas

Macroconidias escasas de pared delgada lisa, claviformes.

Microconidias abundantes, laterales, habitualmente claviformes.

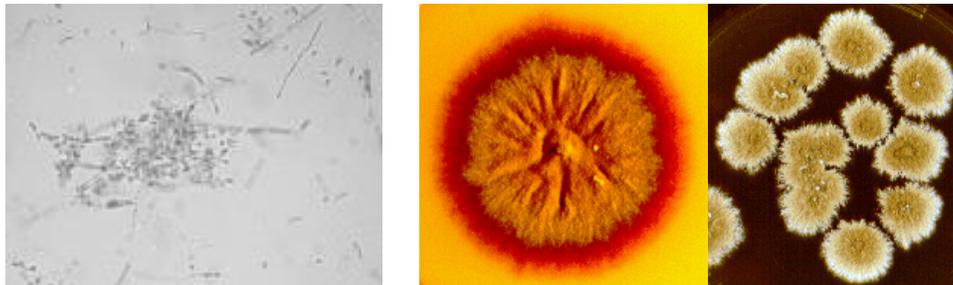


Fig. 2.14 – Ejemplo de imagen microscópica de *T. tonsurans* y cultivos

2.8. Recolección y manipulación de las muestras

La calidad del diagnóstico de las micosis depende de la calidad y cantidad del material recogido del paciente, de las condiciones de envío, conservación, transporte y procesamiento, y de la pericia del micólogo.

Las muestras deben ser representativas, abundantes, libres de contaminantes exógenos o endógenos, y de sustancias que inhiban o alteren la viabilidad de los hongos.

Para asegurar la calidad de la muestra es preferible que lo realice un micólogo, la inocule en los medios de cultivo, y la procese para la observación microscópica. En caso de no ser posible, se debe entrenar al personal de manera tal que pueda elegir adecuadamente el sitio más típico y activo de la lesión, y obtener la muestra mediante

técnicas e instrumental específico para cada caso, evitando las contaminaciones ambientales por esporas de hongos saprófitos (exógenas) y/o por gérmenes “habituales” de diferentes áreas del cuerpo humano (endógenas).

Es conveniente recoger los especímenes en recipientes estériles irrompibles y tamaño adecuado, y conservarlas a temperatura ambiente. Si las muestras fueron recogidas en un laboratorio periférico para ser enviadas a otro centro, es conveniente adjuntar una pequeña ficha con tipo de espécimen, edad, sexo y características sobresalientes de la lesión.

Métodos de obtención de muestras

Existen tres tipos de técnicas básicas de recolección de especímenes:

Raspado – Se utiliza para las lesiones descamativas de la piel glabra (sin pelo). Se raspa el borde activo de la lesión con un bisturí estéril colocado perpendicularmente a la superficie de la piel; si las lesiones son vesiculosas, se remueve el techo de las vesículas con el bisturí; cuando la lesión afecta los pliegues interdigitales, el material se toma del borde de las lesiones junto a la piel sana de los dedos, evitando áreas maceradas.

Cinta adhesiva transparente – Consiste en cortar un fragmento de cinta de aproximadamente 10 cm de largo, que se aplica sobre la lesión, raspando con el borde lateral de la uña para que se adhieran las escamas. Se aplica la cinta sobre una lámina portaobjetos limpia, rebatiendo los extremos.

Depilación – Se utiliza en las lesiones de cuero cabelludo y otras áreas pilosas para recoger el vello de la piel cuando el folículo está inflamado. La muestra se toma con pinza depilatoria estéril aplicada perpendicularmente a la superficie de la piel y siguiendo el sentido del pelo. Se deben extraer aproximadamente 20 pelos enfermos y las escamas circundantes.



Fig. 2.15 – Recolección de muestras de especímenes en lesiones

Medidas de bioseguridad

Para la manipulación de las muestras se requiere una serie de medidas de bioseguridad:

- 1- Todos los hongos aislados de muestras clínicas provenientes de pacientes bajo sospecha de micosis sistémica deben ser procesados dentro del equipo de seguridad biológica Clase II. Nunca se trabajan fuera del equipo hongos aislados de muestras de sangre, esputo, biopsias, orina y otros fluidos corporales, pus, exudados, secreciones o líquidos de drenaje, hayan sido enviados o no para diagnóstico micológico.
- 2- Nunca oler los cultivos, ni examinarlos abriendo las cajas de Petri. Siempre se destapará la caja dentro del equipo de bioseguridad.
- 3- Nunca pipetear con la boca, utilizar siempre propipetas.
- 4- Autoclavar todos los especímenes y cultivos antes de descartarlos y no conservar cultivos innecesarios.
- 5- Utilizar tubos con tapa a rosca o con tapón de algodón para preparar los medios para aislamiento primario, subcultivo o conservación de cepas de colección.
- 6- Desinfectar diariamente las áreas de trabajo y las mesadas. No dejar acumular polvillo en las esquinas y/o rendijas donde puedan acumularse esporas.
- 7- No tener macetas con plantas en el laboratorio donde se procesen los especímenes clínicos o cepas. Los hongos pueden crecer en la tierra, incluso los patógenos primarios.

Aislamiento en cultivo

Independientemente del resultado de la observación localizada, debe procederse al cultivo, empleándose más habitualmente el agar dextrosa-peptona de Sabouraud pH 5,6 suplementado con cicloheximida (actidiona), para la inhibición de hongos saprobios y con diferentes antibacterianos termoestables como como cloranfenicol o gentamicina. Otros medios son el agar DTM para muestras con una elevada contaminación microbiana o los medios comerciales como agar Mycosel y agar Mycobiotic.

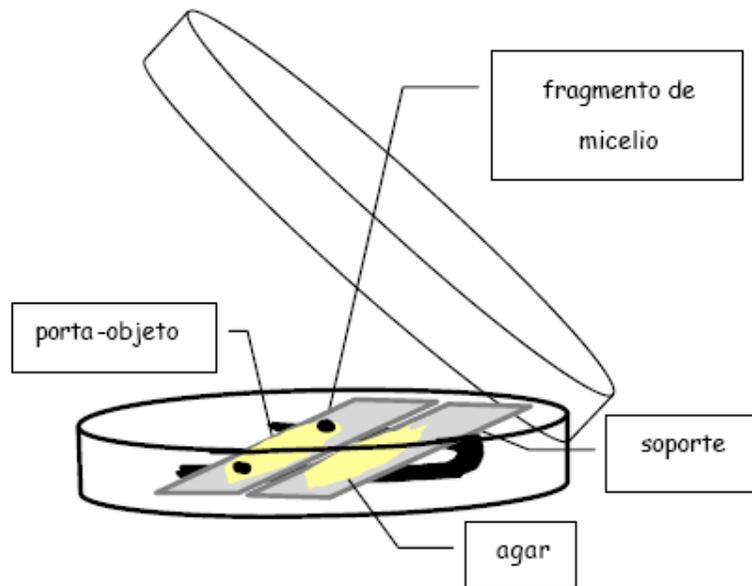


Fig. 2.16 – Preparación de cultivo [DAV04]

2.9. Diagnóstico

Los dermatofitos tienen características microbiológicas y fisiológicas muy similares. Las dermatomicosis son las infecciones fúngicas que más comúnmente afectan al hombre. Han presentado importantes variaciones en el tiempo, en relación con los cambios culturales, higiénicos o migratorios registrados en las distintas épocas. En la actualidad, las especies aisladas con mayor frecuencia muestran una distribución cosmopolita.

Las distintas afecciones se asocian a factores fisiológicos (secreciones sebáceas u hormonales), ocupacionales (deportistas, agricultores) o higiénicos (tipo de vestimenta, uso de calzado cerrado), además de procesos nosológicos en tratamientos con corticoesteroides o diabetes.

La fuente de infección es diversa (suelo, animales domésticos, personas infectadas) a través de un contacto directo o mediante diferentes fómites (peine, toalla, calzado).

El diagnóstico de laboratorio es fundamentalmente directo y por lo tanto se basa en la visualización, aislamiento e identificación del agente causal.

El aspecto y localización de la lesión o lesiones debe tenerse en cuenta para la correcta orientación diagnóstica, pero independientemente de la ubicación principal, debe procederse a la inspección de otras áreas temporales.

Identificación

La mayoría de los dermatofitos presentan características reconocibles a las dos semanas de incubación de los cultivos. La identificación es fundamentalmente morfométrica y se basa en la realización de exámenes macroscópicos y microscópicos.

El examen macroscópico determina la velocidad de crecimiento, la forma o aspecto, textura, consistencia, color y pigmentación de la colonia. El examen microscópico tras la obtención directa de un fragmento de la colonia o mediante el uso de cinta transparente, permite el reconocimiento de características diferenciales.

Algunas especies presentan características macro y microscópicas muy similares que exigen para su diferenciación el uso de métodos específicos como el estudio de características nutricionales o el estudio enzimático.

2.10. Antifúngicos

En la actualidad [PON02], existe un grupo relativamente reducido de fármacos útiles para el tratamiento de las micosis, denominados antifúngicos. La mayoría actúan sobre la membrana citoplásmica, aunque existen antifúngicos que actúan en el citoplasma, núcleo o pared celular.

La anfotericina B y la nistatina son antifúngicos poliénicos que actúan uniéndose al ergosterol de la membrana celular fúngica produciendo una alteración de su permeabilidad. La anfotericina B es el antifúngico más utilizado en las micosis severas pero en las células humanas puede unirse al colesterol, produciendo una alta toxicidad cuando se utilizan dosis elevadas o usada en tratamientos prolongados. Esta toxicidad se ha reducido con el desarrollo de nuevas presentaciones farmacológicas que integran a este antifúngico en liposomas o lo asocian a lípidos. Los azoles constituyen una amplia familia de antifúngicos que actúan inhibiendo la síntesis del ergosterol mediante el bloqueo de la acción de las enzimas dependientes del citocromo P450.

Existen azoles de uso tópico (como clotrimazol, miconazol, econazol, bifonazol, tioconazol y sertaconazol) y de uso sistémico (como el ketoconazol, fluconazol, itraconazol y voriconazol). Otros antifúngicos son la griseofulvina, las equinocandinas, las neumocandinas y las nikomicinas.



Fig. 2.17 - Adhesión de conidios de *Trichophyton mentagrophytes* al estrato córneo

Parte II: Métodos de Data Mining

3. Análisis univariado y multivariado

Data Mining o Minería de Datos tiene como objetivo descubrir patrones interesantes en grandes cantidades de datos que pueden estar almacenados en bases de datos, u otros tipos de repositorios. Es un campo interdisciplinario joven, en el que participan sistemas de bases de datos, *data warehousing*, estadística, aprendizaje automático, visualización de datos y computación de *alta performance*.

Su objetivo principal es obtener la mayor cantidad de información de un conjunto de datos sin hipótesis previas.

Como áreas de interés figuran por ejemplo el reconocimiento de patrones, el procesamiento de señales, con campos de aplicación tales como la industria, la economía y la bioinformática.

El análisis de los datos consiste de una secuencia iterativa de pasos:

- Limpieza de los datos – remoción de ruido y datos inconsistentes
- Integración de datos – combinación de varias fuentes de datos
- Transformación de datos – transformación o consolidación de datos en la forma apropiada para la aplicación de los distintos métodos
- Data Mining – aplicación de métodos para extracción de patrones de datos
- Evaluación de patrones – identificación de patrones de interés que representan el conocimiento extraído
- Presentación del conocimiento – aplicación de técnicas de visualización y representación del conocimiento para presentar la información al usuario

3.1. Análisis univariado – Estadísticos descriptivos

La estadística descriptiva propone una serie de indicadores que permiten tener una percepción rápida de lo que ocurre en un fenómeno.

La *teoría de muestreo* es un estudio de las relaciones existentes entre una *población* y *muestras* extraídas de la misma. Permite estimar cantidades desconocidas de la población (tales como la media poblacional, la varianza, etc.) a partir del conocimiento de las correspondientes cantidades muestrales.

Medidas de tendencia central

Son indicadores estadísticos que muestran hacia qué valor o valores se agrupan los datos.

a) **Mediana**

Sean x_i los elementos del vector x ordenados crecientemente. La mediana es el valor que deja a cada lado la mitad de los valores de la muestra. Siendo n el número de datos, la mediana será el elemento $x_{(n+1)/2}$ si n es impar. Si no, $(x_{n/2} + x_{(n+1)/2}) / 2$.

b) **Media aritmética**

También llamada promedio es el valor resultante que se obtiene al dividir la sumatoria de un conjunto de datos sobre el número total de datos. Se define como

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i$$

que puede ser estimada como

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

c) **Moda**

Indica el valor que más se repite o la clase que posee mayor frecuencia. En el caso de que dos valores presenten la misma frecuencia, el conjunto de datos se denomina bimodal. Para más de dos modas, el conjunto de datos es multimodal.

Medidas de dispersión

Indican la distancia promedio de los datos respecto a las medidas de tendencia central.

a) **Varianza**

Es el resultado de la división de la sumatoria de las distancias existente entre cada dato y su media aritmética elevadas al cuadrado y el número total de datos. Se define como:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

que puede ser estimada como

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

b) **Desviación estándar**

La desviación estándar o desviación típica es el resultado de hallar la raíz cuadrada de la varianza.

3.2. Análisis multivariado

Una matriz de datos multivariados tiene la forma típica de

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1c} \\ x_{21} & x_{22} & \dots & x_{2c} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nc} \end{pmatrix}$$

Cada vector representa a uno de los c individuos con los n atributos observados. Se pueden representar como filas o columnas.

Se considera

c – número de registros o individuos

n – número de variables o atributos de cada individuo

En este contexto, el **vector media** se representa como

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}$$

donde los μ_i resultan de calcular la media al vector fila.

De la misma manera, el **vector varianza** se define como

$$\sigma^2 = \begin{pmatrix} \sigma_1^2 \\ \sigma_2^2 \\ \vdots \\ \sigma_n^2 \end{pmatrix}$$

donde σ^2 es

$$\sigma_i^2 = \frac{1}{c-1} \sum_{k=1}^c (x_{ik} - \mu_i)^2$$

3.3. Covarianza

La **covarianza** de dos variables x_i y x_j se representa como

$$\sigma_{ij} = \frac{1}{c-1} \sum_{k=1}^c (x_{ik} - \mu_i)(x_{jk} - \mu_j)$$

Si $i = j$ entonces representa la varianza de x_i

Matriz de covarianzas

La matriz de covarianzas, también llamada de varianzas-covarianzas o matriz de dispersión, se forma de la siguiente manera

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_{nn} \end{pmatrix}$$

En algunos casos, la matriz Σ es estimada como S de la forma

$$S = \frac{1}{c-1} \sum_{i=1}^c (x_i - \mu)(x_i - \mu)^T$$

3.4. Correlación de las variables

Para que el Análisis de Componentes Principales dé mejores resultados, es conveniente que las variables involucradas en el problema estén altamente correlacionadas.

Es por esto que previo al análisis, se estudia la correlación de las variables. Para ello, se pueden utilizar dos indicadores:

- Gráficamente, puede utilizarse un gráfico *de dispersión o scatterplot*
- Analíticamente, existe el *coeficiente de correlación de Pearson*

Gráfico de Dispersión

El gráfico de dispersión permite visualizar las posibles relaciones entre dos o tres variables. También conocido como nube de puntos, muestra los valores de cada atributo en un sistema de ejes cartesianos. Si la relación es del tipo lineal, estamos en presencia de una correlación.

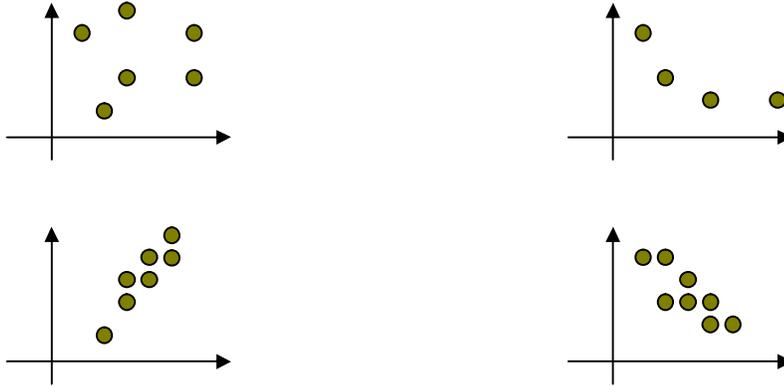


Fig. 3.1 – Se observan distintos tipos de relación entre dos variables en forma gráfica. En el primer caso, aparentemente no habría una relación entre las variables. En el segundo caso, la relación no es lineal. En los casos siguientes se ven correlaciones altas, positiva y negativa respectivamente.

Coefficiente de Pearson

Para el análisis de correlaciones en forma analítica de variables cuantitativas, se puede calcular el *coeficiente de correlación de Pearson*.

El coeficiente se calcula como:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

El valor varía en el intervalo (-1 ; 1):

Si $r = 0$, no existe ninguna correlación. Indica una independencia total entre las dos variables, es decir, que la variación de una de ellas no influye en absoluto en el valor que pueda tomar la otra.

Si $r = 1$, existe una correlación positiva perfecta. El índice indica una dependencia total entre las dos variables denominada relación directa: cuando una de ellas aumenta, la otra también lo hace en idéntica proporción.

Si $0 < r < 1$, existe una correlación positiva

Si $r = -1$, existe una correlación negativa perfecta. El índice indica una dependencia total entre las dos variables llamada relación inversa: cuando una de ellas aumenta, la otra disminuye en idéntica proporción.

Si $-1 < r < 0$, existe una correlación negativa

Dado un conjunto de variables es posible armar una matriz de correlaciones ρ donde cada posición ρ_{ij} representa el coeficiente de correlación entre las variables i y j .

4. Análisis de Componentes Principales

El Análisis de Componentes Principales (PCA) comprende un procedimiento matemático que transforma un conjunto de variables correlacionadas en un conjunto menor de variables no correlacionadas llamadas *componentes principales*.

Uno de los objetivos del método es descubrir la verdadera dimensionalidad de los datos y cuando es menor que la de las variables originales, éstas se pueden reemplazar por un número menor de variables subyacentes, con poca pérdida de información.

A menudo revela relaciones entre las variables que previamente no se sospechaban y permite interpretaciones con resultados particulares. También posibilita la detección de valores atípicos u *outliers* que no aparecen en los análisis básicos. Los outliers, una vez detectados requieren de un tratamiento especial (análisis de supresión, rastreo del dato original, reemplazo por la media de la variable, etc.) dado que pueden alterar los resultados o denotar errores al momento de la extracción de la muestra.

Las nuevas variables resultan útiles para el cribado de datos, verificación de hipótesis, verificación de agrupaciones, preprocesamiento de los datos para otros métodos como regresión, compresión de datos, etc.

4.1. Obtención de las nuevas componentes

Algebraicamente, las componentes principales son una combinación lineal de las p variables del conjunto original. Geométricamente, esta combinación lineal representa la selección de un nuevo sistema de coordenadas obtenido por la rotación del sistema original.

Los nuevos ejes representan las direcciones con máxima variabilidad y proveen una descripción más simple de la estructura de la covarianza.

El primer paso consiste en hallar la matriz de covarianzas y de ésta obtener sus autovalores y autovectores.

En el caso de estandarizarse previamente los datos, podría utilizarse para los cálculos la matriz de correlaciones, dado que en ese caso resulta igual que la matriz de covarianzas.

4.2. Autovalores y Autovectores

El siguiente paso consiste en hallar los autovalores y autovectores de la matriz de covarianza, que conforman el nuevo sistema de coordenadas.

Los autovalores o eigenvalores, también llamados raíces características o raíces latentes, de Σ son las raíces de la ecuación polinomial dada por

$$|\Sigma - \lambda \mathbf{I}| = 0$$

En términos de determinantes es una ecuación polinomial en λ , de p -ésimo grado.

A cada autovalor de Σ le corresponde un vector diferente de cero llamado autovector o eigenvector (también conocido como vector característico o vector latente) que satisface

$$\Sigma c_i = \lambda_i c_i, \text{ para } i = 1, 2, \dots, p.$$

Si Σ es una matriz simétrica de números reales, entonces sus autovalores y autovectores también consistirán en números reales.

Los autovalores de Σ se denotan por $\lambda_1 \cdot \lambda_2 \cdot \dots \cdot \lambda_p$

Los autovectores no son únicos, de modo que se normalizan de tal forma que $c_i' c_i = 1$

Cuando dos autovalores no son iguales, sus autovectores correspondientes serán ortogonales. Cuando dos o más autovalores de Σ son iguales, se pueden elegir los autovectores correspondientes de modo que sean ortogonales entre sí.

La traza de una matriz simétrica es igual a la suma de sus autovalores, es decir

$$\text{tr}(\Sigma) = \sum \lambda_i$$

El determinante de una matriz simétrica siempre es igual al producto de sus autovalores, es decir $|\Sigma| = \prod \lambda_i$

Una matriz simétrica Σ es positiva definida si y sólo si $\lambda_i > 0$, para cada i .

Una matriz simétrica Σ es positiva semidefinida si y sólo si $\lambda_i \geq 0$, para cada i y por lo menos un autovalor es igual a cero.

Si Σ es una matriz no negativa de rango m , entonces habrá exactamente m autovalores diferentes de cero.

Si Σ es una matriz simétrica, existe una matriz ortogonal C , tal que $C' \Sigma C = \Delta$, donde Δ es una matriz diagonal. Los elementos de Δ son los autovalores de Σ y las columnas de C son sus autovectores correspondientes.

Esto implica

$$\Sigma = C \Delta C' = \sum \lambda_i c_i c_i'$$

que se conoce como descomposición espectral de Σ .

4.3. Autovalores y autovectores generalizados

Dadas dos matrices A y B , se denominan autovalores y autovectores generalizados los pares α_k y β_k de escalares y los vectores x_k que cumplan:

$$\beta_k A x_k = \alpha_k B x_k \text{ para } k = 1, 2, \dots, n$$

Los escalares $\lambda_k = \alpha_k / \beta_k$ son los autovalores generalizados y los x_k los autovectores generalizados. Si la matriz B tiene inversa los vectores x_k serán los vectores propios de $B^{-1}A$.

4.4. Proyecciones en los nuevos ejes

Finalmente, se calculan las proyecciones de las variables originales en el nuevo espacio, obteniéndose así nuevas variables que conforman las componentes principales.

Los autovalores del sistema se ordenan $\lambda_1 \cdot \lambda_2 \cdot \dots \cdot \lambda_p$ de modo de que el primero explique el mayor porcentaje de variabilidad del sistema, el segundo autovalor el siguiente porcentaje, etc.

Esto permite descartar los últimos siguiendo varios criterios:

- se toman en cuenta los primeros autovalores hasta alcanzar un alto porcentaje de explicación de la variabilidad, por ejemplo 80%
- se considera la cantidad de autovalores hasta que un punto de inflexión de la serie representada por éstos
- si se utiliza la matriz de correlaciones, se consideran sólo los autovalores mayores a 1

Las nuevas componentes se hallan multiplicando la matriz de variables originales A por una matriz C cuyas columnas están formadas por los autovectores seleccionados.

$$\left[\begin{array}{c} \text{Variables} \\ \text{originales} \end{array} \right] \left[\begin{array}{c} \text{Autovectores} \end{array} \right] = \left[\begin{array}{c} \text{Nuevas} \\ \text{componentes} \end{array} \right]$$

$$CP_n = e_{n1} * X_1 + e_{n2} * X_2 + \dots + e_{np} * X_p$$

con variables originales X_1, X_2, \dots, X_p y autovectores e_1, e_2, \dots, e_p

El ACP para ordenar observaciones se basa en la descomposición espectral de la matriz de covarianzas o de correlación entre variables de dimensión $p \times p$. La selección entre el estimador insesgado y el estimador máximo-verosímil de la matriz de covarianza poblacional es irrelevante, ya que produce las mismas componentes principales muestrales.

Con los autovectores como vectores de coeficientes para la combinación lineal se puede demostrar que las componentes principales son combinaciones lineales no correlacionadas cuyas varianzas son máximas.

La j -ésima componente principal (CP $_j$) es algebraicamente una combinación lineal de las p variables originales. Las nuevas variables usan información contenida en cada una de las variables originales, algunas variables pueden contribuir más a la combinación lineal que otras. Además se satisface que entre dos componentes cualesquiera, la covarianza es nula.

Los coeficientes de cada variable original estandarizados para una CP, permiten identificar las variables con mayor contribución en la explicación de la variabilidad entre observaciones en el eje asociado a la CP correspondiente.

4.5. Propiedades de PCA

Resultado 1. Sea $\mathbf{X}' = [X_1, X_2, \dots, X_p]$ con matriz de covarianzas Σ y pares de autovalores y autovectores de Σ $(\lambda_1, e_1), (\lambda_2, e_2), \dots, (\lambda_p, e_p)$, donde $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$. Sea $Y_1 = e_1' \mathbf{X}, Y_2 = e_2' \mathbf{X}, \dots, Y_p = e_p' \mathbf{X}$ las componentes principales. Entonces

$$\sigma_{11} + \sigma_{22} + \dots + \sigma_{pp} = \sum_p \text{Var}(X_i) = \lambda_1 + \lambda_2 + \dots + \lambda_p = \sum_p \text{Var}(Y_i)$$

Prueba. La traza de Σ es $\sigma_{11} + \sigma_{22} + \dots + \sigma_{pp} = \text{tr}(\Sigma)$. Dada la descomposición espectral de los autovectores, podemos escribir $\Sigma = \mathbf{P} \Delta \mathbf{P}'$

donde Δ es la matriz diagonal de autovalores y $\mathbf{P} = [e_1, e_2, \dots, e_p]$ así que

$\mathbf{P} \mathbf{P}' = \mathbf{P}' \mathbf{P} = \mathbf{I}$. Tenemos entonces

$$\text{tr}(\Sigma) = \text{tr}(\mathbf{P} \Delta \mathbf{P}') = \text{tr}(\Delta \mathbf{P}' \mathbf{P}) = \text{tr}(\Delta) = \lambda_1 + \lambda_2 + \dots + \lambda_p$$

Así

$$\sum_p \text{Var}(X_i) = \text{tr}(\Sigma) = \text{tr}(\Delta) = \sum_p \text{Var}(Y_i)$$

La varianza total poblacional = $\sigma_{11} + \sigma_{22} + \dots + \sigma_{pp} = \lambda_1 + \lambda_2 + \dots + \lambda_p$

En consecuencia, la proporción de varianza total explicada por la k -ésima componente principal es

$$\% \text{Variancia explicada}_k = \lambda_k / (\lambda_1 + \lambda_2 + \dots + \lambda_p)$$

con $k = 1, 2, \dots, p$

La mayor parte (por ejemplo el 80 al 90%) de la varianza total poblacional, puede ser atribuida a la primeras una, dos o tres componentes, por lo que dichas componentes podrían reemplazar las p variables originales sin mucha pérdida de información.

Cada componente del vector de coeficientes $e_1' = [e_{11}, \dots, e_{1k}, \dots, e_{1p}]$ amerita inspección. La magnitud de e_{ik} mide la importancia de la k -ésima variable a la i -ésima

componente principal, más allá de las otras variables. En particular e_{ik} es proporcional al coeficiente de correlación entre Y_i y X_k .

Resultado 2. Si $Y_1 = e'_1 \mathbf{X}$, $Y_2 = e'_2 \mathbf{X}$, ..., $Y_p = e'_p \mathbf{X}$ son las componentes principales obtenidas de la matriz de covarianzas Σ , entonces

$$\rho_{Y_i X_k} = \frac{e_{ik} \sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}$$

con $i, k = 1, 2, \dots, p$, son los coeficientes de correlación entre las componentes Y_i y las variables X_k . Aquí (λ_1, e_1) , (λ_2, e_2) , ..., (λ_p, e_p) son los pares de autovalores-autovectores de Σ .

Prueba. Si $a'_k = [0, \dots, 0, 1, 0, \dots, 0]$ con $X_k = a'_k \mathbf{X}$ y

$\text{Cov}(X_k, Y_i) = \text{Cov}(a'_k \mathbf{X}, e'_i \mathbf{X}) = a'_k \Sigma e_i$. Como $\Sigma e_i = \lambda_i e_i$, $\text{Cov}(X_k, Y_i) = a'_k \lambda_i e_i = \lambda_i e_{ik}$.

Entonces $\text{Var}(Y_i) = \lambda_i$ y $\text{Var}(X_k) = \sigma_{kk}$ por lo que

$$\rho_{Y_i X_k} = \frac{\text{Cov}(Y_i, X_k)}{\sqrt{\text{Var}(Y_i)} \sqrt{\text{Var}(X_k)}} = \frac{\lambda_i e_{ik}}{\sqrt{\lambda_i} \sqrt{\sigma_{kk}}} = \frac{e_{ik} \sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}$$

con $i, k = 1, 2, \dots, p$

Aunque las correlaciones de las variables con las componentes principales a menudo ayudan a interpretar las componentes, miden sólo la contribución univariada de una X a una componente Y . Esto es, ellas no indican la importancia de una X a una componente Y en presencia de otras X 's. Por esta razón, algunos estadísticos recomiendan que sólo los coeficientes e_{ik} y no las correlaciones, sean usadas para interpretar las componentes. En la práctica, las variables con coeficientes relativamente grandes (en valor absoluto) tienden a tener grandes correlaciones, así las dos medidas de importancia, la primera multivariada y la segunda univariada, frecuentemente dan resultados similares.

Resultado 3. La i -ésima componente principal de variables estandarizadas $\mathbf{Z}' = [Z_1, Z_2, \dots, Z_p]$ con $\text{Cov}(\mathbf{Z}) = \mathbf{p}$, está dada por

$$Y_i = e'_i \mathbf{Z} = e'_i (\mathbf{V}^{1/2})^{-1} (\mathbf{X} - \mu) \quad i = 1, 2, \dots, p$$

Más aún

$$\sum_p \text{Var}(X_i) = \sum_p \text{Var}(Y_i) = p$$

y

$$\rho_{Y_i X_k} = e_{ik} \cdot \lambda_i \quad i, k = 1, 2, \dots, p$$

En este caso $(\lambda_1, e_1), (\lambda_2, e_2), \dots, (\lambda_p, e_p)$ son los pares de autovalores-autovectores de ρ , con $\lambda_1 \cdot \lambda_2 \cdot \dots \cdot \lambda_p \neq 0$.

Prueba. La varianza total poblacional (variables estandarizadas) es simplemente p , la suma de los elementos de la diagonal de la matriz ρ . Usando lo visto en el **Resultado 1** con \mathbf{Z} en lugar de \mathbf{X} , encontramos la proporción de la varianza explicada por la k -ésima componente principal de \mathbf{Z} como λ_k/p , con $k = 1, 2, \dots, p$, donde λ_k son los autovalores de ρ .

4.6. PCA para compresión de datos

Supongamos que los datos a ser comprimidos consisten de N tuplas o vectores de datos, de k dimensiones. El Análisis de Componentes Principales busca c vectores ortogonales de dimensión k , los cuales sean los que mejor representan la información, siendo $c \leq k$. Los datos originales son proyectados en un espacio mucho más pequeño, reduciendo la dimensionalidad, dando como resultado una compresión de los mismos. Sin embargo no se reduce el número de variables originales, sino que se obtiene un conjunto de nuevas variables (con menos variables que el original) que conserva la misma información principal.

El procedimiento básico consiste de los siguientes pasos:

1. Los datos originales pueden ser normalizados, de modo que los valores de todos los atributos estén en un mismo rango. Esto ayuda a que los atributos con dominios más grandes no sean dominantes frente a las otras variables.
2. PCA computa c vectores ortonormales que provee una base para los datos originales normalizados. Éstos son vectores unitarios cuyos puntos están en una dirección perpendicular a los otros. Estos vectores son denominados *componentes principales*.
3. Las componentes principales son ordenadas según su "significancia" en orden decreciente. Esencialmente sirven como un nuevo conjunto de ejes para los datos, proveyendo de más información sobre la varianza, desde el primero al último. Esta información ayuda a identificar grupos o patrones en el conjunto de datos.
4. Dado que las componentes son ordenadas según un orden decreciente de la "significancia", el tamaño de los datos puede ser reducido eliminando las componentes más débiles, es decir las que explican menor varianza.

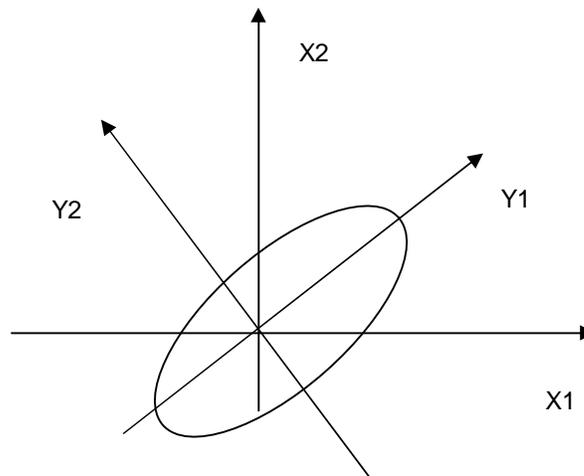


Fig. 4.1 – PCA conforma un nuevo sistema de ejes. X1 y X2 conforman las variables originales y Y1 y Y2 las nuevas componentes

Usando las componentes principales más fuertes, es posible reconstruir una buena aproximación de los datos originales.

PCA no es costoso en términos computacionales, puede ser aplicado para ordenar atributos y es útil para el manejo de información dispersa en un conjunto de datos. Información de más de dos dimensiones puede ser transformada para reducir el problema a dos dimensiones.

5. Análisis Discriminante

Otro método de Data Mining es el *Análisis Discriminante*. Permite describir algebraicamente las relaciones entre dos o más poblaciones (grupos) de manera tal que las diferencias entre ellas se maximicen o se hagan más evidentes. Se realiza frecuentemente con fines predictivos relacionados a la clasificación, en una de las poblaciones existentes, de nuevas observaciones u observaciones sobre las cuales no se conoce a qué grupo pertenecen.

Una observación nueva, la cual no fue utilizada para la construcción de la regla de clasificación, se asignará al grupo en el cual tienen más probabilidad de pertenecer en base a sus características medidas. Para tal asignación es necesario definir una regla de clasificación. La función discriminante lineal puede ser usada para este fin. Además, puede ser usado con el objetivo de encontrar el subconjunto de variables que mejor explica la variabilidad entre grupos.

5.1. Reglas de clasificación

El LDA permite caracterizar grupos diferenciados, mediante la generación de reglas de clasificación. Luego, es posible predecir a qué grupo correspondería una nueva observación de la cual se desconoce el origen. Para esto pueden utilizarse varios métodos, como por ejemplo los descriptos a continuación.

Supongamos que se tienen dos poblaciones normales multivariadas Π_1 y Π_2 y se sabe que un nuevo vector de observaciones \mathbf{x} proviene de Π_1 o de Π_2 . Se necesita una regla que se pueda usar para predecir de cuál de las dos poblaciones es más probable que provenga \mathbf{x} .

a) Regla de función discriminante lineal

Cuando dos poblaciones normales multivariadas tienen iguales matrices de varianzas-covarianzas (es decir, cuando $\Sigma_1 = \Sigma_2$) la regla puede consistir en

$$\text{Elija } \Pi_1 \text{ si } \mathbf{b}'\mathbf{x} - k > 0 \text{ y de lo contrario elija } \Pi_2 \text{ donde } \mathbf{b} = \Sigma^{-1} (\mu_1 - \mu_2) \text{ y}$$
$$k = 1/2 (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 + \mu_2)$$

La función $\mathbf{b}'\mathbf{x}$ se llama *función discriminante lineal* de \mathbf{x} y es la función lineal única de los elementos en \mathbf{x} que resume toda la información contenida en este vector, de la que se dispone para realizar una discriminación efectiva entre dos poblaciones normales multivariadas que tienen iguales matrices de varianzas-covarianzas.

b) Regla de distancia de Mahalanobis

Cuando dos poblaciones normales multivariadas tienen iguales matrices de varianzas-covarianzas, la regla es equivalente a

$$\text{Elija } \Pi_1 \text{ si } d_1 < d_2, \text{ donde } d_i = (\mathbf{x} - \mu_i)' \Sigma^{-1} (\mathbf{x} - \mu_i) \text{ para } i = 1, 2.$$

La cantidad d_1 , es en cierto sentido una medida de lo lejos que está \mathbf{x} de μ_i y d_i se llama *cuadrado de la distancia de Mahalanobis* entre \mathbf{x} y μ_i . Esta medida de la distancia toma en consideración las varianzas y covarianzas de las variables medidas. La regla de la distancia de Mahalanobis al cuadrado clasifica una observación en la población cuya media esté "más próxima".

c) Regla de probabilidad posterior

Cuando las matrices de varianzas-covarianzas son iguales, la cantidad $P(\Pi_i | \mathbf{x})$ definida por

$$P(\Pi_i | \mathbf{x}) = \exp[(-1/2) d_i] / \{ \exp[(-1/2) d_1] + \exp[(-1/2) d_2] \}$$

Se llama *probabilidad posterior* de la población Π_i , dado \mathbf{x} . En realidad la probabilidad posterior no es una probabilidad verdadera, porque no se está considerando ningún evento aleatorio. La observación pertenece a una de las poblaciones y la incertidumbre proviene de la capacidad para elegir la población correcta.

Por lo tanto, una regla discriminante basada en probabilidades posteriores sería:

$$\text{Elija } \Pi_1 \text{ si } P(\Pi_1 | \mathbf{x}) < P(\Pi_2 | \mathbf{x}), \text{ y de lo contrario elija } \Pi_2$$

5.2. Estimaciones de las probabilidades de una clasificación errónea

Cuando se realiza un análisis discriminante se necesita poder determinar las probabilidades de las clasificaciones correctas de las nuevas observaciones. Algunos métodos para estas estimaciones son:

Restitución – consiste en probar la clasificación con los mismos datos utilizados para la generación de las reglas. La mayor desventaja de este método es que estima en exceso las probabilidades de una clasificación correcta

Datos propuestos – consiste en dividir los datos disponibles en dos grupos, uno de entrenamiento para la generación de las reglas de clasificación y uno de prueba para estimar las probabilidades de clasificaciones erróneas

Validación cruzada – se quita del conjunto la primera observación y se generan las reglas de clasificación con los datos restantes. Luego se reincorpora esta observación y se elimina la segunda, generando nuevas reglas. Y así siguiendo con el resto de los datos, para finalmente crear una matriz resumen para estas estimaciones validadas en forma cruzada

5.3. Funciones discriminantes canónicas

Fisher introdujo la idea de análisis discriminante canónico, en donde se crean nuevas variables al tomar combinaciones lineales especiales de las variables originales. Las variables canónicas se crean de modo que contengan toda la información útil que se encuentra en un conjunto de variables originales. En cierto sentido son semejantes a las componentes principales, aunque su cálculo difiere.

La cantidad de funciones es igual a la cantidad de grupos a discriminar menos 1. Salvo el caso especial donde la cantidad de variables sea menor a la cantidad de grupos; en ese caso la cantidad de funciones es igual a la cantidad de variables menos 1.

A partir de las funciones canónicas es posible por ejemplo obtener un gráfico con la dispersión de los grupos en pocas dimensiones, ubicándose el centroide de cada uno de ellos. Por ejemplo, esto puede utilizarse para predecir la correspondencia de un nuevo individuo, según la cercanía de éste a cada uno de los centroides.

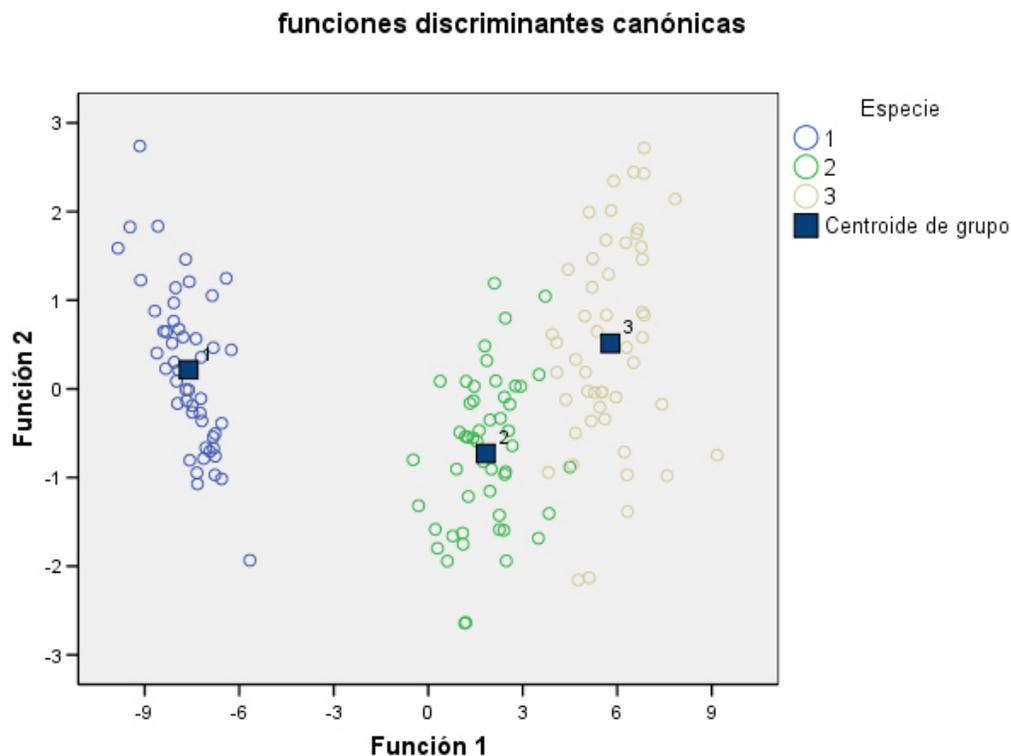


Fig. 5.1 – Gráfico de dispersión de las funciones discriminantes canónicas de las especies de Iris con los centroides de cada uno de los grupos (Anexo C – Ejemplos de PCA y LDA)

Parte III: Reconocimiento de Rostros

6. Tratamiento de Imágenes

La visión es uno de los mecanismos sensoriales de percepción más importantes en el ser humano. El procesamiento de imágenes digitales se centra en:

- ▀ Mejora de la calidad para la interpretación humana
- ▀ Procesamiento de los datos para la percepción de las máquinas en forma autónoma

La imagen se presenta digitalizada en forma de matriz con una resolución de $M \times N$ elementos. Cada elemento de la matriz se denomina píxel (picture element) y se corresponde con el nivel de luminosidad del punto correspondiente. La representación de una imagen en tonos de grises es diferente a la de una en colores.

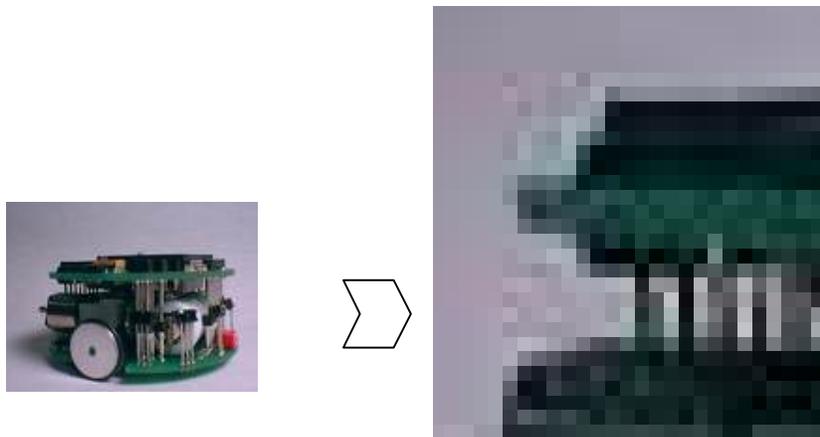
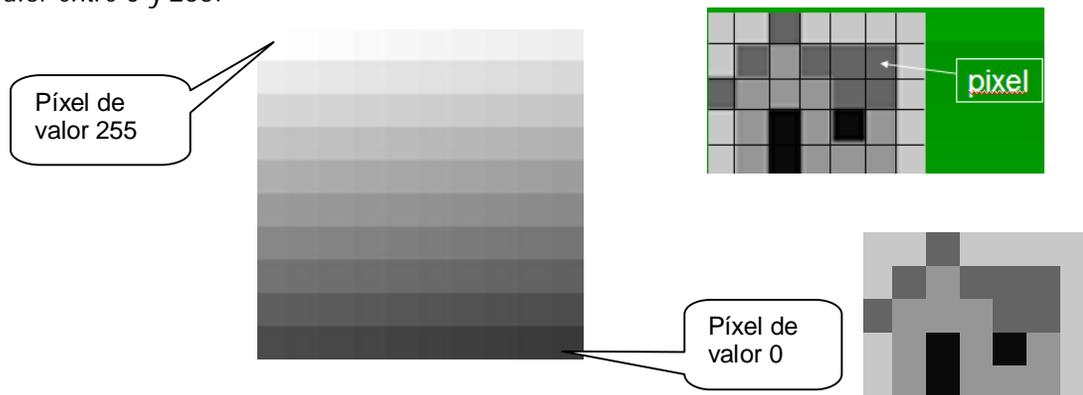


Fig. 6.1 - Se observa al ampliar la imagen que la misma está conformada de píxeles de diferentes colores

La imagen en blanco y negro (en realidad en tonos de grises) tiene en cada punto un valor entre 0 y 255.



A cada imagen la podemos transformar en una matriz, donde cada número es el valor de cada pixel.

```

200 200 100 200 200 200 200
200 100 150 100 100 100 200
100 150 150 150 100 100 200
200 150 10 150 10 150 200
200 150 10 150 150 150 200

```

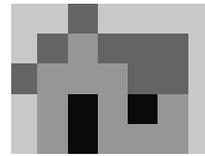
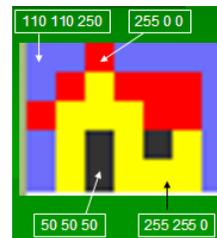
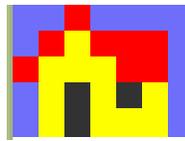


Fig. 6.0.1 – Ejemplo de imagen en tonos de grises y matriz de valores de píxeles

En el caso de una imagen en colores, hay distintas representaciones, pero básicamente cada pixel tiene 3 valores entre 0 y 255: uno para el rojo (R), uno para el verde (G) y uno para el azul (B).

También podemos armar una matriz con los valores de los pixels.



```

110 110 250 110 110 250 255 0 0 110 110 250 110 110 250 110 110 250 110 110 250
110 110 250 255 0 0 255 255 0 255 0 0 255 0 0 255 0 0 110 110 250
255 0 0 255 255 0 255 255 0 255 255 0 255 0 0 255 0 0 110 110 250
110 110 250 255 255 0 50 50 50 255 255 0 50 50 50 255 255 0 110 110 250
110 110 250 255 255 0 50 50 50 255 255 0 255 255 0 255 255 0 110 110 250

```

Fig. 6.0.2 – Ejemplo de imagen en color y matriz de valores de píxeles

Si consideramos el valor del píxel como una tercera dimensión de los datos, podemos obtener una imagen tridimensional de la superficie de la misma.

Por ejemplo, podemos ver las superficies de estas dos imágenes, obtenidas con el software ImageJ [IMA07] (*Anexo A – Software utilizado*).

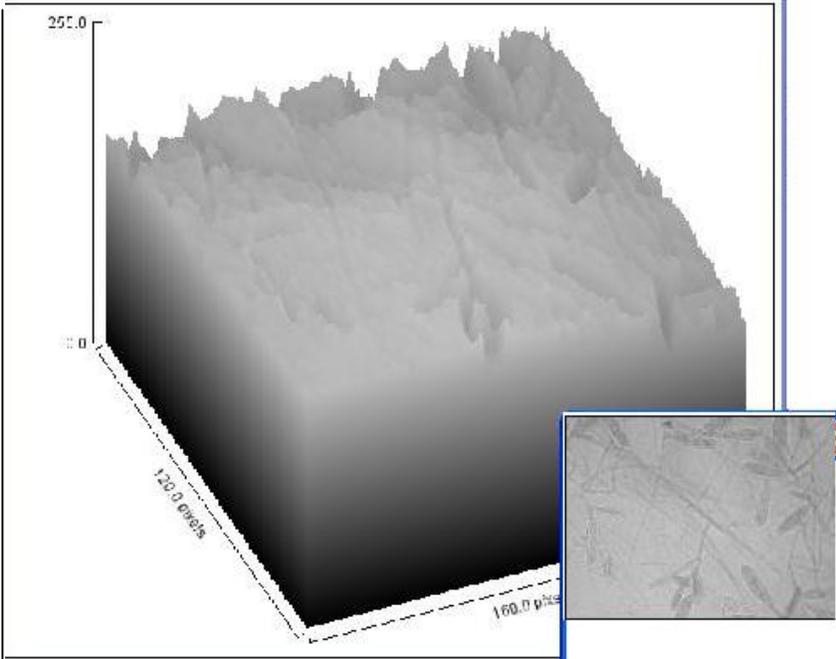


Fig. 6.0.3 – Superficie de una imagen de un dermatofito E. floccosum

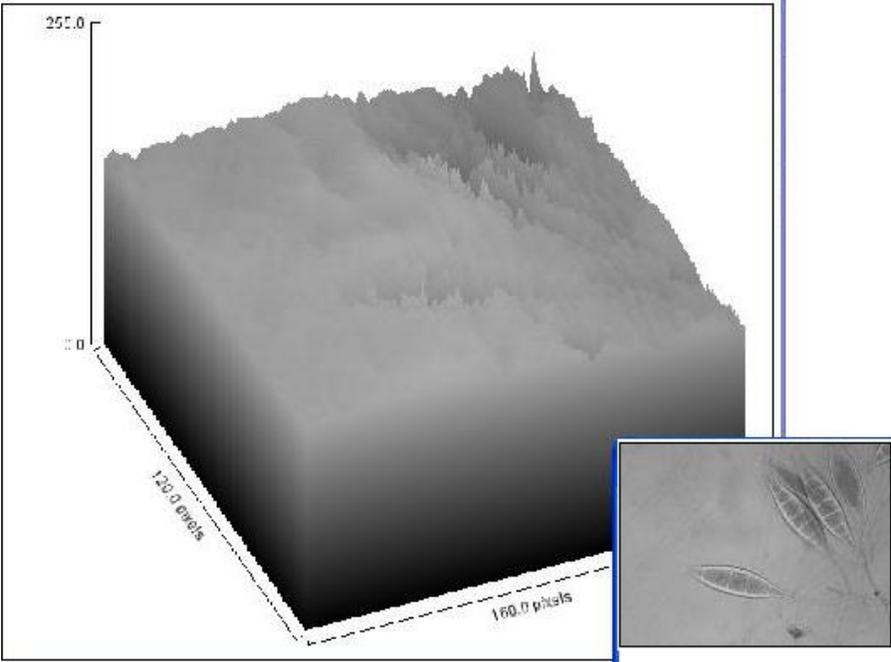


Fig. 6.0.4 – Superficie de una imagen de un dermatofito M. canis

6.1. Histograma de una imagen

Supongamos dada una imagen en niveles de grises, siendo el rango de 256 colores (de 0 a 255). El histograma de la imagen consiste en una gráfica donde se muestra el número de píxeles de cada nivel de gris que aparecen en la imagen.

Por ejemplo, la siguiente imagen tiene como histograma (imágenes extraídas de [FOT05]):

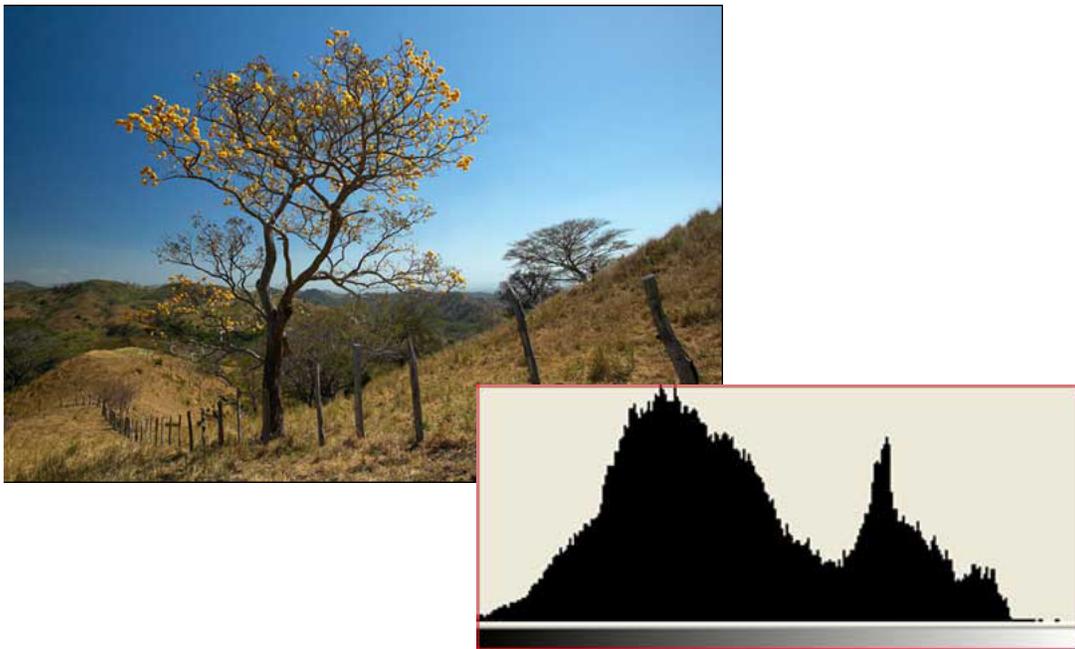


Fig. 6.0.5 – Ejemplo de histograma de imagen

El análisis estadístico derivado del histograma puede servir para comparar contrastes e intensidades entre imágenes. El histograma podría ser alterado para producir cambios en la imagen.

Por ejemplo, el histograma es utilizado para binarizar una imagen digital, es decir, convertirla en una imagen en blanco y negro, de tal manera que se preserven las propiedades "esenciales" de la imagen. La forma usual de binarizar una imagen es eligiendo un valor adecuado L dentro de los niveles de grises, tal que el histograma forme un "valle" en ese nivel. Todos los niveles de grises menores que L se convierten en 0 (negro), y los mayores que L se convierten en 255 (blanco).



Fig. 6.0.6 – Ejemplo de histograma de imagen oscura [FOT05]



Fig. 6.0.7 – Ejemplo de histograma de imagen clara [FOT05]

6.2. Filtros y Convolución

Los *filtros espaciales* [IMG07] tienen como objetivo modificar la contribución de determinados rangos de frecuencias a la formación de la imagen. El término *espacial* se refiere al hecho de que el filtro se aplica directamente a la imagen y no a una transformada de la misma, es decir, el nivel de gris de un píxel se obtiene directamente en función del valor de sus vecinos.

Los filtros espaciales pueden clasificarse basándose en su linealidad: *filtros lineales* y *filtros no lineales*. A su vez los filtros lineales pueden clasificarse según las frecuencias que dejen pasar: los *filtros paso bajo* atenúan o eliminan las componentes de alta frecuencia a la vez que dejan inalteradas las bajas frecuencias; los *filtros paso alto* atenúan o eliminan las componentes de baja frecuencia con lo que agudizan las componentes de alta frecuencia; los *filtros paso banda* eliminan regiones elegidas de frecuencias intermedias.

El tratamiento de imágenes más empleado y conocido, es el tratamiento espacial también conocido como *convolución*. Las convoluciones discretas son muy usadas en el procesado de imagen para el suavizado de imágenes, el afilado de imágenes, detección de bordes, y otros efectos. Mediante este proceso se calcula el valor de un determinado punto en función de su valor y del valor de los puntos que le rodean, aplicando una simple operación matemática en función de la cual se obtendrá un valor resultante para el punto en cuestión.

La idea es definir una máscara (una matriz cuadrada) que será aplicada a cada punto de la imagen, obteniéndose una nueva imagen transformada.

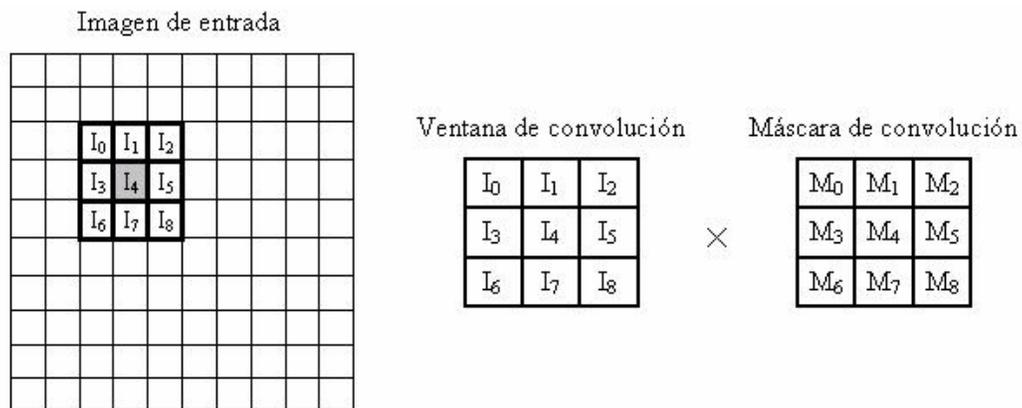


Fig. 6.0.8 –Definición de una máscara de convolución

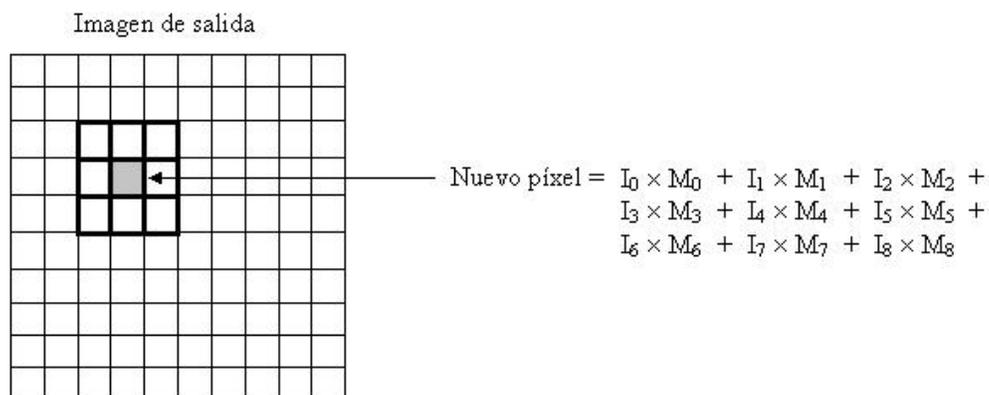


Fig. 6.0.9 –Aplicación de una convolución píxel a píxel

6.3. Transformación de las imágenes en datos

Para las pruebas del presente trabajo, cada imagen color fue transformada a niveles de grises (pasaje a blanco y negro) y se cambió el tamaño al requerido para su estudio.

Luego se tomaron cada una de las imágenes originales y se armó una variable con cada una de ellas, trasladando cada fila de píxeles.

De esta manera, cada imagen pasa a ser una variable de una gran base de datos, a la cual pueden aplicarse los métodos estadísticos multivariados correspondientes. Una imagen de dimensiones $m \times n$ píxeles se transforma en un vector de longitud $m \cdot n$.

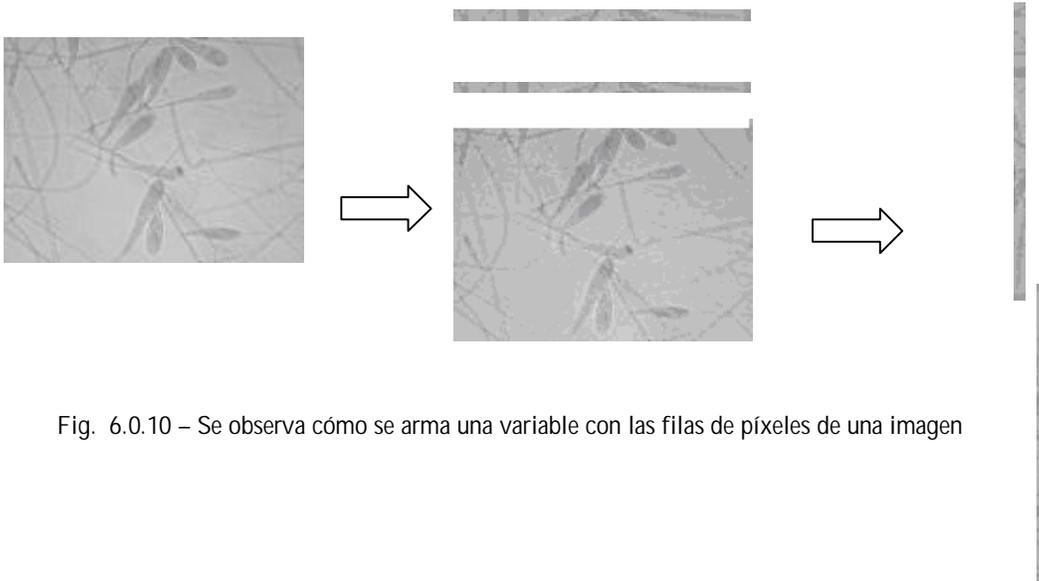


Fig. 6.0.10 – Se observa cómo se arma una variable con las filas de píxeles de una imagen

6.4. Transformada de Hotelling

La transformada de Hotelling, también llamada de Karhunen y Loève, está basada en propiedades estadísticas de representaciones vectoriales. Representa la aplicación del método de Análisis de Componentes Principales en el área de imágenes.

Para M vectores procedentes de una muestra aleatoria con vectores x_i , la media m_x y la matriz de covarianza C_x , se calculan como [PAJ02]:

$$m_x = \frac{1}{M} \sum_{k=1}^M m_k$$

$$C_x = \frac{1}{M} \sum_{k=1}^M x_k x_k^t - m_k m_k^t$$

Sean e_i y λ_i con $i = 1, 2, \dots, n$, los autovectores y autovalores de C_x , ordenados en orden descendente, de modo que $\lambda_j \geq \lambda_{j+1}$, para $j = 1, 2, \dots, n-1$. Por definición, los autovectores y autovalores de una matriz C de dimensión $n \times n$ satisfacen la relación $Ce_i = \lambda_i e_i$ para $i = 1, 2, \dots, n$. Sea \mathbf{A} una matriz cuyas filas están formadas por los autovectores de C_x , ordenados de forma que la primera fila de \mathbf{A} es el autovector asociado con el autovalor de mayor valor y la última fila es el autovector asociado con el autovalor más pequeño.

Supongamos que \mathbf{A} es una matriz de transformación que transforma los vectores \mathbf{x} en vectores \mathbf{y} como sigue:

$$\mathbf{y} = \mathbf{A} (\mathbf{x} - \mathbf{m}_x)$$

Esta ecuación se denomina *transformada de Hotelling*. La media de los vectores \mathbf{y} resultantes de esta transformación es cero, esto es, $\mathbf{m}_y = \mathbf{0}$.

Y la matriz de covarianza de los vectores \mathbf{y} se puede obtener en términos de \mathbf{A} y C_x por medio de

$$C_y = \mathbf{A} C_x \mathbf{A}^t$$

Además, C_y es una matriz diagonal cuyos elementos a lo largo de la diagonal principal son los autovalores de C_x , esto es,

$$C_y = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_n \end{bmatrix}$$

El efecto de aplicar esta transformación es establecer un nuevo sistema de coordenadas cuyo origen es el centroide de la muestra y cuyos ejes se sitúan en la dirección de los autovectores de C_x . Este alineamiento elimina la correlación de los datos.

Otra propiedad de la transformada de Hotelling es la reconstrucción de \mathbf{x} a partir de \mathbf{y} , puesto que las filas de \mathbf{A} son vectores ortonormales $\mathbf{A}^{-1} = \mathbf{A}^t$ y cualquier vector \mathbf{x} puede recuperarse a partir de su correspondiente \mathbf{y} y utilizando la relación $\mathbf{x} = \mathbf{A}^t \mathbf{y} + \mathbf{m}_x$

6.5. Compresión de imágenes

La compresión de imágenes implica la reducción de la dimensión de los archivos de imágenes mientras se retiene la información necesaria.

Actualmente, resulta cada vez más importante, dado el crecimiento de los sistemas multimedia, aplicaciones de video y desarrollos Web, que requieren de grandes cantidades de espacio, tanto para su almacenamiento como su transmisión.

Los algoritmos de compresión se desarrollan teniendo en cuenta la redundancia presente en los datos de las imágenes. Existen tres tipos básicos de redundancia: a) de códigos, b) interpíxeles y c) psicovisual. La redundancia de códigos ocurre cuando los datos que se utilizan para representar la imagen no se usan de manera óptima. La redundancia interpíxeles ocurre cuando los píxeles adyacentes tienden a estar altamente correlacionados. El tercer tipo se refiere al hecho de que alguna información es más importante para el sistema de visión humano que otros tipos de información.

Los métodos primarios de compresión de imágenes se clasifican en: los que preservan los datos, donde la imagen pueden ser reproducida exactamente luego de la descompresión, y los métodos en los que se pierde parte de la información.

Dadas las características del Análisis de Componentes Principales en los que se basa la Transformada de Hotelling, este método resulta útil para la compresión de imágenes con pérdida de información.

7. Técnicas de reconocimiento de rostros

Actualmente el reconocimiento de rostros en forma automática está en auge debido a la importancia de los sistemas de seguridad, en particular para el ingreso de personal autorizado o por ejemplo en aeropuertos o control de aduanas. Por este motivo existen diversos estudios sobre cómo detectar una cara en una imagen y cómo autenticar a los individuos involucrados.



Fig. 7.1 - Ejemplo de rostro humano

También es importante su estudio para casos de interfaz hombre-máquina, ya sea para identificación del individuo al registrarse, detecciones de cansancio de conductores de automóviles o para interactuar, mediante movimientos del cursor gestuales o por ejemplo casos de robots que identifican al usuario.

Los seres humanos reconocemos las caras de familiares, amigos y conocidos de manera inmediata casi sin poder definir sus diferencias. Tomamos en cuenta características adicionales como la forma del cabello o el tono de voz, pero seguimos conociendo a cada persona, aunque use anteojos, cambie el color o el corte de pelo o se deje crecer barba.

Existen diversas técnicas para la identificación de rostros humanos, las cuales podrían clasificarse en dos grandes grupos: la identificación por características y las aproximaciones estadísticas también llamadas *eigenfaces*.

7.1. Técnicas por características

Los rostros humanos comparten ciertas características como la ubicación de los ojos, el color de los labios, el marco de las cejas, etc.

Estas técnicas de reconocimiento toman en cuenta en las imágenes modelos colorímetros y proporciones geométricas de la disposición de los componentes del rostro, curvaturas de huesos, etc.

La tarea de identificación de rostros humanos, consta de dos etapas principales:

- **Detección:** dentro de la imagen detectar la presencia de una cara humana y separarla del fondo de la misma
- **Reconocimiento:** reconocer la identidad de la persona identificada, entre las identidades almacenadas en el sistema.

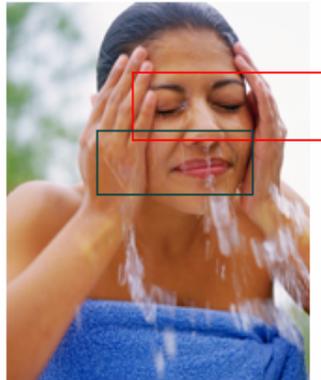


Fig. 7.2 – La detección de una cara en una imagen es el paso previo al reconocimiento del individuo

7.2. Problemática

Para la detección se basan de características de bajo nivel como son los bordes, los niveles de gris, el color y el movimiento. La distribución de mínimos locales de niveles de gris puede señalar la presencia de cejas, pupilas y labios.

Si bien las caras están conformadas de manera de representar una simetría bilateral, ambas partes no son exactamente iguales.

Se analizan expresiones faciales, comportamiento de músculos y huesos (por ejemplo apertura de la boca, movimiento de cejas) (*Fig. 7.3*)

Hay diversos factores que hacen diferir una misma cara en diversos instantes:

- Variaciones en la iluminación de la imagen
- Cambios de expresiones faciales, según contexto, situación actual, emociones, etc.
- Cambios de pose – vista de frente, movimientos de la cabeza, posturas corporales, etc.
- Cambios de forma y color – anteojos, maquillaje, barba, peinado
- Cambios de textura – adelgazamiento/engorde, envejecimiento, etc.



Fig. 7.3 – Diferentes poses de una misma persona [GRO01]

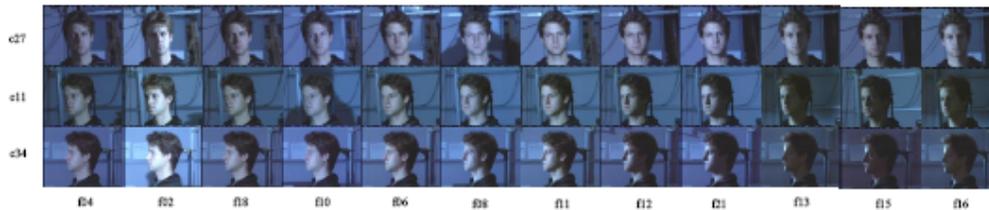


Fig. 7.4 – Variaciones de iluminación [GRO01]

7.3. Análisis de bajo nivel

Bordes: Uno de los rasgos más primitivos que tiene cualquier figura es su contorno. Los trabajos que utilizan esta idea extraen los bordes de la cara tanto externos como internos. Luego son sometidos al análisis de forma y posición. También se puede utilizar para detectar si el individuo lleva anteojos.

Niveles de gris: Los rasgos faciales como las cejas, pupilas y los labios aparecen generalmente más oscurecidos que las regiones de su alrededor. Algunos algoritmos de extracción de rasgos faciales buscan mínimos locales dentro de regiones faciales segmentadas. La posición relativa de los ojos y su detección también puede ser descubierta con este tipo de métodos.



Fig. 7.5 – En este ejemplo se observan dos rostros con distintos rasgos faciales (imágenes extraídas de [FAC02])

Color: El color de la piel humana (en sus diferentes variantes) ha permitido desarrollar algunas técnicas que detectan la raza. También se utiliza el color para ubicar la boca o los ojos del individuo.



Fig. 7.6 – Se dificulta el reconocimiento de la boca (también en el caso de los ojos) debido a que las diferentes posiciones de los labios hacen cambiar la luminosidad, colores y geometría de la misma (imágenes extraídas de [FAC02])

7.4. Análisis de rasgos

Búsqueda de rasgos: Estas técnicas buscan rasgos prominentes que permiten localizar rasgos menos prominentes partiendo de hipótesis geométricas. Por ejemplo una pequeña área sobre una alargada puede corresponderse al escenario (cabeza sobre los hombros) y un par de regiones oscuras encontradas en el área facial incrementa la probabilidad de que aquello sea realmente una cara. Los rasgos más usados son los ojos, contorno de la cabeza y el cuerpo (bajo la cabeza).

Análisis de constelaciones: se basa en el uso de un modelo probabilístico que estudia la posición espacial de los rasgos faciales, intentando buscar patrones que se asemejen a una cara.

7.5. Modelos de silueta activa

Snakes: Se usan comúnmente para localizar el contorno de la cabeza. Se inicializa una *snake* en las proximidades de una cara, y posteriormente se va fijando a los contornos que encuentra hasta asumir la forma de la cara. Cuando la *snake* llegue al equilibrio se habrá ajustado a la forma de la cara.

Plantillas deformables: Es el siguiente paso, usando las *snakes* para encontrar más rasgos faciales además del contorno de la cara. Por ejemplo, se pueden encontrar los ojos usando para las *snakes* un mecanismo de deformación que incluye el cálculo de gradientes de una función de energía.

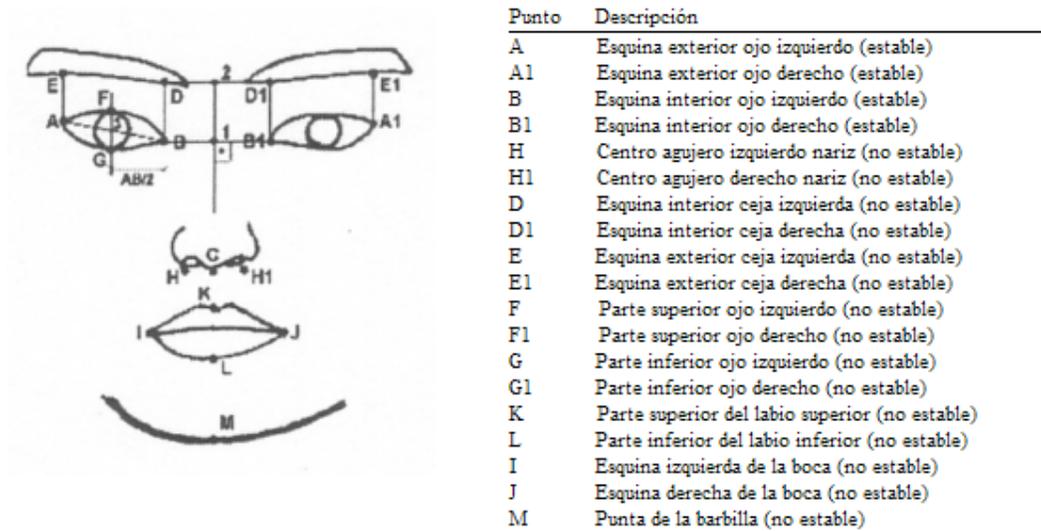


Fig. 7.7 – Puntos clave para el modelo frontal (extraído de [SEV06])

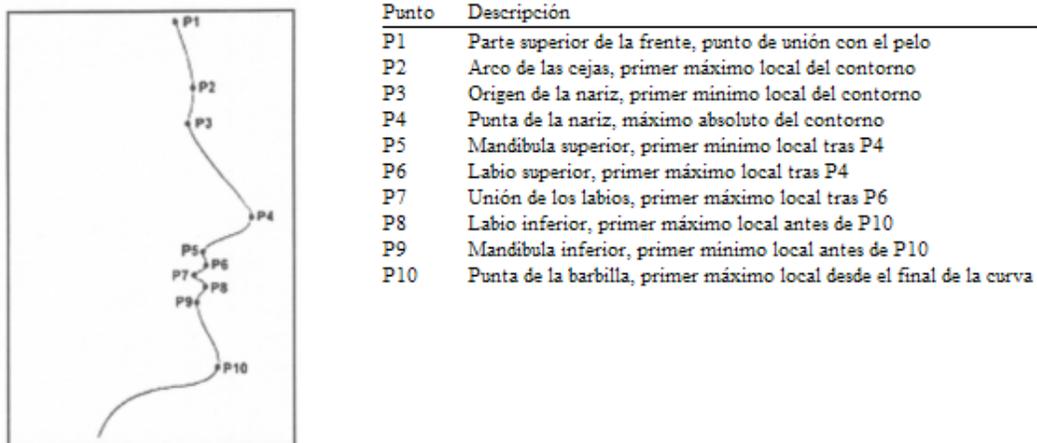


Fig. 7.8 – Puntos clave para el modelo lateral (extraído de [SEV06])

Parte IV: Método de reconocimiento *Eigenfungi*

8. Desarrollo del método

En este trabajo desarrollamos un método automático para el reconocimiento de especies de hongos microscópicos, que denominamos *eigenfungi*. Está basado en la metodología para reconocimiento de rostros denominado *eigenfaces*, al que se le introducen varias modificaciones que mejoran su exactitud en el análisis de imágenes microscópicas de hongos.

El método de *eigenfaces* presentado por Turk y Pentland [TRK91] es utilizado para el reconocimiento de rostros de personas. Se basa en la metodología de la Transformada de Hotelling. *Eigenfaces* proviene del prefijo alemán *eigen* que significa propio y proviene de la característica del método de encontrar un nuevo sistema de coordenadas con los *autovectores* de la matriz de covarianzas del conjunto original, en inglés *eigenvectors*.

En este caso, se toma una muestra de fotos de los individuos que se quieren reconocer (por ejemplo las personas autorizadas en una empresa) y se arma un conjunto de nuevas imágenes denominadas *eigenfaces*. Éstas contienen la información principal de las imágenes originales.

Luego se obtiene la distancia de cada foto a las *eigenfaces*. Se agrupan las fotos por individuo y se calcula la distancia promedio del grupo.

Al intentar reconocer a una persona, se le saca una foto y se calcula la distancia de ésta a las *eigenfaces*. Finalmente se compara esta distancia con la de cada grupo, siendo la mínima la que identifica la persona analizada.

Veamos ahora los pasos que componen el método de *eigenfaces*. Las imágenes que se muestran corresponden a la base "The ORL Faces" del Cambridge University Computer Laboratory.

1. Primero se obtiene un conjunto de varias imágenes por persona, en lo posible con diferentes expresiones y condiciones de iluminación. Sea M el conjunto de las imágenes $\Gamma_1, \Gamma_2, \dots, \Gamma_M$



Fig. 8.1 - Ejemplos de imágenes de entrenamiento

2. Se calcula la imagen media del conjunto como

$$\psi = \frac{1}{M} \sum_{i=1}^M \Gamma_i$$

3. Luego se resta la imagen media Ψ a cada imagen del conjunto de entrenamiento

$$\phi_i = \Gamma_i - \psi$$

4. Se arma una matriz A con las imágenes resultantes

$$A = \{\Phi_1, \Phi_2, \dots, \Phi_M\}$$

5. A partir de A se calcula la matriz C de covarianzas

$$C = \frac{1}{M} \sum_{i=1}^M \phi_i \phi_i^t = AA^t$$

6. Se calculan los autovalores y autovectores v de C .

7. A partir de los autovectores encontrados y las imágenes (menos la imagen media), se calculan las *eigenfaces* U

$$U_i = \sum_{k=1}^M v_{ik} \phi_k$$

con $i = 1, \dots, M$



Fig. 8.2 - Ejemplos de eigenfaces

8. Posteriormente se halla la distancia de cada imagen original a cada eigenface y con eso se arma un vector de distancias para cada imagen

$$s_i = U\phi_i$$

con $i = 1, 2, \dots, M$

9. Se agrupan todas las imágenes de un mismo individuo y se calcula el promedio de los vectores distancia. Este nuevo vector, que representa a la persona, se denomina *vector de clase*

10. Cuando se tiene una nueva imagen j , se le resta la media y se halla su vector de distancia

$$s_j = U\phi_j$$

11. Finalmente, se compara el nuevo vector de distancias con los vectores de clase de cada individuo. La nueva imagen entonces, se corresponde con el individuo cuyo vector de clase sea el más cercano

8.1. Diferencias entre rostros y hongos

Las imágenes microscópicas de los dermatofitos tienen características diferentes a las imágenes de rostros:

Tópico	Rostros	Hongos
Cantidad de objetos a reconocer	Un único objeto (la cara)	Varios objetos (conidias, hifas)
Importancia de los objetos de fondo	Los objetos de fondo son eliminados	Todos los objetos de la imagen son importantes
Normalización de los objetos	Los objetos pueden ser normalizados de modo de homogeneizar el tamaño de las cabezas, posición de los ojos, etc.	Los objetos no pueden ser normalizados



Fig. 8.3 - Ejemplos de imágenes de persona, de *M. canis* y de *T. tonsurans*

Esto haría pensar que el método de las *eigenfaces* sería incompatible a la hora de identificar hongos microscópicos. Sin embargo, con unas modificaciones que permitan adaptar el método a este tipo de imágenes, la exactitud de la clasificación es muy buena, requiriéndose conjuntos de pocas imágenes para el entrenamiento.

8.2. Descripción del método

Básicamente el método *eigenfungi* consta de los siguientes pasos:

Entrenamiento

- obtención del conjunto de imágenes de las especies a identificar
- cálculo de la imagen media
- cálculo de la matriz de covarianza de las imágenes
- cálculo de las *eigenfungi*
- obtención de la distancia (o peso) de cada imagen original a cada *eigenfungi*

Utilización

- cálculo de la distancia de cada nueva imagen a cada *eigenfungi*
- comparación de las distancias obtenidas respecto de las distancias de las imágenes originales, para hallar la más similar e identificar la especie

8.3. Cálculo de los *eigenfungi*

Una vez obtenido el conjunto de imágenes, se procede como en el caso de las *eigenfaces*, obteniéndose los *eigenfungi*.



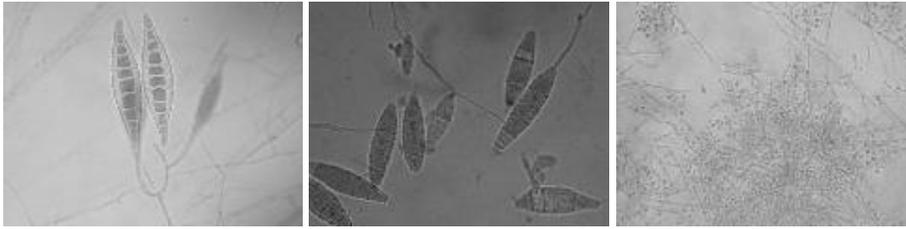


Fig. 8.4 - Ejemplos de imágenes originales

1. Se calcula la imagen media del conjunto como

$$\psi = \frac{1}{M} \sum_{i=1}^M \Gamma_n$$

2. Luego se resta la imagen media Ψ a cada imagen del conjunto de entrenamiento

$$\phi_i = \Gamma_i - \psi$$

3. Se arma una matriz A con las imágenes resultantes

$$A = \{\Phi_1, \Phi_2, \dots, \Phi_M\}$$

4. A partir de A se calcula la matriz C de covarianzas

$$C = \frac{1}{M} \sum_{i=1}^M \phi_n \phi_n^t = AA^t$$

5. Se calculan los autovalores y autovectores v de C .

6. A partir de los autovectores encontrados y las imágenes (menos la imagen media), se calculan los *eigenfungi* U

$$U_i = \sum_{k=1}^M v_{ik} \phi_k$$

con $i = 1, \dots, M$

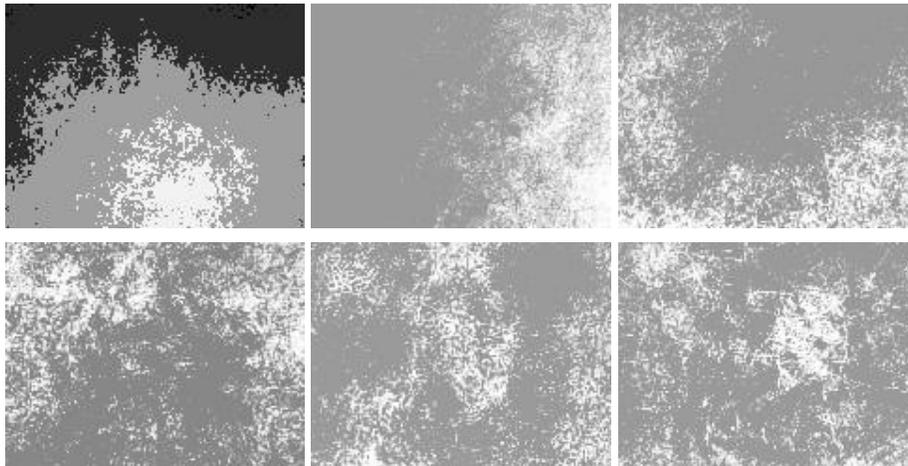


Fig. 8.5 - Ejemplos de eigenfungi

7. Posteriormente se halla la distancia de cada imagen original a cada *eigenfungi* y con eso se arma un *vector de distancias* para cada imagen

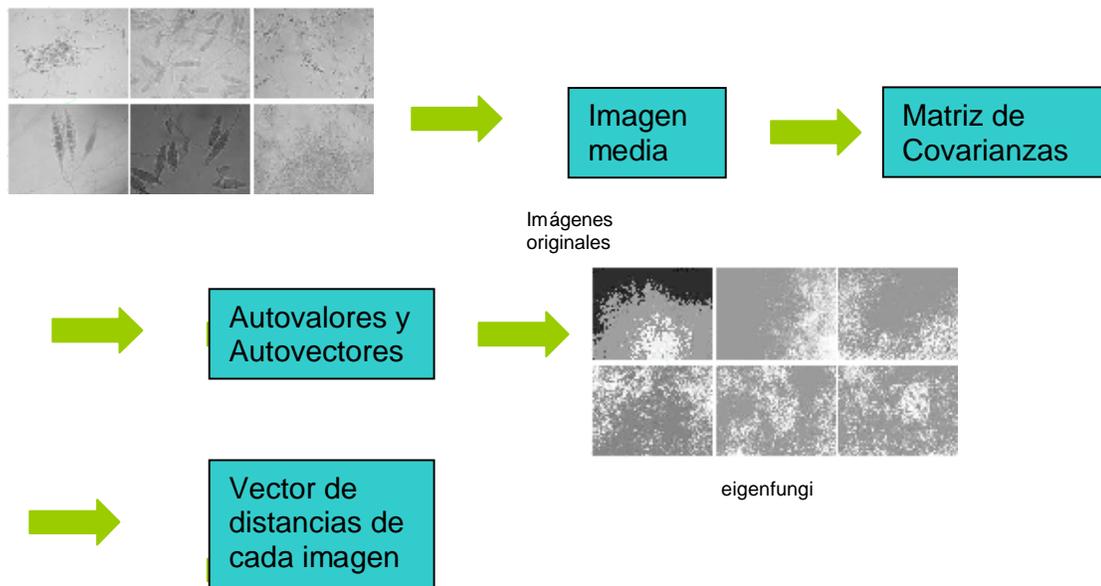
La diferencia principal aparece al momento de comparar las distancias de las imágenes: en lugar de comparar con el *vector de clase* de cada especie, se compara con cada imagen del conjunto de entrenamiento.

Esto da mejores resultados, debido a que hay especies muy similares y existen detalles en las micro o macronidias, tales como tabiques internos que son difíciles de distinguir si se utiliza el promedio.

8. Cuando se tiene una nueva imagen, se le resta la media y se halla su *vector de distancias*

9. Finalmente, se compara el nuevo vector de distancias con el vector de distancias de cada imagen original

Veamos ahora un resumen del método



A cada imagen a reconocer se le calcula su vector de distancias y se compara con cada vector de las imágenes originales. El vector más cercano indica a qué individuo pertenece la nueva imagen.

Parte V: Pruebas experimentales

9. Características de las pruebas

Para la elaboración y validación de la metodología, se estudiaron imágenes de hongos microscópicos de las seis especies principales de dermatofitos, obtenidas de muestras provistas por el Departamento de Micología del Instituto Nacional de Enfermedades Infecciosas (INEI), ANLIS "Carlos G. Malbrán".

Fueron tomadas con un aumento de 400x y originalmente medían 1600x1200 píxeles. Luego de varias pruebas, se determinó que el tamaño de las imágenes no influía en los resultados por lo que se decidió disminuirlas a 160x120 píxeles.

También se pasaron a escalas de grises, mediante la rutina específica provista por Matlab.

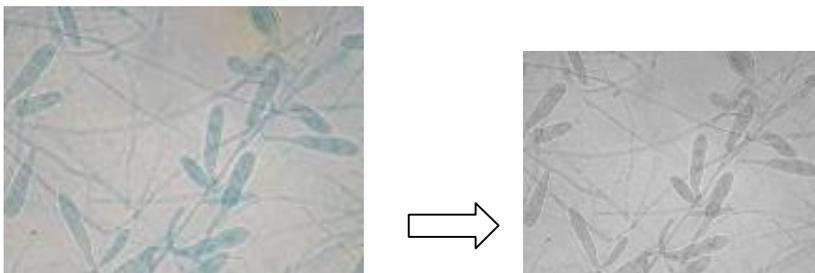


Fig. 9.1 – Ejemplo de cambio de tamaño y paso de color a tonos de grises

Las imágenes identifican micro y macro conidias, e hifas, según la especie.

Para las primeras pruebas (una muestra) se utilizaron 6 imágenes de entrenamiento y 6 imágenes de prueba por cada especie por cada muestra. Haciendo un total de 36 imágenes de entrenamiento y 36 imágenes de prueba por cada muestra.

9.1. Software desarrollado para las pruebas

Se desarrolló un software para realizar las pruebas a nivel binarias y totales. Este programa se realizó en Matlab, utilizándose rutinas obtenidas de la Universidad Drexel Philadelphia, USA [DRE07] (*Anexo B – Software Desarrollado*).

9.2. Tipos de pruebas

Se realizaron dos tipos de pruebas:

- Pruebas binarias

Se dispusieron los objetos o especies de a pares, entrenando y reconociendo dos cada vez

Por ejemplo, *E. floccosum* versus *M. canis*

- Pruebas totales

Se entrenó y testeó con todos los objetos o especies a la vez.

Por ejemplo, se intenta que el sistema reconozca a cuál de las 6 especies pertenece una imagen

10. Primeras pruebas - *eigenimages*

Dado que las imágenes de hongos microscópicos difieren mucho de las imágenes de caras, primero se estudió el método de *eigenfaces* en imágenes de objetos cotidianos.

Éstas comparten con los rostros las características de:

- hay un único objeto en la imagen que debe ser analizado
- los objetos del fondo son eliminados
- algunos objetos en las imágenes podrían normalizarse, en cuanto a dimensiones de los mismos

En algunos trabajos como [VIC02] al aplicar el método de *eigenfaces* en imágenes de objetos cotidianos, denominan la técnica como *eigenimages*.

Las imágenes utilizadas para estas pruebas corresponden a la base de imágenes de objetos COIL-20 (Columbia Object Image Library). La librería COIL-20 está formada por 1440 imágenes de 20 objetos diferentes (72 imágenes por cada objeto rotados 360°).

Se seleccionaron 6 objetos y se tomaron 10 imágenes de cada uno para el entrenamiento y 10 imágenes de cada uno para el testeo. Haciendo un total de 60 imágenes de entrenamiento y 60 de prueba.

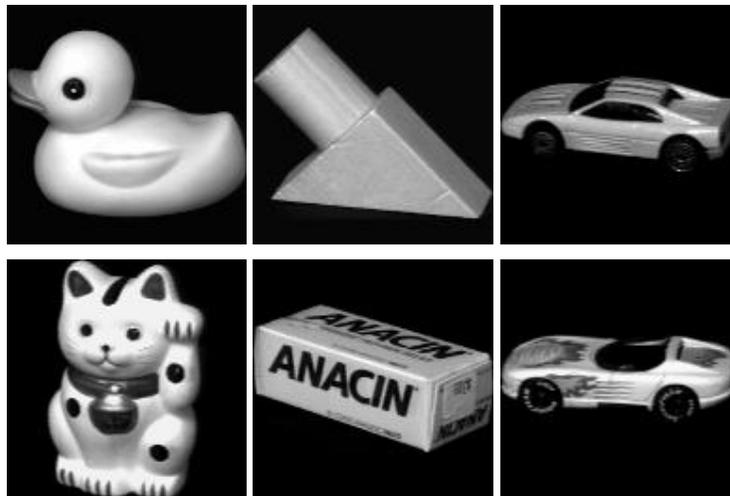


Fig. 10.1- Ejemplos de imágenes de entrenamiento

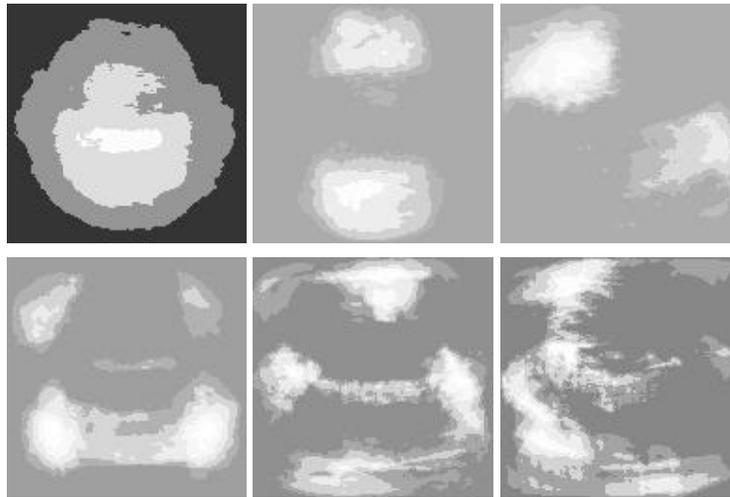


Fig. 10.2 - Ejemplos de eigenimages

Los porcentajes de acierto, con el método de *eigenfaces* original fueron los siguientes:

Binarias	Objeto1	Objeto2	Objeto3	Objeto4	Objeto5	Objeto6
Objeto1		80 %	85%	100%	90%	85%
Objeto2			100%	65%	85%	100%
Objeto3				100%	85%	70%
Objeto4					100%	100%
Objeto5						50%
Objeto6						
Totales	86,67%	<i>eigenfaces</i>				

Tabla 10.0.1 – Porcentajes de acierto eigenfaces con objetos

Los porcentajes de acierto son muy altos en general, por ejemplo de un 100% entre los objetos:

Objeto 1 vs Objeto 4

Objeto 3 vs Objeto 4

Objeto 2 vs Objeto 3

Objeto 4 vs Objeto 5

Objeto 2 vs Objeto 6

Objeto 4 vs Objeto 6

Pero existen pares con porcentajes bajos como el Objeto 2 vs el Objeto 4 con un 65% y el par Objeto 5 vs Objeto 6 directamente no fue reconocido (con un 50% de aciertos).

10.1. Aplicación de los *eigenfungi* a los objetos

Si se analiza la forma y características de los objetos estudiados, en algunos casos se puede ver cierta similitud a nivel superficial (forma, tamaño) entre los objetos, como por ejemplo entre el Objeto 3 y el Objeto 6, dado que los dos son autos. La diferencia entre ambos se ve a nivel de detalles *dentro* de la forma. Dado que lo que caracteriza al método *eigenfungi* es su capacidad de “respetar” los detalles internos de los objetos (como los tabiques de las macroconidias que hacen diferir a algunas especies), se probó el método con las imágenes de los objetos.

Se ve que efectivamente los porcentajes de acierto se incrementaron, llegando prácticamente a un 100% en todos los pares:

Binarias	Objeto1	Objeto2	Objeto3	Objeto4	Objeto5	Objeto6
Objeto1		90%	100%	100%	95%	100%
Objeto2			100%	95%	95%	100%
Objeto3				100%	90%	90%
Objeto4					100%	100%
Objeto5						90%
Objeto6						
Totales	86,67%	<i>eigenfungi</i>				

Tabla 10.1.1 – Porcentajes de acierto eigenfungi con objetos

El porcentaje de pruebas totales no se modificó, pero igual fue alto. Y el par Objeto 5 vs Objeto 6 subió de un 50% a un 90% de acierto.

11. Pruebas con dermatofitos

Una vez utilizado el método en imágenes de objetos, se realizaron pruebas con las imágenes de hongos microscópicos.

Se hicieron dos tipos de pruebas:

- binarias, probando el reconocimiento de las especies de a pares
- totales, probando el reconocimiento con todas las especies a la vez

11.1. Pruebas binarias: *E. floccosum* versus *M. canis*

Primero se tomaron las imágenes de los dermatofitos y se realizaron pruebas con el método *eigenfaces* original y luego se lo comparó con *eigenfungi*.

Por ejemplo, comenzamos con las imágenes de las especies *E. floccosum* y de *M. canis*.

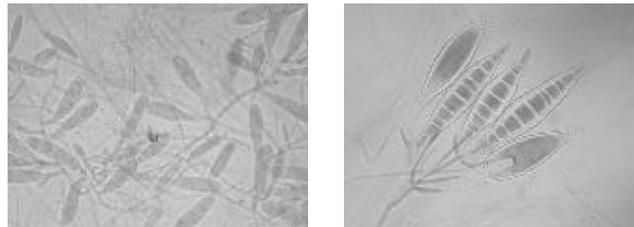


Fig. 11.1 - Ejemplos de imágenes de entrenamiento de la especie *E. floccosum* y de *M. canis*

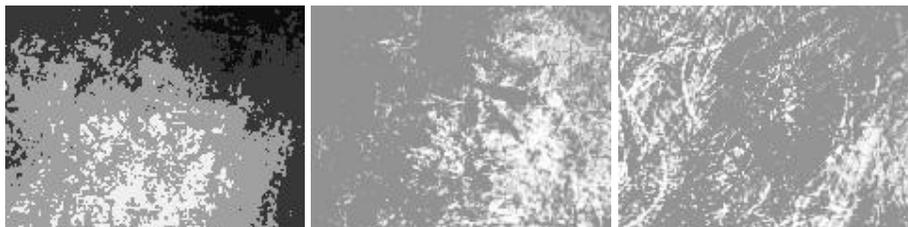


Fig. 11.2 - Ejemplos de eigenfungi de la prueba *E. floccosum* versus *M. canis*

El porcentaje obtenido con el método *eigenfaces* fue del 91,67% de acierto. O sea que se reconocieron casi todas las imágenes del conjunto de testeo, salvo una imagen.

Luego, se probó con el método de *eigenfungi*, y el porcentaje obtenido alcanzó un 100%.

11.2. Pruebas con todas las especies: *eigenfaces*

Se organizaron todas las especies por pares y se realizaron pruebas tanto con las *eigenfaces* como los *eigenfungi*. Luego se ingresaron todas las especies a la vez y se analizaron los porcentajes de acierto totales.

Probando con las *eigenfaces*, los porcentajes de acierto fueron los siguientes:

Binarias	floccosum	canis	gypseum	mentagro	rubrum	tonsurans
floccosum		91,67%	83,33%	75%	100%	91,67%
canis			100%	100%	100%	100%
gypseum				100%	50%	91,67%
mentagro					83,33%	75%
rubrum						83,33%
tonsurans						
Totales	80,56%	<i>eigenfaces</i>				

Tabla 11.2.1 – Porcentajes de acierto eigenfaces con dermatofitos

Se obtuvieron varios aciertos del 100% de reconocimiento de las imágenes, por ejemplo:

- *E. floccosum* vs *T. rubrum*
- *M. canis* con el resto de las especies (con *E. floccosum* 91,67%)
- *M. gypseum* vs *T. mentagrophytes*

En general los porcentajes de acierto fueron bastantes altos. En el caso de *M. gypseum* vs *T. rubrum*, sin embargo, no hubo reconocimiento entre las especies, dado que el porcentaje fue del 50%.

Si se ingresan todas las especies a la vez, el porcentaje es bastante alto, del 80,56%.

11.3. Pruebas con todas las especies: *eigenfungi*

Con los *eigenfungi*, vemos que los porcentajes de acierto en general se incrementan:

Binarias	floccosum	canis	gypseum	mentagro	rubrum	tonsurans
floccosum		100%	83,33%	100%	83,33%	50%
canis			100%	100%	100%	91,67%

gypseum				91,67%	100%	100%
mentagro					91,67%	91,67%
rubrum						91,67%
tonsurans						
Totales	80,56%	<i>eigenfungi</i>				

Tabla 11.3.1 – Porcentajes de acierto eigenfungi con dermatofitos

Los porcentajes de acierto se incrementaron con respecto a los resultados anteriores en 7 de los pares:

- *E. floccosum* vs *M. canis* de 91,67% a 100%
- *E. floccosum* vs *T. mentagrophytes* de 75% a 100%
- *M. gypseum* vs *T. rubrum* de 50% a 100%
- *M. gypseum* vs *T. tonsurans* de 91,67% a 100%
- *T. mentagrophytes* vs *T. rubrum* de 83,33% a 91,67%
- *T. mentagrophytes* vs *T. tonsurans* de 75% a 91,67%
- *T. rubrum* vs *T. tonsurans* de 83,33% a 91,67%

Hubo una pequeña reducción en el caso de *E. floccosum* vs *T. rubrum* de 100% a 83,33%, pero el porcentaje igualmente siguió siendo alto.

Si bien ahora se reconoce el par *M. gypseum* vs *T. rubrum* (y con un 100%), se detecta que hay un par no reconocido: *E. floccosum* vs *T. tonsurans* con un 50% de acierto (que sí se reconocía en el caso de las *eigenfaces*).

El porcentaje total siguió siendo alto (de un 80,56%) aunque no se modificó entre la aplicación de cada método.

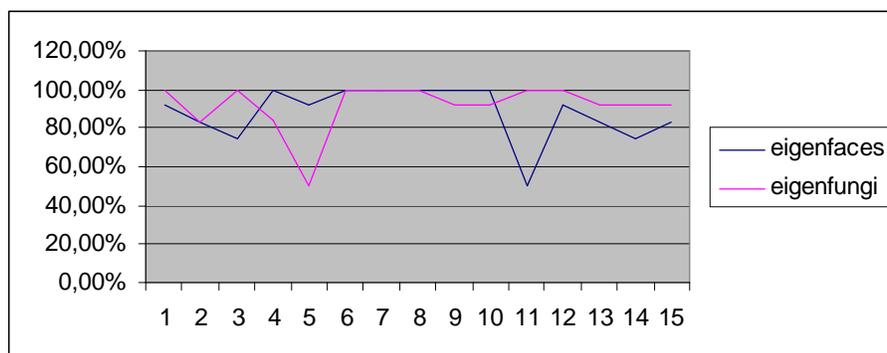


Fig. 11.1 – Comparación de porcentajes de acierto entre los métodos eigenfaces y eigenfungi

Posteriormente a estas pruebas, la idea fue buscar un preprocesamiento, que combinado con el método de *eigenfungi*, incrementara los porcentajes de acierto y además permitiera el reconocimiento de *todos* los pares de especies.

11.4. Pruebas con preprocesamientos

Tipos de preprocesamiento estudiados

A fin de conseguir un incremento en la exactitud de la clasificación de imágenes del método *eigenfungi*, se combinó con una serie de preprocesamientos. Para transformar las imágenes según los preprocesamientos correspondientes, fue utilizado el programa ImageJ [IMA07] (*Anexo A – Software utilizado*).

a) Detección de contornos

Los puntos de *borde* son píxeles alrededor de los cuales la imagen presenta una brusca variación en los niveles de gris. El objetivo de la *detección de contornos* o *extracción de bordes* es localizar en una imagen los bordes más probables generados por los elementos de la escena.

Los bordes son cadenas conectadas de puntos que conforman fragmentos de contorno. Son generados por los objetos sólidos de la escena, las marcas en las superficies, las sombras, etc. La filosofía básica de muchos de estos algoritmos se basa en el concepto de cómputo de derivadas locales (primera y segunda).

La primera derivada es cero en todas las regiones de intensidad constante y tiene un valor constante en toda la transición de intensidad. La segunda derivada, en cambio, es cero en todos los puntos, excepto en el comienzo y el final de una transición de intensidad. Por tanto, un cambio de intensidad se manifiesta como un cambio brusco en la primera derivada y presenta un paso por cero, es decir se produce un cambio de signo en su valor en la segunda derivada.



Fig. 11.2 – Ejemplo de detección de contornos en una imagen de *M. canis*

b) Imágenes binarias

La *binarización* de imágenes consiste en determinar un umbral de luminosidad px_k y entonces, los píxeles cuyo valor excedan px_k se transforman en 255 y los menores en 0.

$$B(px_i) =$$

si $px_i > px_k$ entonces $px_i = 255$

si $px_i \leq px_k$ entonces $px_i = 0$



Fig. 11.3 – Ejemplo de imagen de *E. floccosum* luego de aplicarle la binarización

c) Corrección de histograma

Una de las técnicas más utilizadas para la mejora del contraste de la imagen original es la de igualación o ecualización de histogramas. Se trata de una técnica que realza la imagen original mediante una determinada transformación o modificación del histograma que expande la distribución de los niveles de gris.

Dicha expansión debe ser lo más suave posible en el sentido de que idealmente debería haber el mismo número de píxeles por niveles de gris. Dado que el número de píxeles en una imagen de dimensión $N \times M$ es precisamente este producto y el número de niveles de gris sobre el cual se va a realizar la expansión es N_g , un histograma ideal sería plano, con el mismo número de píxeles en cada nivel de gris, es decir,

$$\text{Número ideal de píxeles en cada nivel de gris} = N \times M / N_g$$



Fig. 11.4 – Ejemplo de imagen de *T. tonsurans* con corrección de histograma

d) Suavizado de bordes

Las operaciones de suavizado son útiles para reducir el ruido y otros efectos no deseados que pueden estar presentes en una imagen digital como resultado del muestreo, cuantización y transmisión, o bien por perturbaciones en el sistema tales como partículas de polvo en el sistema óptico.

El caso de suavizado por el promedio del entorno de vecindad, es una técnica directa en el dominio espacial. Dada una imagen $g(i, j)$ se obtiene una imagen suavizada $f(i, j)$ cuya intensidad para cada punto (i, j) se obtiene promediando los valores de intensidad de los píxeles de g incluidos en el entorno de vecindad predefinido de (i, j) . Es decir, la operación puede sintetizarse en la siguiente expresión

$$f(i, j) = 1/P \sum g(m, n) \quad \text{con } (m, n) \in S$$

para todos los píxeles (i, j) de $g(I, j)$. S es el conjunto de coordenadas de los puntos situados en el entorno de vecindad de (i, j) incluido el propio (i, j) y P es el número total de puntos del entorno de vecindad.

Por ejemplo, si el entorno de vecindad de S es una ventana de dimensión 3×3 la operación realizada sería la misma que si la imagen hubiese sido filtrada con el núcleo del filtro pasa bajo siguiente,

$$h = \begin{vmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{vmatrix} \quad 1/9$$



Fig. 11.5 – Ejemplo de imagen de *M. canis* con suavizado de bordes

e) Transformada de Fourier

Una imagen con alta *frecuencia espacial* cambia periódicamente el valor de los niveles de intensidad o niveles de gris en un intervalo espacial pequeño, o lo que es lo mismo, en distancias pequeñas de la imagen. Por tanto, los niveles de gris cambian de forma más o menos abrupta de un nivel de gris a otro. Por el contrario, las bajas frecuencias espaciales corresponden a cambios más lentos en la variación de los niveles de gris donde los cambios ocurren gradualmente de una posición a otra de la imagen.

La formalización de la *Transformada de Fourier* pasa por el correspondiente *análisis de Fourier*. Esta transformada para una imagen $f(x, y)$ está definida por la siguiente expresión

$$F(u, v) = T \{ f(x, y) \} = \iint f(x, y) \exp(-2\pi i (ux + vy)) dx dy$$

De manera análoga, la *Transformada inversa de Fourier* es

$$f(x, y) = T \{ F(u, v) \} = \iint F(u, v) \exp(2\pi i (ux + vy)) du dv$$

Para cada par de valores de las frecuencias espaciales u y v se tiene una exponencial en la suma generalizada, dicha exponencial está multiplicada por el coeficiente de peso $F(u, v)$. Por tanto, puede verse como los coeficientes de peso de la expansión de la función de intensidad f en una suma de exponenciales.

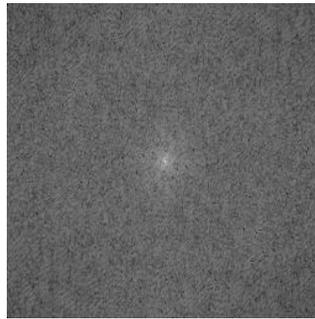


Fig. 11.6 – Ejemplo de imagen de *E. floccosum* transformada por la Transformada de Fourier

f) Desenfoque Gaussiano

En el *desenfoque* o *suavizado Gaussiano* el núcleo para la convolución es una función Gaussiana 2-D, de media 0 y desviación estándar σ .

$$G(i, j) = e^{-(i^2 + j^2) / 2 \sigma^2}$$

El suavizado Gaussiano puede implementarse eficientemente gracias al hecho de que su núcleo es separable. Esto significa que la convolución de una imagen I con el núcleo Gaussiano 2-D se puede realizar por la convolución de todas las filas y luego todas las columnas con una Gaussiana 1-D con idéntica σ , lo cual disminuye el costo computacional.

Gracias a la separabilidad del núcleo Gaussiano, se pueden considerar sólo máscaras 1-D. Para construir una máscara discreta Gaussiana, se muestrea una Gaussiana continua.

Para lo cual se debe determinar el ancho de la máscara discreta w , dado el núcleo Gaussiano que se quiere utilizar, o a la inversa, el valor de σ de la Gaussiana continua dado el ancho w de la máscara deseada.



Fig. 11.7 – Ejemplo de imagen de *M. canis* con desenfoque Gaussiano

11.5. Resultados obtenidos

Se combinó el método de *eigenfungi* con los distintos preprocesamientos a fin de incrementar los porcentajes de acierto obtenidos con las imágenes originales.

Luego se compararon los resultados obtenidos con el método *eigenfungi* puro (Tabla 11.3.1) con cada combinación.

a) Detección de contornos

La detección de contornos empeoró considerablemente los resultados.

Binarias	floccosum	canis	gypseum	mentagro	rubrum	tonsurans
floccosum		50%	50%	50%	50%	50%
canis			75%	50%	41,67%	58,33%
gypseum				50%	83,33%	91,67%
mentagro					50%	50%
rubrum						50%
tonsurans						
Totales	33,33%	<i>eigenfungi</i>	<i>contornos</i>			

Tabla 11.5.1 – Porcentajes de acierto *eigenfungi* combinado con detección de contornos

Sólo 3 de todos los pares analizados registraron un valor de porcentaje alto:

- *M. canis* vs *M. gypseum* 75%
- *M. gypseum* vs *T. rubrum* 83,33%
- *M. gypseum* vs *T. tonsurans* 91,67%

El resto de las pruebas dio 50%, o sea que no hubo reconocimiento de las especies y en el caso de *M. canis* vs *T. rubrum*, el porcentaje fue aún más bajo, de un 41,67%.

Las pruebas totales también fueron muy bajas con un 33,33% (contra un 80,50% del método puro).

b) Transformada de Fourier

Binarias	floccosum	canis	gypseum	mentagro	rubrum	tonsurans
Floccosum		75%	83,33%	100%	100%	66,67%
Canis			66,67%	91,67%	100%	66,67%
Gypseum				58,33%	58,33%	83,33%
Mentagro					66,67%	100%
Rubrum						100%
Tonsurans						
Totales	30,55%	<i>eigenfungi</i>	<i>fft</i>			

Tabla 11.5.2 – Porcentajes de acierto eigenfungi combinado con transformada de Fourier

En el caso de la Transformada de Fourier, se ve que algunos reconocimientos fueron muy altos, como *E. floccosum* vs *T. mentagrophytes* y *E. floccosum* vs *T. rubrum* con un 100%. Pero hubo varios pares con porcentajes bajos o que no fueron reconocidos, como *M. gypseum* vs *T. mentagrophytes* y *M. gypseum* vs *T. rubrum* con un porcentaje de acierto del 58,33%.

Las pruebas totales también fueron muy bajas con un 30,55% (aún más bajo que con detección de contornos) contra un 80,50% del método puro.

c) Imágenes binarizadas

Con las imágenes binarizadas, los porcentajes son similares a los obtenidos con la Transformada de Fourier.

Binarias	floccosum	Canis	gypseum	mentagro	rubrum	tonsurans
floccosum		100%	100%	100%	75%	66,67%
canis			91,67%	100%	100%	66,67%
gypseum				83,33%	58,33%	58,33%
mentagro					75%	75%
rubrum						91,67%
tonsurans						
Totales	61,11%	<i>eigenfungi</i>	<i>binariz</i>			

Tabla 11.5.3 – Porcentajes de acierto eigenfungi combinado con imágenes binarizadas

Hubo varios pares con 100% de acierto como *E. floccosum* vs *M. gypseum* y *E. floccosum* vs *T. mentagrophytes*, sin embargo otros no fueron reconocidos, como *M. gypseum* vs *T. rubrum* o *M. gypseum* vs *T. tonsurans* con un 58,33% de acierto.

El porcentaje total fue de un 61,11%. Aunque fue mayor que los casos de contornos y Fourier, fue mucho menor que el 80,56% del original.

11.6. Preprocesamiento seleccionado

Finalmente, el preprocesamiento que combinado con el método de *eigenfungi* fue el que mejor resultados produjo, fue el *suavizado de bordes con una posterior corrección del histograma*.

Los porcentajes obtenidos fueron los siguientes:

Binarias	floccosum	canis	gypseum	mentagro	rubrum	tonsurans
floccosum		91,67%	91,67%	100%	100%	83,33%
canis			100%	100%	100%	91,67%
gypseum				100%	100%	91,67%
mentagro					100%	100%
rubrum						100%
tonsurans						
Totales	86,11%	<i>eigenfungi</i>	<i>suavizado</i>	<i>histograma</i>		

Tabla 11.6.1 – Porcentajes de acierto eigenfungi combinado con suavizado de bordes y corrección de histograma

Comparando con el método *eigenfungi* puro (Tabla 11.3.1), los porcentajes de acierto se incrementaron y todos los pares de especies fueron reconocidos. Se obtuvo prácticamente un 100% de acierto en todas las pruebas (solamente un poco menor en el caso de *E. floccosum* vs *T. tonsurans* con un 83,33%).

También subió el porcentaje a nivel total, de un 80,56% a un 86,11%.

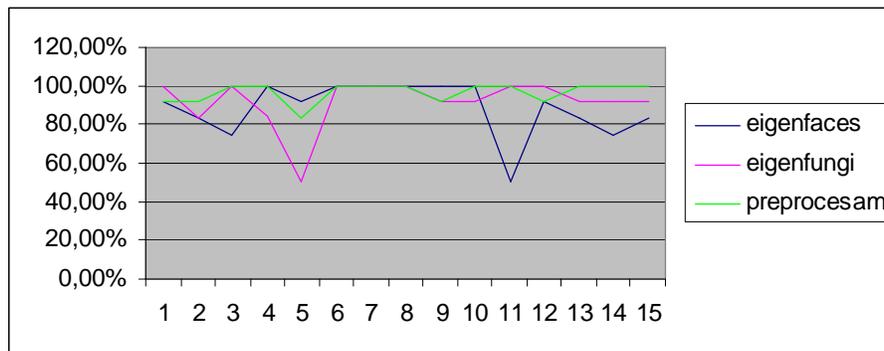


Fig. 11.10 – Comparación de porcentajes de acierto entre los métodos eigenfaces, eigenfungi y eigenfungi con preprocesamiento de suavizado de bordes y ecualización de histograma.

11.7. Otros preprocesamientos con porcentajes altos

Hubo otros dos preprocesamientos que también consiguieron altos porcentajes al combinarlos con el método de *eigenfungi*. Pero en ambos casos existieron pares que no fueron reconocidos, por lo que se prefirió el preprocesamiento de suavizado de bordes con ecualización de histograma.

Suavizado de bordes

Por ejemplo, un suavizado de bordes (sin la corrección del histograma), registró los siguientes resultados:

Binarias	floccosum	canis	gypseum	mentagro	rubrum	tonsurans
Floccosum		100%	83,33%	100%	100%	50%
Canis			100%	100%	100%	91,67%
Gypseum				100%	100%	66,67%
Mentagro					100%	100%
Rubrum						91,67%
Tonsurans						
Totales	91,67%	<i>eigenfungi</i>	<i>suavizado</i>			

Tabla 11.7.1 – Porcentajes de acierto eigenfungi combinado con suavizado de bordes

Los valores fueron altos, en forma similar al preprocesamiento de suavizado cuando se incluye la ecualización del histograma. Incluso las pruebas totales fueron mejores (de un 86,11% se eleva a un 91,67%). Pero no se reconocen todos los pares de especies, sino que falla en el par *E. floccosum* versus *T. tonsurans* con un 50% de acierto.

Desenfoque gaussiano con ecuación de histograma

Otro de los preprocesamientos que dio altos resultados fue el desenfoque gaussiano con ecuación del histograma.

Binarias	floccosum	canis	gypseum	mentagro	rubrum	tonsurans
Floccosum		100%	83,33%	100%	100%	53,33%
Canis			100%	100%	100%	91,67%
Gypseum				100%	100%	75%
Mentagro					100%	100%
Rubrum						91,67%
Tonsurans						
Totales	91,67%	<i>eigenfungi</i>	<i>desenfoque</i>	<i>histograma</i>		

Tabla 11.7.2 – Porcentajes de acierto eigenfungi combinado con desenfoque gaussiano y corrección de histograma

Si bien las pruebas fueron muy buenas y el total se incrementó de un 86,11% a un 91,67%, se ve que el par *E. floccosum* versus *T. tonsurans* no fue reconocido y además hubo una caída en el reconocimiento de *M. gypseum* versus *T. tonsurans* a un 75%.

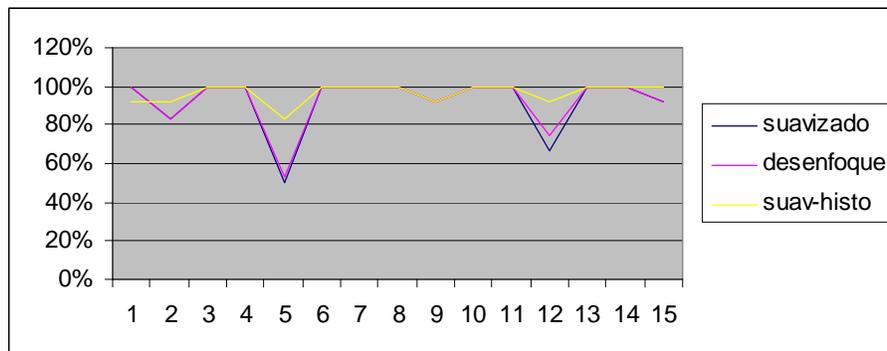


Fig. 11.12– Comparación de porcentajes de acierto entre el preprocesamiento de suavizado de bordes y ecuación de histograma, con suavizado solo y desenfoque gaussiano.

12. Pruebas de Robustez

Para verificar la robustez del método, se degradaron las imágenes con dos tipos de ruido y luego se les aplicó el método de reconocimiento *eigenfungi* combinado con el preprocesamiento de suavizado de bordes y corrección de histograma. Para la aplicación de ruido a las imágenes, se utilizó el programa ImageJ [IMA07] (*Anexo A – Software utilizado*).

12.1. Ruido gaussiano

El ruido es una información no deseada que contamina la imagen. Aparece en imágenes procedente de una gran variedad de fuentes. En cada paso de proceso de adquisición de una imagen digital, se presentan fluctuaciones originadas por fenómenos naturales que añaden un valor aleatorio al valor exacto de la intensidad para un determinado píxel.

En las imágenes típicas, el ruido puede modelarse como una distribución *gaussiana* (normal), *uniforme* o “*sal y pimienta*” (impulso).

Para esta prueba se aplicó a las imágenes un *ruido gaussiano* de media 0 y σ 25.

$$H_s = \frac{1e^{-(s-m)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}$$

Con g nivel de gris del ruido, m valor medio y σ desviación estándar.



Fig. 12.1 – Imagen de *M. canis* a la cual se le aplicó ruido gaussiano

Luego de alterada, a cada imagen se le aplicó el preprocesamiento de suavizado de bordes con una corrección del histograma, que había registrado los mejores resultados (Tabla 11.6.1).



Fig. 12.2 – Imagen de *M. canis* a la cual se le aplicó ruido gaussiano y luego suavizado de bordes con corrección de histograma

Los porcentajes de acierto, con el método de *eigenfungi* combinado con el preprocesamiento seleccionado fueron los siguientes:

Binarias	floccosum	canis	gypseum	mentagro	Rubrum	tonsurans
Floccosum		100%	91,67%	91,67%	91,67%	83,33%
Canis			100%	100%	100%	100%
Gypseum				100%	100%	83,33%
Mentagro					100%	100%
Rubrum						100%
Tonsurans						
Totales	88,89%	<i>eigenfungi</i>	<i>Ruido</i>	<i>suavizado</i>	<i>histograma</i>	

Tabla 12.1.1 – Porcentajes de acierto eigenfungi combinado con suavizado de bordes y corrección de histograma, previa degradación con ruido

Se observa que a pesar de haber degradado las imágenes con el ruido, igualmente los porcentajes fueron muy altos. Todos los pares fueron reconocidos y casi el 70% de las pruebas dio un porcentaje de acierto del 100%.

También fue alto el porcentaje de las pruebas totales, con un porcentaje de acierto del 88,89%.

12.2. Ruido “sal y pimienta”

En el ruido impulsivo o de “sal y pimienta” hay sólo dos posibles valores a y b y la probabilidad de cada uno es típicamente menor que el 0,1 del total del histograma de la imagen, con valores mayores el ruido puede dominar la imagen. Para una imagen de 256 niveles de gris, el valor típico de la pimienta es 0 y de la sal 255.

Puede generarse mediante la función dada por la ecuación

$$f_{sp}(x, y) = \begin{cases} \text{si } r < 1 & \text{entonces } f(x, y) \\ \text{si } r \cdot 1 & \text{entonces } g_{\min} + s (g_{\max} - g_{\min}) \end{cases}$$

Este tipo de ruido se presenta en las imágenes como un rociado de puntos luminosos y oscuros y puede ser causado por errores de transmisión o ruido externo que contamina la conversión analógica digital.

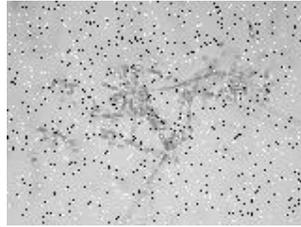


Fig. 12.3 – Imagen de *T. tonsurans* a la cual se le aplicó ruido “sal y pimienta”

Se aplicó a las imágenes alteradas el preprocesamiento de suavizado de bordes con una corrección del histograma y se probó el reconocimiento.

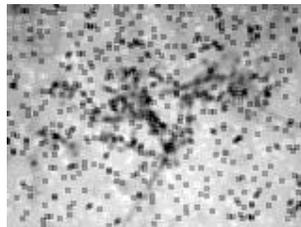


Fig. 12.4 – Imagen de *T. tonsurans* a la cual se le aplicó ruido “sal y pimienta” y luego suavizado de bordes con corrección de histograma

Los porcentajes de acierto de las imágenes degradadas, con el método de *eigenfungi* con el preprocesamiento fueron los siguientes:

Binarias	Floccosum	canis	gypseum	mentagro	rubrum	tonsurans
floccosum		100%	83,33%	100%	100%	91,67%
canis			91,67%	100%	100%	91,67%
gypseum				100%	75%	100%
mentagro					100%	100%
rubrum						83,33%
tonsurans						
Totales	86,11%	<i>eigenfungi</i>	<i>ruido syp</i>	<i>suavizado</i>	<i>histograma</i>	

Tabla 12.2.1 – Porcentajes de acierto eigenfungi combinado con suavizado de bordes y corrección de histograma, previa degradación con ruido “sal y pimienta”

Al haber alterado las imágenes con el ruido “sal y pimienta” se puede observar la robustez del método, dado que casi todas las pruebas dieron un 100% de acierto.

Solamente se detectó un porcentaje no tan alto entre *M. gypseum* y *T. rubrum* a un 75%, pero las pruebas fueron muy buenas, incluyendo las pruebas totales con un 86,11% de acierto del reconocimiento.

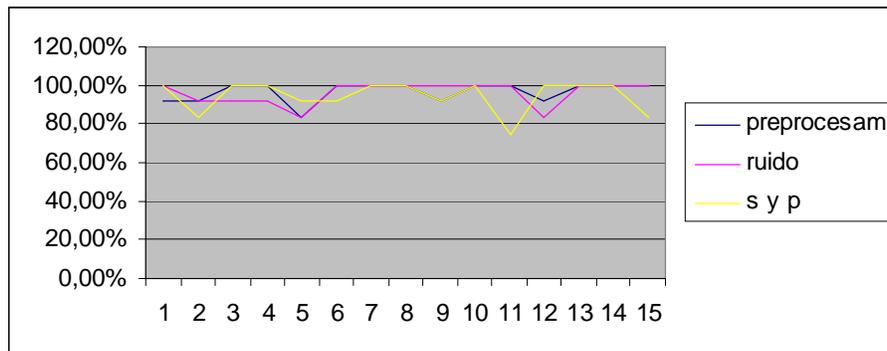


Fig. 12.5 – Comparación de porcentajes de acierto entre imágenes sin degradar, con ruido y con ruido sal y pimienta

13. Pruebas con dos muestras

Las personas, pueden presentar cambios temporales como peinados o anteojos, pero es siempre el mismo individuo el que se intenta reconocer.

En el caso de los hongos microscópicos, en cambio, el reconocimiento es a nivel de especies, por lo que podrían existir diferencias entre las imágenes si fueron tomadas a partir de diferentes muestras.

Es por esto que se realizaron pruebas con dos muestras diferentes de cada especie, duplicando la cantidad de imágenes tanto de entrenamiento como de testeo. Es decir, que considerando 6 imágenes de entrenamiento y 6 de testeo por especie y por muestra, se utilizaron 72 imágenes de entrenamiento y 72 para las pruebas.

La idea fue verificar si el método podía “aprender” más información al agregar imágenes que si bien pertenecen a la misma especie, tienen ciertas diferencias que podrían “confundir” al sistema.

13.1. Pruebas con todas las especies: *eigenfaces*

Se organizaron las especies por pares y se realizaron pruebas tanto con las *eigenfaces* como los *eigenfungi*. Luego se ingresaron todas las especies a la vez y se analizaron los porcentajes de acierto en el reconocimiento de cada una.

Probando con las *eigenfaces*, los resultados fueron los siguientes:

Binarias	floccosum	canis	gypseum	mentagro	rubrum	tonsurans
floccosum		62,5%	50%	79,17%	83,33%	70,83%
canis			79,17%	62,5%	66,67%	54,17%
gypseum				87,5%	50%	87,5%
mentagro					70,83%	66,67%
rubrum						75%
tonsurans						
Totales	<50%	<i>eigenfaces</i>	<i>2muestr</i>			

Tabla 13.1.1 – Porcentajes de acierto eigenfaces con dos muestras

Estas pruebas son complicadas, pues podrían hacer que el sistema “desaprenda” por las diferencias entre muestras. Sin embargo, a pesar de ello, se obtuvieron algunos porcentajes muy altos en varios pares, por ejemplo:

- *E. floccosum* vs *T. rubrum* con 83,33%

- *M. gypseum* vs *T. mentagrophytes* con 87,5%
- *M. gypseum* vs *T. tonsurans* con 87,5%

Por otro lado, hay algunos pares que no fueron reconocidos (con un 50% de acierto):

- *E. floccosum* vs *M. gypseum*
- *M. gypseum* vs *T. rubrum*

Con respecto a las pruebas totales, el porcentaje de reconocimiento fue muy bajo, menor al 50%.

13.2. Pruebas con todas las especies: *eigenfungi*

Con el método de *eigenfungi* los resultados obtenidos con dos muestras fueron los siguientes:

Binarias	floccosum	canis	gypseum	mentagro	rubrum	tonsurans
floccosum		79,17%	54,17%	95,83%	70,83%	50%
canis			79,17%	75%	79,17%	62,5%
gypseum				95,83%	95,83%	87,5%
mentagro					95,83%	70,83%
rubrum						87,5%
tonsurans						
Totales	<50%	<i>eigenfungi</i>	2muestr			

Tabla 13.2.1 – Porcentajes de acierto eigenfungi con dos muestras

Comparando los resultados con los obtenidos al aplicar *eigenfaces*, se registró un incremento en casi todas las pruebas, por ejemplo:

- *E. floccosum* vs *M. canis* pasó del 62,5% a casi el 80%
- *E. floccosum* vs *T. mentagrophytes* pasó del 79% a casi el 100%
- *T. mentagrophytes* vs *T. rubrum* pasó del 70% a casi el 100%
- *T. rubrum* vs *T. tonsurans* subió de un 75% a un 87,5%

El par *M. gypseum* vs *T. rubrum* que en el caso de las *eigenfaces* no había sido reconocido, con los *eigenfungi* mostró un 95,83% de acierto.

Por otro lado, el par *E. floccosum* vs *M. gypseum* siguió sin ser reconocido (subió apenas de un 50% a un 54,17%) y se agregó el par *E. floccosum* vs *T. tonsurans* con un 50% (cuando con las *eigenfaces* había llegado a un 70,83%).

Las pruebas totales continuaron siendo menores al 50%.

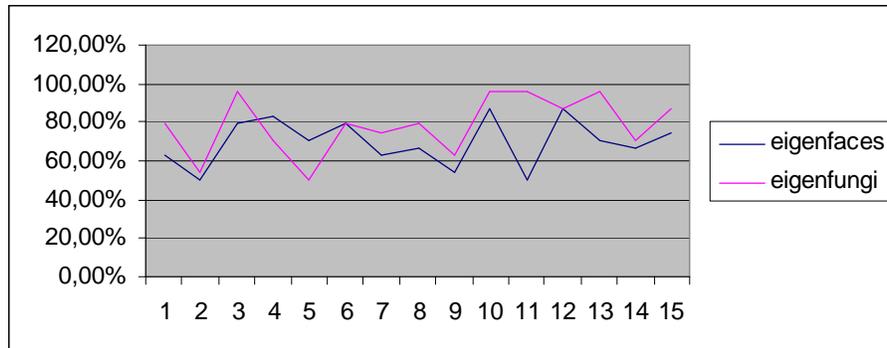


Fig. 13.1 – Comparación de porcentajes de acierto entre los métodos eigenfaces y eigenfungi

A fin de mejorar estos resultados, y de la misma manera que se hizo en el caso de una muestra, se repitieron estas pruebas combinando el método con distintos preprocesamientos.

13.3. Pruebas con preprocesamientos

a) Detección de contornos

Se extrajeron los contornos de cada imagen y se aplicó el método de *eigenfungi*. Los porcentajes de acierto fueron los siguientes:

Binarias	floccosum	canis	gypseum	mentagro	rubrum	tonsurans
floccosum		50%	50%	45,83%	50%	50%
Canis			54,17%	50%	41,67%	54,17%
Gypseum				50%	58,33%	70,83%
mentagro					50%	50%
Rubrum						54,17%
tonsurans						
Totales	<50%	<i>eigenfungi</i>	<i>contornos</i>		<i>2mustr</i>	

Tabla 13.3.1 – Porcentajes de acierto eigenfungi con dos muestras combinado con detección de contornos

Tal como ocurrió en el caso de una muestra, se ve que los porcentajes de acierto no mejoraron sino que bajaron significativamente. Salvo el caso del par *M. gypseum* vs *T. tonsurans* con un 70,83%, el resto de los pares en general no fue reconocido, con un 50% o menos de acierto.

Las pruebas globales, también fueron muy bajas con un porcentaje menor al 50%.

b) Transformada de Fourier

Los porcentajes de acierto, luego de aplicada la Transformada de Fourier a las imágenes, fueron los siguientes:

Binarias	floccosum	canis	gypseum	mentagro	rubrum	tonsurans
floccosum		66,67%	58,33%	91,66%	91,66%	66,67%
Canis			75%	79,17%	83,33%	58,33%
Gypseum				58,33%	50%	58,33%
mentagro					79,17%	95,83%
Rubrum						83,33%
tonsurans						
Totales	< 50%	<i>eigenfungi</i>	<i>Fft</i>		<i>2mustr</i>	

Tabla 13.3.2 – Porcentajes de acierto eigenfungi con dos muestras combinado con transformada de Fourier

Si se comparan los resultados con los porcentajes obtenidos con *eigenfungi* sin preprocesamientos (Tabla 13.2.1), se ve que algunos se incrementaron, pero muchos se redujeron.

Por ejemplo, se ve un aumento en los casos de:

- *M. canis* vs *T. mentagrophytes* de un 75% a un 79,17%
- *M. canis* vs *T. rubrum* de un 79,17% a un 83,33%
- *T. mentagrophytes* vs *T. tonsurans* de un 70,83% a un 95,83%

Pero en general el resto de los porcentajes es más bajo que con el método puro y además varios pares no fueron reconocidos, como en los casos de:

- *E. floccosum* vs *M. gypseum* (58,33%)
- *M. canis* vs *T. tonsurans* (58,33%)
- *M. gypseum* vs *T. mentagrophytes* (58,33%)
- *M. gypseum* vs *T. rubrum* (50%).

Las pruebas totales continuaron registrando un porcentaje menor al 50%.

c) Imágenes binarizadas

El método aplicado a imágenes binarizadas, dio los siguientes resultados:

Binarias	floccosum	canis	gypseum	mentagro	rubrum	tonsurans
floccosum		75%	91,67%	91,67%	70,83%	62,5%
Canis			70,83%	83,33%	75%	50%
Gypseum				91,67%	58,33%	50%
mentagro					95,83%	75%
Rubrum						66,67%
tonsurans						
Totales	< 50%	<i>eigenfungi</i>	<i>binarias</i>		<i>2muestr</i>	

Tabla 13.3.3 – Porcentajes de acierto eigenfungi con dos muestras combinado con imágenes binarizadas

En este caso, se lograron bastantes pares con porcentajes altos, como *E. floccosum* vs *M. gypseum* y *E. floccosum* vs *T. mentagrophytes*, con un 91,67% de aciertos. Pero hubo 3 pares de especies no reconocidos:

- *M. canis* vs *T. tonsurans* con 50%
- *M. gypseum* vs *T. rubrum* con 58,33%
- *M. gypseum* vs *T. tonsurans* con 50%

Para las pruebas con todas las especies a la vez, no se consiguió un aumento del porcentaje de acierto.

d) Suavizado de bordes

Binarias	floccosum	canis	gypseum	mentagro	rubrum	tonsurans
floccosum		79,17%	87,5%	100%	87,5%	58,33%
Canis			75%	70,83%	75%	62,5%
Gypseum				95,83%	87,5%	70,83%
mentagro					100%	62,5%
Rubrum						95,83%
tonsurans						
Totales		<i>eigenfungi</i>	<i>suavizado</i>		<i>2muestr</i>	

Tabla 13.3.4 – Porcentajes de acierto eigenfungi con dos muestras combinado con suavizado de bordes

Aplicando un preprocesamiento de suavizado de bordes existieron dos pares con el 100% de acierto: *E. floccosum* vs *T. mentagrophytes* y *T. mentagrophytes* vs *T. rubrum*.

Además, todos los pares fueron reconocidos salvo *E. floccosum* vs *T. tonsurans* con el 58,33% de aciertos.

Comparando estos resultados con los obtenidos con el método *eigenfungi* puro (Tabla 13.2.1), hubo algunos porcentajes que se incrementaron, como *T. rubrum* vs *T. tonsurans* que subió de 87,5% a 95,83%, pero la mayoría fue similar y hasta hubo algunos porcentajes que resultaron menores, como *M. gypseum* vs *T. tonsurans* que pasó de 70,83% a 62,83% y *M. gypseum* vs *T. rubrum*, que se redujo de 95,83% a 87,5%.

13.4. Preprocesamiento seleccionado

El preprocesamiento que presentó mejores resultados, fue el de suavizado de bordes y corrección de histograma, al igual que en el caso de una muestra sola. Los porcentajes obtenidos fueron los siguientes:

Binarias	floccosum	canis	gypseum	mentagro	rubrum	tonsurans
floccosum		83,33%	83,33%	100%	87,5%	70,83%
Canis			70,83%	70,83%	75%	66,67%
Gypseum				95,83%	95,83%	75%
mentagro					100%	66,67%
Rubrum						100%
tonsurans						
Totales	55,56%	<i>eigenfungi</i>	<i>suavizado</i>	<i>Histograma</i>	<i>2muestr</i>	

Tabla 13.4.1 – Porcentajes de acierto eigenfungi con dos muestras combinado con suavizado de bordes y corrección histograma

En comparación con el método de *eigenfungi* puro (Tabla 13.2.1), 8 de los pares se vieron incrementados y el resto se mantuvo o bajó muy poco como *M. canis* vs *T. rubrum* que bajó de 79,17% a 75%.

Salvo *M. canis* vs *T. tonsurans* y *T. mentagrophytes* vs *T. tonsurans*, ambos con 66,67% de acierto, las demás pruebas tienen porcentajes mayores al 70% y hay 5 de ellos con 95,83% o 100%. Todos los pares fueron reconocidos.

El porcentaje total no se modificó significativamente, ya que superó el 50% pero igualmente fue muy bajo, con un 55,56%.

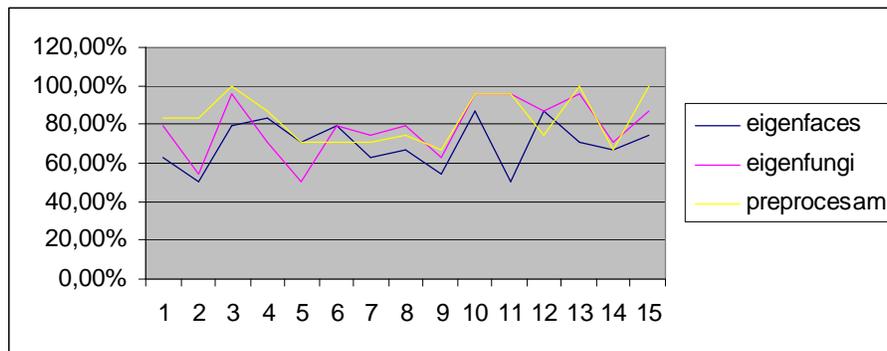


Fig. 13.2 – Comparación de porcentajes de acierto entre los métodos eigenfaces, eigenfungi y eigenfungi con preprocesamiento de suavizado de bordes y ecualización de histograma.

13.5. Otros procesamientos con valores altos

Probando con desenfoque gaussiano y corrección de histograma, los resultados fueron similares al caso de suavizado con corrección de histograma (Tabla 13.4.1), pero resultó menor en el reconocimiento del par *E. floccosum* versus *T. tonsurans* (de 70,83% a 62,5%).

Binarias	floccosum	canis	gypseum	mentagro	rubrum	tonsurans
Floccosum		79,17%	87,5%	100%	91,67%	62,5%
Canis			75%	70,83%	75%	66,67%
Gypseum				95,83%	87,5%	66,67%
mentagro					100%	62,5%
Rubrum						95,83%
tonsurans						
Totales	52,78%	<i>eigenfungi</i>	<i>desenfoque</i>	<i>histograma</i>	<i>2muestr</i>	

Tabla 13.5.1 – Porcentajes de acierto eigenfungi con dos muestras combinado con desenfoque gaussiano y corrección histograma

13.6. Resumen de preprocesamientos

Si se reunieran en una tabla los mejores resultados de cada par de especies obtenidos con cada tipo de reprocesamiento, se podrían ver los siguientes porcentajes:

Binarias	floccosum	canis	gypseum	mentagro	rubrum	tonsurans
Floccosum		83,33%	91,67%	100%	91,67%	70,83%
Canis			75%	83,33%	83,33%	66,67%

Gypseum				95,83%	95,83%	75%
mentagro					100%	95,83%
Rubrum						100%
tonsurans						
Totales		<i>eigenfungi</i>	<i>Varios</i>	resumen	<i>2muestr</i>	

Tabla 13.6.1 – Resumen de porcentajes de acierto eigenfungi con dos muestras combinado con varios preprocesamientos

Salvo 4 casos con porcentajes entre 66,67% y 75%, el resto es mayor al 83% llegando varios al 100%.

Para el caso de pruebas totales no se encontraron preprocesamientos que lograran mejoras sustanciales. Se consiguieron mejoras si no se utilizaban todas las especies a la vez, sino que se sacaba alguna y se hacían pruebas, por ejemplo 5 con especies juntas. Específicamente, las pruebas realizadas sin la especie *T. tonsurans*, llegaron a un porcentaje del 70,83%.

14. Errores de origen

Las imágenes analizadas para este estudio fueron utilizadas directamente desde su obtención. Es decir, que no hubo una preselección de las imágenes para descartar problemas o dejar sólo las mejores.

Si embargo, el método resultó robusto ante errores propios de las imágenes en el caso de una muestra, como por ejemplo:

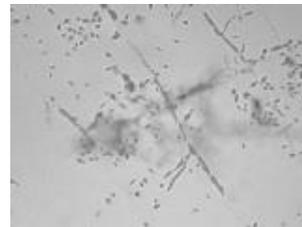
- Manchas
- imágenes borrosas
- elementos incompletos

En el caso de 2 muestras los porcentajes de reconocimiento no fueron tan altos como los obtenidos en las pruebas con sólo una muestra, en especial con las pruebas con todas las imágenes a la vez. En esto influye la calidad de las imágenes utilizadas, que reflejan muestras comunes, sin mejoramientos de los preparados, ni descartes de imágenes, ni recortes o identificación de elementos por parte de expertos humanos.

Como ejemplo, las siguientes imágenes fueron utilizadas a pesar de presentar errores de origen:



M. canis con mancha en la parte inferior izquierda de la imagen



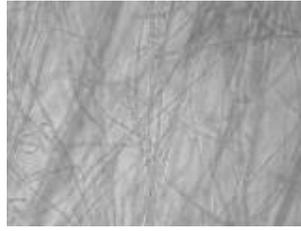
T. tonsurans con microconidias superpuestas



M. canis con macroconidias incompletas



Imagen de E. floccosum con mancha en la parte superior izquierda



M. canis con ausencia de conidias

Las manchas (dependiendo de la dimensión y tonalidad) pueden provocar falsos positivos al sistema, que las interpreta como elementos característico de la imagen. La ausencia de micro y macroconidias podría confundir también a un experto humano que podría no identificar la especie de una imagen que presenta sólo hifas.

Los métodos de reconocimiento automático resultan una herramienta muy útil a médicos e investigadores. Aunque no identifiquen exactamente la especie en todos los casos, ayudan al humano en la identificación, brindando pistas sobre las características de la muestra estudiada, especialmente en el caso de personas sin gran experiencia o especialización en el área.

15. Pruebas de predicción

Para estas pruebas se agregaron al conjunto de imágenes las correspondientes a una segunda muestra de dermatofitos, pero solamente al momento del testeo.

Considerando 6 imágenes de entrenamiento y 6 de testeo por especie y por muestra, se utilizaron 36 imágenes de entrenamiento y 72 imágenes de prueba.

La idea fue verificar si el método podía “predecir” más información que la suministrada al momento del entrenamiento.

15.1. Pruebas con todas las especies: *eigenfaces*

Se organizaron las especies por pares y se realizaron pruebas tanto con las *eigenfaces* como los *eigenfungi*. Luego se ingresaron todas las especies a la vez y analizando los porcentajes de acierto.

Probando con las *eigenfaces*, los resultados fueron los siguientes:

Binarias	floccosum	canis	gypseum	mentagro	rubrum	tonsurans
Floccosum		75%	75%	70,83%	75%	54,17%
Canis			79,17%	54,17%	66,67%	54,17%
Gypseum				87,5%	50%	66,67%
mentagro					66,67%	58,33%
Rubrum						66,67%
tonsurans						
Totales	<50%	<i>eigenfaces</i>	<i>test2</i>			

Tabla 15.1.1 – Porcentajes de acierto eigenfaces para predicción

A pesar de no haber entrenado con la segunda muestra, igualmente el sistema pudo predecir varios de los pares, obteniéndose 6 casos con un porcentaje mayor al 70%, siendo uno de ellos del 87,5% en *M. gypseum* vs *T. mentagrophytes*.

Además, hubo 5 de los pares que registraron un porcentaje menor al 60%:

- *E. floccosum* vs *T. tonsurans* con 54,17%
- *M. canis* vs *T. mentagrophytes* con 54,17%
- *M. canis* vs *T. tonsurans* con 54,17%
- *M. gypseum* vs *T. rubrum* con 50%
- *T. mentagrophytes* con *T. tonsurans* con 58,33%

El porcentaje de las pruebas con todas las especies a la vez resultó menor al 50%.

15.2. Pruebas con todas las especies: *eigenfungi*

Los porcentajes de acierto, con el método de *eigenfungi* fueron los siguientes:

Binarias	floccosum	canis	Gypseum	mentagro	rubrum	tonsurans
Floccosum		79,17%	70,85%	83,33%	70,83%	37,5%
Canis			79,17%	58,33%	70,83%	45,83%
Gypseum				91,67%	75%	83,33%
mentagro					87,5%	62,5%
Rubrum						70,83%
tonsurans						
Totales	<50%	<i>eigenfungi</i>	<i>Test2</i>			

Tabla 15.2.1 – Porcentajes de acierto eigenfungi para predicción

Varios de los porcentajes subieron comparados con las pruebas de *eigenfaces* (Tabla 15.1.1); 11 de los pares tuvieron un resultado mayor al 70% (5 pares más que en el caso de las *eigenfaces*) y 4 de ellos fueron mayores al 80%:

- E. floccosum vs T. mentagrophytes con 83,33%
- M. gypseum vs T. mentagrophytes con 91,67%
- M. gypseum vs T. tonsurans con 83,33%
- T. mentagrophytes vs T. rubrum con 87,5%

Sin embargo 3 de los pares no fueron reconocidos y el caso de *E. floccosum* vs *T. tonsurans* bajó de 54,17% a 37,5%.

El porcentaje de las pruebas con todas las especies a la vez continuó siendo menor al 50%.

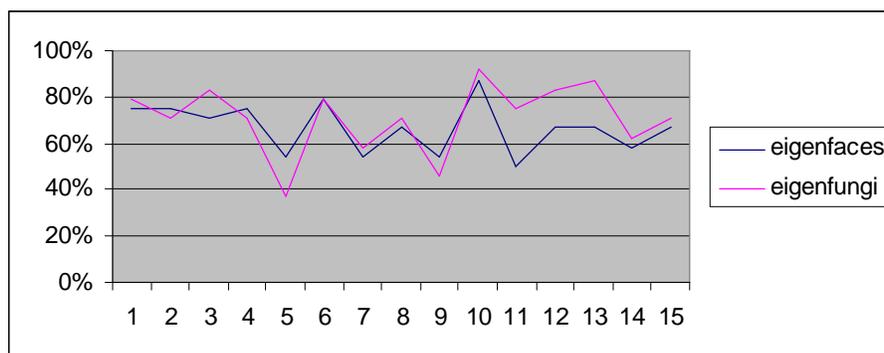


Fig. 15.1 – Comparación de porcentajes de acierto entre los métodos eigenfaces y eigenfungi

15.3. Pruebas con preprocesamientos

A fin de mejorar los resultados obtenidos, y de la misma manera que se hizo en los casos de una y dos muestras, se repitieron las pruebas combinando el método con distintos preprocesamientos.

a) Detección de contornos

Los porcentajes de acierto en el caso de las imágenes con detección de contornos fueron los siguientes:

Binarias	floccosum	canis	gypseum	mentagro	rubrum	tonsurans
floccosum		50%	50%	45,83%	50%	50%
Canis				50%	54,17%	50%
Gypseum				50%	75%	62,5%
mentagro					50%	50%
Rubrum						50%
tonsurans						
Totales		<i>eigenfungi</i>	<i>contornos</i>		<i>Test2</i>	

Tabla 15.3.1 – Porcentajes de acierto eigenfungi para predicción combinado con detección de contornos

Al igual que en el caso de una y dos muestras, la previa detección de contornos no mejoró los resultados sino que los porcentajes fueron muy bajos. Solamente 2 pares fueron mayores al 60%:

- *M. gypseum* vs *T. rubrum* 75%
- *M. gypseum* vs *T. tonsurans* 62,5%

Y nuevamente el porcentaje de pruebas totales fue menor al 50%.

b) Ecuilización de histograma

Probando el método sobre imágenes con corrección de histograma, los porcentajes de acierto con respecto al método *eigenfungi* puro (Tabla 15.2.1) fueron similares, por lo que no se consiguió una mejora sustancial con este preprocesamiento.

Binarias	floccosum	canis	gypseum	Mentagro	rubrum	tonsurans
floccosum		79,17%	66,67%	83,33%	75%	37,5%
Canis			79,17%	58,33%	70,83%	45,83%
Gypseum				91,67%	70,83%	83,33%
mentagro					87,5%	62,5%
Rubrum						70,83%

tonsurans						
Totales		<i>eigenfungi</i>	<i>Histograma</i>		<i>Test2</i>	

Tabla 15.3.2 – Porcentajes de acierto eigenfungi para predicción combinado con corrección de histograma

Solamente un par se incrementó *E. floccosum* vs *T. rubrum* de 70,83% a 75%. Las pruebas totales no superaron el 50%.

c) Transformada de Fourier

En el caso de las imágenes con Transformada de Fourier, los porcentajes se incrementaron significativamente.

Binarias	floccosum	canis	gypseum	mentagro	rubrum	tonsurans
Floccosum		62,5%	79,17%	91,67%	100%	54,17%
Canis			70,83%	75%	87,5%	62,5%
Gypseum				62,5%	45,83%	83,33%
Mentagro					70,83%	91,67%
Rubrum						100%
Tonsurans						
Totales		<i>eigenfungi</i>	<i>fft</i>		<i>Test2</i>	

Tabla 15.3.3 – Porcentajes de acierto eigenfungi para predicción combinado con transformada de Fourier

En este caso 4 pares llegaron a más del 90% de acierto, dos de ellos con 100%:

- *E. floccosum* vs *T. rubrum*
- *T. rubrum* vs *T. tonsurans*

Sin embargo, hubo dos pares que no fueron reconocidos:

- *E. floccosum* vs *T. tonsurans* (54,17%)
- *M. gypseum* vs *T. rubrum* (45,83%)

d) Imágenes binarizadas

Si bien hubo un par cuyo porcentaje subió significativamente de 70,85% a 91,67% (*E. floccosum* vs *M. gypseum*), el resto de los porcentajes se mantuvo prácticamente igual que con el método puro (Tabla 15.2.1).

Binarias	floccosum	Canis	gypseum	mentagro	rubrum	Tonsurans
floccosum		79,17%	91,67%	83,33%	70,83%	58,33%

Canis			66,67%	62,5%	75%	58,33%
Gypseum				87,5%	58,33%	54,17%
mentagro					79,17%	62,5%
Rubrum						70,83%
Tonsurans						
Totales		<i>eigenfungi</i>	<i>binarias</i>		<i>Test2</i>	

Tabla 15.3.4 – Porcentajes de acierto eigenfungi para predicción combinado con imágenes binarizadas

e) Suavizado de bordes y corrección de histograma

Los porcentajes de acierto, con el método de *eigenfungi* combinado con un preprocesamiento de suavizado de bordes y corrección de histograma, fueron los siguientes:

Binarias	floccosum	Canis	gypseum	mentagro	rubrum	Tonsurans
Floccosum		75%	83,33%	87,5%	79,17%	45,83%
Canis			75%	58,33%	75%	50%
Gypseum				87,5%	79,17%	66,67%
Mentagro					95,83%	70,83%
Rubrum						75%
Tonsurans						
Totales	52,78%	<i>eigenfungi</i>	<i>suavizado</i>	<i>histograma</i>	<i>test2</i>	

Tabla 15.3.5 – Porcentajes de acierto eigenfungi para predicción combinado con suavizado de bordes y corrección de histograma

Se registraron muchos casos altos en las pruebas binarias. El método serviría entonces en la predicción de la mayoría de los pares como por ejemplo, *T. mentagrophytes* versus *T. rubrum* con un porcentaje de aciertos del 95,83%.

Sin embargo esto no es así con algunos casos como *E. floccosum* versus *T. tonsurans* con un 45,83% y *M. canis* vs *T. tonsurans* con un 50%. A nivel de pruebas totales, el porcentaje fue bajo, aunque superior al 50% (es del 52,78%).

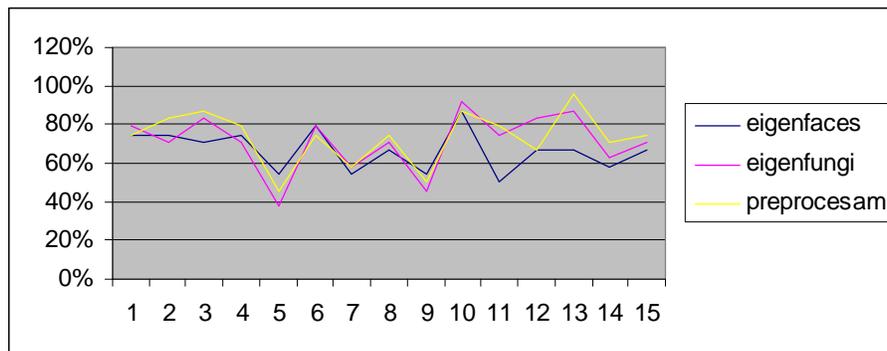


Fig. 15.2 – Comparación de porcentajes de acierto entre los métodos eigenfaces, eigenfungi y eigenfungi con preprocesamiento de suavizado de bordes y eualización de histograma.

f) Desenfoque gaussiano y corrección de histograma

Al probar con desenfoque gaussiano con corrección de histograma, los porcentajes fueron similares al método puro (Tabla 15.2.1), pero disminuyeron un poco, en particular en el caso del reconocimiento del par *E. floccosum* versus *T. tonsurans* (de 37,5% a 29,17%).

Binarias	floccosum	canis	gypseum	mentagro	rubrum	Tonsurans
Floccosum		79,17%	70,83%	83,33%	79,17%	29,17%
Canis			75%	54,17%	70,83%	50%
Gypseum				87,5%	75%	70,83%
mentagro					87,5%	70,83%
Rubrum						70,83%
tonsurans						
Totales	52,78%	<i>eigenfungi</i>	<i>desenfoque</i>	<i>histograma</i>	<i>Test2</i>	

Tabla 15.3.6 – Porcentajes de acierto eigenfungi para predicción combinado con desenfoque gaussiano y corrección de histograma

15.4. Preprocesamiento seleccionado

Comparando los valores obtenidos con cada preprocesamiento, la combinación mejor con los *eigenfungi* en el caso de entrenamiento con una muestra y testeo con dos muestras, resultó ser con el preprocesamiento de Transformada de Fourier (Tabla 15.3.3). En este caso 4 pares llegaron a más del 90%, dos de ellos con 100%. Y además, solamente dos pares no fueron reconocidos.

El caso de suavizado de bordes con corrección de histograma también dio resultados altos (algunos más altos que con Transformada de Fourier), pero solamente 1 par llega a más del 90% y hay 3 pares que no fueron reconocidos.

15.5. Resumen de preprocesamientos

Si se reúnen en una misma tabla los mejores resultados de cada par de especies obtenidos con cada tipo de reprocesamiento, se podrían ver los siguientes porcentajes:

Binarias	floccosum	canis	gypseum	mentagro	rubrum	tonsurans
Floccosum		79,17%	91,67%	91,67%	100%	58,33%
Canis			75%	75%	87,5%	62,5%
Gypseum				87,5%	79,17%	83,33%
mentagro					95,83%	91,67%
Rubrum						100%
tonsurans						
Totales		<i>eigenfungi</i>	<i>varios</i>	resumen	<i>Test2</i>	

Tabla 15.5.1 – Resumen de porcentajes de acierto eigenfungi para predicción combinado con varios preprocesamientos

Salvo *E. floccosum* vs *T. tonsurans* (58,33%) y *M. canis* vs *T. tonsurans* (62,5%) todos los pares resultan mayores al 75% de acierto con 6 entre el 90 y 100%.

Hay que destacar que no se encontró un preprocesamiento adecuado que combinado con el método *eigenfungi* logre un porcentaje de acierto alto para el par *E. floccosum* vs *T. tonsurans* (se llegó al 58,33%).

Para el caso de pruebas totales no se consiguieron preprocesamientos que logran mejoras sustanciales al 50% de acierto. El porcentaje únicamente subió en el caso de no usar todas las especies a la vez, sino sacando alguna, por ejemplo 5 especies juntas. De esta manera, sin *M. canis* el porcentaje total subió a 58,33%. Y sin *M. canis* y sin *T. tonsurans*, a 70,83%.

Parte VI: Comparaciones con otros métodos de Data Mining

16. Variantes de PCA

A fin de encontrar mejoras al método de reconocimiento desarrollado, se probaron algunas variantes del método de las *eigenfungi*. Sin embargo, se decidió no aplicarlas al estudio de los dermatofitos dado que, a pesar de registrarse algunos pares con porcentajes altos, los resultados de acierto en forma global fueron más bajos en comparación con el método *eigenfungi* presentado (en particular en las pruebas a nivel totales) y algunos pares no fueron reconocidos.

Describimos entonces dos variantes desarrolladas

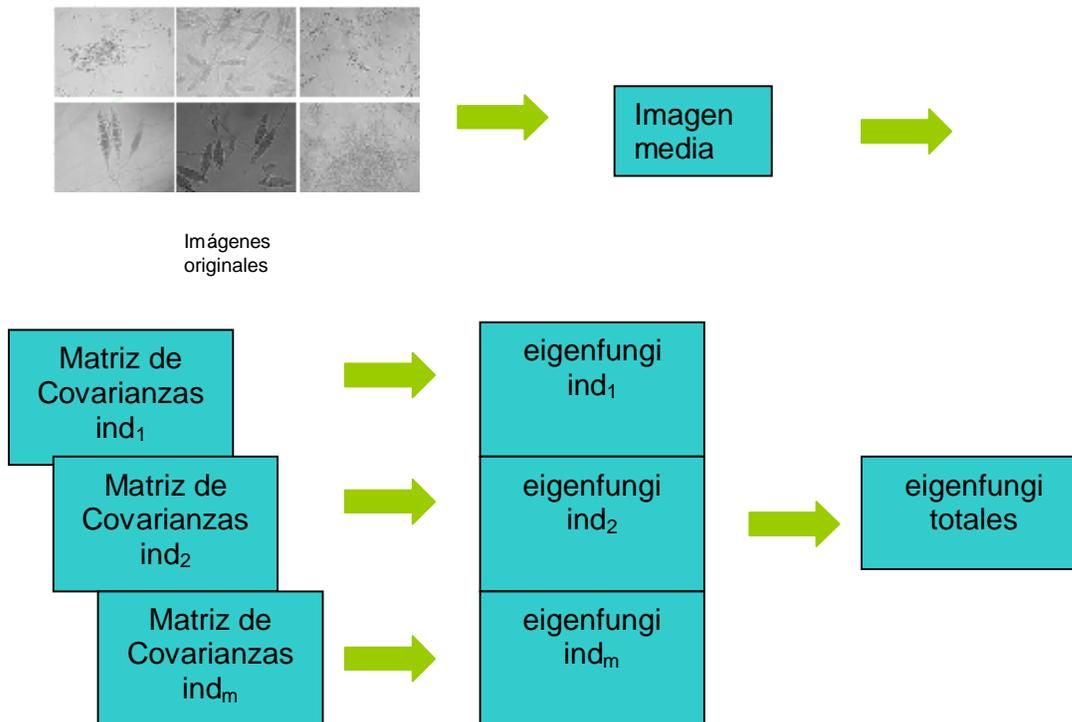
- cálculo de multiespacios
- utilización de distancia Manhattan

16.1. Multiespacios

La variante de Multiespacios plantea que, al momento de generar las *eigenfungi*, no se utilicen todas las imágenes originales, sino solamente las de cada individuo cada vez.

Supongamos que se consideran m individuos, con k imágenes de cada uno para realizar el entrenamiento. Cuando se calcula la matriz de covarianzas, no se toman las $m*k$ imágenes, sino que se obtienen m matrices de covarianzas (una por individuo).

A partir de éstas se arman m conjuntos de *eigenfungi*. Luego se juntan todas las *eigenfungi* obtenidas y se prosigue con el método, calculándose los vectores de distancias correspondientes.



Las pruebas realizadas con dermatofitos con generación de Multiespacios dieron los siguientes resultados de porcentajes de acierto.

Binarias	floccosum	canis	gypseum	mentagro	rubrum	tonsurans
floccosum		91,67%	75%	50%	50%	50%
canis			66,67%	66,67%	100%	66,67%
gypseum					50%	91,67%
mentagro					50%	100%
rubrum						50%
tonsurans						
Totales		<i>Eigenfungi</i>	<i>Multiesp</i>			

Tabla 16.1.1 – Porcentajes de acierto variante eigenfungi multiespacios

Salvo un par de especies como *E. floccosum* versus *M. canis* con 91,67% de acierto, y por ejemplo *M. canis* con *T. rubrum* con un 100%, los porcentajes en general fueron muy bajos y varios pares no fueron reconocidos, como *E. floccosum* versus *T. rubrum*.

16.2. Distancia Manhattan

Para hallar los vectores de distancias que caracterizan cada imagen de entrenamiento, se compara cada una con las *eigenfungi* obtenidas. Para esta comparación se calcula la distancia de cada imagen a cada *eigenfungi*, siendo utilizada para esto la distancia Euclídea.

Dados 2 vectores t y w de longitud M , la distancia Euclídea entre ellos se define como:

$$\text{Euclidea}^2(t, w) = \sum_{i=1}^M (t_i - w_i)^2$$

En vez de utilizar esta distancia, se probó el método modificado por la aplicación de la distancia Manhattan, que se define como:

$$\text{Manhattan}(t, w) = \sum_{i=1}^M |t_i - w_i|$$

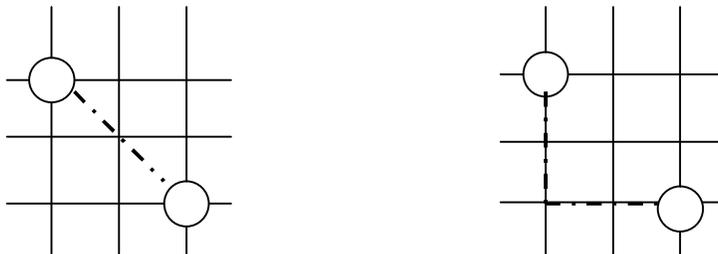


Fig. 16.1 – Ejemplos gráficos de las distancias entre dos puntos, Euclídea en el primer caso y Manhattan en el segundo.

Las pruebas realizadas con dermatofitos con la distancia Manhattan dieron los siguientes resultados de porcentajes de acierto.

Binarias	floccosum	canis	gypseum	mentagro	rubrum	tonsurans
Floccosum		75%	91,67%	100%	75%	50%
Canis			100%	91,67%	100%	91,67%
Gypseum				83,33%	100%	100%
mentagro					75%	75%
Rubrum						100%
tonsurans						
Totales	50%	<i>Eigenfungi</i>	<i>Manhattan</i>			

Tabla 16.2.1 – Porcentajes de acierto variante eigenfungi distancia Manhattan

Se observaron altos porcentajes de acierto entre varias especies como *E. floccosum* versus *T. mentagrophytes* con un 100%. Pero no se reconoció el par *E. floccosum* versus *T. tonsurans* con un 50% y el porcentaje de las pruebas totales fue muy bajo, también del 50%.

16.3. Combinación de ambas variantes

Se armó una tercera variante combinando la construcción de varios espacios para las *eigenfungi* con distancia Manhattan para hallar los vectores de distancia.

Los resultados obtenidos en este caso, fueron los siguientes:

Binarias	floccosum	canis	gypseum	mentagro	rubrum	tonsurans
Floccosum		100%	66,67%	91,67%	83,33%	50%
Canis			91,67%	100%	100%	83,33%
Gypseum				66,67%	75%	50%
mentagro					66,67%	50%
Rubrum						100%
tonsurans						
Totales	47,2%	<i>Eigenfungi</i>	<i>Manhattan</i>	<i>Multiesp</i>		

Tabla 16.3.1 - Porcentajes de acierto variante eigenfungi combinando multiespacio y distancia Manhattan

Se registraron varios porcentajes altos como *E. floccosum* versus *M. canis* con un 100%. Pero no se reconocieron varios pares, como *M. gypseum* versus *T. tonsurans*. El porcentaje de las pruebas totales fue muy bajo, del 47,2%.

16.4. Combinación de ambas variantes con preprocesamientos

También se probó esta combinación con algunos preprocesamientos, no obteniéndose mejoras sustanciales en los resultados. Por ejemplo, combinando con suavizado de bordes, los resultados de porcentajes de acierto fueron:

Binarias	floccosum	Canis	gypseum	mentagro	rubrum	tonsurans
Floccosum		91,66%	83,33%	100%	100%	61,67%
Canis			91,67%	91,67%	100%	100%
Gypseum				83,33%	58,33%	66,67%
mentagro					75%	75%
Rubrum						100%
tonsurans						

Totales	41,67%	<i>Eigenfungi</i>	<i>Manhattan</i>	<i>Multiesp</i>	<i>Suaviz</i>	
----------------	--------	-------------------	------------------	-----------------	---------------	--

Tabla 16.4.1 - Porcentajes de acierto variante eigenfungi combinando multiespacio y distancia Manhattan con preprocesamiento de suavizado de bordes

Se registraron varios porcentajes altos como *E. floccosum* versus *T. mentagrophytes* con un 100%. Pero no se reconocieron varios pares, como *E. floccosum* versus *T. tonsurans* y *M. gypseum* versus *T. rubrum*. El porcentaje de las pruebas totales bajó al 41,67%.

17. Método de *Fisherfaces*

En 1997, Belhumeur, Hespanha y Kriegman [BEL97] desarrollaron un método similar a las *eigenfaces*, pero que en lugar de basarse en el *Análisis de Componentes Principales*, se basa en el *Análisis Discriminante Lineal* (LDA).

La primera aproximación al problema de discriminación lineal para $k=2$ grupos fue sugerida por Fisher (1936) quien abordó el problema desde una óptica univariada usando una combinación lineal de las características observadas.

Es por esto que el método fue denominado *fisherfaces*.

17.1. Cálculo de las *fisherfaces*

Sea un conjunto de N imágenes $\{x_1, x_2, \dots, x_n\}$ y se asume que cada imagen pertenece a una de las c clases $\{X_1, X_2, \dots, X_c\}$. Sea también una transformación lineal que mapea el espacio original de imágenes n -dimensional en un espacio de características m -dimensional, donde $m < n$. Los nuevos vectores de características $y_k \in \mathbf{R}^m$ se definen por la siguiente transformación lineal

$$y_k = W^t x_k \quad k = 1, 2, \dots, N$$

donde $W^t \in \mathbf{R}^{n \times m}$ es una matriz de columnas ortonormales.

El método del *Análisis Discriminante Lineal* de Fisher selecciona un W de tal manera que la razón entre la dispersión *entre* clases y la dispersión *dentro* de cada clase es maximizada.

Sea la matriz de dispersión (o matriz de varianzas-covarianzas) *entre* clases definida como

$$S_B = \sum_{i=1}^c N_i (\mu_i - \mu) (\mu_i - \mu)^T$$

y la matriz de dispersión *dentro* de clases definida como

$$S_W = \sum_{i=1}^c \sum_{x_k \in X_i} (x_k - \mu_i) (x_k - \mu_i)^T$$

donde μ_i es la imagen media de la clase X_i y N_i es el número de muestras en la clase X_i . Si S_W es no singular, la proyección óptima W_{opt} es elegida como la matriz con columnas ortonormales (una base ortonormal es aquella que además de ser ortogonal, la norma de cada elemento que la compone es igual a 1).

Ésta maximiza la razón del determinante de la matriz de dispersión *entre* clases de las muestras proyectadas al determinante de la matriz de dispersión *dentro* de clases, por ejemplo

$$\begin{aligned} W_{opt} &= \arg \max_W \frac{|W^T S_B W|}{|W^T S_W W|} \\ &= [\mathbf{w}_1 \ \mathbf{w}_2 \ \dots \ \mathbf{w}_m] \end{aligned}$$

donde $\{\mathbf{w}_i \mid i = 1, 2, \dots, m\}$ es el conjunto de los autovectores generalizados de S_B y S_W correspondientes a los m autovalores generalizados mayores $\{\lambda_i \mid i = 1, 2, \dots, m\}$, por ejemplo

$$S_B \mathbf{w}_i = \lambda_i S_W \mathbf{w}_i \quad i = 1, 2, \dots, m$$

Nótese que existen al menos $c-1$ autovalores que no son nulos y además $c-1$ es un límite superior sobre m , con c el número de clases.

En el problema del reconocimiento de rostros existe la dificultad de que la matriz de dispersión $S_W \cdot \mathbf{R}^{n \times m}$ es siempre singular (es decir, no tiene inversa). Esto proviene del hecho de que por ejemplo el número de imágenes N en el conjunto de entrenamiento es mucho menor que el número de píxeles en cada imagen n . Esto significa que es posible elegir la matriz W tal que la distribución dentro de clases de las muestras proyectadas puede ser exactamente cero.

El método de las *Fisherfaces* evita este problema proyectando el conjunto de imágenes a un espacio de menos dimensiones así la matriz de dispersión *dentro* de clases resultante S_W es no singular. Esto se logra usando el Análisis de Componentes Principales (PCA) la dimensión del espacio de características a $N - c$ y entonces aplicar el Análisis Discriminante Lineal (LDA) para reducir la dimensión a $c-1$. Más formalmente W_{opt} está dado por

$$W_{opt}^T = W_{fld}^T W_{pca}^T$$

donde

$$\begin{aligned} W_{pca} &= \arg \max_W |W^T S_T W| \\ W_{fld} &= \arg \max_W \frac{|W^T W_{pca}^T S_B W_{pca} W|}{|W^T W_{pca}^T S_W W_{pca} W|} \end{aligned}$$

17.2. Comparación entre las *eigenfaces* y las *fisherfaces*

A continuación se describen algunos aspectos sobre los cuales se comparan los métodos de *eigenfaces* y *fisherfaces*.

a) Eficiencia

Los resultados a través de varios trabajos de comparación (por ejemplo [GRO01] [RUI05] [CHE05]) indican que ambos métodos clasifican en forma similar.

b) Método base

Las *eigenfaces* se basan en el Análisis de Componentes Principales (PCA) y las *fisherfaces* en el Análisis Discriminante Lineal de Fisher (LDA).

c) Sensibilidad

El método de las *fisherfaces* es menos sensible a variaciones de las condiciones de iluminación y distintas expresiones faciales [BEL97].

d) Metodología

En las *eigenfaces* se buscan los vectores que reducen la dimensión del espacio de imágenes y mejor describen los datos para su codificación. En el caso de las *fisherfaces* se buscan los vectores que proporcionan mejor discriminación entre clases después de la proyección.

e) Complejidad

El algoritmo de *eigenfaces* es mucho más sencillo que el de *fisherfaces* [BRO04]. Además, el de *fisherfaces* requiere del de *eigenfaces* en su proceso.

f) Vigencia

Según Zhang y otros en su trabajo “*Diagonal Principal Component Analysis for Face Recognition*” de 2006 [ZHA06] expresan: “Especialmente para el problema de contar con sólo una imagen por persona para el entrenamiento, la mayoría de los métodos de reconocimiento tales como [...] las *fisherfaces*, fallan, mientras que las variantes de PCA todavía se siguen utilizando. Ésta es una de las razones por las cuales el reconocimiento de caras basado en PCA sigue todavía muy activo aunque haya sido estudiado por décadas.”

17.3. *Fisherfaces* aplicado a los dermatofitos

Se utilizó el software *FisherFaces for Face Recognition* de Luigi Rosa [ROS06] desarrollado en Matlab (*Anexo A – Software utilizado*) y se realizaron pruebas similares a las efectuadas con el método de *eigenfungi* con las imágenes de dermatofitos:

■ Pruebas binarias

Se dispusieron los objetos o especies de a pares, entrenando y reconociendo dos cada vez

Por ejemplo, *E. floccosum* versus *M. canis*

■ Pruebas totales

Se entrenó y testeó con todos los objetos o especies a la vez.

Por ejemplo, se intenta que el sistema reconozca cuál de las 6 especies pertenece una imagen

Según se analizó en el punto anterior, los resultados de clasificación de las *fisherfaces* son muy parecidos a las *eigenfaces*, con diferencias a nivel de variaciones de iluminación (según la fuente de luz) o de expresiones faciales. Estas diferencias son aplicables a casos de rostros humanos, no así en el caso de los hongos microscópicos.

Es por esto que, como fue previsto, los resultados obtenidos entre las *fisherfaces* y los *eigenfungi* aplicados a las imágenes de los dermatofitos fueron muy similares.

Por ejemplo, se hicieron pruebas binarias entre la especie *E. floccosum* contra el resto y los porcentajes de acierto fueron:

Binarias	canis	gypseum	mentagro	rubrum	tonsurans
Eigenfungi	100%	83,33%	100%	83,83%	50%
Fisherfaces	75%	100%	83,33%	83,33%	91,67%

Tabla 17.3.1 – Comparación porcentajes de acierto eigenfungi y fisherfaces, especie *E. floccosum* versus el resto

En este caso, se ve que los valores en general fueron un poco más bajos que en el caso de *eigenfungi*, por ejemplo contra *M. canis* se consigue un 100% con los *eigenfungi* y un 75% con las *fisherfaces*. En el caso de *E. floccosum* vs *T. tonsurans*, en cambio, se consigue un 91,67% con las *fisherfaces* y un 50% con los *eigenfungi*; esta particularidad también se detectó al comparar los *eigenfungi* con las *eigenfaces*, pero queda “salvado” combinando los *eigenfungi* con preprocesamientos.

En el caso de las pruebas totales, se registró un porcentaje menor con el método de las *fisherfaces*:

Totales	
Eigenfungi	80,56%
Fisherfaces	72%

Tabla 17.3.2 – Comparación porcentajes de acierto eigenfungi y fisherfaces, pruebas totales

17.4. Fisherfaces con preprocesamientos

Se realizaron pruebas con el preprocesamiento de suavizado de bordes y ecualización de histograma (seleccionado en el caso del método *eigenfungi*). De la misma manera que sucedió con los *eigenfungi*, los porcentajes de acierto se incrementaron.

Por ejemplo, en el caso de *E. floccosum* contra el resto de las especies, los porcentajes fueron:

Binarias	canis	gypseum	mentagro	rubrum	tonsurans
Fisherfaces sin prep	75%	100%	83,33%	83,33%	91,67%
Fisherfaces con prep	100%	83,33%	91,67%	100%	100%

Tabla 17.4.1 – Comparación porcentajes de acierto fisherfaces y fisherfaces combinado con preprocesamiento suavizado de bordes y corrección de histograma, especie *E. floccosum* versus el resto

Se observó un incremento en general de los valores, por ejemplo *E. floccosum* vs *T. tonsurans*, subió de un 75% a un 100% de acierto. Se notó una reducción en el caso de *E. floccosum* vs *M. gypseum*, de un 100% a un 83,33% pero en general el resto de los porcentajes subió.

De todos modos, los valores obtenidos con los *eifgenungi* combinados con preprocesamientos fueron muy altos (casi un 100% en todos los casos), por lo que se prefiere ese método, además de las ventajas antedichas, como la simplicidad del algoritmo.

Parte VII: Conclusiones y trabajos futuros

En este trabajo desarrollamos un método automático para el reconocimiento de especies de hongos microscópicos, que llamamos *eigenfungi*. Está basado en la metodología para reconocimiento de rostros denominado *eigenfaces*.

La base matemática se sustenta en el *Análisis de Componentes Principales*, que es un método estadístico de análisis que descompone datos multidimensionales a un subespacio de menor dimensión pero preservando las características esenciales de los datos tratados.

Se obtuvieron imágenes de las 6 principales especies de hongos dermatofitos:

<i>Epidermophyton floccosum</i>	<i>Trichophyton mentagrophytes</i>
<i>Microsporium canis</i>	<i>Trichophyton rubrum</i>
<i>Microsporium gypseum</i>	<i>Trichophyton tonsurans</i>

Se desarrolló un software en Matlab y se realizaron 2 tipos de pruebas:

■ Pruebas binarias

Se dispusieron los objetos o especies de a pares, entrenando y reconociendo dos cada vez

Por ejemplo, *E. floccosum* versus *M. canis*

■ Pruebas totales

Se entrenó y testeó con todos los objetos o especies a la vez.

Por ejemplo, se intenta que el sistema reconozca a cuál de las 6 especies pertenece una imagen

Las pruebas dieron resultados con porcentajes altos y se mejoraron con la combinación del método con preprocesamientos. Con el tratamiento de las imágenes de un suavizado de bordes con corrección de histograma, se obtuvo casi un 100% de porcentaje de acierto, según se observa en la siguiente tabla:

Binarias	floccosum	canis	gypseum	mentagro	rubrum	tonsurans
floccosum		91,67%	91,67%	100%	100%	83,33%
canis			100%	100%	100%	91,67%
gypseum				100%	100%	91,67%
mentagro					100%	100%
rubrum						100%
tonsurans						
Totales	86,11%	<i>eigenfungi</i>	<i>suavizado</i>	<i>Histograma</i>		

Se repitieron estas pruebas con imágenes degradadas por ruido y también se obtuvieron porcentajes altos, observándose la robutez del método.

Luego se hicieron pruebas con 2 muestras tanto en el entrenamiento como el testeo y luego con 2 muestras sólo en el testeo (sin darle al sistema la segunda muestra para entrenar).

Las pruebas binarias combinadas con preprocesamientos resultaron con altos porcentajes de acierto (aunque no tan altos como en el caso de una única muestra). Por ejemplo, en el caso de 2 muestras con suavizado de bordes y corrección de histograma, se ven los siguientes resultados:

Binarias	floccosum	canis	gypseum	mentagro	rubrum	tonsurans
floccosum		83,33%	83,33%	100%	87,5%	70,83%
canis			70,83%	70,83%	75%	66,67%
gypseum				95,83%	95,83%	75%
mentagro					100%	66,67%
rubrum						100%
tonsurans						
Totales	55,56%	<i>eigenfungi</i>	<i>suavizado</i>	<i>Histograma</i>	<i>2muestr</i>	

Y en el caso de 2 muestras para testeo (pruebas de predicción) los valores obtenidos con el preprocesamiento de la Transformada de Fourier fueron:

Binarias	floccosum	canis	gypseum	mentagro	rubrum	tonsurans
floccosum		62,5%	79,17%	91,67%	100%	54,17%
canis			70,83%	75%	87,5%	62,5%
gypseum				62,5%	45,83%	83,33%
mentagro					70,83%	91,67%
rubrum						100%
tonsurans						
Totales		<i>eigenfungi</i>	<i>fft</i>		<i>Test2</i>	

En ambos casos de 2 muestras no se logró un alto porcentaje de pruebas totales (55,56% en 2 muestras al entrenar y testear y <50% en las pruebas de predicción). Pero este porcentaje aumentó si no se usaban todas las especies a la vez, sino que se sacaba alguna y se hacían pruebas con por ejemplo 5 especies juntas. Específicamente, las pruebas realizadas sin la especie *T. tonsurans*, llegaron a un porcentaje del 70,83% con 2 muestras entrenando y testeando.

En las pruebas de predicción no encontramos un preprocesamiento adecuado que combinado con el método *eigenfungi* lograra un porcentaje de acierto alto para el par *E. floccosum* vs *T. tonsurans* (se llegó al 58,33%).

Luego se compararon estos resultados con variantes del algoritmo PCA (generación de multiespacios, utilización de distancia Manhattan, combinación con preprocesamientos). Sin embargo, se decidió no aplicarlas al estudio de los dermatofitos dado que, a pesar de

registrarse algunos pares con porcentajes altos, los resultados de acierto en forma global fueron más bajos en comparación con el método *eigenfungi* presentado (en particular en las pruebas a nivel totales) y algunos pares no fueron reconocidos.

Por último, se comparó con otro método de reconocimiento de rostros con técnicas de Data Mining, denominado *fisherfaces* que se basa en el Análisis Discriminante. Los resultados fueron similares (aunque un poco menores a nivel de pruebas totales). Sin embargo, los valores obtenidos con los *eifgenungi* combinados con preprocesamientos fueron muy altos (casi un 100% en todos los casos), por lo que se prefiere ese método, que presenta ventajas con respecto a las *fisherfaces*, como por ejemplo la simplicidad del algoritmo.

17.5. Características del método

Además del hecho de no haber encontrado en la bibliografía métodos automáticos para reconocer imágenes de micología, el método presentado cuenta con una serie de ventajas al momento del reconocimiento de las especies:

■ Imágenes como una base de datos

Este método no se basa en la forma tradicional de búsqueda de patrones en imágenes. Transforma las mismas y aplica técnicas propias de Data Mining, considerando al conjunto de imágenes como una base de datos con registros y atributos

■ No necesita de recortes manuales de las imágenes

Algunos métodos de reconocimiento de patrones requieren que el experto humano, antes del análisis, recorte la imagen para centrarla en determinados elementos (como por ejemplo separar posibles bacterias) o que se trabajen manualmente delimitando formas a identificar

■ No necesita de normalización de las imágenes

Si bien la aplicación de preprocesamientos (como suavizado de bordes) incrementan la exactitud de la clasificación, las imágenes pueden utilizarse sin necesidad de transformaciones específicas, ni normalizaciones (por ejemplo, en el caso de los rostros humanos, es conveniente previo al análisis homogeneizar el tamaño de las cabezas, ubicación de los ojos y de las bocas, etc.)

■ Pocas imágenes de entrenamiento

No se requiere la recolección de grandes cantidades de imágenes. Esto incrementa la velocidad de entrenamiento y el almacenamiento necesario para las imágenes y las matrices que conforman el método

■ No se requieren imágenes de gran tamaño

Las imágenes recién extraídas son generalmente de gran tamaño, dificultándose su almacenamiento y manipulación. El método da buenos resultados aún con las imágenes reducidas en tamaño

■ Robustez

Imágenes degradadas con ruido presentaron un alto porcentaje de exactitud de acierto en el reconocimiento

■ Es rápido tanto al momento del entrenamiento y el uso

Tanto el entrenamiento como el uso del método conllevan una duración del orden de los segundos de ejecución (dependiendo del tamaño y cantidad de imágenes elegidos)

■ El algoritmo es matemática y computacionalmente simple

Matemáticamente, se utilizan conceptos estadísticos clásicos, como matriz de varianzas-covarianzas, autovalores y autovectores. Las imágenes son tratadas como una gran base de datos. No se requieren búsquedas de patrones tradicionales (que requieren análisis de formas y relaciones entre píxeles)

En trabajos posteriores podría trabajarse sobre la aplicación de este método a otros tipos de hongos microscópicos como los *Aspergillus*, que producen una serie de complicaciones como neumonías y dilataciones bronquiales.

También podría analizarse la utilización en el reconocimiento de colonias. Las imágenes de colonias representan otra visualización de los preparados. Consisten en muestras de cultivos en *cápsulas de Petri*. Para caracterizar las especies se utilizan las diferencias de tamaño, colores y texturas de cada muestra.

Otros trabajos podrían buscar mejoras en la predicción a nivel de pruebas totales. En el caso de identificar imágenes de 2 muestras, no se obtuvieron porcentajes de acierto alto al intentar reconocer las 6 especies a la vez. Y en las pruebas de predicción no se pudo reconocer el par par *E. floccosum* vs *T. tonsurans*. Se podría buscar una mejora al método o intentar con otros tipos de preprocesamientos.

También se podrían implementar y comparar la aplicación en hongos microscópicos de otros métodos para el reconocimiento de rostros, como *Independent Component Analysis* o *ICA* de Bartlett y otros [BAR98], que proyectan los datos sobre vectores básicos estadísticamente independientes, *Mixture of Principal Component* de Deepak y otros [DEE02], que usa una mezcla de eigen-espacios para capturar variaciones en los

datos, *PCA Diagonal* de Zhang y otros [ZHA06], que busca los vectores de proyección óptimos desde imágenes diagonalizadas, etc.

Bibliografía¹

- [BAL04] Miguel Ángel González Ballester, Xavier Pennec, Marius George Linguraru, Nicolás Ayache – “Generalized image models and their application as statistical models of images” – *Medical Image Analysis* 8 pp 361-369 – 2004
- [BAR98] Marian Stewart Bartlett, Terrence J. Sejnowski – “Independent component representations for face recognition” – *Proceedings of the SPIE: Conference on Human Vision and Electronic Imaging III*, vol. 3299, pp. 528-539 – 1998
- [BAR03] Marian Stewart Bartlett, Javier R. Movellan, Terrence J. Sejnowski – “Face Recognition by Independent Component Analysis” – *IEEE Transactions on Neural Networks*, Vol 13 N° 6 – Noviembre 2003
- [BEL97] Peter N. Belhumeur, Joao P. Hespanha, David J. Kriegman – “Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection” – *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, n° 7, pp. 711-720 – Julio 1997
- [BRO04] Alan Brooks, L. Gao – “Face Recognition: Eigenface and Fisherface Performance Across Pose” – *Final Project Report, University of Northwestern, USA* – Junio 2004
- [BUS04] Oscar H. Bustos, Alejandro C. Frery, Mario A. Lamfri e Carlos M. Scavuzzo – “Técnicas Estadísticas en Teledetección Espacial” – *CLATSE VI, Concepción, Chile* – Noviembre 2004
- [CAR01] Cárdenas Aurora, Tincopa Luis, Fernández Wilberto, Valverde Jenny, Agip Hernán – “Tiña capitis, frecuencia de agentes etiológicos” – *Dermatología peruana Vol. 11 N° 1* – Junio 2001
- [CHE04] Songcan C. Chen, Yulian L. Zhu – “Subpattern-based principal component analysis” – *Pattern Recognition* 37 (1) pp1081-1083 – Enero 2004
- [CHE05] S.C. Chen, Y.L. Zhu, D.Q. Zhang, J.Y. Yang – “Feature extraction approaches based on matrix pattern: MatPCA and MatFLDA” – *Pattern Recognition Letters* 26(8) 1157-1167 – 2005
- [CUE01] María Soledad Cuétara – “Procesamiento de las muestras superficiales” – *Revista Iberoamericana de Micología* – 2001
- [DEL06] Kresimir Delac, Mislav Grgic, Sonja Grgic – “Independent Comparative Study of PCA, ICA, and LDA on FERET Data Set” – *University of Zagreb, Croatia* – Febrero 2006
- [DAV04] Graciela O. Davel, Cristina E. Canteros, Laura L. Rodero – *Diagnóstico de micosis superficiales – Manual Teórico Práctico* – Instituto Nacional de Enfermedades Infecciosas ANLIS “Dr. Carlos G. Malbrán” – Agosto 2004
- [DEB07] Natalia Debandi, Ana S. Haedo, Marcelo Soria – “Reconocimiento y clasificación de hongos dermatofitos usando Máquinas de Soporte Vectorial (SVM) Tesis de Licenciatura” – *Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires* – Mayo 2007

¹ Las páginas citadas fueron verificadas en Julio 2007

- [DEE02] Deepak S. Turaga, T. Chen – “Face recognition using mixtures of principal components” – *IEEE ICIP, Rochester* – Setiembre 2002
- [DOR00] Thorsten Dorge, Jens Michael Carstensen, Jens Christian Frisvad – “Direct Identification of pure *Penicillium* species using image analysis” – *Journal of Microbiological Methods, Volume 41, Number 2, pp. 121-133(13)* – Julio 2000
- [DRE07] Santiago Serrano – “Eigenface Tutorial” – *Robotics Lab Drexel University, Philadelphia, USA*
<http://www.pages.drexel.edu/~sis26/Eigenface%20Tutorial.htm>
- [FAC02] “FACS - Facial Action Coding System” – *Robotics Institute, Carnegie Mellon University* – 2002
<http://www.cs.cmu.edu/afs/cs/project/face/www/facs.htm>
- [FAR05] Gonzalo Farías, Matilde Santos, Sebastián Dormido-Canto – “Desarrollo de una aplicación para la integración de técnicas de reconocimiento de patrones” – *XXVI Jornadas de Automática, España* – 2005
- [FOR98] Manuel Guillermo Forero – “Reconstrucción tridimensional de la cabeza y el cerebro a partir de IRM” – *Memorias 1er Congreso Latinoamericano de Ingeniería Biomédica, México* – Noviembre 1998
- [FOT05] “Tutorial Histograma” – *Asociación Profesional de Fotoperiodistas Asturianos, España* – 2005
<http://www.fotoperiodistas.org/recursos/histograma.pdf>
- [GAO03] Y. Gao – “Efficiently comparing face images using a modified Hausdorff distance” – *IEEE VISIP (150), N° 6 pp 346-350* – Diciembre 2003
- [GON04] Miguel Ángel González Ballester, Xavier Pennec, Marius George Linguraru, Nicholas Ayache – “Generalized image models and their application as statistical models of images” – *Medical Image Analysis 8 pp 361-369* – 2004
- [GOU97] V. Gouaillier, L. Gagnon – “Ship Silhouette Recognition Using Principal Components Analysis” – *SPIE Proc. #3164, pp. 59-69* – 1997
- [GRO01] R. Gross, J. Shi, J. Cohn – “Quo vadis Face Recognition? - The current state of the art in Face Recognition” – *Third Workshop on Empirical Evaluation Methods in Computer Vision, December* – 2001
- [GUO00] Guodong Guo, Stan Z. Li, Kapluk Chan – “Face Recognition by Support Vector Machines” – *Proceedings of the fourth IEEE International Conference on Automatic FGR'2000 pp 96-201* – 2000
- [GUP02] Himaanshu Gupta, Amit K. Agrawal, Tarun Pruthi, Chandra Shekhar, Rama Chellappa – “An Experimental Evaluation of Linear and Kernel-Based Methods for Face Recognition” – *IEEE Workshop on the Application of Computer Vision (WACV), Florida USA* – 2002
- [HAN01] Jiawei Han, Micheline Kamber – *Data Mining: Concepts and Techniques* – Morgan Kaufmann Publishers – 2001

[HEI03] Bernd Heisele, Thomas Serre, Sam Prentice, Tomaso Poggio – “Hierarchical classification and feature reduction for fast face detection with support vector machines” – *Pattern Recognition 36 pp 2007-2017* – 2003

[HLH] “Medical Pictures from CDC” – *Hardin Library for the Health Sciences, University of Iowa* – Marzo 2006

<http://www.lib.uiowa.edu/hardin/md/cdc/2938.html>

[IBO06] Marcelo J. Armengot Iborra – “Análisis comparativo de métodos basados en subespacios aplicados al reconocimiento de caras” – *Universidad de Valencia, España* – Setiembre 2006

[IMA07] “ImageJ - Image Processing and Analysis in Java” – *National Institutes of Health, USA*

<http://rsb.info.nih.gov/ij/>

[IMG07] “Técnicas de procesado de imagen” – *Departamento de Electrónica y Sistemas, Universidade da Coruña, España*

http://www.des.udc.es/~adriana/TercerCiclo/CursoImagen/curso/web/Filtrado_Espacial.html

[INF07] “Infostat – Software estadístico” – *Universidad de Córdoba, Argentina* – 2007

<http://www.infostat.com.ar/>

[ING01] Iain M. Inglis, Alison J. Gray – “An Evaluation of Semiautomatic Approaches to Contour Segmentation Applied to Fungal Hyphae” – *Biometrics 57, 232-239* – Marzo 2001

[IRI05] “Iris Data” – *Statistic Laboratory, University of Heidelberg, Germany* – Octubre 2005

<http://www.statlab.uni-heidelberg.de/data/iris/>

[JOH92] Johnson R.A., Wichern Dean W. – *Applied Multivariate Statistical Analysis* – Prentice Hall Inc. USA. – 1992

[JOH98] Dallas E. Johnson – *Métodos multivariados aplicados al análisis de datos* – International Thomson Editores – 1998

[KAR00] S. A. Karkanis, D.K. Iakovidis, D. E. Maroulis, G. D. Magoulas, N. G. Thefanous – “Tumor Recognition in Endoscopic Video Images using Artificial Neural Network Architectures” – *Proceedings of The 26th EUROMICRO Conference (EUROMICRO'00) Volume 2* – Setiembre 2005

[KIN97] Irwin King, Lei Xu – “Localized Principal Component Analysis Learning for Face Feature Extraction and Recognition” – *Proceedings to the Workshop on 3D Computer Vision '97 pp124-128* – 1997

[KLT03] “Transformada de Karhunen-Loeve (KLT) y Transformada Discreta de Coseno (DCT)” – *Instituto de Ingeniería Eléctrica, Universidad de la República, Uruguay* – Octubre 2003

[KRE04] Jon Krueger, Marshall Robinson, Doug Kochelek, Matthew Escarra – “Face Detection Using Eigenfaces” – *The Connexions Project, Rice University, Houston, USA* – Diciembre 2004

- [KRU04] Jon Krueger, Marshall Robinson, Doug Kochelek, Matthew Escarra – “Obtaining The Eigenface Basis” – *The Connexions Project, Rice University, Houston, USA* – Diciembre 2004
- [LIA04] Yongmin Lia, Shaogang Gongb, Jamie Sherrahc, Heather Liddellb – “Support vector machine based multi-view face detection and recognition” – *Image Vision Computing 22 pp 413-427* – 2004
- [LIU01] J. Liu, F.B. Dazzo, O. Glagoleva, B. Yu, A.K. Jain – “CMEIAS: A Computer-Aided System for the Image Analysis of Bacterial Morphotypes in Microbial Communities” – *Microbial Ecology* – Febrero 2001
- [LUX05] Xiaoguang Lu, Anil K. Jain – “Integrating Range and Texture Information for 3D Face Recognition” – *Proc. 7th IEEE Workshop on Applications of Computer Vision (WACV'05) pp 156-163* – 2005
- [LUX06] Xiaoguang Lu, Anil K. Jain – “Automatic Feature Extraction for Multiview 3D Face Recognition” – *Proc. 7th IEEE International Conference on Automatic Face and Gesture Recognition (FC2006) pp 585-590* – Abril 2006
- [MAN02] Mandell, Douglas, Bennet – *Enfermedades infecciosas* – Editorial Médica Panamericana 5ta edición – 2002
- [MIC07] Ricardo Leal – “Micosis Superficiales” – *El Salvador*
<http://www.geocities.com/ralv7/micosup.htm>
- [MYC07] “Mycology Online” – *School of Molecular & Biomedical Science, University of Adelaide, Australia* – Junio 2007
<http://www.mycology.adelaide.edu.au/gallery/photos/Efloccosum1.html>
- [NUG06] Conor Nugent, Pádraig Cunningham – “Object Reconition and Active Learning in Microscope Images” – *Proceedings of 17th Irish Conference on Artificial Intelligence and Cognitive Science* – 2006
- [PAJ02] Gonzalo Pajares Martinsanz, Jesús M. de la Cruz García – *Visión por computador* – Alfaomega – 2002
- [PAZ00] M. Pazouki, T. Panda – “Understanding the morphology of fungi” – *Bioprocess Engineering 22 pp 127-143* – 2000
- [PER03] C. Pérez , M. A. Vicente, C. Fernández, O. Reinoso – “Aplicación de los diferentes espacios de color para detección y seguimiento de caras” – *Actas de las XXIV Jornadas de Automática, ISBN: 84-931846-7-5* – 2003
- [PIN98] F. A. C. Pinto, J. F. Reid – “Heading angle and offset determination using principal component analysis” – *ASAE Paper N° 98-3113* – 1998
- [PON02] José Pontón, Ma. Dolores Moragues, Josepa Gené – “Hongos y actinomicetos alergénicos” – *Revista Iberoamericana de Micología* – 2002
- [PUJ01] Albert Pujol, Jordi Vitrià, Felipe Lumbreras, Juan José Villanueva – “Topological principal component analysis for face encoding and recognition” – *Pattern Recognition Letters 22(6/7): 769-776* – 2001

- [QUE07] Víctor Manuel Quesada Ibarguien, Juan Carlos Vergara Schmalbach – *Estadística Básica con aplicaciones en Ms Excel* – Universidad de Cartagena, España – 2007
- [REY99] René Reynaga, Ramiro Aguilar, Homero Bañados, Rubén Cuarite – “Reconocimiento de Patrones a Partir de Imágenes Aéreas” – *Universidad de Salamanca, España* – 1999
- [RIC04] Marcela L. Riccillo, Ana S. Haedo, Natalia Debandi, Daniel Vazquez V. – “Comparación de Softwares Estadísticos” – *CLATSE VI Congreso Latinoamericano de Sociedades de Estadística – SAE Sociedad Argentina de Estadística, SOCHE Sociedad Chilena de Estadística Concepción, Chile* – Noviembre 2004
- [ROS06] Luigi Rosa – “FisherFaces for Face Recognition” - *Advanced Source Code .Com* – Enero 2006
<http://www.advancedsourcecode.com/facefaces.asp>
- [ROW98] H. A. Rowley, S. Baluja, T. Kanade – “Neural Network-based face detection” – *Pattern Analysis and Machine Intelligence IEEE Transactions on Volume 20 23-38* – Enero 1998
- [RUI05] J. Ruiz-del-Solar, P. Navarrete – “Eigenspace-based face recognition: a comparative study of different approaches” – *IEEE Transactions on Systems, Man and Cybernetics, Part C, Vol. 35, Issue 3, pp. 315-325* – Agosto 2005
- [SAN07] Francisco Sánchez Santaella, Luz María Roldán Vílchez – “Detección y Reconocimiento de Caras” – *Modelos de Inteligencia Artificial, Universidad de Granada* – 2007
<http://decsai.ugr.es/>
- [SEV06] “Reconocimiento de rostros” – *Departamento Matemática Aplicada 1, Universidad de Sevilla, España* – 2006
<http://alojamientos.us.es/gtocoma/pid/pid10/deteccioncaras.htm>
- [SHL05] Jonathon Shlens – “A Tutorial on Principal Component Analysis” – *Systems Neurobiology Laboratory, Institute for Nonlinear Science, University of California, San Diego, USA* – Diciembre 2005
- [SIL05] Víctor Silva – “Presente y Futuro en el Diagnóstico de las Micosis Invasivas” – *Laboratorio de Micología Médica, Programa de Microbiología y Micología, ICBM Facultad de Medicina, Universidad de Chile* – Agosto 2005
http://www.medwave.cl/cursos/Micologia2004/5/1.act?tpl=im_ficha_cursos.tpl
- [SIM05] J.P. Simmons, D.M. Dimiduk, M. De Graef – “Automatic particle coordination recognition using principal component analysis and kohonen neural nets” – *Microscopy and Microanalysis, vol. 11 (supplement 2), pp. 1634-1635* – 2005
- [SMI02] Lindsay I. Smith – “A Tutorial on Principal Components Analysis” – *University of Otago, New Zealand* – Febrero 2002
- [SPI70] Murray R. Spiegel – *Teoría y Problemas de Estadística* – Libros McGraw-Hill – 1970
- [SPS07] “SPSS - Statistical Product and Service Solutions” – *SPSS Inc., Chicago USA*

<http://www.spss.com>

[TKL05] “Transformada de Karhunen-Loeve” – *Departamento de Ingeniería Audiovisual y Comunicaciones, Universidad Politécnica de Madrid, España* – 2005

[TOR03] Antonio Torralba – “Statistic of Natural Image Categories” – *Network: Computation in Neural Systems Volume 14 Number 3* – Agosto 2003

[TRK91] M. Turk, A. Pentland – “Eigenfaces for recognition” – *Journal of Cognitive Neuroscience 3 (1): 71–86* – 1991

[TRP91] M. Turk, A. Pentland – “Face recognition using eigenfaces” – *Proc. IEEE Conference on Computer Vision and Pattern Recognition: 586–591* – 1991

[VAL94] Dominique Valentin, Hervé Abdi, Alice J. O’Toole, Garrison W. Cottrell – “Connectionist model of face processing: A survey” – *Pattern Recognition, N° 27 pp1209-1230* – 1994

[VAS05] M. Alex O. Vasilescu, Demetri Terzopoulos – “Multilinear Independent Components Analysis” – *Proc. Computer Vision and Pattern Recognition Conf. (CVPR ’05)* – 2005

[VBE07] “Companion Animal Health Care” – *Van Beek Global, USA*

<http://www.vanbeekglobal.com/text/univField.htm>

[VER98] K. Verpoulos, C. Campbell, G. Learmonth, B. Knight, J. Simpson – “The Automated Identification of Tubercle Bacilli using Image Processing and Neural Computing Techniques” – *Proceeding of the 8th International Conference on Artificial Neural Networks, vol2, pp 797-802* – 1998

[VIC02] M. A. Vicente, O. Reinoso, C. Pérez, J.A. Sabater, J.A. Azorín – “Reconocimiento de Objetos 3D Mediante Análisis PCA” – *XXIII Jornadas de Automática Tenerife ISBN: 84-699-8916-2* – 2002

[VID02] César Vidal, Juan Miguel García-Gómez, Luis Martí-Bonmatí, Alfons Juan, Montserrat Robles – “Clasificación de estirpes histológicas de tumores de partes blandas mediante reconocimiento de patrones a partir de imágenes de RM” – *Sociedad Española de Ingeniería Biomédica, CASEIB 2002, Libro de Actas (ISBN 84-600-9818-4) pp 439-442* – 2002

[VIL00] P. Rayón Villela – “Arquitectura para el Reconocimiento de Formas por Indexado en una Gran Base de Modelos: Aplicación al Reconocimiento de Rostros” – *Computación y Sistemas Vol. 3 N° 3 pp 214-219 ISSN 1405-5546* – 2000

[VIL06] Dr. J.J. Vilata Corell – *Micosis Cutáneas* – Editorial Médica Panamericana – 2006

[WID02] Kenneth W. Widmer, Kevin H. Oshima, Suresh D. Pillai – “Identification of Cryptosporidium parvum Oocysts by an Artificial Neural Network Approach” – *American Society for Microbiology Appl Environ Microbiol. 68 (3): 1115-1121* – Marzo 2002

[XIA04] Rong Xiao, Lei Zhang, and Hong-Jiang Zhang – “Feature Selection on Combinations for Efficient Learning from Images” – *Proc. of Asian Conference on Computer Vision (ACCV), Jeju Island, Korea* – 2004

-
- [YAN03] Jian Yang, Jing-yu Yang, Alejandro F. Frangi – “Combined Fisherfaces framework” – *Image and Vision Computing* 21 1037-1044 – 2003
- [YAN04] Jian Yang, David Zhang, Alejandro Frangi, Jing-yu Yang – “Two-Dimensional PCA: A New Approach to Appearance-Based Face Representation and Recognition” – *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol 26 No 1 131-137 – Enero 2004
- [ZHA03] W. Zhao, R. Chellappa, A. Rosenfeld, P.J. Phillips – “Face Recognition: A Literature Survey” – *ACM Computing Surveys*, pp. 399-458 – 2003
- [ZHA06] Daoqiang Zhang, Zhi-Hua Zhou, Songcan Chen – “Diagonal Principal Component Analysis for Face Recognition” – *Pattern Recognition Volume 39*, 140-142 – Enero 2006

Anexo A – Software utilizado

ImageJ

Para el tratamiento de las imágenes procesadas en este trabajo, en la elaboración de los preprocesamientos combinados con el método de *eigenfungi*, se utilizó el software *ImageJ Image Processing and Analysis in Java* [IMA07].

ImageJ es un programa público de procesamiento de imágenes inspirado por NIH Image para Macintosh.

Permite visualizar, editar, analizar, procesar, guardar e imprimir imágenes de 8, 16 y 32 bits. Puede leer imágenes en varios formatos incluyendo TIFF, GIF, JPEG, BMP, DICOM, FITS y RAW.

Soporta stacks, una serie de imágenes en una misma ventana a las que se le puede procesar y luego guardar en una sola operación.

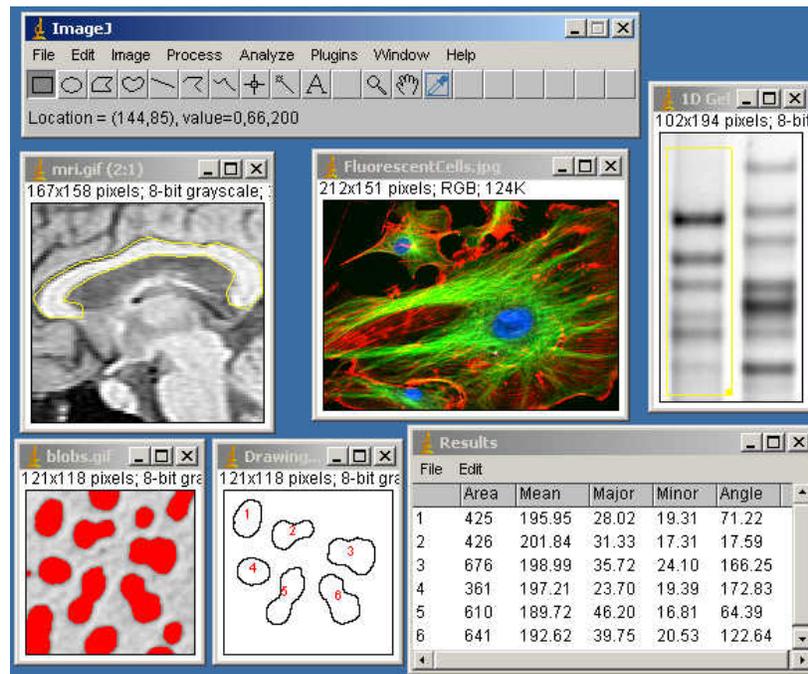


Fig. A.1 – Ejemplos de menú y ventanas de ImageJ

Las opciones del menú incluyen tratamiento de histograma, obtención de contornos, aplicación de filtros, Transformada de Fourier, transformación de imágenes (binarias, color), distorsión con ruido, etc.

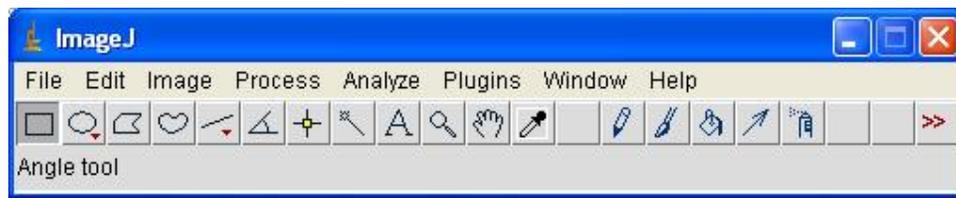


Fig. A.2 – Barra de menú de ImageJ

También permite el desarrollo de plugins en Java que permite su expansión a otros métodos de análisis de imágenes.

Fisherfaces

Para las pruebas con las fisherfaces se utilizó el software FisherFaces for Face Recognition de Luigi Rosa [ROS06] desarrollado en Matlab.

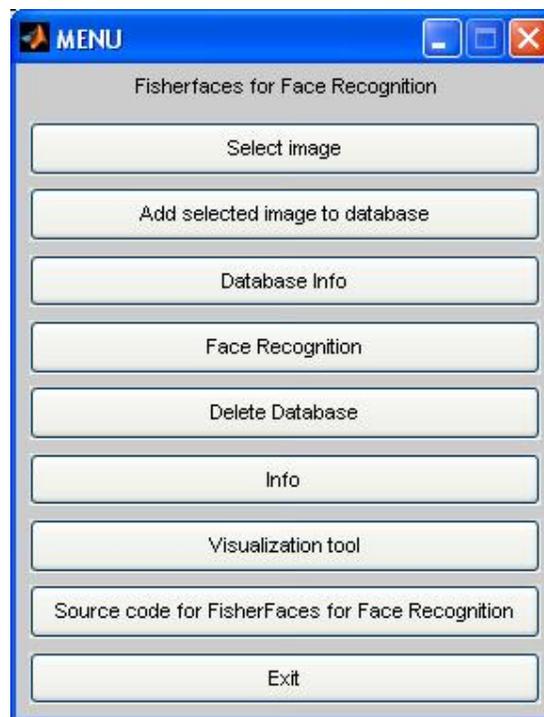


Fig. A.3 – Menú de opciones de Fisherfaces

Este software va creando una base a medida que se incorporan registros indicando con números reales la clase a la que pertenece cada uno. Por ejemplo, se puede considerar la especie *E. floccosum* como clase 1, *M. canis* como clase 2, etc.



Fig. A.4 – Ejemplo de información de composición de una base

Luego es posible reconocer la clase correspondiente a imágenes que no pertenecían al conjunto de entrenamiento, eligiendo la imagen y seleccionando “Face Recognition”. En pantalla indica la clase (en este caso la especie) correspondiente.

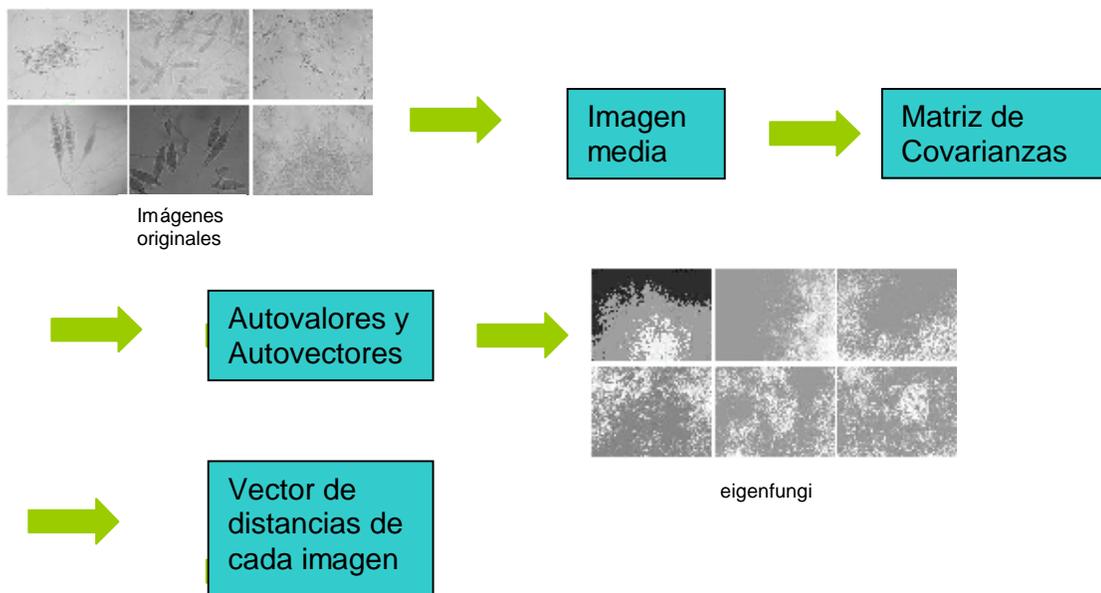
Resumen de Funciones

Opción	Descripción
Select image	Lee la imagen a a agregar a la base o a reconocer
Add selected image to database	La imagen seleccionada es agregada a la base y será utilizada durante el entrenamiento
Database Info	Muestra la información sobre las imágenes en la base
Face Recognition	Determina a qué clase pertenece la imagen seleccionada
Delete Database	Elimina la base armada
Info	Muestra la información sobre el las opciones del software
Visualization tool	Herramienta de visualización
Source code for Fisherfaces for Face Recognition	Redirige a la página Web del software
Exit	Sale del programa

Anexo B – Software desarrollado

Se desarrolló un software para realizar las pruebas a nivel binarias y totales. Este programa se realizó en Matlab, utilizándose rutinas obtenidas de la Universidad Drexel Philadelphia, USA [DRE07].

Veamos un resumen del método *Eigenfungi*



Para determinar la especie de una nueva imagen:

A cada imagen a reconocer se le calcula su vector de distancias y se compara con cada vector de las imágenes originales. El vector más cercano indica a qué individuo pertenece la nueva imagen.

1.1. Características de las imágenes

Para la elaboración y validación de la metodología, se estudiaron imágenes de hongos microscópicos de las seis especies principales de dermatofitos, obtenidas de muestras provistas por el Departamento de Micología del Instituto Nacional de Enfermedades Infecciosas (INEI), ANLIS “Carlos G. Malbrán”.

Las imágenes con extensión JPG originalmente medían 1600x1200 píxeles. Luego fueron pasadas a 160x120 píxeles y a escalas de grises.

1.2. Armado de directorios

Se requieren 2 directorios: *Inbox* y *Test*. Dentro de cada uno se arma un directorio por cada especie a reconocer.

Name	Size	Type
Inbox		File Folder
Test		File Folder

Name	Size	Type
E_floccosum_1		File Folder
M_canis_1		File Folder
M_gypseum_1		File Folder
T_mentagrophytes_1		File Folder
T_rubrum_1		File Folder
T_tonsurans_1		File Folder

Fig. B.1 – Distribución de directorios

1.3. Descripción del programa

Entrenamiento

Se definió una función *InputModel* que recorre los directorios dentro de *Input*, leyendo los archivos de las imágenes. De esta manera se arma:

- una matriz S con las imágenes con las cuales se entrenará el sistema
- un vector y_{clasif} que es el vector de clasificación que indica para cada imagen a qué especie pertenece. Cada especie es identificada con un número consecutivo desde 0

Luego se obtiene m_{image} que es la imagen media del conjunto.

Se transforma la matriz de píxeles S en la matriz dbx (tipo de dato *double*) y se obtiene la matriz de covarianzas C calculando $C = dbx * dbx'$.

En base a la matriz de covarianzas C , se halla un vector d de autovalores y uno v de autovectores.

Los autovalores se reordenan en forma ascendente, y en base a éstos se ordenan los autovectores correspondientes. Con los autovectores normalizados, se arma la matriz *Eigenvectors* con las *eigenfungi*.

Se toma cada imagen del conjunto de entrenamiento (que son parte de la matriz dbx) y se arma un vector de distancias, multiplicándola por cada *eigenfungi*. Con los vectores de distancias de todas las imágenes se arma una matriz *model*.

Testeo

Se utiliza la función *InputModel* que recorre los directorios dentro de *Test*, leyendo los archivos de las imágenes. De esta manera se arma:

- una matriz T con las imágenes a testear el sistema
- un vector *clasif_esperada* que es el vector de clasificación que indica para cada imagen a qué especie pertenece. Cada especie es identificada con un número consecutivo desde 0

Se transforma la matriz de píxeles T en la matriz $dbxT$ (tipo de dato *double*)

Se toma cada imagen del conjunto de Testeo (que son parte de la matriz $dbxT$) y se arma un vector de distancias, multiplicándola por cada *eigenfungi*. Con los vectores de distancias de todas las imágenes se arma una matriz *pesosT*.

Luego se compara cada vector de distancias de las imágenes de testeo con los vectores de distancias de las imágenes de entrenamiento, y se busca la más cercana (mediante distancia euclídea). De esta manera, dada una imagen de testeo, se le asigna en *vector_clasif* la especie correspondiente de la imagen de entrenamiento más cercana.

Para analizar la tasa de aciertos, se imprime en un archivo de texto la identificación de cada imagen de testeo, junto a su clasificación obtenida según *vector_clasif* y la especie real, según *clasif_esperada*.

1.4. Variantes del programa

Se armaron varias modificaciones del programa para realizar las diferentes pruebas:

Eigenfaces

Es el método original. Una vez hallado el vector de distancias de cada imagen de entrenamiento, se calcula el promedio de los vectores distancia de las imágenes pertenecientes a cada especie. Este vector se denomina *vector de clase*.

Multiespacios

Supongamos que se consideran m individuos, con k imágenes de cada uno para realizar el entrenamiento. Cuando se calcula la matriz de covarianzas, no se toman las $m*k$ imágenes, sino que se obtiene una matriz de covarianzas con el grupo de imágenes de cada individuo .

A partir de éstas se arman m conjuntos de *eigenfungi*. Luego se juntan todas las *eigenfungi* obtenidas y se prosigue con el método.

Manhattan

Al momento de comparar los vectores de distancias de las imágenes de testeo contra los de las imágenes de entrenamiento, en vez de utilizar distancia euclídea, se utiliza distancia Manhattan.

Multiespacios y Manhattan

Se combinan ambas variantes, armando un conjunto de eigenfungi por especie y comparando los vectores de distancia mediante distancia Manhattan.

1.5. Código Matlab *Eigenfungi*

Vemos a continuación el código utilizado para hallar y testear las *eigenfungi*.

```
function [distancias dist_min vector_clasif clasif_esperada nombres] =

pca_train(pathI nbox, pathTest);
um=100;
ustd=80;

%Se levantan las imagenes y se construye la matriz S con pixels.

[S y nombres width_img]=inputModel(pathI nbox,'PCA');
irow=width_img;
icol=size(S,1)/irow;
M=size(S,2);

%Muestra y (y_clasif vector de clasificación que indica a qué clase pertenece cada
%imagen).
y

%Normalización de las imágenes con respecto a condiciones de iluminación.
for i=1:size(S,2)
    temp=double(S(:,i));
    m=mean(temp);
    st=std(temp);
    S(:,i)=(temp-m)*ustd/st+um;
end

%Obtención de la imagen media m (m_image)
m=mean(S,2);

%Muestra la imagen media
tmimg=uint8(m);
img=reshape(tmimg,icol,irow);
img=img';
%imshow(img);
```

```

%title('Mean Image', 'fontsize',18)

%Transformación de la imagen
dbx=[];
for i=1:M
    temp=double(S(:,i));
    dbx=[dbx temp];
end

%Obtención matriz de Covarianza C=AA'y autovalores
A=dbx';
C=A*A';
% vv son los autovectores de C
% dd son los autovalores de C=dbx'*dbx;
[vv dd]=eig(C);

% Elimina autovalores los que sean cero
v=[];
d=[];
for i=1:size(vv,2)
    if(dd(i,i)>1e-4)
        v=[v vv(:,i)];
        d=[d dd(i,i)];
    end
end

%Reordena autovectores de manera ascendente
[B index]=sort(d);
ind=zeros(size(index));
dtemp=zeros(size(index));
vtemp=zeros(size(v));
len=length(index);
for i=1:len
    dtemp(i)=B(len+1-i);
    ind(i)=len+1-index(i);
    vtemp(:,ind(i))=v(:,i);
end
d=dtemp;
v=vtemp;

%Normalización de los autovectores
for i=1:size(v,2)
    kk=v(:,i);
    temp=sqrt(sum(kk.^2));
    v(:,i)=v(:,i)./temp;
end

%matriz de eigenfungi U (eigenvectors)
u=[];
for i=1:size(v,2)
    temp=sqrt(d(i));
    u=[u (dbx*v(:,i))./temp];
end

```

```

    for i=1:size(u,2)
        kk=u(:,i);
        temp=sqrt(sum(kk.^2));
        u(:,i)=u(:,i)/temp;
    end

    % Encuentra el peso de cada imagen de entrenamiento y arma el vector model
    model = [];
    for h=1:size(dbx,2)
        WW=[];
        for i=1:size(u,2)
            t = u(:,i)';
            WeightOfImage = dot(t,dbx(:,h)');
            WW = [WW; WeightOfImage];
        end
        model = [model WW];
    end;

    %Eigenfungi
    eigenvectors=u;
    %Imagen media.
    m_image=m;
    %vector de clasificación
    y_clasif=y;

    %function [distancias dist_min vector_clasif
    %clasif_esperada]=pca_use(pathTest,model,eigenvectors,m_image,y_clasif);

    %Se levantan las imágenes de testeo en la matriz T
    % y es el vector de clasificación (clasif_esperada)
    [T y nombres width_img]=inputModel(pathTest,'PCA');
    irow=width_img;
    icol=size(T,1)/irow;
    nTest=size(T,2);

    %Se normalizan condiciones de iluminación
    for i=1:size(T,2)
        temp=double(T(:,i));
        me=mean(temp);
        st=std(temp);
        T(:,i)=(temp-me)*ustd/st+um;
    end

    % Transformación de imágenes
    dbxT=[]; % A matrix
    for i=1:nTest
        temp=double(T(:,i));
        dbxT=[dbxT temp];
    end

    distancias=[];
    dist_min=[];

```

```

vector_clasif=[];

%Obtiene los pesos de cada una de las imagenes de test y arma pesosT
for j=1:nTest
    InputImage=T(:,j);

    NormImage=dbxT(:,j);
    Difference=NormImage-m_image;

    p = [];
    aa=size(eigenvectors,2);
    for i = 1:aa
        pare = dot(NormImage',eigenvectors(:,i));
        p = [p; pare];
    end

    pesosT = [];
    for i=1:size(eigenvectors,2)
        t = eigenvectors(:,i)';
        WeightOfInputImage = dot(t,Difference');
        pesosT = [pesosT; WeightOfInputImage];
    end

    ll = 1:size(eigenvectors,2);

% Compara los pesos de las imágenes de test con las de entrenamiento
e=[];
for i=1:size(model,2)
    q = model(:,i);
    DiffWeight = pesosT-q;
    mag = norm(DiffWeight);
    e = [e mag];
end

distancias=[distancias e'];

%Determina la especie a la que pertenece cada imagen de test según la de
%entrenamiento más cercana

    dist_min=[dist_min min(e)];
    MaximumValue=max(e)
    MinimumValue=min(e)
    imagenparecida = find(e == min(e));
    grupo=y_clasif(imagenparecida);
    vector_clasif=[vector_clasif;grupo];

end
clasif_esperada=y;

```

Anexo C – Ejemplos de PCA y LDA

Análisis de Componentes Principales (PCA)

En esta sección presentamos un ejemplo práctico de un caso de aplicación del Análisis de Componentes Principales.

Para el ejemplo se procesaron los datos con un software de análisis estadístico llamado Infostat [INF07], desarrollado por la Universidad de Córdoba.

La base de datos utilizada *IRIS* [IRI05] representa un conjunto de características de tres especies de la flor iris: largo y ancho de los sépalos, largo ancho de los pétalos.

Las 3 especies estudiadas son:



iris setosa



iris versicolor



iris virginica

Las variables que conforman la base son:

- Especie – Setosa, Versicolor, Virginica
- SepalLen – largo del sépalo
- SepalWid – ancho del sépalo

- PetalLen – largo del pétalo
- PetalWid – ancho del pétalo

Para caracterizar las variables calculamos las medidas de dispersión

Variable	N	Media	Desviación estándar	Mínimo	Máximo
SepalLen	151	5,84	0,83	4,30	7,90
SepalWid	151	3,06	0,44	2,00	4,40
PetalLen	151	3,74	1,77	1,00	6,90
PetalWid	151	1,19	0,76	0,10	2,50

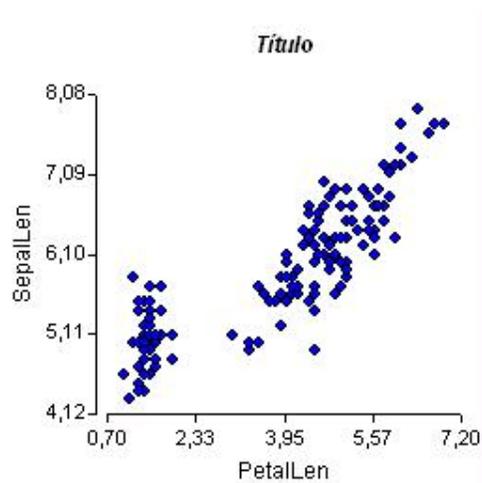
La primera parte del método consiste en estudiar las correlaciones, dado que el método resulta más eficiente ante la presencia de altas correlaciones. Para eso obtenemos el *coeficiente de correlación de Pearson* de cada par de variables

	SepalLen	SepalWid	PetalLen	PetalWid
SepalLen	1,00			
SepalWid	-0,12	1,00		
PetalLen	0,87	-0,44	1,00	
PetalWid	0,82	-0,38	0,96	1,00

Vemos que existen altas correlaciones:

PetalLen – SepalLen con una correlación de 0,87. También PetalWid – SepalLen con 0,82. Además PetalWid con PetalLen con 0,96.

Si analizamos las correlaciones a través de nubes de puntos (o gráficos de dispersión) vemos estas relaciones en forma gráfica. Por ejemplo, SepalLen con PetalLen:



Si aplicamos el método de Componentes Principales con las 4 variables numéricas, obtenemos los 4 autovalores correspondientes

Lambda	Valor	Proporción	Prop. Acum.
1	2,92	0,73	0,73
2	0,91	0,23	0,96
3	0,15	0,04	0,99
4	0,02	0,01	1,00

El primer autovalor explica el 73% de la variabilidad del sistema. Entre los dos primeros autovalores explican el 96% de la variabilidad.

Según esto y el criterio que indica que uno puede preservar los autovalores que expliquen más del 80%, nos quedaríamos con los dos primeros ejes, descartando el resto de las componentes.

A partir de los autovalores, se calculan los autovectores (o *eigenvectors*, por lo que se representan como “ei”).

Variables	e1	e2	e3	e4
SepalLen	0,52	0,39	0,72	-0,26
SepalWid	-0,27	0,92	-0,25	0,12
PetalLen	0,58	0,03	-0,14	0,80
PetalWid	0,56	0,07	-0,63	-0,52

Los autovectores representan la nueva base de ejes de coordenadas, donde se proyectarán los datos originales.

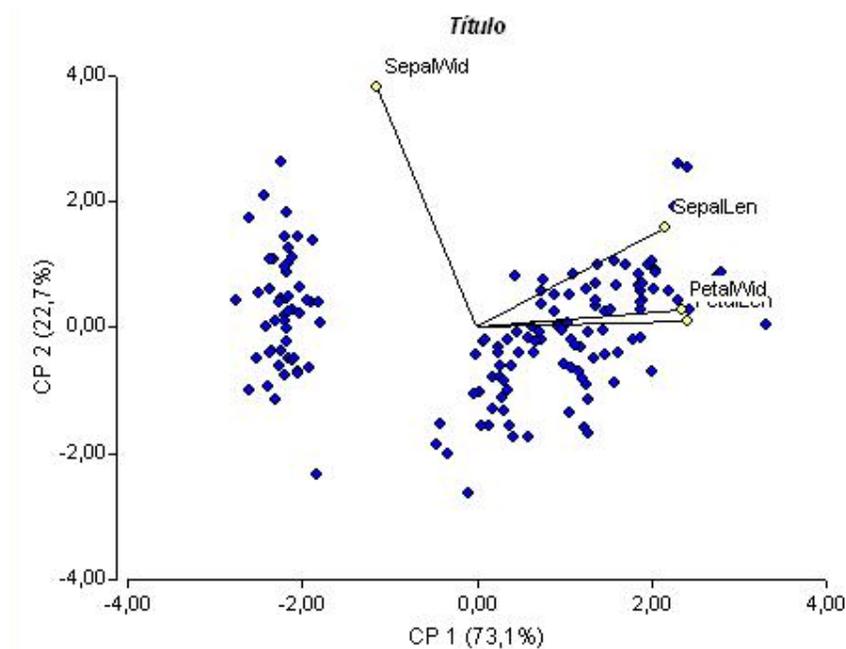
Podemos ver que las variables que más aportan al primer eje, según el valor del coeficiente de cada autovector, son SepalLen, PetalLen y PetalWid. Al segundo eje es SepalWid quien más contribuye.

Para hallar las nuevas componentes, se utilizan las siguientes fórmulas:

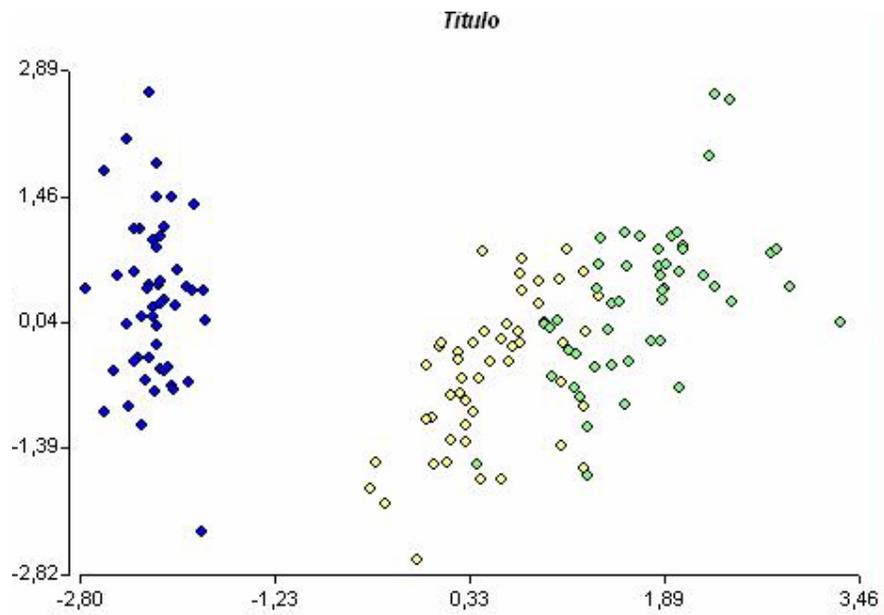
$$CP1 = 0,52 * SepalLen - 0,27 * SepalWid + 0,58 * PetalLen + 0,56 * PetalWid$$

$$CP2 = 0,39 * SepalLen + 0,92 * SepalWid + 0,03 * PetalLen + 0,07 * PetalWid$$

Para las 2 primeras nuevas componentes, obtenemos el siguiente gráfico



Podemos diferenciar en este gráfico las 3 especies de flores analizadas, para ver si hay o no relaciones entre las variables.



Podemos ver gráficamente las agrupaciones de datos según las especies. Esto hubiera sido imposible en un gráfico con las variables originales, dado que son más de 3 dimensiones.

Análisis Discriminante Linear de Fisher (LDA)

En esta sección presentamos un ejemplo práctico de un caso de aplicación del Análisis Discriminante Lineal.

Para el ejemplo se procesaron los datos con el software de análisis estadístico SPSS [SPS07].

La base de datos utilizada *IRIS* [IRI05] representa un conjunto de características de tres especies de la flor iris: largo y ancho de los sépalos, largo ancho de los pétalos.

Con respecto a las funciones discriminantes canónicas, observamos 2 autovalores

Autovalores				
Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	32,192(a)	99,1	99,1	,985
2	,285(a)	,9	100,0	,471

a Se han empleado las 2 primeras funciones discriminantes canónicas en el análisis.

Siendo que el primero explica un 99,1% de la varianza. El segundo autovalor aporta poca información discriminante.

Los coeficientes de las funciones canónicas son:

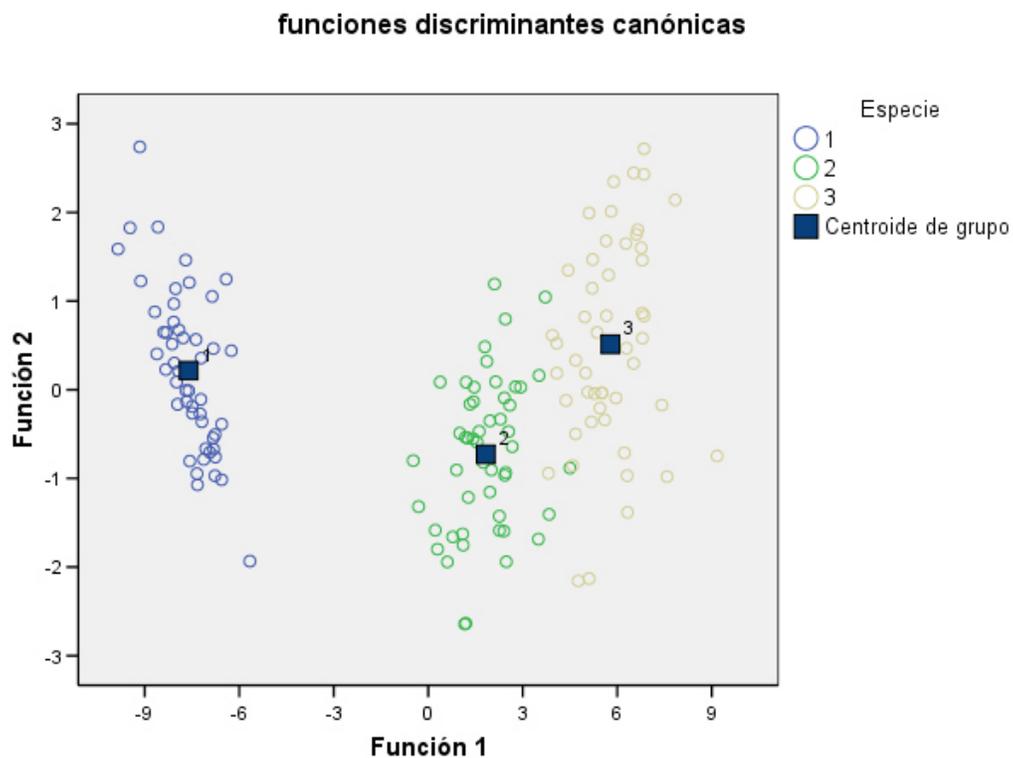
	Función	
	1	2
	SepalLen	-,427
SepalWid	-,521	,735
PetalLen	,947	-,401
PetalWid	,575	,581

Según esas funciones, los centroides de cada grupo son los siguientes

Funciones en los centroides de los grupos	
Especie	Función

	1	2
1	-7,608	,215
2	1,825	-,728
3	5,783	,513

Vemos esta información gráficamente



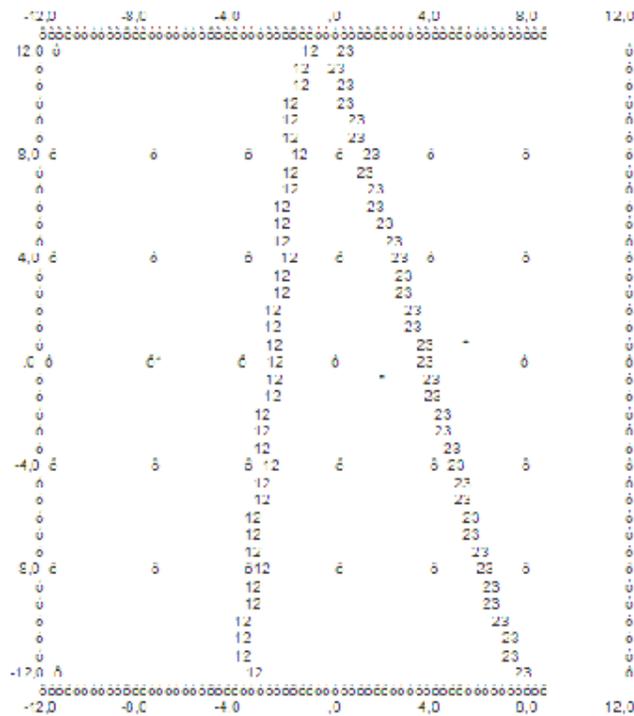
Se observan los tres grupos bien diferenciados. El grupo 1 correspondiente a la especie Setosa está más separado de los otros dos.

Para clasificar una flor que no pertenecía al conjunto de entrenamiento, utilizamos las funciones de clasificación, una por grupo

Coeficientes de la función de clasificación	
	Especie

	1	2	3
SepalLen	23,544	15,698	12,446
SepalWid	23,588	7,073	3,685
PetalLen	-16,431	5,211	12,767
PetalWid	-17,398	6,434	21,079
(Constante)	-86,308	-72,853	-104,368
Funciones discriminantes lineales de Fisher			

Según estas funciones, podemos visualizar la distribución de los puntos en un mapa territorial, donde se definen en forma gráfica las fronteras entre grupos.



Los resultados de la clasificación se visualizan en una matriz de confusión, donde se puede analizar el porcentaje de individuos bien clasificados.

Resultados de la clasificación(a)						
		Especie	Grupo de pertenencia pronosticado			Total
			1	2	3	

Original	Recuento	1	50	0	0	50
		2	0	48	2	50
		3	0	1	49	50
	%	1	100,0	,0	,0	100,0
		2	,0	96,0	4,0	100,0
		3	,0	2,0	98,0	100,0
a Clasificados correctamente el 98,0% de los casos agrupados originales.						

Vemos que quedaron mal clasificadas 2 flores del grupo 2 y una del grupo 3. Por lo que el porcentaje de clasificación correcta del grupo 2 es del 96% y del grupo 3 del 98%. Haciendo un total de aciertos del 98% de los casos agrupados originales.