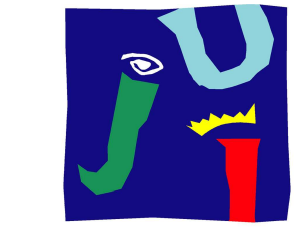


Departamento de Lenguajes y Sistemas Informáticos

Face Gender Classification under Realistic Conditions

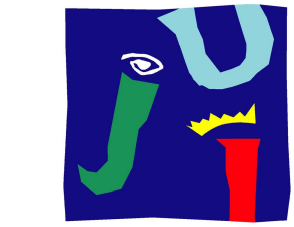
Dealing with Neutral, Expressive and Partially Occluded Faces

Yasmina Andreu Cabedo

Supervised by: Dr. Pedro García Sevilla
Dr. Ramón A. Mollineda Cárdenas

A thesis submitted
to the Universitat Jaume I
for the degree of Doctor of Philosophy

September, 2014



Departamento de Lenguajes y Sistemas Informáticos

Clasificación Facial de Género en Condiciones Realistas

Tratando con Imágenes de Caras Neutras, Expresivas y Parcialmente Ocluidas

Yasmina Andreu Cabedo

Dirigida por: Dr. Pedro García Sevilla
Dr. Ramón A. Mollineda Cárdenas

Septiembre, 2014

To my parents and sister
for their unconditional support.

A mis padres y hermana
por su apoyo incondicional.

Contents

List of Figures

List of Tables

Acknowledgements

Over the years dedicated to the completion of this work, I have had the pleasure of meeting many people who I have a lot to thank for. I would like to dedicate this section to all those people who have helped me grow professionally as well as personally and who have made possible all the good memories I have from my time as a PhD student.

First, I would like to thank my supervisors, Dr. Pedro García Sevilla and Dr. Ramón A. Mollineda. Ramón, thanks for believing in me and giving me the opportunity to work by your side and be part of the research group since the beginning of my Master's final project. Without that chance, quite possibly I wouldn't be where I am now. Pedro, you not only have taught me about computer vision and teaching, but you have always encouraged me and showed me the bright side of everything. Going into your office thinking "what awful results!" and coming out smiling and eager to continue working was a regular occurrence. If all the supervisors were like you, PhD Comics would not make sense.

To all my colleagues in the Vision group, without you all this time would have passed much more slowly. Olga, thanks for all the shared laughs and chit-chats and also for making my life much easier regarding paperwork. To Javi, Eva, Raúl and all the others who made coffee time one of the best moments of the day. A special mention is deserved for my officemate of the last 4 years, Rubén. After all those hours together and all the conversations from which I have learned and improved as a person, I made a true friend.

Thanks to Prof. Aleix M. Martínez and Dr. Neill Campbell for hosting my research stays and allowing me to be part of their research groups. I would like to especially name those who made me feel at home when in fact I was in Ohio: Sam, Fabian, Felipe and Paulo, or in Bristol: Alex and Jason. Jason, thanks for being so patient and answering my more than a few questions about the English language, you sure are a good friend. Alex, thanks for our endless conversations from which I always learned something. You are a great friend who, I know, will always be there for me.

Dani and Silvana, thanks for all those good moments that allowed me to disconnect. To Paula and Natalia, for making possible those partying times which seemed impossible for a PhD student to have. To Dalia, for always making me feel like a superwoman. To Jose María, for showing me that most things could be seen from another perspective and

for always being there when I needed him.

Finally, my most sincere gratitude goes to my family. Mum and Dad, I would never be able to thank you enough for all that you have done for me. Thank you so much for always believing in me and supporting me in all my decisions. Without you both, I would not be who I am today. Patri, thanks for being my best friend, for always encouraging me and making me think I can do anything and for taking part in our silliness since we were kids. You are the best sister one could wish for. Thanks also to Lia who has made me laugh and enjoy much more that I could have ever imagined.

Financial Support

Thanks to the University Jaume I for funding the last four years of my work, and to the Ministry of Science and Innovation and the Foundation Caixa Castelló-Bancaixa for financially supporting the previous years.

Agradecimientos

Durante los años dedicados a la realización de este trabajo, he tenido el placer de conocer a muchas personas a las que les tengo mucho que agradecer. Esta sección quiero dedicarla a todos aquellos que me han ayudado a crecer tanto profesional como personalmente y han hecho posible que tenga muy buenos recuerdos de mi época como estudiante de doctorado.

En primer lugar quisiera dar las gracias a mis directores, Dr. Pedro García Sevilla y Dr. Ramón A. Mollineda. Ramón, gracias por haber confiado en mi y darme la oportunidad de trabajar contigo y formar parte del grupo de investigación desde el Trabajo Final de Máster. Muy posiblemente, sin esa oportunidad no estaría ahora donde estoy. Pedro, no sólo me has enseñado sobre visión y docencia, sino que siempre me has animado y mostrado el lado bueno de las cosas. Entrar a tu despacho pensando "¡qué malos son estos resultados!" y salir con una sonrisa y ganas de continuar era de lo más habitual. Si todos los directores de tesis fueran como tu, la tira cómica PhD Comics no tendría ningún sentido.

A todos mis compañeros del grupo de visión, sin vosotros todo este tiempo habría pasado mucho más despacio. Olga, gracias por las risas y buenos ratos que compartimos y por haber hecho que los temas de papeleo fueran pan comido. A Javi, Eva, Raúl y todos los demás que hicieron que el momento del café fuera uno de los mejores del día. Se merece una mención especial mi compañero de despacho de los últimos 4 años, Rubén. Después de tantas horas y tantas conversaciones que me han hecho aprender y mejorar como persona, me llevo un gran amigo.

Gracias a Prof. Aleix M. Martínez y a Dr. Neill Campbell por supervisar mis estancias en el extranjero y permitirme formar parte de sus grupos de investigación. Especialmente, me gustaría nombrar a los que me hicieron sentir como en casa cuando en realidad me encontraba en Ohio: Sam, Fabian, Felipe y Paulo, o en Bristol: Alex y Jason. Jason, thanks for being so patient and answering my more than a few questions about the English language, you sure are a good friend. Alex, thanks for our endless conversations from which I always learned something. You're a great friend who, I know, will always be there for me.

Dani y Silvana, gracias por todos esos ratos que me han permitido desconectar. A Paula y Natalia, por hacer posible momentos de risas y cachondeo que parece que un doctorando no pueda tener. A Dalia, por siempre hacerme sentir superwoman. A Jose María, por

hacerme ver las cosas de otra manera y estar ahí siempre que le he necesitado.

Por último, mi más sincero agradecimiento va para mi familia. Mamá y papá, nunca os podré agradecer lo suficiente todo lo que habéis hecho por mi. Muchísimas gracias por siempre confiar en mi y apoyarme en todas mis decisiones. Sin vosotros, no sería quien soy. Patri, gracias por ser mi mejor amiga, por siempre darme ánimos y hacerme pensar que puedo con todo y por todas las tonterías que compartimos desde niñas. Eres la mejor hermana que se puede desear. Gracias también a Lia que me ha hecho reir y disfrutar mucho más de lo que jamás hubiera podido imaginar.

Financiación

Gracias a la Universidad Jaume I por haber financiado los últimos cuatro años de mi trabajo y al Ministerio de Ciencia e Innovación y a la Fundación Caixa Castelló-Bancaixa por la financiación de años anteriores.

Abstract

This thesis is focused on face gender classification under more realistic scenarios than those traditionally considered in the literature. In real environments, many problems can arise due to the lack of control over the subjects and their surroundings. Moreover, the individuals' characteristics, such as age or race, can significantly vary. At the same time, the subjects can express their emotions by means of facial expressions as well as wear pieces of clothing covering their faces both of which would result in face images with distortions. These are the main problems that we tackle in this thesis together with some other complications related to having different illumination conditions and inaccurate face detections.

Firstly, we study the possibility of classifying gender from individual face parts, such as the eyes, the nose, the mouth and the chin. From experimental results obtained over two datasets, we concluded that the eyes are the most reliable region and that different face parts provide complementary information about the gender of the person.

Secondly, we propose a novel type of local features (Ranking Labels) and a new classification approach based on local neighbourhoods. The proposed features, which describe local regions of the image, are based on contrast values and they maintain a certain amount of the spatial information. The classification method consists in an ensemble of classifiers where each base learner specialises in a specific region of the image. We provide a comprehensive analysis of the behaviour of the proposed techniques and some state-of-the-art methods when utilising images of neutral and expressive faces and also faces covered by sun glasses and scarves in different training and testing combinations. The empirical results indicated that all the solutions performed similarly when the training and test images had the same characteristics. However, when the training and test sets contained different types of images, our methods showed a more robust behaviour than the rest.

Additionally, we perform a statistical study on the influence of the image resolution in face gender classification using ten image resolutions going from an extremely low resolution to the highest resolution available in the datasets. The optimal image resolutions were found in the range 22×18 to 90×72 pixels. However, an image resolution as low as 3×2 pixels provided useful information to distinguish between genders.

Resumen

Esta tesis se centra en la clasificación de género a partir de imágenes faciales tratando el problema con un enfoque más realista que el tradicionalmente utilizado en la literatura. En entornos reales, pueden surgir varios problemas debido a la falta de control sobre los sujetos y su entorno. Además es probable que, las características de los individuos, como son su edad y raza, varíen significativamente. Al mismo tiempo, los sujetos pueden manifestar sus emociones mediante expresiones faciales así como llevar puestos complementos que cubran partes de su cara, lo cual provoca que las imágenes faciales contengan ciertas distorsiones. Estos son los principales problemas, junto con otras complicaciones como las causadas por cambios de iluminación y detecciones imprecisas de la cara, que abordamos en este trabajo.

Comenzamos estudiando la posibilidad de clasificar el género dadas partes de la cara, como son los ojos, la nariz, la boca y el mentón. A partir de los resultados experimentales que se obtuvieron utilizando dos bases de datos de imágenes faciales, concluimos que los ojos eran la región de la cara que proporcionaba resultados más robustos y que distintas partes de la cara contienen información complementaria sobre el género de la persona.

Seguidamente, propusimos un nuevo tipo de características locales y un método de clasificación basado en vecindades. Las características propuestas se basan en valores de contraste locales, aunque manteniendo información espacial. El método de clasificación consiste en una combinación de clasificadores donde cada clasificador base se especializa en una región concreta de la cara. Ambas propuestas se compararon con las técnicas más utilizadas en este campo mediante un completo análisis experimental utilizando imágenes de caras neutras y expresivas y también imágenes de caras con gafas de sol y bufandas. Los resultados empíricos indican que todas las soluciones resuelven la tarea de forma estadísticamente equivalente cuando las imágenes de entrenamiento y test tienen las mismas características. Sin embargo, cuando los conjuntos de entrenamiento y test contienen imágenes de distintos tipos, nuestras propuestas muestran un comportamiento más robusto que el resto.

Por último, presentamos un estudio estadístico de la influencia de la resolución de las imágenes en la clasificación de género. Los resultados mostraron que las resoluciones óptimas están entre 22×18 y 90×72 píxeles. Sin embargo, imágenes de sólo 3×2 píxeles proporcionan información útil para comenzar a distinguir entre géneros.

Sinopsis de la Tesis

This chapter fulfils a requirement of the Spanish PhD regulation RD 99/2011, which states the criteria to obtain international doctorate certification. In particular, it specifies that part of the thesis has to be written in one of the official languages of Universitat Jaume I, which are Spanish and Valencian. Thus, the aim of the following sections is to summarize in Spanish the previous chapters that have been reported in English, including motivation and objectives, contributions, conclusions and future work lines.

Este capítulo cumple con la normativa de los estudios de doctorado regulados por el RD 99/2011, que establece los criterios necesarios para obtener la mención internacional en el título de Doctor. Concretamente, ésta establece que parte de la tesis debe ser escrita en un idioma oficial de la Universitat Jaume I. Por tanto, ya que la tesis se ha redactado en inglés, el objetivo de este capítulo es presentar una visión global de la tesis en español, incluyendo motivación y objetivos, contribuciones, conclusiones y trabajo futuro.

Introducción

Nuestras caras proporcionan una gran cantidad de información sobre nosotros mismos. Aparte de nuestra identidad, también indican muchos otros rasgos demográficos como son nuestro género, edad y raza. En la actualidad, conseguir de forma automática este tipo de información puede ser de gran utilidad en incontables tareas. En particular, el conocer el género de las personas puede utilizarse en estudios dinámicos de mercado, en sistemas de vigilancia y seguridad, en la interacción de máquinas y humanos y en servicios personalizados en un gran número de negocios. Por otro lado, esta información también puede servir de filtro en sistemas de reconocimiento automático. Por ejemplo, en sistemas de reconocimiento biométrico el conocimiento del género del individuo podría reducir el espacio de búsqueda a la mitad. Estos son tan solo algunos ejemplos de las muchas aplicaciones posibles de los sistemas automáticos de clasificación de género.

Aunque reconocer el género de una persona a partir de una imagen facial puede resultar relativamente sencillo para los humanos, puede resultar una tarea compleja en el caso de sistemas automáticos. Estos sistemas se enfrentan a un número considerable de dificultades que a los humanos nos pasan desapercibidas. Nosotros podemos reconocer el género de caras que vemos tanto de cerca como de lejos, es decir de caras de distintos tamaños. Además, no nos importa si las proporciones de la cara no son exactamente las habituales. Por ejemplo, podemos seguir distinguiendo el género de una persona con la nariz más grande o los ojos más pequeños de lo habitual. Nuestra habilidad tampoco se ve afectada por los cambios de iluminación, ya que podemos distinguir hombres y mujeres tanto de día como de noche. Incluso podemos realizar esta tarea cuando las personas muestran emociones, es decir, con caras alegres, disgustadas, etc., o cuando partes de las caras están ocultas debido al uso de accesorios como las gafas de sol o bufandas.

Todas las anteriores complicaciones deben ser tratadas por los sistemas automáticos de reconocimiento de género. En esta tesis nos centramos en mejorar los principales problemas que surgen al reconocer el género automáticamente a partir de imágenes faciales en situaciones más realistas de lo habitual en este campo de investigación.

Motivación y Objetivos

El problema de reconocimiento automático de género a partir de imágenes faciales sigue sin estar completamente resuelto en escenarios reales. La mayoría de los trabajos publicados abordan este problema dando por hecho que se cumplen una serie de condiciones que rara vez están presentes en situationes reales. Principalmente, estos trabajos asumen que la iluminación es similar en todas las imágenes y que la mayoría de los individuos pertenecen a los mismos grupos demográficos. Además, gran parte de estos estudios no tienen en cuenta imágenes donde las caras pueden mostrar expresiones o estar parcialmente ocluidas, lo cual ocurre muy comúnmente en entornos reales.

El principal objetivo de esta tesis es mejorar el reconocimiento automático de género en escenarios donde pueden darse las complicaciones anteriores. Con este propósito, se han revisado los pasos del proceso de clasificación y se han propuesto mejoras para cada uno de ellos. Este fin general puede ser dividido en los siguientes objetivos más específicos:

- Estudiar el rol de las diferentes partes de la cara en el reconocimiento de género y la posibilidad de abordar el problema con únicamente la información proporcionada por algunas de esas partes.
- Definir nuevos tipos de características para describir imágenes faciales que proporcionen información fiable, incluso cuando la iluminación de las imágenes varía.
- Diseñar nuevos métodos de clasificación que sean robustos en presencia de imprecisiones en la detección de las caras y de proporciones faciales variables.
- Analizar cómo influye la resolución de la imagen en los resultados de los sistemas automáticos de reconocimiento de género.

- Comprobar la precisión de las soluciones propuestas con imágenes faciales de personas de muy variada demografía (es decir, personas que cubren un amplío rango de edades y razas) y además con caras expresivas y parcialmente ocluidas.
- Porporcionar conclusiones respaldadas por tests estadísticos.

Contribuciones

El trabajo presentado en esta tesis se centra en el reconocimiento automático de género a partir de imágenes faciales. Con la finalidad de mejorar el rendimiento general de estos sistemas, hemos tratado los principales problemas encontrados en cada uno de los pasos del proceso. Para ello, hemos diseñado nuevos métodos para describir el contenido de las imágenes y para clasificar el género de nuevas caras en condiciones realistas. Para comprobar si las técnicas propuestas son adecuadas, se presentan estudios experimentales comparando dichas propuestas con soluciones ampliamente utilizadas en este campo de investigación. Seguidamente, se detallan las contribuciones de nuestro trabajo.

- Evaluamos rigurosamente el rol de ocho partes de la cara en la clasificación de género. La capacidad discriminante de estas partes es estudiada usando varios clasificadores y dos bases de datos para comprobar si su comportamiento es consistente.
- Proporcionamos resultados empíricos que muestran la existencia de información complementaria entre distintas partes de la cara.
- Proponemos nuevas características, llamadas *Ranking Labels*, que describen regiones locales de las imágenes. Estas características son robustas con respecto a cambios en la iluminación ya que se basan en valores de contraste local. Por tanto, proporcionan una descripción independiente de los niveles de gris de la imagen. De forma adicional, la caracterización vectorial usando *Ranking Labels* retiene información espacial lo cual permite hacer frente a imprecisiones en la detección de la cara.
- Presentamos una nueva metodología de clasificación basada en vecindades locales que proporciona cierto nivel de tolerancia hacia caras con distintas proporciones faciales, caras mal alineadas e imprecisiones en la detección de la cara.
- Comparamos experimentalmente mediante el uso de test estadísticos varias técnicas de clasificación de género (incluyendo las propuestas en esta tesis) utilizando imágenes de caras sin expresión, expresivas y parcialmente ocluidas en todas las combinaciones posibles de entrenamiento y test.
- Estudiamos exhaustivamente la influencia de la resolución de la imagen facial en la clasificación automática de género. Se incluye en el estudio un amplio rango de resoluciones de imagen partiendo de resoluciones extremadamente bajas hasta llegar a las más altas posibles.

- Mostramos resultados empíricos que revelan cuál es el menor tamaño de imagen que contiene información útil para diferenciar entre géneros.

- Apoyamos las conclusiones extraídas de nuestros estudios en tres test estadísticos: el estadístico de Iman-Davenport, el método de Holm y la prueba de los rangos con signo de Wilcoxon.

- Evaluamos la robustez de las soluciones presentadas con respecto a cambios en las condiciones de adquisición de las imágenes y de las variables demográficas de los sujetos mediante experimentos cruzando bases de datos.

Resumen del Trabajo Desarrollado

Esta tesis ha revisado los pasos del proceso de clasificación de género a partir de imágenes faciales y ha propuesto mejoras a cada uno de ellos teniendo en cuenta que se quiere abordar el problema en un entorno más realista de lo que es habitual en los trabajos relacionados. En esta sección, se resumen las soluciones presentadas así como los resultados de los distintos estudios empíricos realizados.

El principal objetivo de esta tesis ha sido mejorar la clasificación automática de género en escenarios razonablemente realistas. Bajo estas condiciones pueden surgir problemas debido a la falta de control sobre el entorno y sobre las personas cuyo género se quiere reconocer. Principalmente, nos hemos centrado en las dificultadas encontradas cuando las imágenes faciales presentan distorsiones causada por expresar emociones o llevar accesorios que cubren amplias zonas de la cara como son las gafas de sol y las bufandas. Para afrontar las complicaciones indicadas, hemos presentado nuevas técnicas para realizar cada uno de los pasos del proceso de clasificación junto con pruebas empíricas de la idoneidad de los métodos propuestos. Además, hemos estudiado experimentalmente cómo varios factores (como son la resolución de la imagen o las imprecisiones en la detección de la cara) afectan a las tasas de reconocimiento de género.

Nuestro primer estudio se centró en detectar qué partes de la cara contienen información más discriminante a la hora de distinguir entre géneros. Hemos estudiado el rol de ocho partes de la cara distintas comparando el rendimiento de varios clasificadores entrenados con partes individuales (ojos, nariz, boca y mentón). En esta comparación, también se han incluído clasificadores basados en características holísticas (en concreto, la zona interna de la cara, la externa y la cara completa). Los resultados mostraron que las partes individuales proporcionan suficiente información para reconocer el género de la persona, aunque las características holísticas siempre condujeron a mejores resultados. De todas las partes individuales, los ojos fueron la región que proporcionó los resultados más robustos ya que siempre consiguieron las tasas de reconocimiento más altas considerando dos bases de datos distintas. El hecho de que los clasificadores basados en una única parte de la cara fueran capaces de distinguir entre géneros hizo que nos plantearamos si distintas partes de la cara contienen información complementaria. Para abordar esta cuestión, utilizamos combinacio-

nes de clasificadores donde cada clasificador base aprendía de una parte diferente de la cara. Los resultados experimentales mostraron que las combinaciones de clasificadores basadas en 3 partes alcanzaban resultados similares a aquellas que utilizaban 5 partes. Estos datos sugieren que los problemas de clasificación de género podrían ser resueltos satisfactoriamente incluso cuando no toda la cara es visible.

Seguidamente nos propusimos tratar de resolver el problema en situaciones donde la demografía de los individuos es diversa, las caras pueden presentar distorsiones locales y la detección de las caras puede ser imprecisa. Por un lado, diseñamos un nuevo tipo de características locales (Ranking Labels) para describir las imágenes faciales. Estas características representan la imagen por regiones utilizando para ello valores de contraste local y manteniendo a su vez información espacial. Las Ranking Labels se compararon con otros métodos de caracterización ampliamente utilizados en este campo, concluyendo que eran un tipo de característica adecuado para abordar tareas de reconocimiento de género. Por otro lado, presentamos un nuevo enfoque de clasificación basado en vecindades locales. Éste consiste en un método de combinación de clasificadores diseñado para ser usado junto con descripciones locales de la cara. De este modo, cada miembro de la combinación se especializa en una región concreta de la cara. El hecho de utilizar vecindades permite que el método tenga un cierto nivel de flexibilidad en casos donde las caras pueden no estar alineadas o la detección de las mismas no es precisa. Para comparar el método propuesto con otras técnicas altamente efectivas, utilizamos un amplio conjunto de experimentos tanto cruzando como sin cruzar bases de datos. Los experimentos cruzando bases de datos nos permiten simular escenarios donde la variabilidad demográfica es considerable. Los resultados obtenidos mostraron que el método de clasificación propuesto junto con Ranking Labels es una solución tan fiable como aquellas basadas en enfoques globales cuando tratamos el problema de reconocimiento de género sobre imágenes faciales frontales con expresión neutra.

Para comprobar que las mejoras propuestas podrían utilizarse en situaciones razonablemente realistas, realizamos un estudio experimental con imágenes que presentan una dificultad mayor a las utilizadas hasta el momento. Para ello, al conjunto de imágenes faciales frontales con expresión neutra se añadieron imágenes de caras expresivas e imágenes de caras parcialmente ocluidas. Todas estas imágenes se combinaron en distintos conjuntos de entrenamiento y test en experimentos sobre una única base de datos y también cruzando tres bases de datos distintas. Una completa comparación estadística de los resultados obtenidos por nuestra propuesta así como por varios métodos muy extendidos en el campo del análisis facial reveló varios datos interesantes. Si el conjunto de entrenamiento contiene el mismo tipo de imágenes que el conjunto de test, tanto los enfoques locales como los globales obtienen resultados satisfactorios. Sin embargo, en situaciones donde uno de los conjuntos contiene imágenes con mayor dificultad (caras expresivas o parcialmente ocluidas) el rendimiento de las soluciones locales supera significativamente al obtenido por métodos globales. Los resultados empíricos también indicaron que las características propuestas, Ranking Labels, proporcionaban información más discriminante que el resto de descriptores considerados. Con respecto a los clasificadores, mientras que los SVM globales alcanzaron los mejores resultados únicamente en las tareas menos complejas, la técnica

propuesta conseguía buenos resultados en todos los casos.

Dejando a un lado el principal objetivo de la tesis, también realizamos un estudio exhaustivo sobre la influencia de la resolución de la imagen en las tasas de reconocimiento de género. En este estudio se consideraron diez tamaños de imagen distintos, partiendo de un tamaño extremadamente pequeño (2×1 píxeles) hasta llegar al máximo tamaño disponible en las bases de datos (329×264 píxeles). Los resultados obtenidos indicaron que una resolución de imagen moderada, entre 22×18 y 90×72 píxeles, es óptima para abordar tareas de clasificación de género. Además, de este estudio concluimos que una imagen de una resolución tan baja como 3×2 píxeles proporciona información discriminante para distinguir entre géneros. Este último dato puede ser de un gran valor en aquellos casos en los que no es posible adquirir una imagen de un tamaño razonable por causas ajenas al sistema automático.

En resumen, en esta tesis se han revisado todos los pasos del proceso de clasificación automática de género a partir de imágenes faciales, proponiendo nuevas soluciones para cada uno de ellos. Todas estas mejoras pueden ser empleadas tanto en conjunto como de forma independiente.

Conclusiones y Trabajo Futuro

Conclusiones

Los estudios realizados como parte de la tesis han abarcado todos los pasos del proceso de clasificación automática de género presentando mejoras para cada uno de ellos.

Se ha estudiado la posibilidad de reconocer el género cuando tan sólo una zona de la cara (ojos, nariz, boca o mentón) es visible, concluyendo que los ojos son la parte que contiene más información sobre el género de la persona. Además, se obtuvieron resultados indicando que distintas partes de la cara proporcionan información complementaria.

Para mejorar las tasas de clasificación, hemos propuesto un tipo nuevo de características (Ranking Labels) para describir imágenes de forma local así como un nuevo método de clasificación basado en vecincades locales. Por un lado, las Ranking Labels permiten obtener caracterizaciones que son razonablemente independientes de los niveles de gris de la imagen lo que proporciona mayor robustez cuando las condiciones de iluminación son distintas entre imágenes. Por otro lado, el nuevo método de clasificación permite una mayor tolerancia en situaciones donde la detección de la cara no es totalmente precisa o las caras no han sido alineadas correctamente.

Analizamos el comportamiento de distintas soluciones (entre las que se incluían las mejoras propuestas) a la hora de clasificar el género de imágenes faciales que mostraban caras expresivas y con oclusiones parciales. Para ello se utilizaron imágenes de tres bases de datos distintas con el fin de simular escenarios donde la demografía de los sujetos varía considerablemente. Además se consideraron situaciones donde las imágenes usadas para el entrenamiento y para el test no eran del mismo tipo (caras neutras, expresivas o parcialmente ocluidas). Tras un estudio estadístico exhaustivo de los resultados obtenidos

se concluyó que cuando los conjuntos de entrenamiento y test contienen imágenes de similares características todos las soluciones alcanzan, estadísticamente hablando, los mismos resultados. En cambio, si estos conjuntos presentan distintos niveles de dificultad, nuestras propuestas fueron las que mostraron un comportamiento mejor y más robusto que el resto.

Por último, estudiamos la influencia de la resolución de las imágenes en los resultados de clasificación. Realizamos experimentos con diez tamaños de imagen distintos, concluyendo que una tamaño moderado proporcionaba los mejores resultados. Sin embargo, resoluciones tan pequeñas como 3×2 píxeles aportan información útil para distinguir entre géneros.

Trabajo Futuro

En esta tesis se han presentado importantes mejoras aplicables a sistemas automáticos de reconocimiento de género. Sin embargo, la efectividad de dichos sistemas puede verse limitada en ciertos escenarios reales. Por ello, aún sigue habiendo trabajo por hacer para poder llegar a considerar que el problema está completamente resuelto. En esta sección se describen varias líneas de investigación que deberían tratarse para alcanzar ese objetivo, además de algunas líneas de trabajo que se comenzaron durante la tesis y que no han sido completadas.

Durante la tesis estuvimos trabajando simultáneamente en varias líneas de investigación, algunas de las cuales quedaron sin concluir. A continuación detallamos cuáles son las líneas de trabajo que en el momento de finalización de la tesis quedaron abiertas.

- Estamos llevando a cabo un estudio online con el objetivo de comparar la precisión de los humanos y los sistemas automáticos a la hora de clasificar el género de distintas personas a partir de imágenes faciales. En el estudio se pide a los participantes que indiquen cuál creen que es el género de la persona dada una imagen facial. Se considera un extenso conjunto de imágenes, dentro del cual se pueden encontrar caras sin expresión, expresivas y también parcialmente ocluidas. Para reproducir tanto como sea posible las condiciones a las que se enfrentan los sistemas automáticos, las imágenes son mostradas preprocesadas y con un tamaño reducido. Tras recopilar una cantidad razonable de datos, nuestro siguiente paso es analizar estadísticamente qué características comparten las imágenes que fueron clasificadas con mayor dificultad, tanto en el caso de los humanos como en el de los sistemas automáticos. Adicionalmente, hemos pedido a los participantes que proporcionen datos sobre su edad, raza y género con el propósito de estudiar cómo influyen esos factores cuando se reconoce el género de personas del mismo (o diferente) grupo de edad, raza y género.

- Se está implementando un prototipo funcional que muestra en tiempo real el género de las personas que aparecen en escena. Éste consiste en una cámara que captura imágenes en tiempo real y tan pronto como detecta una cara, ésta es enviada al sistema automático de clasificación de género. La predicción del género se muestra en pantalla representada mediante un cuadrado que emmarca la cara cuyo color indica el

género predicho. Este prototipo está prácticamente terminado, únicamente necesita pasar un proceso exhaustivo de depuración de errores.

Respecto al trabajo futuro, varias son las líneas de investigación que pueden mejorar los sistemas automáticos de clasificación de género actualmente disponibles. En particular, nos vamos a centrar en algunas de las extensiones que podrían ser fácilemente aplicadas a nuestro trabajo.

- En la tesis nos hemos centrado en bases de datos que presentan condiciones de iluminación bastante controladas y en imágenes frontales de la cara. Sin embargo, un sistema de clasificación de género que es fiable en cualquier tipo de situación debería ser capaz de tratar con imágenes con iluminaciones muy diversas. También, debería ser capaz de clasificar caras con diferentes ángulos respecto a la cámara. Por lo tanto, experimentos cruzando bases de datos que contengan estos tipos de imágenes serían el siguiente paso.

- Intuitivamente, las características faciales que indican el género de una persona parecen variar con la edad. Un estudio empírico de clasificación de género separando los individuos de acuerdo a su rango de edad podría proporcionar información a este respecto.

- Según varios estudios psicológicos, los humanos sufrimos del llamado efecto *sesgo interracial* [43]. Éste se refiere a la tendencia de encontrar más dificultad para identificar personas de otras razas distintas a la de uno mismo. Consecuentemente, parece que las características utilizadas para identificar a una persona pueden diferir dependiendo de su raza. Resultaría interesante comprobar si este mismo comportamiento también se presenta en sistemas automáticos. Este efecto podría estudiarse mediante la comparación de la precisión en el reconocimiento de un clasificador genérico (entrenado con caras de distintas razas) y clasificadores específicos para cada raza.

- Los métodos propuestos no han sido diseñados exclusivamente para abordar el problema del reconocimiento de género. Por ello, podrían aplicarse a otros problemas de clasificación. Las aplicaciones más directas serían la clasificación de expresiones faciales, rangos de edad o razas.

- Muchos de los autores del campo del análisis facial proponen utilizar sistemas de reconocimiento de individuos para resolver problemas de clasificación de género, aunque no ha sido probado que ambos problemas puedan solucionarse empleando el mismo enfoque. Por esta razón, creemos que el campo podría enriquecerse con un análisis comparativo concluyente de las técnicas empleadas para abordar ambos problemas.

CHAPTER 1

Introduction

Normally, facial information is used for identification purposes. However, faces indicate many more personal traits, among which gender is found. The knowledge of the gender of a person plays an important role in a variety of computer based applications. In this chapter, we introduce the problem of face gender classification and the issues that can be encountered when automatically addressing it. We then indicate our motivation and objectives for this work, followed by a list of the main contributions and publications. Finally, we provide a chapter-by-chapter summary of the contents of the thesis.

1.1 Introduction to Face Gender Classification

Human faces provide a large amount of information about ourselves. Apart from our identity, our faces indicate other demographic traits such as our gender, age and race. Nowadays, automatically collecting these pieces of information can be useful in countless tasks. Particularly, the knowledge of people's gender can be applied to dynamic market studies, surveillance and security systems, human-computer interaction, personalised services in many businesses, among others. Besides, another application can be to serve as a first filter for recognition tasks. For example, biometric recognition systems could reduce the search by half if the gender of the individual was known.

In the area of face analysis, face recognition [40, 2, 68] and facial expression analysis [28, 69] have been extensively studied compared to gender classification which has been addressed less often. This could partially be due to a general belief that gender classification is similar to a face recognition problem with only two classes. To the best of our knowledge, there are not published studies that support this statement by exploring the performance

differences of automatic systems when dealing with face recognition and gender classification. However, these studies are easily found in the psychology literature [72, 53]. In [72], it is clearly stated that, in order to identify a face, the information that makes it unique has to be encoded. In contrast, to recognise the gender of a face the information encoded must be shared by a group of different faces (male or female). From the point of view of data complexity, gender classification is a two-class problem with a commonly large number of face images from different people per class which results in sparse classes. On the contrary, face recognition is a multi-class problem with usually very few faces per class belonging to the same individual. Therefore, gender classification problems normally have a much higher intra-class variance than face recognition problems.

Although identifying the gender of a person seeing his/her face is a relatively easy task for humans, it becomes a challenging problem when it has to be solved automatically. A considerable number of difficulties encountered by automatic face gender classification systems go unnoticed when humans perform the task. Humans inadvertently deal with faces of different sizes, that is, we can recognise the gender of individuals whose faces we see from up close or in the distance. Besides, we do not mind if the proportions of the face are not how we expect them to be. For instance, we can still tell the gender of a person even if he/she has a larger nose or smaller eyes than the faces we usually see. Additionally, our ability to recognise the gender of a face is barely affected by the lightning conditions. It does not matter if it is daylight or night, we can still distinguish between males and females. These are minor issues for humans, however automatic systems usually require specific techniques to successfully tackle gender classification problems in such situations. An added difficulty for automatic systems is to locate the faces that appear in the input images. This probably seems a trivial task for humans but the available methods for automatically detecting faces are still not entirely accurate.

Apart from the mentioned troubles, there are other factors that might complicate the gender classification task even for humans. These complications, which are easily found in real environments, are related to being presented with faces showing emotions or partially occluded faces. Usually, the individual whose gender is of our interest wears clothing accessories which are beyond our control. These accessories might cover part of the face, as occurs with sunglasses, veils or scarves. In those cases, when the face is not completely visible, not all the information is available which makes harder recognising the gender of the person. In addition, people commonly express their emotions by showing facial expressions and, depending on the emotion, the faces can tremendously vary. For example, an angry face is markedly different from a surprised face. When expressing anger, the main face parts tend to squeeze, whereas if the emotion is surprise the eyebrows raise and the mouth tends to be quite opened. These are important factors that should be taken into account when implementing a gender classification system since it is expected to function satisfactorily in those situations.

Automatic systems can tackle the problem of face gender classification from different perspectives. They can be set out with the aim of recognising the gender of individuals appearing in videos or still images, which could be in 2-D or 3-D. This data could also be

provided in colour or just grey levels. Besides, the system could handle only frontal faces or also faces turned up to an acceptable degree of rotation. Additionally, the faces could appear completely visible or with a certain level of occlusion. All these factors condition the techniques employed for approaching the problem.

1.2 Previous Related Works

In this section, we review the most relevant papers on automatic gender classification. We indicate how other authors address the problem which requires us to mention some of the concepts and techniques that are explained in depth in Chapter 2.

The research on automatic face gender classification goes back to the beginning of the 1990s. The two first attempts to automatically classify the gender of face images were reported by Golomb et al. [30], and Cottrell and Metcalfe [24]. Golomb et al [30] trained a two-layer neural network, which they called "SexNet", to classify face images of 30×30 pixels. As input, the network received the grey level values of the face images (previously equalised and aligned) and provided a gender label as output. "SexNet" was tested with 90 faces (45 males and 45 females), particularly 8 tests were carried out with different sets of 80 images for training and 10 for testing, resulting in an average of 91.90% of correctly classified faces. In addition, they reported that 5 humans classifying the gender of the same 90 faces achieved an average of 88.40% correct answers. On the other mentioned work, Cottrell and Metcalfe [24] designed several neural networks to recognise identity, gender and emotions. For the experiments, they collected a set of 160 face images from 10 male and 10 female subjects which were aligned, reduce to 64×64 pixels and equalised. They reported almost perfect recognition for identity, albeit for only 20 subjects, and perfect gender classification. However, the training and test faces were from the same subjects.

This trend towards Neural Networks continued during the entire decade with some authors proposing new ways of using that classification method. It was also very popular to employ geometric features to describe the faces in the images. These features usually consist of a reasonably sized set of fiducial points or distances between them. Those fiducial points are commonly located at strategic positions in the face, such as the centre of the eyes or the corners of the mouth. Brunelli and Poggio [18] combined both, geometric features and neural networks. They proposed two competing networks trained with geometric features where each network specialised in faces of one gender. They tested these networks with 21 faces images of each gender using a leaving one out technique. As a result, 79% of the faces were correctly classified. Other types of geometric features were also utilised to describe face images. In a study by Wiskott et al. [67], the faces were represented by graphs. They modified a general object recognition system with the aim of addressing face recognition and gender classification problems. The resulting system generated graphs of new faces by means of elastic graph matching. That technique used a face space consisting of manually created model graphs for building new graphs to describe new faces. Given an unseen face, a graph representing it was created and the face was classified as belonging to the individual with the most similar graph. Although this method was mainly employed for face

recognition, they also applied it to gender classification by building a composite of graphs to resemble a given unseen face. Then, if most of the graphs belonged to males/females faces they determined that the given face was male/female. A leaving one out experiment with 112 face images (65% of them being males) resulted in a 90.20% of correct gender classifications. However, Wiskott et al. implied that faces of the test subjects were included in the training set.

In the 2000s, new gender classification methods were proposed, most of which were based on the now well-known Support Vector Machines. In that decade, authors focused most of their attention on appearance features instead of geometric ones. Appearance features can directly consist of pixel values or some transformation applied to those values in order to produce the descriptions. It was also at this time when authors started to automatically detect the position of the face in the image previous to the classification. To the best of our knowledge, the first work where automatic face detection and gender classification were combined was presented by Mogahaddam and Yang [46]. They addressed the gender classification problem with Support Vector Machines (SVM) and showed that the performance of a SVM with Radial Basis Function (RBF) kernel was superior to other pattern recognition techniques, such as, Fisher Linear Discriminant and ensembles of RBF networks. In the experiments, 1496 face images (53% male and 47% female) were used for training and 259 images (51% male and 49% female) for testing. After automatically detecting the area of the image containing the face, and training the classifiers using the pixel values of those areas, a SVM+RBF obtained the best accuracy, that was 96.60%. The authors do not indicate if the face images used in the experiments correspond to different individuals or if there were duplicates.

The first extensive survey on gender classification was published in 2008 by Mäkinen and Raisamo [39] and they utilised automatically detected and aligned faces. By that time, most of the research community considered face alignment essential for improving gender classification. Alignment methods pretend to position faces into a canonical pose, so the location of the most important facial features is the same relative to a fixed coordinate system. The alignment can be performed manually or automatically, and it is done with respect to some given fiducial points. Typically, those points are the centre of the eyes. In Mäkinen and Raisamo's survey, various alignment techniques are employed as well as several gender classification methods. Regarding face alignment, they concluded that better gender classification accuracies are achieved when the faces are not automatically aligned. This same conclusion was drawn from other works [42, 65, 22]. Focusing on gender classification, Mäkinen and Raisamo compared the performances of various classifiers using several appearance-based features. Their study involved a Neural Network with image pixel values as input, SVMs with two different inputs, pixel values and Local Binary Patterns, and a Discrete Adaboost method based on Haar-Like features. They found that SVM with simple pixel values achieved the best classification rate, which was 86.54%.

In recent years, Local Binary Patterns (LBP) have become massively popular in the face analysis field. Although these features were originally defined to describe image textures [52], many authors have used them to describe face images [2]. Alexandre [4] proposed

an ensemble method to tackle gender classification problems using different types of features. He considered texture features (LBPs), shape features (edge directions) and three different image sizes. Each member of the ensemble specialised in one type of feature and image size. He provided a comparison of the performances of the ensemble method and the base classifiers used individually over face images from two databases. As a result, the ensemble achieved better classification accuracies reaching to 99.07% and 91.19% depending on the database. Other authors presented techniques for selecting the most discriminant features, that is the case of Shan [58] and Tapia and Perez [60]. Shan [58] employed Adaboost to choose which LBPs provided more discriminant information (and named them boosted LBPs). He reported the gender classification accuracies achieved by SVMs using pixel values, standard LBPs and boosted LBPs when applied to frontal and near frontal faces. The best rates resulted from the use of boosted LBPs, that was 94.81% of correct classification. Boosted LBPs provided an increase in the accuracy of 1.4% and 3.5% when compared to standard LBPs and pixel values, respectively. Tapia and Perez [60] presented an approach for selecting a reduced set of features based on mutual information. They described the faces by fusing different types of features considering several image scales. The features involved were pixel values, edge directions and LBPs. The fusion of features was passed to the feature selection method and the chosen features were fed to the classifier. The features were combined fixing one factor, either the type of feature or the image size. They reported the highest classification rates when using the best of all the features considered. Those best results were 99.13%, 98.01% and 94.01%, depending on the image database. Although it is not explicitly detailed in the text, it seems that the test set was employed to select the best features. Then, that set of images was again used to test the classifiers based on the combination of best features.

Most of the published works on gender classification involve face images from one database as opposed to using one database for training and another for testing. The work by Bekios-Calfa et al. [13] is one of the very few published studies presenting cross-database experiments to test their proposed approach. They reviewed linear discriminant techniques using single- and cross-database experiments involving three datasets containing frontal and completely visible face images. They reported that single-database experiments were optimistically biased and all the considered methods achieved similar results (above 90% of accuracy) in those cases. However, in cross-database experiments depending on the amount of training data and the demography of the individuals, different methods achieved the best classification rates (ranging from 71.50% to 91.03%).

1.3 Motivation and Objectives

Automatic face gender classification still remains unsolved in real scenarios. Most published works addressing this type of problems assume many conditions which would not prevail in most realistic situations. The main assumptions are that the illumination conditions are similar in all images, and the demography of the individuals is quite controlled. Besides, most of those works do not consider images of expressive or partially occluded faces which

are extremely common in real environments. In addition, most of the related papers drew conclusions from comparing classification accuracies without any statistical support.

The main goal of this thesis is to improve automatic face gender classification in scenarios where the previously mentioned complications can be encounter. With that purpose, each step of the classification process is revised and improvements are proposed. This general goal is divided into the following specific objectives:

- To study the role of different face parts in gender classification and the suitability of addressing the problem with only the information provided by some of those parts.
- To define a new type of features for characterising face images which provides reliable information even considering various illumination settings.
- To design a new classification approach that is robust against inaccuracies in the face detection and faces with different facial proportions.
- To analyse the influence of the face image resolution in the classification results.
- To test the solutions with faces from a broad demography (that is, considering a wide range of ages and races) and also with expressive and partially occluded faces.
- To provide statistically supported conclusions.

1.4 Contributions of the Thesis

The work of this thesis is focussed on solving face gender classification problems. In order to improve the overall performance, we deal with the main issues that arise in each step of the process. With that purpose, new methods are designed for describing the content of images and for classifying the gender of unseen faces in realistic conditions. To check the suitability of the proposed techniques, experimental studies are presented comparing them with other widely employed approaches in the field. Following, the specific contributions of the thesis are outlined.

- A detailed evaluation of the role of eight different face parts in gender classification. The discriminant capabilities of each of the face parts is studied using diverse classifiers and two databases in order to check if their behaviours are consistent (Chapter 3).
- Empirical data showing the existence of complementary information among various face parts (Chapter 3).
- New features, named *Ranking Labels*, which characterise local regions of the images. These are designed to be more robust to changes in illumination by encoding information about local contrast, making the description independent of the actual pixel values. Additionally, the face descriptions derived from *Ranking Labels* maintain spatial information to better cope with inaccuracies in the face detection (Chapter 4).

- A novel classification approach based on local neighbourhoods which has a certain level of tolerance towards faces with different proportions, misaligned faces and inaccuracies in the face detection (Chapter 5).

- A thorough experimental comparison of gender classification techniques (including those proposed in the thesis) when using neutral, expressive and partially occluded faces in all possible combinations of training and testing roles (Chapter 6).

- A comprehensive statistical study on the influence of the resolution of the face images in automatic gender classification. A wide range of image sizes is considered going from extremely low resolutions to the highest possible resolution (Chapter 7).

- Empirical results revealing the smallest face image size that provides useful information to distinguish between genders (Chapter 7).

- Statistically supported conclusions drawn from experimental studies. In order to detect differences among performances of several classification models, three statistical test are applied: Iman-Davenport's statistic, Holm's method and Wilcoxon's Signed Rank test (Chapters 5, 6 and 7).

- Reliable assessments of the robustness of the presented methods to changes in acquisition conditions and demographic variables, such as age and ethnicity, by performing cross-database experiments (Chapters 5, 6 and 7).

1.5 Publications Resulting from the Thesis

The research work developed during the thesis has been validated with several international peer-reviewed conferences and a journal paper. A brief description of the content of each publication is provided.

Andreu, Y., García-Sevilla, P., and Mollineda, R. A. Face gender classification: A statistical study when neutral and distorted faces are combined for training and testing purposes. *Image and Vision Computing 32*, 1 (2014), 27–36

> This paper presents a thorough study of gender classification methodologies performing on neutral, expressive and partially occluded faces, when they are used in all possible arrangements of training and testing roles. A comprehensive comparison of two representation approaches, three types of features, three classifiers and two performance measures is provided over single- and cross-database experiments. Experiments revealed that when training and test sets contain different types of faces, our local models using the 1-NN rule outperform global approaches. However, with the same type of faces, even if the acquisition conditions are diverse, statistical evidence indicated that global SVMs and local 1-NNs perform equally.

Andreu, Y., López-Centelles, J., Mollineda, R. A., and García-Sevilla, P. Analysis of the effect of image resolution on automatic face gender classification. In 22^{nd} *International Conference on Pattern Recognition* (2014) (to appear)

This paper presents a thorough study into the influence of the image resolution on automatic face gender classification. The images involved range from extremely low size (2×1 pixels) to the highest possible resolution. A comprehensive comparison of the performances achieved by two classifiers using ten different image sizes is provided by means of two performance measures. Single- and cross-database experiments are designed over three well-known face databases. A detailed statistical analysis of the results revealed that a face as small as 3×2 pixels carries some useful information for distinguishing between genders. However, in situations where higher resolution face images are available, moderately sized faces from 22×18 to 90×72 pixels are optimal for this task.

Andreu, Y., Mollineda, R. A., and García-Sevilla, P. Assessing the effect of crossing databases on global and local approaches for face gender classification. In 15^{th} *International Conference on Computer Analysis of Images and Patterns* (2013), vol. 8047 of *Lecture Notes in Computer Science*, pp. 204–211

This paper presents a comprehensive statistical study of the suitability of global and local approaches for face gender classification from frontal non-occluded faces. A realistic scenario is simulated with cross-database experiments where acquisition and demographic conditions considerably vary between training and test images. The performances of three classifiers using two types of features are compared for the two approaches. Supported by three statistical tests, the main conclusion is that if training and test faces are acquired under different conditions from diverse populations, no significant differences exist between global and local solutions.

Andreu, Y., García-Sevilla, P., and Mollineda, R. A. Gender classification from neutral and expressive. In 6^{th} *International Conference on Machine Vision* (2013), vol. 9067 of *SPIE Proceedings*, pp. 906723–26

This paper presents a statistical study of local versus global approaches for classifying gender from neutral and expressive faces. A cross-dataset evaluation is provided by using different databases for training and testing, as well as several well-known classifiers and widely used features for facial description. Three statistical tests supported the hypothesis that local approaches are more suitable than global ones for solving gender classification problems over expressive faces when training with non-expressive faces. However, if a large set of expressive faces is available for training, global solutions outperform local ones.

Andreu, Y., García-Sevilla, P., and Mollineda, R. A. Dealing with inaccurate face detection for automatic gender recognition with partially occluded faces. In 14^{th} *Iberoamerican Congress on Pattern Recognition* (2009), vol. 5856 of *Lecture Notes in Computer Science*, pp. 749–757

Gender classification has not been extensively studied when the face cannot be accurately detected and it can also be partially occluded. In this paper, we propose a new type of appearance-based features, called *Ranking Labels*, which are designed to better cope with the previously mentioned circumstances by providing spatial information. An experimental comparison of the proposed features with two widely used characterisation techniques (Local Binary Patterns and Local Contrast Histograms) is provided. The results of several experiments prove that *Ranking Labels* description is the most reliable among those considered.

Andreu, Y., Mollineda, R. A., and Garcia-Sevilla, P. Gender recognition from a partial view of the face using local feature vectors. In 4^{th} *Iberian Conference on Pattern Recognition and Image Analysis* (2009), vol. 5524 of *Lecture Notes in Computer Science*, pp. 481–488

This paper presents an empirical assessment of the robustness of *Ranking Label* descriptors to address gender classification problems with different levels of accuracy in the face detections. Descriptions based on *Ranking Labels* are employed to characterise top half faces. Due to the fact that only the top half of the face is used, this is a feasible approach in those situations where the bottom half is hidden. The experimental results indicated that *Ranking Label* features are robust towards face detections with different degrees of inaccuracy.

Andreu, Y., and Mollineda, R. A. The role of face parts in gender recognition. In *International Conference on Image Analysis and Recognition* (2008), vol. 5112 of *Lecture Notes in Computer Science*, pp. 945–954

This paper compares the discriminant capabilities of different face parts for gender classification purposes. It goes beyond previous works with respect to the number of face parts and classifiers considered. The experimental results indicate that individual face parts offer enough information to allow discrimination between genders.

Andreu, Y., and Mollineda, R. A. On the complementarity of face parts for gender recognition. In 13^{th} *Iberoamerican Congress on Pattern Recognition* (2008), vol. 5197 of *Lecture Notes in Computer Science*, pp. 252–260

This paper evaluates the expected complementarity between the most prominent parts of the face for addressing gender classification. Several ensembles of classifiers based on various single face parts are designed using different combination strategies. The experimental results show that, as expected, ensembles perform significantly better than plain classifiers based on single parts.

1.6 Structure of the Thesis

This document presents a comprehensive review of the published papers resulting from the work of the thesis in chronological order of publication. In Chapter 2, we introduce the basic concepts necessary to understand gender classification problems and explain in detail the classification methodology followed in all the experimental studies of the thesis. This methodology has five main steps: 1) face detection and preprocessing, 2) face description, 3) gender classification, 4) performance evaluation and 5) statistical analysis. The rest of the chapters are closely related to a particular step within the general framework.

In Chapter 3, we present an empirical study of the roles of different face parts. This study goes beyond previous related works with respect to the number of parts considered and the diversity of the classifiers employed. In addition, the complementarity of the information provided by several face parts is tested by means of ensembles of classifiers. The main aspects analysed in this chapter are related to the first step of the methodology, since the division of the face in parts would take place in the preprocessing stage. This chapter explains in detail the work published in [10] and [9].

In Chapter 4, we propose a new type of features (named *Ranking Labels*) for face characterisation which are designed to provide local contrast values while keeping some structural information. We present various experiments whose results show that *Ranking Labels* are more reliable than other local features that are widely used in the field. The work reviewed in this chapter, which was published in [5] and [11], is focused on the second step of the general methodology, that is, face description.

In Chapter 5, we propose a classification method based on local neighbourhoods. It could be seen as an ensemble of local classifiers, where each classifier specialises in a particular region of the image. The idea of neighbouring regions is developed to provide a certain level of tolerance towards misaligned faces and faces with different facial proportions. We present empirical data suggesting that our method is as suitable as global approaches in situations where the latter are believed to be more appropriate. The main contribution of this work, published in [12], is directly related to the third step of the methodology which is gender classification.

In Chapter 6, we present and exhaustive comparison of the performances of a diverse set of classification models when neutral, expressive and partially occluded face images are considered. The classification models compared differ from one another with respect to the approach (global or local) adopted to address the task, the type of feature chosen to describe the face images and the classifier employed. The conclusions drawn from the empirical comparison are supported by several statistical tests applied to two performance measures. Therefore, this work, published in [7], could be placed within the fourth and fifth steps of the general methodology.

In Chapter 7, we perform a detailed experimental analysis of the influence of the image resolution on face gender classification. We consider images going from extremely low resolutions to the highest size provided with the databases. Empirical data indicating which is the smallest image size containing discriminant information is also provided. This

work is related to the first step of the methodology, where the image size needs to be decided.

In Chapter 8, we conclude the thesis by presenting the main findings of these investigations. In addition, we summarise several open lines of research and other possible lines of future work.

CHAPTER 2

Gender Classification Methodology

One type of pattern recognition problem is classification, which attempts to assign each input value to one of the given classes. This is exactly what automatic face gender classification aims to achieve, given a face image assign a gender label to it. In this chapter, we present the general experimental methodology commonly followed in classification experiments. In order to contextualise the explanation of the process, we introduce an example of a realistic gender classification system.

2.1 General Methodology

In this section, we describe the general methodology which is followed by many classification systems, including the empirical studies presented in the thesis. In order to illustrate the explanation of this methodology, an example of a gender classification system is provided.

Imagine we are building a system for showing *smart adverts* on screens placed around a shopping centre. By *smart advert* we mean that the advert will specifically target the individuals looking at it. This system could be focusing in many aspects of the individuals (such as, gender, race or age range), here we will focus on their gender. For that, we would need to automatically identify the gender of those people looking at the screen before selecting the advert to show. One way to address this would be installing a camera on the screen facing the individuals, so it could capture visual information from them, for example a photograph. Then, a gender classification system would get this image (containing one or more faces) and would identify the gender of each person. Knowing the viewers' gender, an advert that is considered interesting for most of them (males or females) would be shown.

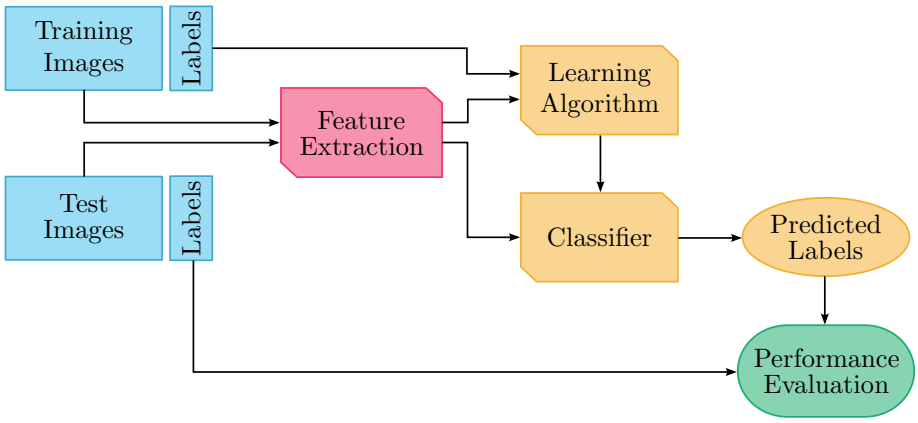


Figure 2.1: Steps of the classification methodology followed in the experimental studies of the thesis.

Before the system is capable of distinguishing between genders, it has to *learn* how female and male faces look like and which are the facial characteristics shared by people of the same gender. However, this is an arduous and difficult task that we would prefer to make automatically. In order to let the system learn by itself, we should provide it with a big enough set of face images from both genders and a methodology for extracting information (also known as *features*) from those images. Once the system has learned how to differentiate female and male faces, we could feed into it a face image in order to obtained a prediction of the gender of the person in the image. From what it has been explained so far, our system has to *learn* how to distinguish male and female faces previously to be operative. This process of learning is commonly referred to as *training*, and the face images from which the system learns are the *training images* or *training set*. Then, given a previously unseen face image, the system *classifies* it as male or female.

Normally, before installing the system in the shopping centre, we would test whether it works as expected. For this purpose, we should use images which are not involved in the *training* process, otherwise we would not know if the system has properly learned the differences between males and females or has just memorised the faces it has seen. The images used in this *test* process are referred to as *test images* or *test set*. To evaluate how well the system works, we would need to quantify its performance; therefore, we would define a *performance measure*. For instance, a straightforward measure would be the percentage of *test images* whose gender label the system has correctly *classified*. Figure 2.1 shows all the steps of this classification methodology and their relations.

Summing up, for building a gender classification system, we would need a set of face images that would be divided into subsets, the *training set* and the *test set*. We would also have to define how to extract *features* that describe the faces that appear in the images. Then, a *learning algorithm* would create a *model* using the information provided by the *features* extracted from the *training images*. Finally, we would *evaluate* the *model* by

supplying it with the *features* extracted from *test images* and computing a *performance measure*. Once the assessment is satisfactory, the system would be ready to be placed in the shopping centre.

All those tasks can be arranged in four groups, which coincide with the four steps of the methodology: 1) face detection and preprocessing, 2) face description, 3) gender classification, and 4) performance evaluation. In the following sections the processes involved in each of these steps are detailed.

2.2 Face Detection and Preprocessing

Although, it has not been explicitly indicated, before describing a face, we need to know where exactly the faces are located in the image. Once the image has been acquired from the camera, we would use a *face detector* to localise the faces in it. Afterwards, each area of the image containing a face would be extracted and preprocessed. Following, we describe how the face detectors work and what preprocessing techniques are applied to the images.

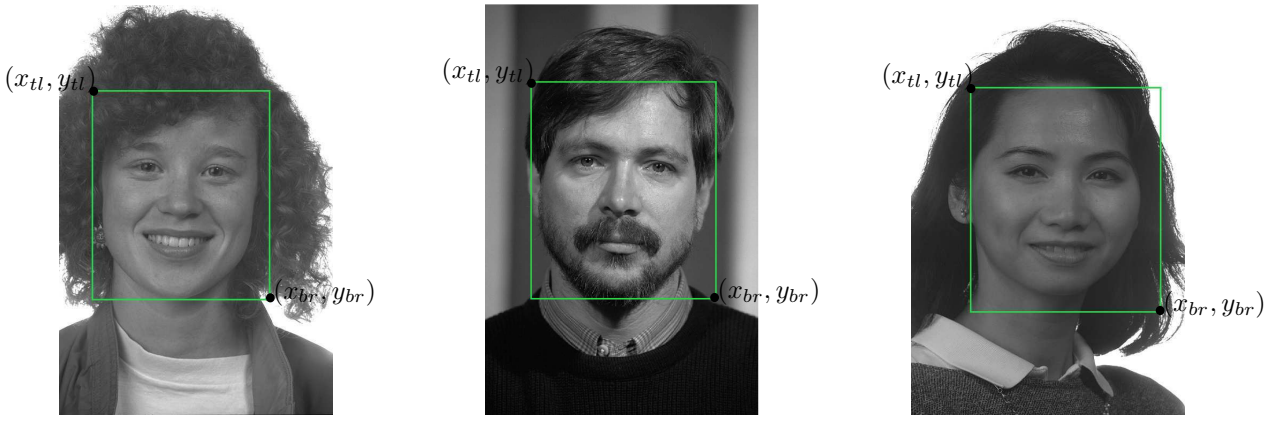
2.2.1 Face Detector Methods

A face detector can follow many different approaches, which produce different results with respect to the geometry of the area containing the face (mainly, shape and size). In this section, we focus on those approaches taken in the experiments included in the thesis.

Detection based on the Proportions of the Face

This first approach, which assumes that the coordinates of the eyes are known, is based on the expected proportions of the human face. Leonardo da Vinci proposed these proportions which are still in use by clinicians to objectively assess facial aesthetics and correct certain disproportions [48]. Da Vinci stated that the face can be divided into three horizontal regions of the same size whose boundaries are the hairline, the eyebrows, the bottom of the nose, and the chin. He also affirmed that the face could be divided into five equal vertical spaces where the distance between the eyes is the same as the distance at the side of the eyes. These rules of proportion are a good guide, but faces are not commonly this perfect, neither are completely symmetrical. Therefore, taking into account these rules and having available the coordinates of the centre of the eyes in the image, we performed an empirical study to calculate the points which delimit the area of the image that most likely will contain the face. Given the coordinates of the centre of the right eye, (x_r, y_r), and the left eye, (x_l, y_l), and the distance between them, $d = \sqrt{(x_l - x_r)^2 + (y_l - y_r)^2}$, the top left corner of the area of the image containing the face is estimated by,

$$(x_{tl}, y_{tl}) = (x_r - 0.75 \times d, \frac{y_r + y_l}{2} - 1.15 \times d) \tag{2.1}$$



(a) Accurate detection. (b) Accurate detection. (c) Inaccurate detection.

Figure 2.2: Face detections obtained with the method based on the expected proportions of the face. The delimiting points have been calculated using Eq. 2.1 and 2.2.

and the bottom right corner is,

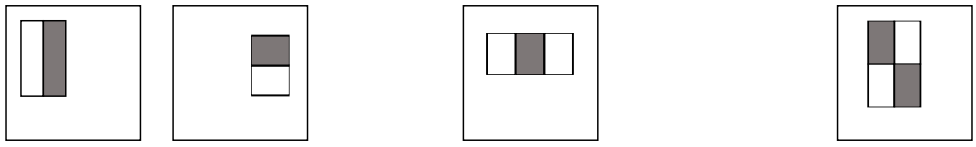
$$(x_{br}, y_{br}) = (x_l + 0.75 \times d, \frac{y_r + y_l}{2} + 1.8 \times d). \tag{2.2}$$

Some examples of the areas detected using Eq. 2.1 and 2.2 are shown in Figure 2.2. Figures 2.2(a) and 2.2(b) show two sufficiently accurate face detections, as the boundaries of the divisions match the expected facial features. Horizontally, those boundaries are the hairline and the bottom of the chin, and, vertically, they are approximately both side of the face. However, this face detection commits some errors with faces that have proportions different to those of the majority. As shown in Figure 2.2(c), the bottom boundary does not coincide with the chin in this case, since the face is smaller than the average.

In order to use this face detector, we need the coordinates of both eyes. This can be provided along with the face images (some face databases have this information) or an eye detector could be use. In the first case, the coordinates of the centre of each eye are almost exact, since they tend to be manually marked. In the second case, having to detect the eyes automatically could introduce some errors that might lead to erroneous face detections. In the experiments where this face detector is involved, we indicate how the coordinates are obtained.

Detection with Viola-Jones algorithm

Viola-Jones algorithm [63] is a widely-used automatic method for real-time object detection. Since it is automatic, it has to learn how the target objects look like previously to be fully functional. In fact, this is a classification problem: there are objects in the image and we



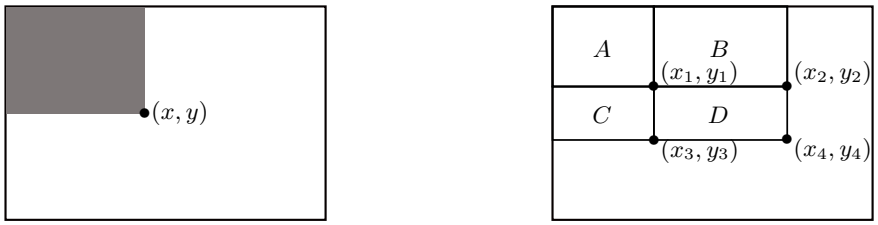
(a) Two-rectangle features (b) Three-rectangle feature (c) Four-rectangle feature

Figure 2.3: Features involved in the Viola-Jones detector [63]. The sum of the pixel values within the white rectangles are subtracted from the sum of those in the grey rectangles.

want to know if the objects of interest are present. To address this problem, the process followed is similar to that of a general classification system. Therefore, for explaining how this algorithm works, we use some concepts that will be explained following in the chapter.

For learning the appearance of the objects of interest, patches are considered over the images and features describing their content are extracted. The value of these features is the difference of the sums of the pixel values within rectangular areas. Four different features are involved regarding the number of rectangles considered and the position with respect to the others. Figure 2.3 shows the distribution of the rectangles in these four features. The value of each feature is calculated by subtracting the sum of the pixel values within the white rectangles from the sum of those in the grey rectangles. In order to quickly calculate the value of these features, an intermediate representation of the original image, called integral image, is employed. The value of each point (x, y) in the integral image is computed by the sum of all the pixels above and to the left from that position in the original image (see Figure 2.4(a)). Building the integral image from the original image only takes a few integer operations. Using the integral image, calculating the value of a feature takes a constant time because each rectangular area in a feature is always adjacent to at least one other rectangle. For example, given the integral image showed in Figure 2.4(b), the sum of the pixels within rectangle D can be seen as the sum of the pixels in rectangles $(A + B + C + D) - (A + B) - (A + C) + A$. Conveniently, $A + B + C + D$ is the integral image's value at point (x_4, y_4), $A + B$ is the value at (x_2, y_2), $A + C$ is the value at (x_3, y_3), and A is the value at (x_1, y_1). Therefore, the sum of the pixels within D is $(x_4, y_4) - (x_2, y_2) - (x_3, y_3) + (x_1, y_1)$.

There are a large number of features associated with each patch, and even though the features can be calculated efficiently, the computation of the whole set of features is prohibitively expensive. Hence, a variant of the AdaBoost algorithm is used to select the most discriminant features. The original AdaBoost algorithm combines a collection of simple learners (also known as, weak classifiers) to form a stronger classifier. This can be interpreted as a greedy feature selection process, where each weak learner is designed to select the single rectangle feature which best separates the positive and negative samples. For each feature, a weak learner determines a threshold, so that feature is chosen if their value is higher than the threshold.



(a) The value of the integral image at point (x, y) is the sum of the pixel values above and to the left (the pixels within the grey area).

(b) The sum of the pixels within rectangle D can be computed by $(x_4, y_4) - (x_2, y_2) - (x_3, y_3) + (x_1, y_1)$.

Figure 2.4: Integral image definition and usage for computing features.

The face detector consists in a cascade of classifiers whose inputs are patches. A positive response from the first classifier triggers the second classifier, and so on. Therefore, if a patch goes through every classifier in the cascade, it is predicted as containing the object of interest. A negative outcome at any point leads to the immediate rejection of the patch. Each layer of the cascade is trained by AdaBoost with the features selected to reach a target detection rate and false positive rate (both set in the learning stage). In other words, each layer only employs those features selected by AdaBoost to decide whether the patch contains the target object.

This detector has been broadly employed for detecting faces since it has been made publicly available together with a very good training set of features. Therefore, using that training set the detector learns how to accurately distinguish faces from non-faces. This public implementation is provided in the OpenCV library [1] and a training set of frontal face images, named "`frontalface_alt2`", is also provided in the mentioned open-source library. Some examples of the face detections using such implementation of the Viola-Jones algorithm are shown in Figure 2.5.

As can be seen comparing Figure 2.5 with Figure 2.2, the area returned by Viola-Jones algorithm is smaller than that obtained by the previously presented detector based on the expected proportions of the face. Viola-Jones detection excludes some parts of the face which are included in the previous detector results. Particularly, the area detected with Viola-Jones algorithm does not include the forehead and part of the chin.

Once we know that there is a face somewhere in the image, we could pass the whole image to the next step or we could use just the region of the image where the face is located. If the gender classification system would learn from full images, it could be misled by focusing the decision on background information. For instance, if there were several different backgrounds shared by many images, the system could learn about that information. Then, we would have a background predictor instead of a gender classifier. In order to avoid this, after locating the face in the image, an image is created by cropping the pixels within the

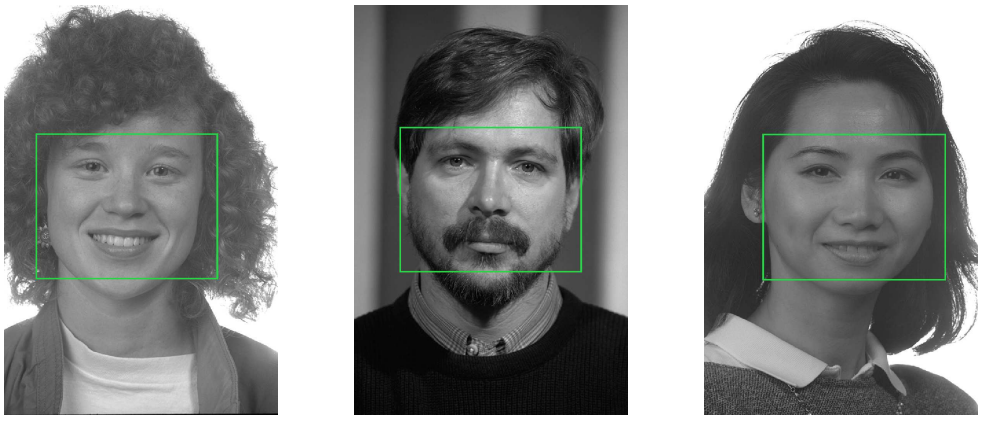


Figure 2.5: Face detections using Viola-Jones detector.

area containing the face. The aim of detecting the face and cropping the original image is to only keep the information provided by the face. After this, in the new image, the number of pixels that do not belong to the face has been considerably decreased compared to the original face image.

2.2.2 Preprocessing Techniques

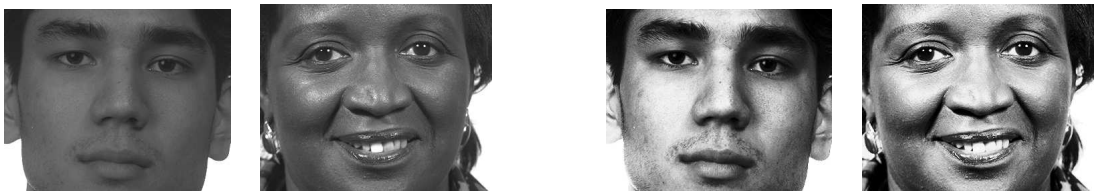
In this preprocessing step, several techniques to try to avoid certain problems are applied. An overview of those techniques and the problems they tackle is given below.

- Histogram equalisation is employed to better deal with images with different illumination conditions.
- Face Alignment methods are usually applied to situate the main facial features in the same position within all face images. All the experimental studies of the thesis only involve unaligned face images.
- Image resizing is utilised to make all images have the same size.

Following, we give further details about these processes and the reasons why they are or are not utilised in the experiments of the thesis.

Histogram Equalisation

An issue that we could encounter is having different lighting conditions when the images were captured. This could result in lighter and darker images and the system could learn about that information. In order to minimise the effect of different illumination conditions, the histogram of the cropped image is equalised. Histograms are usually arrays whose



(a) Before histogram equalisation (b) After histogram equalisation

Figure 2.6: Examples of cropped images to the area of the face, (a) previous to equalising the histogram and (b) after histogram equalisation.

elements (referred to as *bins*) represent the number of occurrences of a range of values. Normally, a histogram of a grey image has one bin per each possible grey level value. As grey levels are in the range $[0, 255]$, the typical histogram of a grey image has 256 bins. Histogram equalisation is a technique for enhancing contrast by evenly using every grey level in the full range. Let H be the histogram of a grey image, then $H(g)$ indicates the number of occurrences of the grey level value g in the image. In order to equalise H, first the cumulative histogram H_c is computed. Given a grey level value g, $H_c(g)$ is,

$$H_c(g) = \sum_{i=0}^{g} H(i). \tag{2.3}$$

Then, the bin associated to the grey level g of the equalised histogram, H_e, is,

$$H_e(g) = \frac{255}{H_c(255)} \times H_c(g). \tag{2.4}$$

In Figure 2.6, images of two individuals from different races (a Caucasian male and an African-American woman) are shown before and after applying histogram equalisation. As can be seen in Figure 2.6(a) some spots in the woman's face shine (for example, the nose and the cheeks), while the man's face does not has those. After equalising the histograms (see Figure 2.6(b)), both faces show lighter regions around the cheeks and nose along with being images with higher contrast which makes details stand out.

Face Alignment

When addressing facial analysis problems, many solutions employ aligned faces. Aligned faces are those face images where most face parts are expected to be in the same position in all images. Mainly, alignment methods focus on the eyes and try to move them to a given position within the image by means of different techniques, such as scaling the image and rotating it. In the literature, there are several works studying the usefulness of this alignment [42, 65, 22] whose results show that aligned faces do not improve the accuracy of the classification. For this reason, the experiments of this thesis do not use aligned faces.

Image Resizing

As was shown in Section 2.2.1, the areas of the image containing the face can vary in size. For instance, in images taken from the distance, the faces would appear smaller than if the camera was closer to the individuals. This would result in detected areas of different sizes. In order to have face images of the same size, the cropped images are scaled to a common size. The interpolation process required for resizing the image uses a three-lobed Lanczos windowed sinc function [62] which keeps the aspect ratio of the cropped face image. Some authors have proved that the image resolution does not affect the gender classification as long as the images have a reasonable size [46, 39, 27]. In Chapter 7, we analyse in depth the effects of different image resolutions in solving gender classification problems.

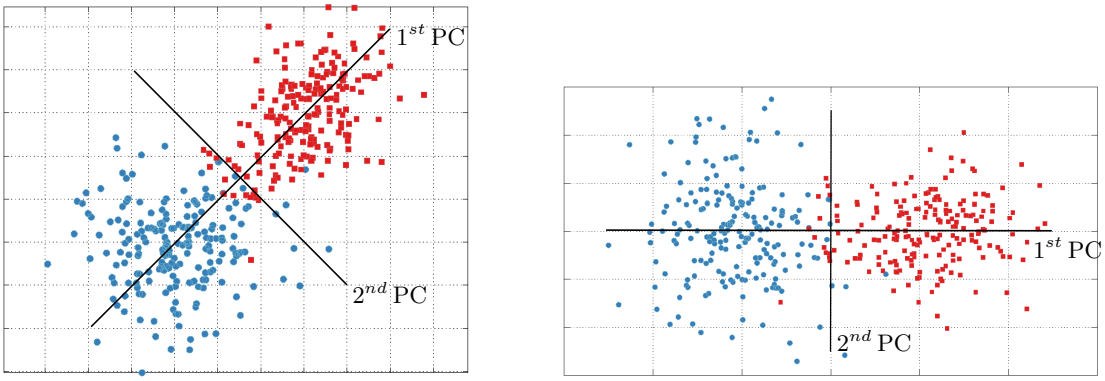
2.3 Face Description

Once we have a preprocessed image containing the face area, the information provided by the face should be described so a learning algorithm can learn from it. This information is generally encoded as a set of numerical data (commonly referred to as *features*). There are many ways of describing the information contained in an image, most of which can be classified either as appearance-based or geometric-based approaches. In appearance-based approaches, the information extracted from the image is based on the values of the pixels in the image. These pixel values can be directly used as features or a transformation can be applied to them to obtain a face description in a different space. In geometric-based approaches, the information extracted is related to geometric characteristic of the face, that is, distances or ratios between particular face parts. For instance, geometric features could be the distance between the eyes, the eyebrows thickness or the ratio of mouth width to thickness of lips. Regarding the extension of the area described, we can have global or local features. Global features are those extracted taking into account the whole image at once, as opposed to local features which are extracted from certain regions of the image. Consequently, global descriptions provide information about the structure of the face (configural information) as well as information about the individual face parts (featural information), while local descriptions only provide information about isolated regions of the face (featural information).

As mentioned above, the description of an image consists in a set of features (numerical data) which is generally handled in the form of an array (known as *feature vector*). Let D be the number of features (elements in such array), then a face is represented in a $D-$dimensional *feature space*. This could be the original representation space, or it could be a *transformed space* where the number of dimensions is usually smaller than D.

2.3.1 Grey Levels

The values of the pixels that form the face image can be directly used as features. In that case, the content of the image is described by means of the grey level values of its pixels. A



(a) Data represented in the original space.

(b) Data represented in the PCA space.

Figure 2.7: Two-dimensional data drawn from two classes before and after applying PCA (PCA basis vectors shown in black).

description using grey levels consists in a vector whose elements are all the pixel values of the image read following a pre-defined order. In the experiments, the process starts from the top left corner pixel to the right and downwards. Even though this seems a very simple and straightforward descriptor, it has successfully been used for addressing face analysis problems, among them gender classification [39].

2.3.2 Principal Component Analysis

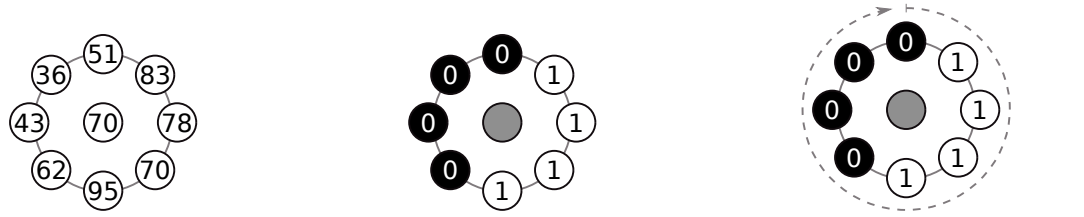
Principal Component Analysis (PCA) is a widely adopted technique for transforming the representation space [32]. By using PCA, our goal is to obtain a new representation space that best describes the variance of our data. Additionally, PCA can also be applied to reduce the dimensionality of the feature space. PCA analyses the data, which possibly consist of several correlated features, and searches for a new space whose basis vectors correspond to those directions in the original space with maximum variance. PCA could be thought of as a procedure for revealing the internal structure of the data in a way that best explains its variance. Figure 2.7 shows some data represented in the original space (Figure 2.7(a)) and the PCA space (Figure 2.7(b)) with the basis vectors of the PCA space in black. As can be seen, the greatest variance is given by the first component and the second component provides the greatest possible variance while being orthogonal to the previous component. This will continue to further components in a higher dimensional data. In PCA, lower components always carry more variance than higher ones. Consequently, selecting those components that account for a reasonably high percentage of the variance, we would reduce the dimensionality of the representation space without much loss of information. In Figure 2.7(b), keeping only the first component would result in a 1-dimensional space where it would be possible to discriminate between both classes.

Figure 2.8: The first 20 *eigenfaces*, that is, the first 20 eigenvectors showed as images. The PCA was calculated with images from FERET database.

Formally, let $\mathbf{W}$ be a linear transformation that maps the original $D-$dimensional space onto a $F-$dimensional space where $D \ll F$. Then, new feature vectors are defined by $y_i = \mathbf{W}^T x_i$ where $x_i \in \Re^D$ and $y_i \in \Re^F$. The columns in $\mathbf{W}$ are the eigenvectors e_i obtained from solving $\lambda_i e_i = \mathbf{A} e_i$, where $\mathbf{A}$ is the covariance matrix and λ_i is the eigenvalue associated to the eigenvector e_i. Before obtaining the eigenvectors, we should subtract the average of all vectors to each of the vectors to ensure that the mean of the data is zero.

In our classification system, the transformation $\mathbf{W}$ is computed from the training data and then new training and test sets are obtained by applying the transformation to the original data. Usually, only those features containing a certain percentage of the variance of the data (typically, 95% or 99%) are kept.

In face analysis, this method has been widely used, mainly for face recognition [57, 64, 50], but also for gender classification [37], leading to very good performances in both cases. When the original data are face images, if the eigenvectors are shown as images, we see that they could be identified as faces (in the literature they are referred to as *eigenfaces*). Some eigenfaces are shown in Figure 2.8. When calculating the PCA transformation, we are looking for a linear combination of these eigenfaces that produces the input face image.



(a) Each circle represents one pixel and the number inside it is its grey level value.

(b) If the pixel-value is higher than the centre's, number 1 is assigned to that pixel, else number 0.

(c) The LBP value associated to the central pixel is 01111000 which is obtained by reading the binary values following the direction of the arrow.

Figure 2.9: Computation of $LBP_{8,R}$.

2.3.3 Local Binary Patterns

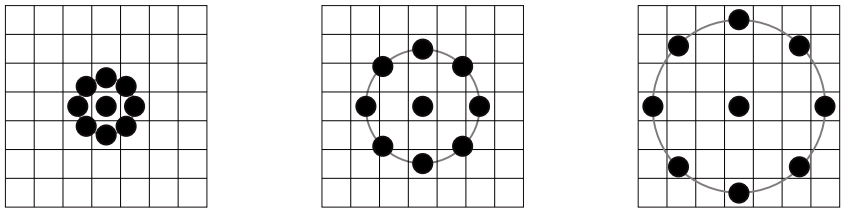
Local Binary Patterns (LBPs) were originally defined to characterise image textures [51]. More recently, they have been used as face descriptors [2] which provide information about the texture of the face.

An LBP value associated to a given pixel is a binary number which is calculated considering a neighbourhood around that pixel. To create this binary number, all neighbours are given either value 1, if they are brighter than the central pixel, or value 0 otherwise. The values assigned to the neighbours are read sequentially in the clockwise direction to form the binary pattern which characterises the central pixel. An example of the computation of LBP values is depicted in Figure 2.9. It is worth mentioning that LBPs do not provide information about contrast since the magnitude of the grey level differences is ignored.

To deal with textures at different scales, LBP can use neighbourhoods of different radii. A local neighbourhood is defined as a set of sampling points spaced in a circle centred at the pixel to be labelled. Hence, the radius of the neighbourhood indicates how far from the centre the pixels considered are. The notation $LBP_{P,R}$ refers to LBPs with a neighbourhood of P sample points on a circle of radius R. Figure 2.10 shows which pixels are considered with neighbourhoods of various radii. In case a sample point does not fall in the centre of a pixel, a bilinear interpolation[1] is used.

The LBP operator can be improved by using the so-called uniform LBP [52]. A uniform pattern is one that has at most two one-to-zero or zero-to-one transitions in the circular binary code. The total number of uniform LBP values ($\text{LBP}^{u}_{P,R}$), when considering a neighbourhood of 8 points, is 58. Although the number of patterns is significantly reduced from 256 (all possible $LBP_{8,R}$) to 58 ($LBP^{u}_{8,R}$), this smaller set of uniform patterns provides

[1] In image processing, a bilinear interpolation considers the four pixels of known values surrounding the location of the unknown point. Then, a weighted average of the known values is computed to obtain the interpolated value. The weights are based on the distances from the unknown point to the known pixels.



(a) Radius $R = 1$ (b) Radius $R = 2$ (c) Radius $R = 3$

Figure 2.10: $LBP_{8,R}$ neighbourhoods for different values of R.

the majority of patterns of texture, sometimes over 90% [51].

As it has been described to this point, the LBP operator provides a representation sensitive to rotation because several different codes can be obtained depending on which is the first neighbour considered when creating the binary code. In Figure **??** the starting pixel is the one on top, but it could be any of them as long as they are read in a clockwise direction. In order to obtain rotationally invariant LBPs [52], all possible binary numbers that can be obtained by starting the sequence from all neighbours in turn are considered. Then, the smallest of the constructed numbers is chosen. If the object (in our case, the face) is slightly inclined in the image, the rotation invariant uniform LBP ($LBP^{ri,u}_{P,R}$) is supposed to provide a description which is not affected by that inclination.

Normally, a representation based on LBPs consists in a histogram where the LBP values obtained for each of the pixel in the image (or area of interest) are accumulated. In the histogram, the bins show how many times the corresponding LBP codes have been produced. The number of bins depends on the type of LBP employed. In the literature, the most common number of sample points (neighbours) is $P = 8$. In such case, if uniform sensitive to rotation LBPs are used, the histogram has 59 bins, that is, 58 possible $LBP^{u}_{8,R}$ values and an extra bin for accumulating all the non-uniform LBPs obtained. When using rotation invariant uniform LBP, the number of bins is reduced to 10 bins, since there are 9 possible $LBP^{ri,u}_{8,R}$ codes and an extra bin for all the other codes.

2.3.4 Local Contrast Histogram

Describing an image using the grey level values of its pixels implicitly provides information about the contrast. However, shifts in the grey scale, which could be due to changes in illumination, strongly affect such type of features. For instance, two face images of the same person could seem quite different just because one is much darker than the other. In cases where illumination cannot be controlled, Local Contrast Histogram (LCH) is a more suitable descriptor, since it provides information about the contrast while being invariant against shifts in the grey scale. LCHs are usually combined with LBPs to overcome the lack of contrast information of the latter.

(a) Each circle represents one pixel and the number inside it is its grey level value.

(b) Those pixels that are darker than the centred pixel marked in black, the brighter ones in white.

(c)
- Brighter pixels add up to 326 and darker ones to 192.
- The contrast value (C) of the central pixel is:
  $C = \frac{1}{4} \times 326 - \frac{1}{4} \times 192$
  $C \approx 115$

The local contrast value assigned to the central pixel is 115 which is obtained after the indicated operations.

Figure 2.11: Computation of $LCH_{8,R}$.

LCH provides information of the contrast of local areas. For this purpose, neighbourhoods are defined in a similar way as for LBPs. Then, the contrast value associated to the central pixel is calculated taking into account all its neighbours. To compute the local contrast of a pixel, the average of the grey level values of those neighbours that are brighter than that pixel is subtracted from the average of the grey level values of the darker ones. An example of the computation of local contrast values is depicted in Figure 2.11.

More formally, let P be the number of sample points in the neighbourhood and $g(i)$ be the grey level value of the i^{th} pixel in the neighbourhood. The contrast around the central pixel, whose grey level value is g_c, is,

$$C = \frac{1}{N_b} \sum_{i=1}^{P} b(g(i) - g_c) \times g(i) - \frac{1}{P - N_b} \sum_{i=1}^{P} b(g_c - g(i)) \times g(i) \tag{2.5}$$

where

$$N_b = \sum_{i=1}^{P} b(g(i) - g_c) \tag{2.6}$$

and

$$b(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}. \tag{2.7}$$

Finally, the local contrast values of all neighbourhoods are accumulated in a histogram to obtain the $\text{LCH}_{P,R}$ descriptor. This notation means that the neighbourhood used has P sample points on a circle of radius R. The number of possible contrast values can be different for each image, therefore the number of bins in the LCH should be set beforehand. In order to make LCHs directly comparable to LBPs, two versions of LCH are considered. In one version, the number of bins is set to 10 (for comparisons with rotation invariant LBPs) and, in the other, it is set to 59 (for comparisons with sensitive to rotation LBPs).

2.4 Gender Classification

Gender classification is a pattern recognition problem, specifically it is an instance of supervised learning. In this type of problems, we have samples from various classes whose class labels are known. Those samples can be divided into two sets, the *training* set and the *test* set. The features extracted from the training set, along with the corresponding class labels, are passed to a *learning algorithm* which uses this information to build a *classifier*. Once the model has been trained, test features can be fed into the classifier in order to obtain a prediction of the class label of that test sample. Unlike in other pattern recognition problems, in gender classification special care should be taken for not including images of the same individual in the same set (training or test). The reason is that if the same subjects appeared in both set, the classifier could learn how to recognise individuals instead of learning about the characteristics of each gender.

Summarising, the aim of this classification step is to build a classifier (also known as model) with the ability to predict the gender of previously unseen faces. This *model* could be expressed in many forms, for example, graphs, algebraic equations or probability functions. In this section, the models involved in the thesis as well as the learning algorithms used for creating them are introduced.

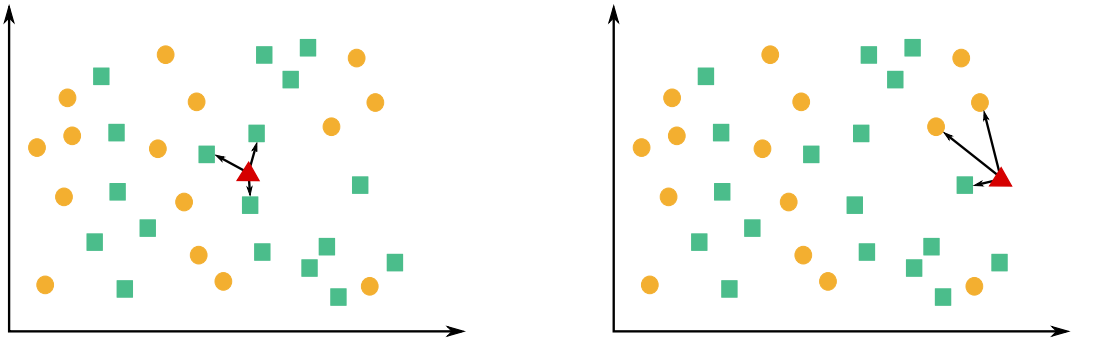
2.4.1 K-Nearest Neighbour Classifier

The k-Nearest Neighbour (k-NN) is a very straightforward classification rule which provides good performances despite its simplicity. In this case, no model is created and therefore, there is no learning process. Its simplest version is 1-NN (where $k = 1$) and it works as follows. Given a test sample $\mathbf{z}$, the classifier looks for which of the training samples is the most similar to $\mathbf{z}$, and predicts $\mathbf{z}$ to have the same label as that sample. In the general case, when taking into account k nearest neighbours, the classifier works as follows:

for *each test sample* $\mathbf{z}$ **do**
  Find the k most similar training samples
  Predict $\mathbf{z}$ to have the most common class label among those k samples
end

In the previous algorithm there is an unspecified detail, how to quantify the similarity between two samples. There are different ways to do that, in our experiments this quantification is done by measuring distances between samples represented in a vectorial space. In particular, the Euclidean distance and the Chi Square distance are used. Being d the number of features, let $\mathbf{z} = (z_1, z_2, ..., z_d)$ and $\mathbf{y} = (y_1, y_2, ..., y_d)$ be the feature vector representing a test sample and a training sample, respectively. Then, the Euclidean distance between them is,

$$Euclidean(\mathbf{z}, \mathbf{y}) = \sqrt{\sum_{i=1}^{d} (z_i - y_i)^2} \tag{2.8}$$



(a) The test sample is classified as "green square". (b) The test sample is classified as "yellow circle".

Figure 2.12: Examples of k-NN classification with $k = 3$. The test sample (red triangle) should be classified as belonging to class "yellow circle" or "green square".

and the Chi Square distance is,

$$\chi^2(\mathbf{z}, \mathbf{y}) = \sum_{i=1}^{d} \frac{(z_i - y_i)^2}{z_i + y_i}. \tag{2.9}$$

In order to choose the k nearest neighbours, the distance from the test sample to all training samples should be computed. Therefore, this algorithm is computationally intensive for large training sets.

Figure 2.12 shows the k-NN classification for two different test samples. It can be seen that it does not matter how far the k nearest neighbours are from the test sample, this classifier always takes into account the number of neighbours indicated.

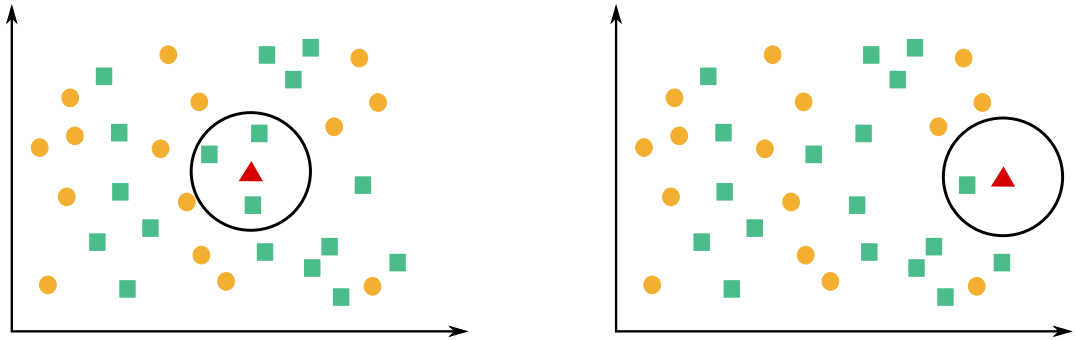
Note that depending on the number of classes and the value of k, there could be a tie for several classes. If those cases, a rule for breaking the tie should be defined. In our experiments, ties are not possible because there are two classes and we always choose k to be an odd number.

2.4.2 Parzen Windows Classifier

As the previous classifier, Parzen Windows [54] does not build a model. Parzen Windows is a non-parametric technique which attempts to estimate the underlying density functions from the training data. Given a test sample $\mathbf{z}$, the training data is used to compute the posterior probabilities of $\mathbf{z}$ belonging to each of the classes. The sample $\mathbf{z}$ is assigned to the class with the maximum posterior probability.

More formally, considering a region R of the feature space which is centred at $\mathbf{z}$, an estimate of the probability of the underlying distribution at point $\mathbf{z}$ is,

$$\hat{p}(\mathbf{z}) = \frac{M}{NV} \tag{2.10}$$



(a) The test sample is classified as "green square". (b) The test sample is classified as "green square".

Figure 2.13: Examples of classification with Parzen Windows. The test sample (red triangle) should be classified as belonging to class "yellow circle" or "green square".

where M is the number of samples that fall within region R, N is the total number of training samples and V is the volume of R. Assuming that R is a $d-$dimensional hypercube centred at $\mathbf{z}$ with side length h and volume V, a Parzen Window, $\varphi(\cdot)$, is defined so that, for all training samples $\mathbf{x}_i$, $\varphi\left(\frac{\mathbf{z}-\mathbf{x}_i}{h}\right)$ is equal to 1 if $\mathbf{x}_i$ falls inside the hypercube and is 0 otherwise. Thus, the total number of samples within the hypercube is given by,

$$M = \sum_{i=1}^{N} \varphi\left(\frac{\mathbf{z}-\mathbf{x}_i}{h}\right). \tag{2.11}$$

By substituting Eq. 2.11 and the volume of R in Eq. 2.10, an estimate for the probability of sample $\mathbf{z}$ is,

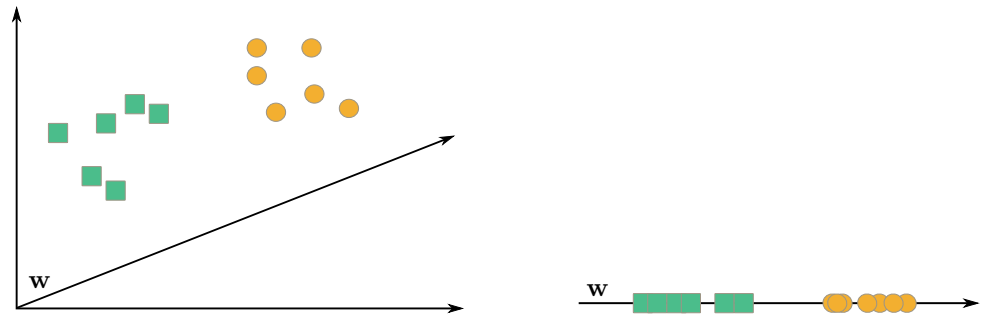
$$\hat{p}(\mathbf{z}) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{V}\,\varphi\left(\frac{\mathbf{z}-\mathbf{x}_i}{h}\right). \tag{2.12}$$

In order to obtain a more general approach to density estimation, other classes of windowing functions rather than the hypercube can be used. A popular choice for the window function is the Gaussian. In that case an estimate for the probability of sample $\mathbf{z}$ is given by,

$$\hat{p}(\mathbf{z}) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{V\sqrt{2\pi}}\, exp\left[-\frac{1}{2h^2}(\mathbf{z}-\mathbf{x}_i)^2\right]. \tag{2.13}$$

For classifying sample $\mathbf{z}$, Equation 2.13 is used to separately calculate the probability of that samples, $\hat{p}(\mathbf{z})$, to belong to each class. That is done by taking into account only the training samples, $\mathbf{x}_i$, that belong to one class at a time. Finally, $\mathbf{z}$ is assigned to the class with higher probability.

Figure 2.13 depicts the Parzen Windows classification of two test samples. As can be seen, the number of training instances that are considering in the classification varies.



(a) Data represented in the original space. The LDA basis vector is indicated by **w**.

(b) One-dimensional space obtained after applying LDA.

Figure 2.14: Two-dimensional data drawn from two classes before and after applying LDA.

Unlike in the k-NN classification, using this classifier we fix the size of the region, not the number of samples.

All training instances are taken into account when estimating the probability of a test sample which makes Parzen Windows computationally expensive with large training sets.

2.4.3 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) [29] is a technique that searches for a linear transformation of the data into a new representation space where the classes are better separated. This is achieved by maximising the ratio of between-class to within-class separation which guarantees maximal separability.

The linear transformation provided by LDA maps the original $D-$dimensional space into an $F-$dimensional space where $F = c - 1$ being c the number of classes. Figure 2.14 shows an example of how LDA will transformed the space of a 2-class problem into a new 1-dimensional space.

Formally, given an original feature vector $\mathbf{x} \in \Re^{\mathbf{D}}$, its projection, $y \in \Re^{F}$, into the new space is defined by $y = \mathbf{w}^T\mathbf{x}$. The learning algorithm for LDA looks for the linear transformation, $\mathbf{w}$, that maximises,

$$J(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{S}_B\mathbf{w}}{\mathbf{w}^T\mathbf{S}_W\mathbf{w}} \tag{2.14}$$

where $\mathbf{S}_B$ and $\mathbf{S}_W$ are respectively the between-class and within-class scatter matrices. LDA is optimal when the classes are normally distributed. Considering a two-class (C_1 and C_2) problem and assuming that $p(\mathbf{x}|C_k) \sim N(\mu_k, \Sigma)$, the linear transformation $\mathbf{w}$ is,

$$\mathbf{w} = \Sigma^{-1}(\mu_1 - \mu_2) \tag{2.15}$$

where μ_1 and μ_2 are the means of the classes C_1 and C_2, respectively, and Σ is the covariance matrix. LDA assumes the same covariance matrix for all classes, that is, $\Sigma_k = \Sigma, \forall k$.

Given a test sample, $\mathbf{z}$, its predicted class is,

$$\widehat{G}(\mathbf{z}) = \arg\max_k \delta_k(\mathbf{z}) \tag{2.16}$$

The linear discriminant function, $\delta_k(\mathbf{z})$, is,

$$\delta_k(\mathbf{z}) = \mathbf{z}^T\hat{\Sigma}^{-1}\hat{\mu}_k - \frac{1}{2}\hat{\mu}_k^T\hat{\Sigma}^{-1}\hat{\mu}_k + \log(\hat{\pi}_k) \tag{2.17}$$

where $\hat{\mu}_k$ is the estimated mean of each class, $\hat{\Sigma}$ is the estimated covariance matrix and $\hat{\pi}_k$ is the estimated prior probability of class C_k. All these values are estimated using the training data.

LDA is most commonly applied to a relatively low-dimensional intermediate space in order to avoid mathematical problems due to the fact that the number of samples per class is usually small with respect to the dimensionality of the original feature space (for details see [14]). Hence, in gender classification problems, LDA is usually applied after reducing the dimensionality with PCA [13].

2.4.4 Quadratic Discriminant Analysis

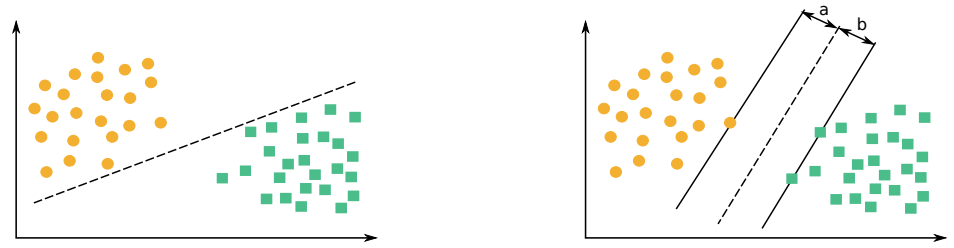
Quadratic Discriminant Analysis (QDA) is an extension of LDA which uses the same normal model but allows each class to have its own covariance matrix. This leads to quadratic decision boundaries instead of linear boundaries which could potentially fit the data better.

The classification rule is that of Equation 2.16, but using a quadratic discriminant function instead of a linear one. The quadratic discriminant function is,

$$\delta_k(\mathbf{z}) = -\frac{1}{2}\log(|\hat{\Sigma}_k|) - \frac{1}{2}(z - \hat{\mu}_k)^T\hat{\Sigma}_k^{-1}(z - \hat{\mu}_k) + \log(\hat{\pi}_k). \tag{2.18}$$

2.4.5 Support Vector Machine

Support Vector Machine (SVM) [23] is a widely used technique for binary classification. The main idea of SVMs is to find a decision boundary (a hyperplane) which optimally separates the $D-$dimensional data into its two classes. Assuming linearly separable classes, there are an infinity number of hyperplanes which correctly separate the samples of each class. Figure 2.15 shows a scatter plot of some data drawn from two different classes and two possible boundaries which perfectly separate both classes. Intuitively, looking at that figure, it seems that the best separation is achieved by the hyperplane that maximises the margins to the closest samples of each class (that is the boundary of Figure 2.15(b)). This is what the learning algorithm for SVMs does, it finds a linear decision boundary with the largest margins.



(a) Non optimal decision boundary.

(b) Optimal decision boundary with maximum margins.

Figure 2.15: Two alternative decision boundaries that perfectly separate the data in two classes.

So far we have assumed that the data is linearly separable, however we can encounter many problems where this assumption does not hold. In order to tackle non-linearly separable problems, SVMs use the "kernel trick". The idea is to map the data into a higher dimensional space where a linear separating boundary can be found. Figure 2.16 shows non-separable 2-dimensional data which can be separated by a linear boundary (a 3-D plane) after mapping it into a 3-dimensional space.

The kernel trick refers to the use of kernel functions which are a class of mathematical functions that allow us to operate in the new space without computing the coordinates of the data in that space. Although there are many kernel functions which can be employed, the most commonly used kernels for addressing gender classification problems are the Polynomial kernel and the Radial Basis Function kernel [46, 39].

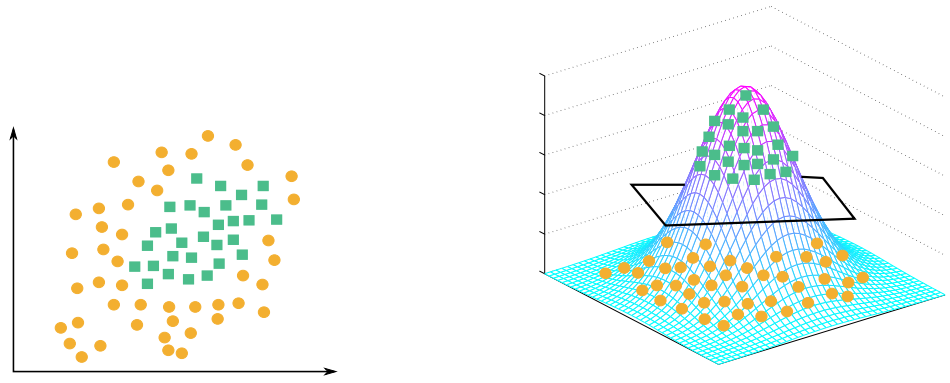
2.4.6 Ensemble Methods

Ensemble methods employ multiple models whose outputs are typically combined to obtain the final prediction [38]. The general architecture of an ensemble of classifiers is depicted in Figure 2.17. The principle underlying ensemble methods is that a decision based on a combination of individual predictions should be more accurate, on average, than any individual prediction.

A key factor in the design of ensembles is the combination strategy adopted. The members of the ensemble could predict class labels, posterior probabilities, or any other quantity. Depending on the type of the outputs, they can be combined following different rules. Two of the most popular rules are the *voting* and the *linear* combiners. The strategies followed for the ensembles used in this thesis are detailed in the corresponding chapter.

When choosing the base learners of the ensemble, the aim is to find classifiers which differ in their decisions so they complement each other. There are several ways in what they can be different from one another:

- Different learning algorithms for building the models.



(a) Non-linearly separable data.

(b) The same data is linearly separable after mapping it into a higher dimensional space.

Figure 2.16: Non-linearly separable problem before and after mapping the data into a higher dimensional space.

- Same model but with different parameter values, for example different k for k-NN classifiers.
- Different training sets.
- Different representations of the same information.

When ensembles are employed in Chapter 3, we choose to create the base learners using different representation spaces of the same faces. In particular, those various representations are provided by means of different face parts.

2.5 Performance Evaluation

Once we have built a model using one of the algorithms described in the previous section, we would need to evaluate the model's performance on some other data which has not been involved in the learning process. To this end, the available dataset should be divided into subsets. In other words, some data should be left out of the learning process and saved for evaluating the classifier. Therefore, in order to assess the classifier's performance, we need to decide how to divide the data. Another decision that should be taken is what performance measure is appropriate. Depending on some characteristics of the problem, for example, how many samples of each class we have, some measures are better indicators of the classification performance than others. In this section, we discuss the techniques adopted for dividing the data into subsets and the performance measures utilised in the thesis.

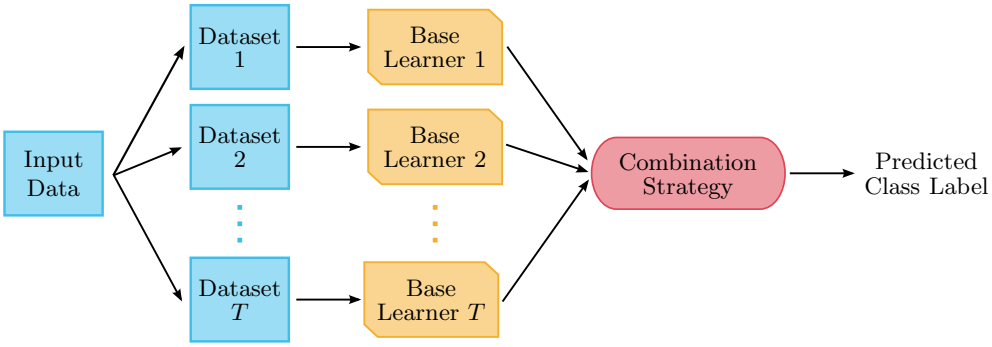


Figure 2.17: Architecture of an ensemble.

2.5.1 Partitioning the Data

In previous sections we have seen a possible solution for partitioning the data, that is, splitting the data into two sets, *training* and *test*. Randomly dividing the data into two sets is one among several possible approaches, however this is usually not the best way to assess how good the model performs. The parameters of the model are optimised to fit the training data, so significantly different models could be obtained when learning from various training sets. Unless we had available an enormous amount of data, in such case (which is rare in real problems) the models would not differ much. In order to obtain a more accurate estimate of how well our classifier would perform in the real world, we could take better advantage of the data we have by using a K-fold cross validation technique for the evaluation.

K-fold Cross-Validation

In *K-fold cross-validation*, the data is randomly divided into K parts, called *folds*. To obtain a pair of datasets, $K-1$ folds are combined to form the training set, while the remaining fold is kept as the test set. Doing this K times, each time taking as the test set a different fold, we get K pairs of training-test sets. Figure 2.18 shows how the data is split for a 5-fold cross-validation. With each of the K pairs of sets, a model learns from the corresponding training set and is evaluated with the test set. Therefore, the model has never seen the test data which simulates the real world.

With K-fold cross-validation, we obtained K set of predicted class labels. In order to calculate the performance measure, the predicted labels in all folds are considered. For example, if the measure consist in the number of correctly classified samples, we would add the successfully predicted labels of the K folds and then divided it by K.

As has been previously mentioned, in gender classification problems, all images of the same subject should be included in the same subset. That way, we avoid contamination effects produced by learning from the same subjects who are going to be classified.

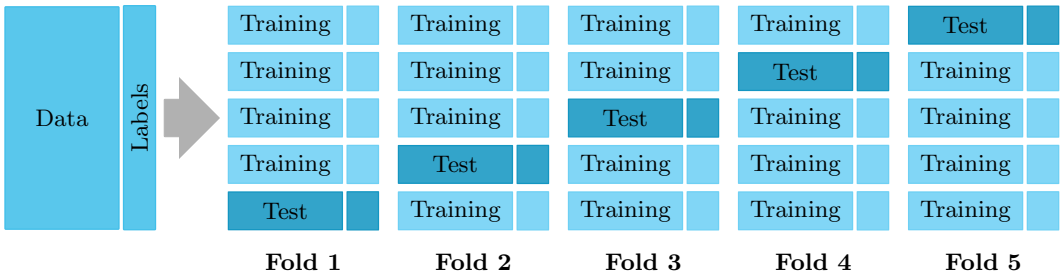


Figure 2.18: Splitting the data into $K = 5$ folds for cross validation.

Table 2.1: Confusion matrix of a two-class problem.

		Predicted Class	
		Positive	*Negative*
True Class	*Positive*	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

2.5.2 Performance Measures

Considering a two-class problem, the classification model predicts one of the two classes which are usually referred to as positive and negative class. The terms positive and negative usually denote the minority and the majority class, respectively. Given a classifier and an instance (face image), there are four possible outcomes. If the instance is positive and it is classified as positive, that counts as a *true positive*; if it is classified as negative, it counts as a *false negative*. If the instance is negative and it is classified as negative, it counts as a *true negative*; if it is classified as positive, it counts as a *false positive*. Given a classifier and a set of instances (the test set), the outputs yielded by the classifier can be represented by a 2×2 *confusion matrix*. Table 2.1 shows the confusion matrix for a two-class problem where the rows indicate the actual class and the columns indicate the predicted class of a set of test samples. The numbers along the main diagonal represent the correct decisions made, and the number of the secondary diagonal shows the errors between the classes.

Several performance measures can be easily derived from combining the four statistics represented in the confusion matrix. Following, we present the measures employed to evaluate the performances of the models utilised in the thesis.

Accuracy

The accuracy is probably the most popular way of evaluating the effectiveness of a model. It computes the percentage of correctly classified samples over the total number of samples.

From the values of the confusion matrix (see Table 2.1), the accuracy is,

$$ACC = \frac{TP + TN}{TP + FN + TN + FP}. \tag{2.19}$$

This measure which is only based on the correct classifications is appropriate as a performance measure when the data set is balanced. However, it is not suitable when there is a considerable imbalance between classes. For example, let us assume that we have a set of face images which is imbalance, there are 90% of males and 10% of females. This set is divided into training and test subsets keeping the ratio of male-female. Then, a model learns from that training set and we evaluate its performance using the test set. Given such conditions, the model probably predicts "male" many more times than "female". If we think about an extreme case, it might only predict "male" because it has seen a much larger number of males than females. As a result, the accuracy could be 90%, since the test set has that amount of males (we kept the ratio male-female of the original set). However, this is not a good assessment because that model lacks the ability of correctly classifying female faces. Empirical and analytical studies have shown that this measure can be biased when the classes are imbalanced.

Geometric Mean

The geometric mean (*G-mean*) indicates the balance between classification performances on the positive and negative classes. This measure takes into account the sensitivity and the specificity of the model. Sensitivity relates to the model's ability to correctly predict the positive class and specificity relates to the model's ability to correctly predict the negative class. Given the confusion matrix, the sensitivity is defined as,

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.20}$$

and the specificity is,

$$Specificity = \frac{TN}{TN + FP}. \tag{2.21}$$

From these statistics, the measure *G-mean* is calculated by,

$$G\text{-}mean = \sqrt{Sensitivity \times Specificity}. \tag{2.22}$$

The value of *G-mean* is low when the model is strongly biased to one of the classes. Therefore, *G-mean* can be considered an unbiased evaluation measure even in problems with imbalance classes which could result in biased classifiers.

D-prime

D-prime (d') is a statistic widely used in Signal Detection Theory which provides information about the bias and sensitivity of a model. The bias relates to the tendency of

misclassifying a negative sample in favour of the positive class, while, as previously mentioned, the sensitivity (also known as hit rate) indicates the model's ability to predict the positive class correctly. The measure *D-prime* is,

$$d' = Z_H - Z_{FA} \tag{2.23}$$

where Z_H and Z_{FA} are the z-scores of the hit rate and the false alarms, respectively. The hit rate and false alarms are two of the metrics that can be calculated from the confusion matrix. The hit rate (or sensitivity) is given by Equation 2.20, and the false alarms are,

$$FA = \frac{FP}{FP + TN}. \tag{2.24}$$

The z-scores are used to compare the values of the two metrics assuming that they follow a normal distribution. The z-score of a value x is,

$$Z_x = \frac{x - \mu}{\sigma} \tag{2.25}$$

since normality is assumed, $\mu = 0$ and $\sigma = 1$. This z-score indicates how many standard deviations away from the mean the sample x lies. However, for computing d' the area under the standard normal curve is needed, which can be obtained by looking up the value of the z-score in a Z table (standard normal table).

The computation of *D-prime* is based on two indices computed separately on the two classes, which penalises biased classifiers. As a result, *D-prime* is robust to skewed classes and a suitable measure for assessing the model in those type of problems.

2.6 Statistical Analysis

For providing a statistical analysis of the performances of various classification models, we employ statistical tests. A statistical test is a procedure to check if a hypothesis holds by analysing the data. Therefore, all the statistical tests are based on a *null hypothesis* which is assumed certain and they search for evidence in the data to reject such hypothesis. In the next sections, we explain the statistical tests that are involved in the thesis. For performing these tests, we employ KEEL data mining software [3].

2.6.1 Iman-Davenport's Test

Iman-Davenport's [34] null hypothesis states that all models are equivalent. Therefore, a rejection of this hypothesis implicates that there are significant differences among the performances of the classification models studied. This statistic is a derivation of Friedman's statistic which ranks the models for each dataset separately according to their performances. The model with the best performance gets the rank 1, the second best gets rank 2, and so on. In case of ties, the average rank is given to those models. Let r_j^i be the rank of the j^{th}

model on the i^{th} experiment and $R_j = \frac{1}{n}\sum_i r_j^i$ be the average rank of the j^{th} model over all experiments. Then, the Friedman's test is,

$$\chi_F^2 = \frac{12n}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right] \tag{2.26}$$

where k is the number of models and n the number of experiments. This statistic follows a Chi-square distribution with $k-1$ degrees of freedom. From Friedman's test we can derive Iman-Davenport's test as,

$$F_F = \frac{(n-1)\chi_F^2}{n(k-1)-\chi_F^2} \tag{2.27}$$

which is distributed according to the F-distribution with $k-1$ and $(k-1)(n-1)$ degrees of freedom. In order to reject the null hypothesis, the F_F statistic should be higher than the corresponding value of the F-distribution. In that case, significant differences exist among the performances of the classification models studied.

2.6.2 Holm's Method

Holm's method [31] is a post-doc test which is applied to find out whether a control classification model presents statistical differences with respect to the remaining models. The control model is usually the best according to Friedman's ranking.

Holm's null hypothesis assumes that the performance of the control model is statistically equivalent to the performance of the other models. It consists in an iterative process that sequentially checks the hypothesis associated with each model (except the control model, since it is not compared with itself). For comparing the i^{th} and j^{th} models, the statistic z is calculated by,

$$z = \frac{(R_i - R_j)}{\sqrt{\frac{k(k+1)}{6n}}}, \tag{2.28}$$

where R_i and R_j are the average rank calculated with Friedman's test for two models, one of which is the control model. Then, this value z is used to obtained the p-value from the normal distribution which will be associated with the hypothesis stating that the performance of model i is equivalent to that of model j.

All the hypotheses are ordered according to their p-values, so that $p_1 \leq p_2 \leq \cdots \leq p_{k-1}$, where p_i is the p-value of the i^{th} hypothesis. Holm's method checks the following condition:

$$p_{(i)} < \frac{\alpha}{k-i} \tag{2.29}$$

where α indicates the confidence level that should be reached to reject the hypothesis. If the condition is met, the i^{th} hypothesis is rejected (i.e. the performance of the i^{th} model is considered statistically worse than the performance of the control model), and the process

continues with p_{i+1}. As soon as one hypothesis cannot be rejected, the process stops and the remaining hypotheses are supported.

Holm's procedure also allow us to compute the adjusted p-values. Unlike the "unadjusted" p-values which only take into account one hypothesis, the adjusted p-value is calculated considering the collection of hypotheses. An adjusted p-value for a particular hypothesis is the smallest overall (that is, "experimentwise") significance level at which that particular hypothesis would be rejected. The adjusted p-value for the i^{th} hypothesis is, $min\{v, 1\}$ where $v = max\{(k-j)\, p_j : 1 \leq j \leq i\}$, k is the total number of models and p_j is the unadjusted p-value of the j^{th} hypothesis.

In the statistical analysis presented in this thesis, the results of Holm's method are presented as shown in Table 2.2. The first column of the table shows the models in ascending order with respect to the adjusted p-values associated to the corresponding hypothesis, the second column shows their adjusted p-values. At the bottom of the table, the control model is marked in bold. The models associated to the rejected hypotheses are shown above the double line.

Table 2.2: Example of the presentation of the results of Holm's method. The hypothesis associated to the models above the double line are rejected with a significance level $\alpha = 0.95$.

Models	P_{Holm}
Model A	adjusted p-value A
Model B	adjusted p-value B
Model C	adjusted p-value C
Model D	adjusted p-value D
Control Model	

2.6.3 Wilcoxon's Signed Rank Test

Wilcoxon's Signed Rank test [66] provides pairwise comparisons, so statistical differences between each pair of classification models can be found. For each pair, Wilcoxon's null hypothesis assumes that both classification models perform equally. This test proceeds by ranking the differences in performance of two models. Let d_i be the difference between the performances of two classification models on the i^{th} experiment. Then, the differences $d_i \; \forall i$ are ranked according to their absolute values, assigning rank 1 to the smallest difference. Let R_+ and R_- be calculated as follows:

$$R_+ = \sum_{d_i > 0} rank(d_i) \tag{2.30}$$

$$R_- = \sum_{d_i < 0} rank(d_i). \tag{2.31}$$

The ranks where $d_i = 0$ are split evenly among R_+ and R_-. When there is an odd number of cases where $d_i = 0$, one of those ranks is ignored. Being $Z = min(R_+, R_-)$, if Z is less or equal than the Wilcoxon distribution for n degrees of freedom, then the null hypothesis stating that both classification models are equivalent is rejected.

In the statistical analysis presented in this thesis, a summary of the results of Wilcoxon's Signed Rank test is presented as shown in Table 2.3. In the table, if the model in the row significantly outperforms the one in the column the symbol "•" is shown, the opposite case is indicated by the symbol "○". Those differences marked above the main diagonal have a significance level $\alpha = 0.90$ and, for those below it, the significance level is $\alpha = 0.95$.

Table 2.3: Example of Wilcoxon's Signed Rank test results. The symbol "•"= the model in the row outperforms that of the column, and "○"= the model in the column outperforms that of the row. Above main diagonal, the significance level is $\alpha = 0.9$, below the main diagonal $\alpha = 0.95$.

	1	2	3	4
Model A (1)	-		○	○
Model B (2)		-	○	○
Model C (3)	•	•		
Model D (4)	•	•		

CHAPTER 3

The Role of Face Parts and Their Complementarity

With the information provided by the whole face, automatic systems can successfully classify the gender of a person, but what are the most useful face parts for distinguish between genders? In this chapter, we study the role of the most prominent face parts in gender classification as well as whether different face parts contain complementary information.

3.1 Motivation and Background

Why could it be interesting to evaluate the effectiveness of using face parts for gender discrimination? A first answer to this question could be related to classification under partial occlusions of the face. In the real world, faces can be partially covered by accessories such as sunglasses, hats, and scarves. Therefore, by evaluating the discriminant capabilities of isolated face parts, we could determined whether it is feasible or not to classify the gender in these situations. One way to proceed would be to use the effectiveness of classifying gender given a visible part as the reliability of the prediction. An extended approach could evaluate the efficacy of two or more visible face parts to jointly predict the gender.

In the literature, some papers studying the importance of face parts in gender classification [36, 19] have been published. Kawano et al. [36] evaluated the differentiation capabilities of the full face, the jaw, the lips/mouth, the nose and the eyes. These regions were manually clipped, represented by an appearance-based method, and classified using Linear Discriminant Analysis. The best classification rates were 93.7% and 89.8% by the full face and jaw, respectively, while the worst accuracy rates were lower than 80% and correspond to the nose and eyes. These results are certainly contrary to general intuition, according to which the eyes or mouth seem more relevant than the jaw to discriminate between women

and men. Apart from subjective judgements, the generalisation of these results is limited when taking into account the fact that the database only contained expressionless Asian faces.

In other related work, Buchala et al. [19] compared the roles of the full face, the eyes region (top half of the face) and the mouth region (bottom half of the face) in gender classification. Subimages with these face parts were extracted and their dimensionality reduced using Principal Component Analysis, Curvilinear Component Analysis and Self Organising Maps. Then, the classification was performed by a Support Vector Machine with a Radial Basis Function kernel. The best gender classification rate was achieved from a joint representation obtained from the full face, eyes and mouth. When only individual parts were considered, the accuracies were 85.5% and 81.25% for the eyes and mouth, respectively. In that study the role of more general face parts (with respect to [36]) was evaluated. However, they did not include the nose and the jaw (since they only considered the face in two halves), and the hair was not removed in full faces. They admit that the hair had a dominating effect on the gender classification based on full faces.

Considering that the results about the discriminant capabilities of face parts obtained in [36, 19] depend on specific components (dataset, classifier, face part description), care should be taken when interpreting their conclusions. For example, the relation between gender classification rates from eyes and mouth is inverted in the two studies, eyes being more discriminant than mouth in [19], and mouth more accurate than eyes in [36].

In this chapter, we go beyond these previous related works, as regards the number of face parts and the diversity of the experimental design. A total number of eight face parts are considered, which are the right eye, both eyes, the nose, the mouth, the chin, the internal face, the external face and the full face. Each one of them is accurately detected in the image in order to not contain extra information that is not provided by the face part in question. For each face part involved, the performances of four different classifiers are compared to check if the discriminant capabilities of the face parts are consistent among classifiers. Besides this comparative study, an analysis of the complementarity of the information provided by different face parts is presented. This complementarity is evaluated by means of ensembles of classifiers where the input to each base classifier is a different face part. Several ensembles are defined varying the number of face parts involved and the method chosen for combining the outputs of the base classifiers.

3.2 The Role of Face Parts

In previous related works, Kawano et al. [36] studied the role of the eyes, right eye, nose, mouth, jaw and full face in gender classification. The face images were cropped manually to obtain images containing only those face parts. In particular, the area considered as eyes consisted in a rectangular region containing both eyes and the space between them, and the nose was a squared area containing the nose and its surroundings. Although, the face parts were produced manually to ensure a correct selection, the definition of these regions was

not accurate enough to provide subimages containing the isolated face part and not much extra information. A more general division of the face in parts was presented by Buchala et al. [19] in a study of the discriminant capabilities of the top and bottom halves of the face.

Our aim is to provide a detailed evaluation of the role of face parts in gender classification, broadening what can be found in the literature. To this goal, eight different face parts are selected. In the following sections, the processes for locating and describing these face parts are explained in depth.

3.2.1 Division of the Face in Parts

In this study the discriminant capabilities of five isolated face parts and three global parts are analysed. Particularly, the isolated face parts are: both eyes (including eyebrows), the right eye, the nose, the mouth, and the chin; and the global parts are: the internal face (eyes, nose, mouth, chin), the external face (hair, ears, contour), and the full face. The process for locating and extracting these face parts from the face images is detailed next.

Given a frontal face image and the coordinates of the two eyes, the regions containing the face parts of interest are defined according to expected proportions of an aesthetic face. The extraction process of face parts is based on an empirical rule about the ideal balance of a human face sketched by Leonardo da Vinci [48]. Da Vinci stated that perfect facial harmony exists when the face can be divided into three equal horizontal sections whose boundaries match with the hairline, the eyebrows, the bottom of the nose and the chin, and it can also be partitioned into five vertical sections that approximate the width of one eye. This rule about the proportions of the face was explained in Section 2.2 as a method for detecting faces. Here, we use a modification of this technique for locating each of the face parts of interest in the image.

A grid that helps to locate the position of each face part of interest is superimposed on the face images. The vertical sections of this grid follow the previous rule, but the horizontal sections are based on a different arrangement, in which some key face features are centred in certain cells of the grid. The layout of this grid is automatically computed from the knowledge of the coordinates of the eyes. Let (x_r, y_r) and (x_l, y_l) be the coordinates of the centres of the right and left eye, respectively. Then, the distance between the eyes is $d = \sqrt{(x_l - x_r)^2 + (y_l - y_r)^2}$ and the middle vertical position between them is $p_m = \frac{y_r + y_l}{2}$. Given this data, all the necessary points for creating the grid are calculated using Equations 3.1 to 3.12.

$$x_1 = x_r - 0.75 \times d \quad (3.1) \qquad x_4 = x_l - 0.25 \times d \quad (3.4)$$

$$x_2 = x_r - 0.25 \times d \quad (3.2) \qquad x_5 = x_l + 0.25 \times d \quad (3.5)$$

$$x_3 = x_r + 0.25 \times d \quad (3.3) \qquad x_6 = x_l + 0.75 \times d \quad (3.6)$$

Figure 3.1: Grid based on the eye coordinates, created from the Equations 3.1 to 3.12, used to locate the face parts in the image (the points calculated with the equations are marked in green).

$$y_1 = p_m - 1.15 \times d \quad (3.7) \qquad y_4 = y_6 - 0.5 \times d \quad (3.10)$$
$$y_2 = y_1 + 0.8 \times d \quad (3.8) \qquad y_5 = y_6 - (0.33 \times (y_1 - y_6)) \quad (3.11)$$
$$y_3 = y_1 + 1.4 \times d \quad (3.9) \qquad y_6 = p_m + 1.8 \times d \quad (3.12)$$

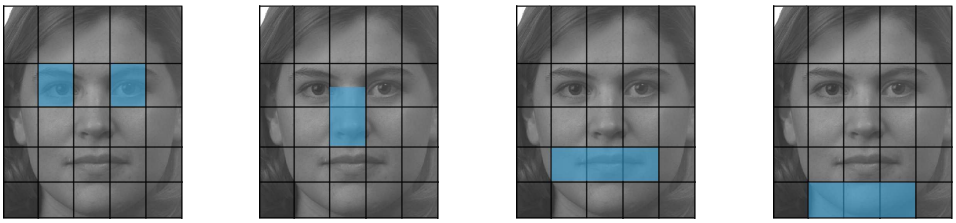
An example of the resulting grid superimposed on a face image is shown in Figure 3.1. The regions of the grid can fully enclose the eyes, but other important parts like the mouth, nose, and chin are delimited only in one direction. For example, the nose is enclosed in the vertical direction, while the mouth is only enclosed horizontally. To fully describe these features, new zones are created by joining adjacent (parts of) cells of the grid. In Figure 3.2 the cells of the grid selected for creating the subimages containing the isolated face parts are marked in blue. Notice that the nose is extracted by joining one cell and half of another and the mouth and chin are enclosed in three adjacent cells.

The global face parts included in our study are:

- The **Full face** which is the whole area enclosed by the previously defined grid.
- The **Internal face** consisting in the area delimited by the eyebrows and the chin without including ears or hair.
- The **External face** which mainly consists of the forehead, ears and part of the hair. This part results of subtracting the internal face from the full face.

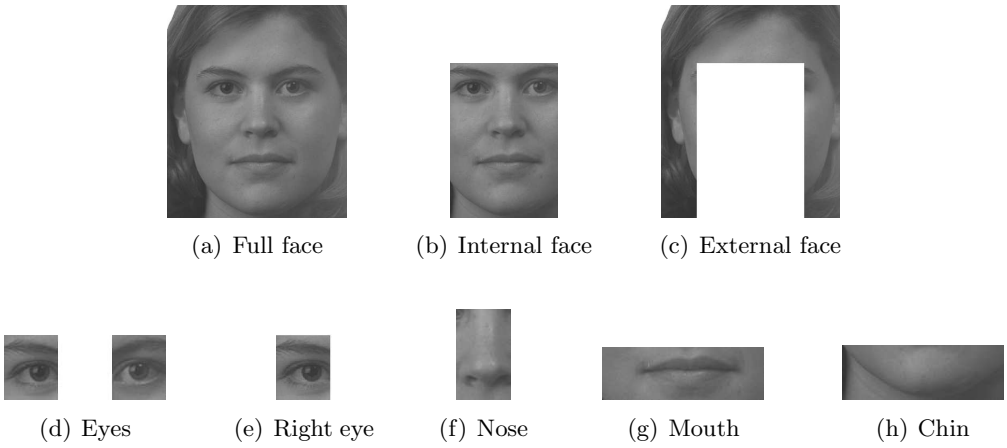
The isolated face parts considered are:

- The **eyes** which only include the eyes and eyebrows but not the area in between.



(a) Eyes (b) Nose (c) Mouth (d) Chin

Figure 3.2: Cells containing the isolated face parts (marked in blue). Notice that to extract a subimage of the nose one cell and half of another are selected.



(a) Full face (b) Internal face (c) External face

(d) Eyes (e) Right eye (f) Nose (g) Mouth (h) Chin

Figure 3.3: The extracted subimages containing the face parts of interest.

- The **right eye** is the right eye contained in the part with both eyes.
- The **nose** which spans from the corner of the eyes to the tip of the nose.
- The **mouth** which is closely delimited vertically but it takes some of the area beyond the corner of the lips (this is due to the high variability of the width of different people's mouth).
- The **chin** which includes some small parts of the neck due to its normal curvature.

Figure 3.3 shows an example of all the eight subimages containing these face parts extracted from a face.

3.2.2 Experimental Methodology

In order to evaluate the effectiveness of the selected face parts for classifying gender, several experiments are performed. For conducting these experiments, the general methodology presented in Chapter 2 is adopted. The steps 2 to 4 of the methodology are performed independently for each of the eight face parts in which the face has been divided. Therefore, this process can be seen as a standard pattern recognition problem. In fact, there are eight classification tasks with the same number of instances and the same set of labels which can be handled as individual problems. Below, the particularities of each step of the methodology are described.

Step 1. Preprocessing

Given a face image, it is converted to grey scale format and the grid enclosing the face parts is computed using the coordinates of the eyes. These coordinates are provided in the face databases. Once the region containing the face is located, this area of the image is equalised which provides better contrast in that area and avoids some problems due to different illumination conditions among the images. Finally, one subimage per face part is extracted as described in Section 3.2.1. An example of the resulting subimages is shown in Figure 3.3.

Step 2. Feature Extraction

Given a subimage containing one of the face parts in which the face has been divided following the process detailed in Section 3.2.1, it is scaled down to a low-resolution image, where new pixels are computed by interpolating the original ones. The new reduced subimage is then represented as a vector of the grey level values of the pixels in the image from the left top corner to the right and downwards.

In order to reduce the dimensionality of the feature space, Principal Component Analysis (PCA) is applied. PCA searches for a feature space which basis vectors correspond to those with the maximum variance in the original space. For obtaining PCA features, first the PCA basis vectors are calculated from the grey level values of the subimages in the training set. Then, this transformation is applied to the grey level features extracted from the training and test sets. A more detailed explanation of this technique is provided in Section 2.2.

Step 3. Classification

Due to the fact that only individual classifiers are considered in this first study, the classification process follows a standard scheme. First, a classifier learns from the features extracted from training face parts. Then, the classification performance of that trained classifier is estimated by using an independent test set of the same face parts that belong to subjects which were not in the training set. In our experiments, four different classifiers are involved:

- Support Vector Machine (SVM),
- K-Nearest Neighbour (k-NN),
- Quadratic Discriminant Analysis Classifier (QDC), and
- Parzen Windows Classifier (Parzen).

For a further understanding of the classification process and these particular classifiers, see Section 2.4.

Step 4. Performance Assessment

The performance of the classifiers is evaluated by means of the correct classification rate.

3.2.3 Face Image Dataset

The experiments involve two standard datasets of face images, FERET and XM2VTS, both described in Sections A.2 and A.4, respectively. From FERET, 2,147 images of 256×384 pixels from 834 subjects were used. From which 1,305 were male faces and 842 were female faces, resulting in a ratio of males to females of 1:0.6. From XM2VTS, 1,378 images of 720×576 pixels from 203 subjects were used. From which 646 were male faces and 732 were female faces, resulting in a ratio of 0.8:1. All these images show a frontal view of faces without occlusions. Both of these databases provide the coordinates of the eyes for each of the face images, which are used in the process of locating the different face parts.

3.2.4 Experimental Setup

In this study, one experiment per each combination of face part, classifier and dataset is performed, resulting in a total of 96 experiments (8 *face parts* $\times$ 6 *classifiers* $\times$ 2 *datasets*).

The classification results of each of the experiments are estimated by a 5-fold cross-validation technique. Special care is taken to have all images of the same individual in the same partition, to avoid the contamination effects that this could produce. Additionally, when creating these partitions, the class balance of the original datasets is maintained.

Regarding the classifiers, the SVM uses a linear-polynomial kernel and the k-Nearest Neighbour classifier uses the Euclidean distance and $k = 1, 5, 9$. The implementations of the classifiers are those available in the PRTools Matlab package [26] with their default parameter values if not indicated otherwise.

Implementation Details

In this section, some implementation details are provide in order to ease the replication of the experiments of this study.

For scaling down the subimages containing the face parts of interest, the method adopted is to represent each cell of the grid with 36 (6×6) features. Consequently, the full face is

(a) Original Resolution (b) Reduced Resolution

Figure 3.4: Example of scaling a chin subimage to a lower resolution image.

Table 3.1: Number of dimensions of the feature space, build from the FERET dataset, before and after applying PCA.

	eyes	nose	mouth	chin	right eye	internal face	external face	full face
before PCA	72	54	108	108	36	432	468	900
after PCA	38	29	48	36	21	62	59	101

reduced to 30×30 features (since it consists of 5×5 cells), the internal face to 24×18 features (4×3 cells), the eyes to 6×12 features (1×2 cells), the nose to 9×6 features (1.5×1 cells), the mouth to 6×18 features (1×3 cells), and the chin to 6×18 features (1×3 cells). The external face is a special case, since it has an empty space were the internal face should be. Then, it results in 5 cells covering the forefront, and 8 cells counting both outer sides of the face which is a total of 13 cells (given that each cell is reduced to 36 features, it results in a total of 468 features). An example of the effect of this scale-down process is shown in Figure 3.4, where a subimage containing a chin is shown before and after this reduction. The dimensionality of the feature space is reduced by applying PCA and retaining 99% of the variance of the training data.

3.2.5 Results

In this section the results obtained are presented and discussed with respect to the image dataset involved. Following, a general discussion of all the obtained results is provided.

Results on FERET database

The number of features extracted from each face part, and the dimensionality of the final feature space after applying PCA to the grey level representations is shown in Table 3.1. The correct classification rates obtained in the experiments are shown in Table 3.2. These same results are shown in Figure 3.5.

The first point to be considered is related to the high discriminant power of the five isolated face parts, most of them led to classification rates above 80%, mainly for the SVM, QDC, 9-NN and Parzen classifiers. With regard to the relative importance of each of them, the nose was the most relevant part in identifying gender for all classifiers except QDC, for

Table 3.2: Classification accuracies per face parts on FERET images.

	eyes	nose	mouth	chin	right eye	internal face	external face	full face
SVM	85.47	86.36	81.61	81.56	81.51	92.37	87.71	95.21
1-NN	77.51	78.39	75.37	76.76	73.04	83.28	84.22	86.54
5-NN	82.07	82.49	79.79	78.53	79.65	87.48	86.17	87.89
9-NN	82.91	83.38	80.68	79.70	80.12	88.03	86.31	87.62
QDC	84.03	81.23	79.46	81.14	82.21	89.11	89.25	92.37
Parzen	80.35	81.88	79.74	78.67	76.81	85.38	85.98	87.44

which learning from eyes led to a better result. Moreover, these two parts (nose and eyes) were more discriminant than the mouth and chin for all classifiers.

As expected, gender classifiers based on global parts of the face were more accurate than those based on individual parts. When SVM and QDC learned from internal, external and full faces, classification rates were very close to or higher than 90%. In particular, the SVM with descriptions of full faces achieved accuracies above 95%. It is also interesting to remark that the external face by itself, dominated by hair, was almost as discriminant as the internal part of the face (see Figure 3.3 for an example of the internal and external face parts). Bearing in mind that the full face is the integration of the internal and the external parts, the effective contribution (in terms of new information) of each part with respect to the other was very small (no greater than 3%).

Results on XM2VTS database

The number of features extracted from each face part, and the dimensionality of the final feature space after applying PCA (retaining 99% of the variance of the original data) is shown in Table 3.3. Using these reduced spaces, the gender classification rates obtained by each classifier are numerically shown in Table 3.4 and, graphically, in Figure 3.6.

These results are clearly lower than those obtained with FERET. In these experiments over XM2VTS images, not all individual parts turned out to be as effective for classifying gender as in FERET. Taking into account their relative importance, the mouth and the

Table 3.3: Number of dimensions of the feature space, build from the XM2VTS dataset, before and after applying PCA.

	eyes	nose	mouth	chin	right eye	internal face	external face	full face
before PCA	72	54	108	108	36	432	468	900
after PCA	42	33	54	48	23	74	76	125

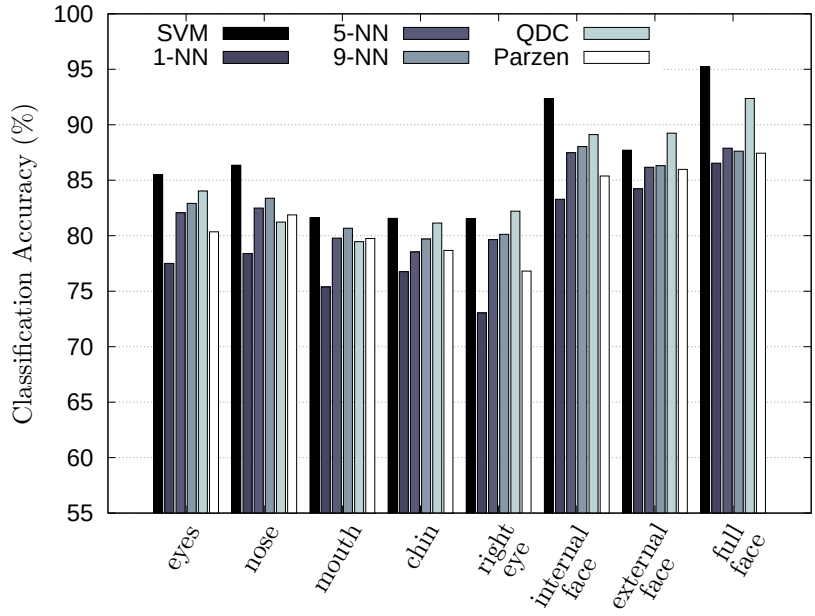


Figure 3.5: Classification accuracies per face parts on FERET images.

eyes (including the single right eye) were the most discriminant parts for all classifiers. Good results were also obtained for the chin dataset, while rates derived from the nose were surprisingly low.

Although global parts of the face are, in general, more effective than individual parts for gender classification, there are important singularities that can be highlighted. Firstly, the differences between global parts and the best individual part (the mouth) is negligible for 4 (similar) classifiers: {1,5,9}-NN and Parzen. Secondly, the internal part of the face appears to be as discriminant as the full face, which suggests that the external part does not contribute any new information. In fact, the external part produces lower recognition rates than the mouth.

3.2.6 Discussion of the Results

In general, the performances achieved using FERET images are superior than the results obtained with XM2VTS faces. The significant differences between both databases could explain the poorer classification results in XM2VTS. Specifically, the number of subjects in FERET is about 4 times larger than the number of individuals in XM2VTS, while the number of images per subject is twice as much in XM2VTS than in FERET. In other words, FERET contains more samples of different female and male faces than XM2VTS which contains more duplicates (images of the same person). Therefore, FERET provides a more varied group of faces than XM2VTS which allows the classifier to learn more general

Table 3.4: Classification accuracies per face parts on XM2VTS images.

	eyes	nose	mouth	chin	right eye	internal face	external face	full face
SVM	82.52	60.38	84.04	76.35	80.34	90.79	84.26	90.64
1-NN	77.58	59.87	83.24	68.44	76.06	83.97	79.18	83.75
5-NN	79.98	63.36	83.10	73.95	78.38	84.62	80.34	84.40
9-NN	79.98	63.65	83.53	73.81	79.25	84.91	80.41	84.55
QDC	80.99	64.52	83.39	76.56	80.12	90.43	83.02	83.39
Parzen	79.10	62.49	83.89	73.15	77.07	84.47	80.19	85.71

differences between genders.

Analysing only the results achieved when using isolated face parts, the eyes resulted in good classifications in both databases, while mouth and chin were also effective. The nose, in spite of being the best individual part in FERET, produced very poor results in XM2VTS. This could be explained by the different illumination conditions between databases. Due to the fact that the nose is usually the most salient part of the face, it is highly exposed to changes in illumination which might produce more unstable descriptions when compared to other face parts.

Due to the fact that global face parts contain much more information about the face than isolated parts, they led to better classification accuracies. As expected, the full face was the most successful global description, closely followed by the internal face. It is worth noting that the external part of the face (composed of part of the hair, the forehead, the ears and the contour) provided valuable discriminant information for predicting gender, since there is a high correlation between gender and traditional cultural patterns such as the length of hair and the use of earrings. However, for this same reason, this external part could lead to misclassification depending on the scenario of the problem, for instance, if the person was in disguise.

As regards the influence of classifiers on the relevance of face parts, the results show a high correlation among the behaviours of the different classifiers which allows to draw robust conclusions about the discriminant capabilities of each face part.

3.3 Complementarity of Face Parts

In the previous section, it was proved the capability of individual face parts to successfully distinguish between genders. The use of these isolated face parts is useful for classifying gender in the presence of partial occlusions of the face. The fact that a subject's face is simultaneously described in different feature spaces, coming from different face parts, has triggered research into the possibility of creating ensembles based on such diversity.

The goals of this section are to analyse whether a potential complementarity among face parts exists and which combinations of those parts provide more discriminant information

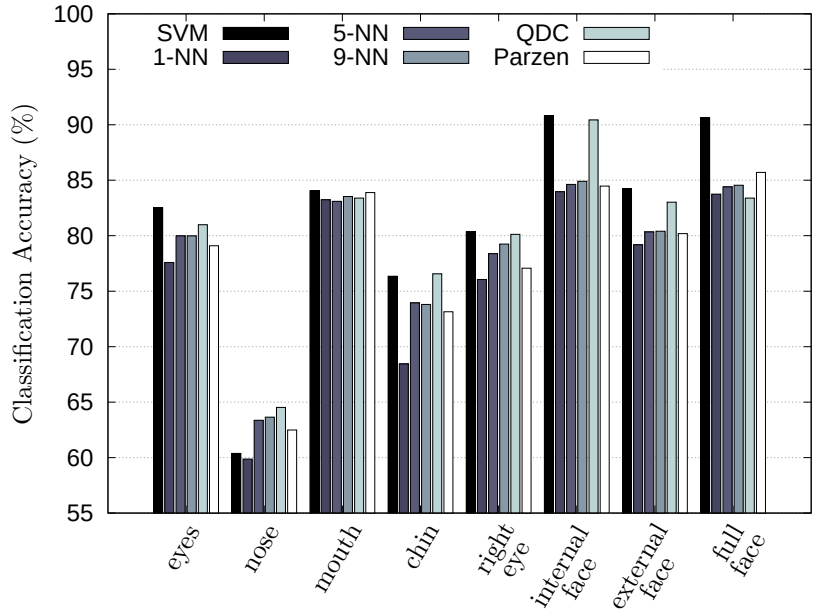


Figure 3.6: Classification accuracies per face parts on XM2VTS images.

for solving gender classification problems. Assuming that not all face parts might be available, ensembles of classifiers based on the visible parts could model the joint contribution of the involved parts to predict the gender. To this end, ensembles of classifiers based on several face parts are designed. These ensembles are then compared with plain classifiers which are trained with isolated face parts.

A First Evidence of Complementary Information between Face Parts

A preliminary study on the potential complementarity among face parts is carried out using the results of the experiments presented in Section 3.2. The disagreements of the SVM classifier (which results were shown in Table 3.2) when learning from each pair of face parts on FERET instances is shown in Table 3.5 (the right eye is excluded from this study for simplicity since it is included in the eyes part). Each number in the table represents the percentage of samples which are successfully classified by the row classifier and erroneously predicted by the column classifier. In other words, each percentage is the potential improvement of the classification (with respect to the column classifier) if both classifiers were used in conjunction. Obviously, the performances of classifiers with lower rates of success have a stronger chance of being enhanced than others with higher accuracies. For example, those classifiers based on mouth and chin could improve their accuracies about 12-13%, while the classifier based on the full face is expected to improve just about 1-2%.

Table 3.5: Percentage of disagreements between SVM classifiers based on different face parts from FERET images. Each number represents the percentage of instances which are successful cases for the row classifier and error cases for the column classifier.

	Eyes	Nose	Mouth	Chin	Internal face	External face	Full face
Eyes	-	8.80	13.97	14.01	3.35	8.01	2.46
Nose	9.68	-	11.73	12.90	3.07	7.49	1.63
Mouth	10.10	6.98	-	8.57	2.65	6.28	1.76
Chin	10.10	8.10	8.52	-	3.16	6.42	1.95
Internal face	10.24	9.08	13.41	13.97	-	9.08	2.14
External face	10.24	8.84	12.38	12.57	4.42	-	1.02
Full face	12.20	10.47	15.37	15.60	4.98	8.52	-

After looking at the percentage of disagreements, it seems that there is a high potential for defining robust mixtures of classifiers to improve the performance of individual classification models. The possibility of creating ensembles which make use of the diversity of different representations (face parts) of a same face image is studied in depth next.

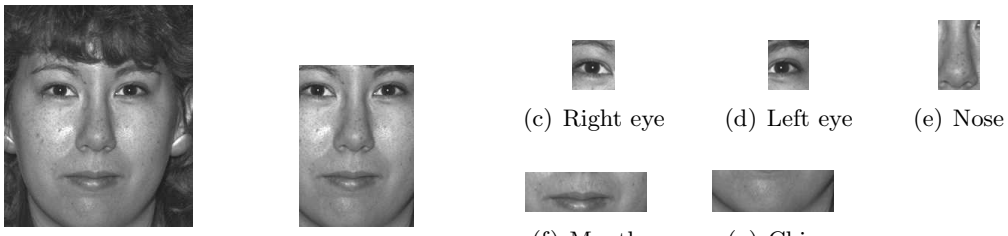
3.3.1 Combining Information from Different Face Parts

Several ensembles of classifiers are used to prove that the joint contribution of separate face parts has more discriminant capabilities than classifications based on individual parts. In particular, ensembles of predictors based on three and five parts are designed, whose decisions are combined using simple and weighted voting, and a SVM classifier. Additionally, the results of these ensembles are compared with those obtained by individual classifiers based on single face parts.

The face parts involved in this complementarity study are two global parts (internal face and full face) and five isolated parts (right eye, left eye, nose, mouth, chin). The eyes participate as two separate entities to find out if one of them is more useful than the other for classifying gender. Unlike in the previous analysis of the role of face parts, the external face is not used in the experiments, since it could lead to erroneous classifications if the subjects were in disguise (for instance, wearing a wig). Figure 3.7 shows an example of these seven face parts extracted from a face image. For extracting the subimages containing each face part, the same method as in the previous study is adopted (for more details see Section 3.2.1).

Ensembles of Classifiers based on Different Face Parts

In order to check whether the information provided by a face part complements that extracted from a different part, several ensembles of classifiers are designed. These ensembles consist of several base predictors where each of them specialises in a particular face part.



(a) Full face (b) Internal Face (c) Right eye (d) Left eye (e) Nose (f) Mouth (g) Chin

Figure 3.7: Example of the subimages containing the face parts of interest.

In other words, each base classifier learns from a specific face part and predicts the gender of a test subject given that face part. Then, those base predictions are combined to obtain the final prediction. In our study, SVM is selected as the base model and three combination strategies are tested. Here, the explanation is focused on the approaches adopted for combining the outputs of the base SVM classifiers.

As mentioned previously, three combination strategies are involved in the experiments. Specifically, two of these strategies are based on voting, simple and weighted voting, while the third one uses a classifier whose inputs are the outputs of the base predictors.

Simple voting (E_{svot})

It is the simplest way of combining the output of the base predictors. This combination rule is a majority vote over the predictions of the members of the ensemble.

Let us denote by $d_j(x)$ the prediction of the base classifier M_j given input x. In a C-class problem, let us assume L base classifiers which produce C outputs each, so a prediction would be referred to as $d_{ji}(x)$ where $i = 1, \ldots, C$ and $j = 1, \ldots, L$. Considering crisp base classifiers which provide binary outputs, $d_{ji}(x) = 1$ if the class predicted for x by M_j is C_i, and $d_{ji}(x) = 0$ otherwise. Combining those predictions by,

$$f_i(x) = \sum_{j=1}^{L} d_{ji}(x) \tag{3.13}$$

where $i = 1, \ldots, C$, we generate C values. Then, we choose the class C_i if $f_i = \max_k f_k$.

In our experiments, we cannot encounter a tie vote because the number of base classifiers is always odd (the proposed ensembles have three or five members) and the decision is made between two classes (male and female).

Weighted voting (E_{wvot})

It is a simple combination strategy consisting in a weighted summation of the outputs of the base predictors per class. In our case, the base predictors (SVM classifiers)

supply additional information about how much they vote for each class (that is, the posterior probabilities) which are used as weights.

Let $w_{ji}(x)$ be the posterior probability of class C_i given the base learner M_j and the input x, that is, $P(C_i|x, M_j)$. The combination of these posterior probabilities is given by,

$$f_i(x) = \frac{1}{L}\sum_{j=1}^{L} w_{ji}(x). \tag{3.14}$$

The final prediction for x would be the class C_i with the maximum value of $f_i(x)$.

A classifier (E_{svm})

It is a combination scheme based on a new input space which is defined by the posterior probabilities of the base classifiers. The idea is that an external classifier (trained with this new space) performs the combination of the base predictor outputs. In order to get a first idea of how this combination strategy works, Figure 3.8 shows the architecture of an ensemble which uses it.

Given a set of training images, the way to proceed to train the combination classifier is collecting the *a posteriori* probabilities of each member of the ensemble. Provided these posterior probabilities, a new training set is created with them (this can be seen as a projection of the original training samples into a new input space). In other words, the input of the combination classifier are the probabilities of each part to be from a female and from a male face. Therefore, the dimensionality of this new space is twice the number of base predictors. The posterior probabilities are obtained by means of a 5-fold cross-validation technique which uses the training images for training and validating the base classifiers. Once the combination classifier has been trained, the base learners are trained with the complete training set.

Given a test face image divided into face parts, each face part goes through the corresponding base learner. The outputs of each base learner (the two probabilities of that part being from a female and a male face) are passed to the trained combination classifier. Finally, a prediction of the gender of that face is obtained.

Several ensembles are designed based on five and three isolated face parts. In particular, all the above combination methods are used in mixtures of classifiers based on five different parts. However, the ensembles using three face parts combine the output of individual classifiers only by using another classifiers (excluding the simple and weighted voting from this part of the study).

3.3.2 Experimental Methodology

The methodology followed to carry out the experiments has the same four steps as the general approach extensively detailed in Chapter 2. Below, some specific details of those steps which are different from the general methodology.

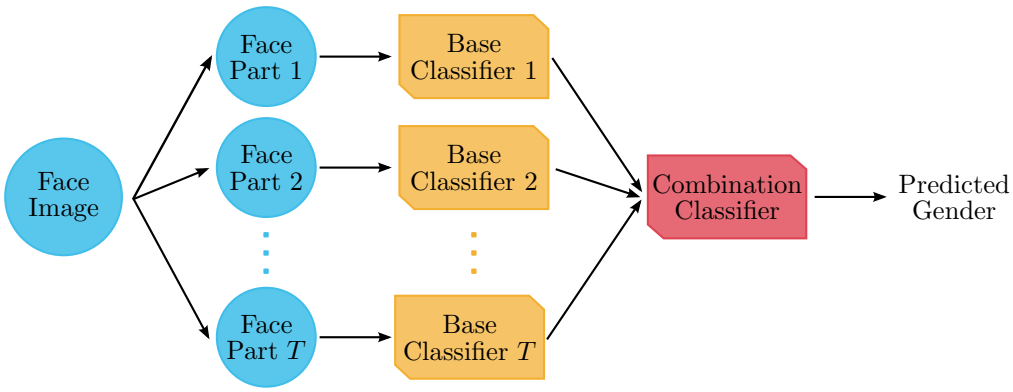


Figure 3.8: Architecture of an ensemble of classifiers with a classifier as a combination strategy.

Step 1. Preprocessing

In this experimental study, the face parts that intervene are the full face and the internal face as holistic face parts, and the right and left eyes separately, the nose, the mouth and the chin as isolated face parts. Figure 3.7 shows an example of each of these seven face parts involved in the experiments.

Step 3. Classification

The classification process is performed by two types of classification models: ensembles of classifiers and individual classifiers. As described in Section 3.3.1, the ensembles of classifiers used in this study have SVMs as base learners whose outputs are combined by three different methods (simple and weighted voting and another classifier). Specifically, these ensembles are evaluated when using the five isolated face parts extracted (right eye, left eye, nose, mouth, chin). Additionally, four extra ensembles are designed using groups of three face parts. In Section 3.3.4, specific details about which face parts are used by each ensemble are provided.

With the purpose of checking if ensembles improve the results achieved by individual classifiers, the ensemble performances are compared to those obtained by seven individual classifiers each based on a different face part.

3.3.3 Face Image Dataset

The experiments are based on the FERET [55] database. The images used are the same images selected from this database in the previous study (see Section 3.2.3).

Table 3.6: Summary of the face parts involved in the ensemble of classifiers.

	Face Parts				
	left eye	right eye	nose	mouth	chin
E^*_{svot}	×	×	×	×	×
E^*_{wvot}	×	×	×	×	×
E^*_{svm}	×	×	×	×	×
E^{enm}_{svm}	×		×	×	
E^{enc}_{svm}	×		×		×
E^{emc}_{svm}	×			×	×
E^{nmc}_{svm}			×	×	×

3.3.4 Experimental Setup

This section describes the experiments performed which involved ensembles of classifiers based on five or three face parts. A summary of which face parts are used per each of the ensembles is shown in Table 3.6. Those ensembles based on five face parts, which correspond precisely to all the isolated face parts extracted (right eye, left eye, nose, mouth, chin), are denoted by E^*_{svot}, E^*_{wvot}, and E^*_{svm}. In this notation, the subscripts refer to the combination strategies for the base predictions, which are simple voting, weighted voting and a SVM classifier, respectively. For the ensembles based on three face parts, only the strategy of combining the base outputs using a SVM classifier is employed. These ensembles are based on all possible combinations of the left eye, nose, mouth and chin. Only one of the eyes is included in this experiments, because it is expected that both eyes provide similar discriminant information. Specifically, the ensembles of three face parts used are: E^{enm}_{svm} based on the eye, nose and mouth; E^{enc}_{svm} based on the eye, nose and chin; E^{emc}_{svm} based on the eye, mouth and chin; and E^{nmc}_{svm} based on the nose, mouth and chin.

A SVM with linear-polynomial kernel was chosen for the base predictor of the ensembles due to its proved effectiveness in solving gender classification problems as it was shown in Section 3.2. The implementation of these SVM classifiers used is the one available in the PRTools Matlab package [26]. These two groups of experiments are designed to compare the performance of individual classifiers and ensembles under the same circumstances. The objective of this comparison is to prove that integrating several face parts provides more discriminant information than single face parts.

The experimental results are computed by averaging five independent runs of a 5-fold cross-validation technique (which gives a total of 25 runs). In each of these runs, the face images of the same individual are only used for training or testing purposes to avoid contamination effects.

Table 3.7: Classification accuracies [and their 95% confidence intervals] on FERET database using individual classifiers and ensembles.

Individual Classifiers

left eye	right eye	nose	mouth	chin	internal face	full face
82.0 [80.6, 83.4]	81.5 [80.1, 82.9]	86.4 [85.1, 87.64]	81.6 [80.2, 83.0]	81.6 [80.1, 82.9]	92.4 [91.4, 93.3]	95.2 [94.4, 96.0]

Ensembles of Classifiers

E^*_{svot}	E^*_{wvot}	E^*_{svm}	E^{enm}_{svm}	E^{enc}_{svm}	E^{emc}_{svm}	E^{nmc}_{svm}
88.4 [87.2, 89.5]	88.9 [87.7, 90.0]	90.5 [89.4, 91.6]	89.7 [88.5, 90.7]	90.6 [89.4, 91.6]	88.5 [87.3, 89.7]	87.6 [86.4, 88.8]

3.3.5 Results and Discussion

The correct classification rates obtained by the ensembles based on five or three face parts are shown in Table 3.7. For comparison, that table also shows the results of the SVM classifier presented in the previous study using FERET database (see Section 3.2) with the addition of two new results using the left and right eyes, separately. Graphically, these results are shown in Figure 3.9.

Most ensembles significantly outperform the plain gender classification of the individual parts with 95% confidence intervals for their average classification rates. These results obtained over the FERET database show the existence of complementary information between face parts, since the performances of the ensembles were better than those achieved by individual classifiers. The only exception is the plain classification based on the nose whose confidence interval slightly overlaps with those of E^*_{svot}, E^{emc}_{svm} and E^{nmc}_{svm}. A comparison between ensembles shows a better behaviour of E^*_{svm} with respect to the other two aggregations of the five parts (E^*_{svot} and E^*_{wvot}).

The ensembles based on three parts, which use a SVM as the combiner, appear to be as discriminant as the combinations of five parts. Particularly, those in which eye and nose coincide (E^{enm}_{svm} and E^{enc}_{svm}) perform better than the ensembles based on the five isolated parts combined by voting (E^*_{svot} and E^*_{wvot}). This is a very good result considering that these ensembles are meant to be useful for gender classification under partial occlusion of the face. In this scenario, no more than 2 or 3 prominent parts are likely to be visible. It is worth noting that the use of a SVM as the combiner was more effective than voting. It is possible that a classifier works as an error-correcting combiner by learning how the base classifiers make errors and how to associate their combinations with correct outputs.

Nevertheless, the ensembles were unable to achieve the rates of plain classifiers based on holistic descriptions of the face (those are the internal and the full face). While the best ensemble is about 2% less accurate than the classifier based on the internal face, the difference between that ensemble and the classifier trained with the full face rises to 5%. There are

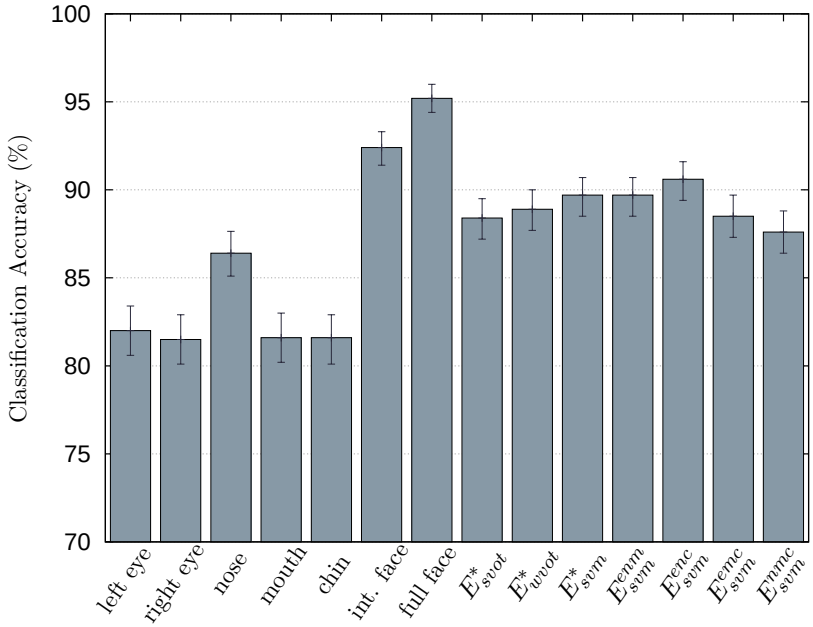


Figure 3.9: Classification accuracies [and their 95% confidence intervals] on FERET database using individual classifiers and ensembles.

two major causes that seem to explain why individual classifiers based on global parts outperform ensembles based on several isolated face parts. Firstly, the holistic representations of the internal and full faces include configural information of the face. It means that the relative position of each face part with respect to the other parts is known, since the whole face is described in those representations. This configural information provides a valuable source of differences between genders (as shown by psychological studies [17, 20]). Secondly, full faces contain other prominent features of the face that have not been considered in this work, like ears, hair and face contour, which have been proved to be discriminant by themselves.

3.4 Conclusions

In this chapter, the relevance of face parts in gender classification as well as the extent to which the most prominent face parts can provide complementary information to distinguish between genders was evaluated. The face parts included in the study are, specifically, both eyes, a single eye, the nose, the mouth, the chin, the internal face, the external face, and the full face. The detection of these parts in the face images was based on an empirical rule about the proportions of the face proposed by Leonardo da Vinci. Two well-known databases of face images and five different classifiers were involved in the experiments.

Empirical evidence showed that individual face parts include enough information to be able to discriminate between genders with success rates above 80%. When their joint contribution is considered (by means of the full face or by just its internal part), the classification rates improved to above 95%. This result is similar to those reported in previous works on gender classification. Our experiments were also designed to evaluate the dependence of the results on the database and the classifier used. A high correlation among classifiers on measuring the relevance of face parts was obtained, however results were strongly dependent on the database. It seems that the significant differences between the two databases were transferred to the results.

A pairwise comparison of the disagreements between SVMs trained with different face parts showed that, as expected, different parts provide diverse discriminant information. This finding led to a study of the complementarity of the information provided by various face parts. In order to detect this potential complementarity, ensembles of classifiers based on several parts were designed. These combined schemes would be suitable solutions to gender classification problems when faces are partially occluded and holistic representations are not possible. The experiments involved ensembles of base classifiers trained with separate descriptions of the left and right eyes, nose, mouth and chin, whose collective decisions were made by simple and weighted voting and by a classifier based on *a posteriori* probabilities. The classification results of these ensembles were compared to those of individual classifiers which used isolated and global parts of the face.

Experiments carried out using images from FERET database showed that the joint contribution of separate parts is more effective for gender classification than isolated parts, but less discriminant than holistic descriptions of faces. Unlike the simple aggregation of parts, the holistic representation includes configural information, whose usefulness at discriminating between genders has been proved by psychological experiments [17, 20].

For addressing gender classification problems, a solution combining only the visible parts of the face seems to be suitable for situations where the face appears partially occluded.

CHAPTER 4

Ranking Labels: A New Type of Local Features

In the literature, many types of features have been proposed for describing face images, which are useful in certain situations. With the main objective of the thesis in mind, a local face characterisation seems appropriate to better deal with the problems that are commonly found in real environments. In this chapter, a new type of features for face characterisation, called *Ranking Labels*, is presented together with empirical evidence of its suitability to describe faces for gender classification purposes.

4.1 Motivation and Background

For automatic face analysis purposes, still face images have been usually characterised using either appearance-based or geometric-based features. Appearance-based features use the value of the pixels in the face image, more commonly after some transformation, to represent the face. Whereas geometric-based approaches are focused on geometric characteristics of the face, such as dimensions of some relevant facial features or distances between them. Besides, face descriptions can also be broadly classified as global or local solutions. Global solutions are those that describe the face as a whole which is usually achieved by means of appearance-based features. On the other hand, local solutions separately describe different regions of the face or isolated facial features. In this last case, appearance-based as well as geometric-based features are possible approaches for locally represent faces.

In recent years, researchers have focused their efforts on solving gender classification problems using different approaches. Some works advocated for appearance-based features, either following a global approach [46, 39, 13] or a local solution [4, 58, 60]. Alternative studies promote geometric-based methods [61], whereas other authors proposed to fuse

both appearance- and geometric-based representations [47]. Which is the best approach to deal with gender classification problems? It all depends on the specific requirements of the problem.

When the face is completely visible, global appearance-based methods report high classification rates [39, 13]. These global representations provide information about the structural relation of the various facial features along with a description covering the whole face. However, in real scenarios, faces could be occluded by clothing or accessories, such as, scarves or sunglasses. In those situations, global approaches may not be the best choice, since they are poorly suited for coping with local variations and occlusions [61]. In such cases, local face descriptions seem more appropriate, although it would be advantageous to additionally have some structural information. It should also be considered that face detectors are not perfect in all situations, albeit they are impressively accurate [63, 33], so it is desirable that the approach taken is able to cope with a certain degree of inaccuracy in the detection. Ideally, a face representation should provide enough discriminant information for successfully addressing gender classification problems when only a partial view of the face is available and the faces are inaccurately detected.

In this chapter, we propose a novel face representation, named *Ranking Labels*, which characterises local areas of the face by means of appearance-based features while keeping some structural information. This representation attempts to provide a more robust characterisation, since it encodes the local contrast while makes it independent from the local intensity values. Two experimental studies are carried out to compare *Ranking Labels* with other local face descriptors and to test its characteristics. In the first study, *Ranking Labels* are compared to several local representations to prove whether they provide more discriminant information for gender classification. In a second study, the experiments are designed to test the robustness of the proposed face characterisation is with respect to the precision of the face detectors.

4.2 Ranking Labels Representation

4.2.1 Characteristics of Ranking Labels

The idea behind *Ranking Labels* representation is to describe the information of the face image by how much the pixel values differ. This description is done by patches, then only the differences among the pixels within an area of the face image are represented. For instance, a *Ranking Labels* description could be interpreted as *all the pixels at the top part of the patch are lighter than the ones at the bottom*, or, in other patch, *all the pixel values are very similar*.

A *Ranking Labels* description can be roughly defined as a vector of ranking positions with respect to the lowest grey level value within a patch. Representing the information contained in each patch by means of ranking positions instead of the actual grey level values makes this representation more robust against variations in illumination. The *Ranking Labels* are relative to the grey levels found within a moderate-sized region. Therefore,

Ranking Labels representations are hardly affected by shifts in the grey scale. For instance, imagine we have an image taken in very low light and another taken outside in the sunlight. The first image will have fairly dark grey levels and the second will contain lighter greys. By describing their content with *Ranking Labels*, since the labels are assigned with respect to the grey levels of the region, the shift in the grey scale between both images will be almost imperceptible. This characterisation in terms of how different the pixel values are (within a patch) is expected to provide a description substantially independent to the given illumination conditions.

For describing the information provided by a face image, not all the differences among pixel values are needed. It is reasonable to consider those values that are significantly close to each other as equals. This way, the information is summarised and the description becomes more general. After summarising the information provided by pixels with similar values, it is possible to obtain vectors with just a few different *Ranking Labels*. This is highly likely when the patches correspond to uniform areas of the face, such as the cheeks or forehead, that usually do not provide much discriminant information. That being the case, those *Ranking Labels* vectors could be discarded. Consequently, the complexity of the final face characterisation would be reduced to the vectors with more information. The method for reducing the complexity of the representation, which is an optional step of the face characterisation proposed, is presented in Section 4.2.3.

Unlike other local descriptions, the feature vectors derived from *Ranking Labels* provide spatial information about the pixels within the patch since each pixel is represented by a *Ranking Label*. At the end of the characterisation process, which is explained in detail in Section 4.2.2, several *Ranking Labels* vectors are extracted from a single face image. The final face description could consist of various feature vectors or could be formed by concatenating all of them. This last option could be considered a semi-global representation. While, *Ranking Labels* supply information about the local contrast of the pixels within a region, the fact of concatenating these features provides details about their location with respect to the other features.

4.2.2 Extraction of Ranking Labels

The extraction of *Ranking Labels* features from a given face image begins with the process of scanning the face with a squared window (the area of the image delimited by the scanning window is referred to as a patch). A vector containing the pixel values within each patch is created, the elements of which correspond to the grey level values of the pixels from the top left corner of the patch to the right and downwards. In order to convert those pixel values to *Ranking Labels*, the grey level values are substituted by their ranking position with respect to the other values in the vector. The process of transforming grey level values into *Ranking Labels* is graphically shown in Figure 4.1. First, a vector is created with the grey level values within the patch. Second, the pixel values are sorted in ascending order. Third, a ranking position is assigned to each value, giving to those values that differ in less than G units the same label. Finally, the *Ranking Labels* are sorted in the same order as

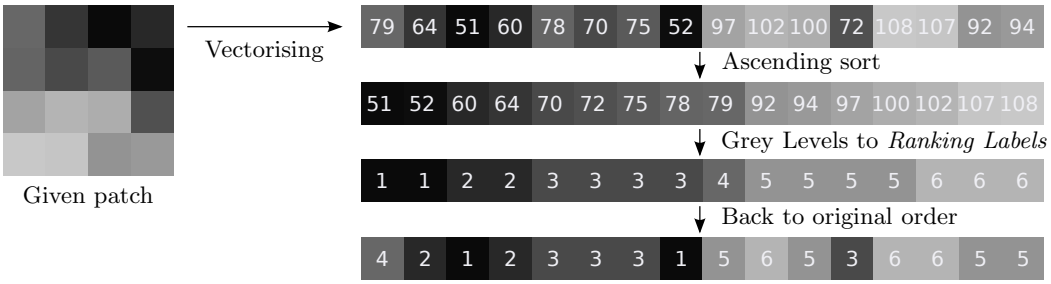


Figure 4.1: Process of transforming grey level values into *Ranking Labels* ($G = 8$).

the original vector of grey levels.

More formally, let $x = (x_1, x_2, ..., x_{N\times N})$ be a vector of the grey level values within a given patch of $N \times N$ pixels and G a threshold that represents the maximum difference between two grey levels to consider them equal. Then, the process continues as follows.

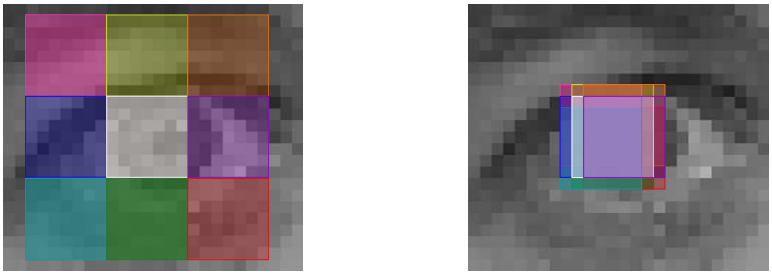
1. Compute the ordered version of x, $x' = (x_{i_1}, x_{i_2}, ..., x_{i_{N\times N}})$ where $x_{i_j} \leq x_{i_{j+1}}$ and $i_j = 1 \ldots N \times N$ is the position of x_{i_j} in the original vector x.

2. Build a vector $r' = (r_{i_1}, r_{i_2}, ..., r_{i_{N\times N}})$ with the ranking positions of the components of x', according to the algorithm:

   Data: x', G
   Result: $r' = (r_{i_1}, r_{i_2}, ..., r_{i_{N\times N}})$
   $k \leftarrow 1$
   $rank \leftarrow 1$
   for $j = 1 : N \times N$ **do**
   if $x_{i_j} - x_{i_k} > G$ **then**
   $k \leftarrow j$
   $rank \leftarrow rank + 1$
   end
   $r_{i_j} \leftarrow rank$
   end

3. Sort the elements in vector r' to build $r = (r_1, r_2, ..., r_{N\times N})$ where $r_j = r_{i_j}$.

The definition of *Ranking Labels* needs two parameters, one of them indicating the size of the patches which will be characterised and the other to indicate the maximum difference between pixel values to considered them as the same value (that is, parameter G). Additionally, the arrangement of patches on the images should be defined. The patches can be overlapped up to a certain extent. This overlapping could go from non-overlapping at all to the maximum overlapping which occurs when the shift from one patch to the next is one pixel. Figure 4.2 shows those two degrees of overlapping exemplified with nine



(a) Non-overlapped patches.

(b) Maximum overlapping. From one patch to the next, there is a 1-pixel shift.

Figure 4.2: Layout of the same nine patches with different overlapping degrees.

patches over an image. As can be seen in Figure 4.2(b), when there is overlapping, the shift from one patch to the next is the same in all directions with respect to a central patch.

4.2.3 Complexity Reduction

Being *Ranking Labels* a representation of local differences in pixel contrast within a medium-sized region, it is possible to obtain a feature vector of very few different *Ranking Labels* if all the pixel values are very similar. These cases would probably correspond to uniform areas in the face, such as the cheeks. The goal of reducing the complexity is to discard those feature vectors that do not provide much relevant information.

The selection of the vectors with less amount of information is performed by comparing their number of different *Ranking Labels* with a given threshold T. Those vectors with less than T different *Ranking Labels* correspond to the regions of the image with the lowest contrast. These parts of the face do not provide too much discriminant information, therefore, the face representation could not include them. This reduction process is an optional step of the face characterisation proposed.

4.3 Comparison with other Local Face Representations

This section presents a comparison between *Ranking Labels* and two well-known local characterisation techniques, Local Binary Patterns and Local Contrast Histograms, to find out which face representation obtains higher performances in different situations.

For this comparison, Local Binary Patterns (LBP) have been chosen because of their proven ability to successfully address facial analysis problems. In the last decade, LBPs have been applied to describe faces in face recognition [2] and gender classification [58] problems achieving noteworthy results. Additionally, we considered that the efficacy of *Ranking Labels* should also be compared with another face characterisation technique based

on the contrast of the pixel values, that is Local Contrast Histograms (LCH). Moreover, we combined the information provided by LBP and LCH as another face characterisation. All these face descriptors are compared when dealing with gender classification problems under different experimental configurations (specific details are given in the next sections).

For checking if these face descriptions provide enough discriminant information when only a partial view of the face is available, only the top half of the face is used in the experiments. This situation could occur when the individuals wear a scarf or other type of clothing that covers the bottom area of the face.

4.3.1 Experimental Methodology

The present study is based on the methodology described in Chapter 2, although with certain particularities which are detailed below.

Step 1. Preprocessing

In this step, the face is detected using the Viola-Jones algorithm. The system does not align the face images or correct the inclination that the face might have. After detecting the face, the top half of the area of the image where the face was detected is extracted, equalised and resized to a pre-established smaller size.

Step 2. Feature Extraction

All the face representations involved in the experiments are based on local features which separately characterised several patches considered over the face image. These patches could have a certain degree of overlapping. In that case, it will be indicated in the description of the experiments. Specifically, the three local descriptors used are:

- Local Binary Patters (LBP),
- Local Contrast Histograms (LCH),
- *Ranking Labels* (RL).

The several feature vectors produced by these methods can be considered individually or can be concatenated to form a single vector. Which of these two approaches is taken will be indicated in the description of the experiments (Section 4.3.3).

Step 3. Classification

The classifier employed is k-Nearest Neighbour (k-NN). Depending on whether several feature vectors or just one vector describe the top half of the face, the classification process is different. If the face description consists of a single vector, the classifier works as a standard k-NN. In case various feature vectors describe the face, its gender is predicted by majority voting of all the predicted labels assigned to those vectors. Two distance metrics,

the Euclidean distance and the Chi square distance, are used in all experiments in order to compare which one is more suitable for addressing our task.

Step 4. Performance Assessment

The classification models are evaluated by their accuracy.

4.3.2 Face Image Dataset

The face images involved in the experiments are from the FERET database (for a detailed description of this database, see Section A.2 in Appendix A). Particularly, only faces in a frontal pose without glasses are used, because glasses could strongly distort the effectiveness of the face for gender classification which might produce misleading conclusions. The experiments are performed over 2,147 images of 256×384 pixels from 834 subjects separated into 842 female faces and 1,305 male faces. Therefore, the ratio of males to females is 1:0.6. These face images are divided into two sets, training and test, consisting of 60% and 40% of the total number of images, respectively. When dividing the images into training and test images, the ratio of male-female faces of the original dataset is maintained. Additionally, as there are several images per subject, this division is carefully implemented in order to assign all the images of the same subject to the same set.

4.3.3 Experimental Setup

This section describes four experiments which have been designed to compare the local face representations previously detailed in Section 4.3.1. The aims of these experiments are to find out which face description provides more information to discriminate between genders and which of them is more suitable for situations where the face is not accurately detected. The details about the four experiments are presented below.

- Experiment 1: The patches considered over the top half faces are not overlapped, consequently the pixels that belong to a patch are never considered in another one. Then, the face description is formed by concatenating all the feature vectors extracted from those non-overlapping patches. Hence, in the classification process, when looking for the nearest neighbour of a test face among the training faces, the test features are compared to those training features that were extracted from the same positions in the images.

- Experiment 2: The patches are overlapped over the image in order to extract more detailed descriptions of the top half faces. As a result, one pixel will belong to several patches and its value will be used to obtain the descriptions of all of them. Although keeping the same size of the face images and of the patches as in the previous experiment, the number of feature vectors is larger due to the overlapping. At the end, the face description consists in the concatenation of all the vectors extracted.

- Experiment 3: Overlapped patches are considered over the face images, as in experiment 2. However, in this case, a face description consists of all the feature vectors extracted from the image. This results in several feature vectors representing a single face. For classifying the gender of the face, first each test vector is assigned the class label of its nearest neighbour found among all the training feature vectors. Then, the predicted gender of the face is obtained by majority voting over all the classes assigned to its vectors.

  Unlike in experiments 1 and 2, each vector is classified by comparing it with the vectors extracted from all training patches independently of the patch's position. Hence, it is expected that inaccuracies in the face detection will not affect as much in this case. This tolerance comes at the expense of having a larger number of vectors which leads to a higher computational cost.

- Experiment 4: Inaccurately detected faces are simulated by shifting the area of the image where the face was detected. The configuration of this experiment is the same as in experiment 3 with the exception that after automatically detecting the face in the image, a random shifting is applied to the area containing the face. The displacement could be at most 10% of the width for the horizontal movement and 10% of the height for the vertical one (considering the dimensions of the area containing the face).

  This experiment allows us to test the face descriptions and the classification methods in a more challenging scenario. Consequently, it could provide more reliable results about which approach would be more suitable for situations where the face detection could not be accurate.

As has been previously indicated, all the face representations involved provide local descriptions. The level of detail of this descriptions is given by several parameters, such as the size of the region from which the local features are extracted, or the number of sample points considered (for LBP and LCH features). The specific parameters of each face description are detailed as follows:

- Uniform LBP with a neighbourhood of 8 sample points and radii 1 ($\mathrm{LBP}^{u}_{8,1}$) or 2 ($\mathrm{LBP}^{u}_{8,2}$).
- Local contrast histograms with a neighbourhood of 8 sample points and radii 1 ($\mathrm{LCH}_{8,1}$) or 2 ($\mathrm{LCH}_{8,2}$).
- *Ranking Labels* with $G = 8$ and no complexity reduction.

Several combinations of this LBP and LCH descriptions are also considered, all of which consist of the concatenation of the feature vectors corresponding to each of the descriptions. Particularly, the combinations used are defined in terms of:

- The type of feature: $\mathrm{LBP}^{u}_{8,1}$ + $\mathrm{LBP}^{u}_{8,2}$ and $\mathrm{LCH}_{8,1}$ + $\mathrm{LCH}_{8,2}$.
- The radius of the neighbourhoods: $\mathrm{LBP}^{u}_{8,1}$ + $\mathrm{LCH}_{8,1}$ and $\mathrm{LBP}^{u}_{8,2}$ + $\mathrm{LCH}_{8,2}$.

- The maximum degree of detail (combining all of the above): $\mathrm{LBP}^{u}_{8,1}$ + $\mathrm{LCH}_{8,1}$ + $\mathrm{LBP}^{u}_{8,2}$ + $\mathrm{LCH}_{8,2}$.

For the face descriptions based on histograms (LBPs and LCHs), the number of bins per histogram defines the number of features. In the case of LBPs, each bin corresponds to one of the possible LBP values. For the rotationally invariant version (RI), there are 10 possible LBP values, resulting in vectors of 10 features. For the sensitive to rotation version (no RI), there are 59 possible LBP values, which will produce vectors of 59 features. In the case of LCHs, there is not a predefined set of possible contrast values, therefore how many bins there are per histogram should be decided. In this study, two versions of LCHs are defined to be comparable to the two versions of LBPs. One version accumulates the contrast values in 10-bin histograms, while the other uses 59-bin histograms.

The size of the patches is 7×7 pixels for all face representations. This indicates that the number of *Ranking Labels* is 49, so each patch is represented by a 49-dimensional vector. For the other two representations, this implies that the 49 (LBP or contrast) values calculated over each pixel within a patch are represented by a 10- or 59-dimensional feature vector.

Implementation Details

Following, specific details needed to replicate the experiments are summarised. The output of the face detector is horizontally split in two halves (only the top half is used in the experiments). Then, the top half of that image is resized to 45×18 pixels. As it has already been mentioned, patches of 7×7 pixels are considered over the top half face image. In the case of non-overlapping patches (experiment 1), there is a total of 12 patches per image. In the overlapped cases (experiment 2, 3 and 4), there is a one pixel shift from one patch to its neighbours. As a result, the number of patches rises up to 468. To test if the classification can benefit from a representation with overlapping patches, the maximum degree of overlapping is chosen as it is the opposite to the non-overlapped scenario.

4.3.4 Results and Discussion

The correct classification rates shown in Table 4.1 correspond to those obtained in each of the experiments presented in Section 4.3.3.

Taking a general look at these results, it can be seen that the Chi square distance succeeded in classifying the gender in more cases than the Euclidean distance. Particularly, in 58 out of 76 experiment the Chi square distance led to higher classification rates than the Euclidean distance (in 9 of those experiments both distances achieved the same accuracies).

As is shown by the results, when only LBP features are involved, the sensitive to rotation descriptions (59 bins) achieved better results than the rotationally invariant versions (10 bins). However, the use of LCH with 59-dimensional vectors resulted in worse accuracies in experiments 1 and 2. This could be explained by the higher dispersion of the data in these cases which leads to a poorer characterisation. When LBP and LCH are combined, the representation is still affected causing lower classification rates.

Table 4.1: Classification accuracies obtained by each of the face representations for the experiments presented in Section 4.3.3. For LBP features, the 10-bin histogram column corresponds to the RI version and the 59-bin column to the no RI version. Marked in bold the best accuracy obtained per experiment.

		Experiment 1		**Experiment 2**		**Experiment 3**		**Experiment 4**	
		10-bin	59-bin	10-bin	59-bin	10-bin	59-bin	10-bin	59-bin
$\mathbf{LBP}^{u}_{8,1}$									
	χ_2	70.88	76.61	74.27	78.48	61.66	71.75	61.08	61.08
	Euclidean	68.30	76.02	73.33	76.37	61.08	70.57	61.08	61.08
$\mathbf{LBP}^{u}_{8,2}$									
	χ_2	68.42	79.06	81.17	78.95	61.43	75.26	61.08	61.08
	Euclidean	68.42	76.73	77.89	75.56	62.02	72.92	62.14	62.14
$\mathbf{LBP}^{u}_{8,1}$ + $\mathbf{LBP}^{u}_{8,2}$									
	χ_2	73.92	80.47	78.13	80.23	62.84	78.55	62.49	62.49
	Euclidean	72.51	78.25	77.43	77.31	62.49	76.32	62.14	62.14
$\mathbf{LCH}_{8,1}$									
	χ_2	75.44	69.36	79.65	74.97	61.08	62.95	61.08	64.36
	Euclidean	73.57	70.64	78.95	72.87	61.08	64.36	61.08	65.06
$\mathbf{LCH}_{8,2}$									
	χ_2	77.89	71.81	79.77	75.79	61.08	63.42	61.08	63.19
	Euclidean	74.27	72.05	76.96	74.50	61.08	63.42	61.08	64.13
$\mathbf{LCH}_{8,1}$ + $\mathbf{LCH}_{8,2}$									
	χ_2	77.89	72.98	79.30	76.26	65.06	64.48	64.83	65.30
	Euclidean	75.44	73.80	77.54	76.73	66.00	63.07	64.48	63.66
$\mathbf{LBP}^{u}_{8,1}$ + $\mathbf{LCH}_{8,1}$									
	χ_2	75.79	79.53	80.23	81.17	66.47	79.95	64.83	79.01
	Euclidean	77.19	77.43	79.65	77.89	67.87	75.15	65.77	73.51
$\mathbf{LBP}^{u}_{8,2}$ + $\mathbf{LCH}_{8,2}$									
	χ_2	80.47	79.88	82.46	81.40	69.05	82.65	69.17	81.71
	Euclidean	77.43	77.66	81.17	77.08	69.40	77.61	69.64	76.08
$\mathbf{LBP}^{u}_{8,1}$ + $\mathbf{LCH}_{8,1}$ + $\mathbf{LBP}^{u}_{8,2}$ + $\mathbf{LCH}_{8,2}$									
	χ_2	**82.69**	81.64	**82.81**	80.82	74.44	85.11	71.16	83.59
	Euclidean	80.70	79.88	81.40	77.19	71.28	78.55	70.81	78.55
Ranking Labels									
	χ_2	78.95		80.12		**88.54**		89.12	
	Euclidean	78.60		79.30		**88.54**		**89.94**	

Concerning the radius of the neighbourhood in histogram based features, radius 2 performs the classification task better than radius 1 in a vast majority of the experiments. Nevertheless, the combination of the same face description using both radii achieves higher rates, but using twice as many features.

Focusing on experiments 1 and 2, LCH features have been proved to be suitable for

discriminating between genders since the classification rates by them are very similar to those achieved by LBP descriptions. LCHs perform better using vectors of 10 features, whereas LBPs obtained slightly higher accuracies with their rotation dependent version (59 features). As expected, using the combination of all LBPs and LCHs configurations increased the classification rates up until 82.69% (experiment 1) and 82.81% (experiment 2), which are the best rates obtained in these experiments. When only LBPs are employed, the accuracies are around 80%, while with just LCH features, the classification rates do not surpass 80%. The *Ranking Labels* description achieved the best results when comparing only individual features (without considering combinations of several types of features). Therefore, *Ranking Labels* representations provide slightly more discriminant information than the other simple characterisations, although less than the combination of LBPs and LCHs. To summarize, experiments 1 and 2 have proved that all the face descriptions are reasonably good to discriminate between genders, since no significant differences were observed in the classification accuracies. As a general rule, the larger the number of features used to describe the faces, the better the classification rates obtained.

Focusing on experiments 3 and 4, *Ranking Labels* reached the highest classification rates, which were close to 90% for both experiments. These results were even better than those obtained in experiments 1 and 2 using this face representation. It should be noted that experiments 1 and 2 relied on totally accurate face detections, since the classification compared patches according to their position in the image. Whereas, in experiments 3 and 4, the classification was tolerant to errors in the face location by searching for a patch's nearest neighbour among all the training patches, independently of their position in the image. However, this substantial improvement only occurs when using *Ranking Labels*. The features based on histograms performed worse in these cases than in experiments 1 and 2. This is probably because *Ranking Labels* keep spatial information about the position of the pixels within the patch, unlike the methods based on histograms.

4.4 Dealing with Inaccurate Face Detections

This section presents a set of experiments designed to prove that the *Ranking Labels* representation is robust to inaccuracies in the face detection and is suitable for situations where the whole face is not visible.

In order to determine the ability of the proposed face characterisation to cope with inaccurately located faces, two different face detection algorithms are employed. The first one needs the coordinates of the eyes as input, whereas the second method is completely automatic. Particularly, different combinations of these two algorithms are applied to the training and test images leading to several experiments.

As in the previous study, only the top half of the detected face is used in order to simulate situations where only a partial view of the face is available.

(a) Detection based on the coordinates of the eyes.

(b) Detection using Viola and Jones algorithm [63].

Figure 4.3: Faces detected by each of the two detectors and the top halves used in the experiments.

4.4.1 Experimental Methodology

This experimental study follows the methodology described in Chapter 2. Below, those details of each step which are specific to the experiments of this section are explained.

Step 1. Preprocessing

Two different face detection algorithms are utilised for testing the performance of the proposed representation with different detection accuracies. The first method defines the area of the face with respect to the separation of the eyes. Let (x_r, y_r) and (x_l, y_l) be the coordinates of the centre of the right and left eyes respectively, and d the distance between them. Then, the top left corner of the area containing the face is given by the coordinates $(x_r - 0.75 \times d, \frac{y_r+y_l}{2} - 1.15 \times d)$ and the bottom right corner is $(x_l + 0.75 \times d, \frac{y_r+y_l}{2} + 1.8 \times d)$. The second face detector is the well-known Viola-Jones detector. Examples of the full and the top half face areas resulting from applying each of these algorithms to the same image are shown in Figure 4.3. These both face detectors are explained with more details in Section 2.2.

Once the area containing the face has been detected, the top half part of that area is extracted, equalised and scaled to a smaller size. The experiments involve only the top half part of the face.

Step 2. Feature Extraction

From the top half of the image resulting from the previous step, *Ranking Label* features are extracted following the process detailed in Section 4.2.2.

Step 3. Classification

The classifier utilised is a k-Nearest Neighbour with Euclidean distance. The patches are classified independently, and only those training patches extracted from the same position as the test patch are considered. The gender of the face is predicted by majority voting of the classes assigned to its patches.

Step 4. Performance Assessment

The evaluation of the performance is done based on the accuracies obtained in each experiment.

4.4.2 Face Image Dataset

The face images involved in the experiments and the division into training and test sets are the same as in the previous study (see Section 4.3.2).

4.4.3 Experimental Setup

As has been previously mentioned, two detection algorithms are used for testing the suitability of the proposed representation under various degrees of accuracy in the face detection. Different combinations of these face detectors are applied to the training and test images leading to the following experiments:

- Experiment 1: The faces in both training and test images are detected using the coordinates of the eyes.
- Experiment 2: The training faces are detected using the coordinates of the eyes and the test faces are automatically detected.
- Experiment 3: The faces in both training and test images are automatically detected.
- Experiment 4: In this case, an inaccurate face detection is simulated. After detecting the faces using the coordinates of the eyes, the area containing the face is shifted. The displacement is a randomly selected value within the range 0 to 15 pixels. A different value is computed for each of the four possible directions (up, down, left, right).

Two series of these four experiments are carried out. In the first series, all the feature vectors extracted from the images are used while, in the second series, the number of feature vectors is reduced before the classification process takes place.

Table 4.2: Classification accuracies obtained using different combinations of the two face detection methods with and without reducing the complexity of the *Ranking Labels* representation. Marked in bold the best result per experiment.

	Tra and Tst coordinates (Exp. 1)	Tra coordinates Tst auto detection (Exp. 2)	Tra and Tst auto detection (Exp. 3)	Shifting the area of the face (Exp. 4)
All RL vectors	79.55	**83.16**	82.57	79.21
Selected RL vectors	**81.07**	82.30	**88.51**	**83.06**

Implementation Details

To replicate the experiments of this study, specific details are given in this section. The top halves of the detected faces are scaled down to a resolution of 30×12 pixels. The face areas resulting from the automatic detector have the same aspect ratio (which is kept when rescaling them). However, this is not always the case for the faces detected using the coordinates of the eyes, since the detection is based on the distance between the eyes. In this case, the scaled images are forced to have the mentioned size. The size of the patches is 9×9 pixels and they are not overlapped. As a result, given the size of the scaled top half face images, 110 feature vectors of 81 components are extracted.

For transforming pixel values into *Ranking Labels*, the parameter which determines the maximum difference between two grey levels to considered them equal is set to $G = 8$. Consequently, a total of 32 *Ranking Labels* are possible given the 256 grey level values. For the process of reducing the complexity of the data (for details about this process see Section 4.2.3), a threshold $T = 16$ is chosen to discard those local feature vectors with poor information for discriminating between genders. It is worth noting that due to the fact that there are 32 possible *Ranking Labels*, by setting $T = 16$, the discarded vectors are those whose number of different *Ranking Labels* is at most half of the total number of labels.

4.4.4 Results and Discussion

The correct classification rates obtained in each of the four experiments detailed in Section 4.4.3 are shown in Table 4.2. The first row of this table shows the results when all the *Ranking Label* vectors are used and the second row shows the results using only the selected *Ranking Label* vectors obtained after applying the reduction process. As can be seen, all the results obtained are very close to or higher than 80% despite the fact that there were no restrictions on how accurate the detection of the face should be.

When keeping all the feature vectors (first row of Table 4.2), the best classification rates were achieved when the automatic face detector was involved either on both sets, or just in the test set (experiments 2 and 3). However, the face detection based on eyes coordinates led to slightly lower performances (experiment 1). This could be due to the differences between the size of the face areas detected (see Figure 4.3). The automatic method selects

(a) Males classified as females. (b) Females classified as males.

Figure 4.4: Examples of misclassified subjects.

a wider area than the algorithm using the eyes coordinates, so it provides more information. This fact together with the classification by patches that provides a certain level of tolerance towards inaccuracies in the detection, contributed to reach better performances.

The classification accuracies achieved when reducing the complexity of the feature space are, in most cases, higher than those obtained with the complete space. Taking into account that a lower number of vectors (those which contained more information) were used to describe the faces, it is possible that the discarded vectors provided information that confused the classifier.

When intentionally shifting the face area (experiment 4), the results obtained are remarkably similar to those achieved in the other experiments. This demonstrates that *Ranking Label* descriptions are robust to variations in the accuracy of face detection.

As a further benchmark to evaluate the results of this work, the experiments presented in Chapter 3 are considered. There, different face parts were globally described using grey level features after accurately detecting the faces using the eyes coordinates. Under such controlled conditions, the best classification rates were 95.21%, 92.37% and 85.47% achieved by a Support Vector Machine trained with the full face, the internal face (which excludes hair, ears and facial contour) and the eyes, respectively. These face parts do not contain the same information as the top half of the face, the full and internal face include a much bigger area of the face and the eyes only consisted of the eyes region. With the *Ranking Labels* representation, the best classification rate was 88.51% which was obtained using about half of the information contained in the internal face. Unlike the results in Chapter 3, the present study was based on a completely automatic extraction of the top half of the face. Therefore, this methodology seems more suitable for being used in real situations.

Regarding missclassified face images, Figure 4.4 shows some examples. In our opinion, these faces are somehow confusing even for human beings, so it can be acceptable that the gender classifier gives erroneous predictions.

4.5 Conclusions

In this chapter, we presented a new type of features which provides face representations fairly independent to illumination variations. This features were proved to contain enough discriminant information to classify the gender from a partial view of the face.

The proposed descriptor, *Ranking Labels*, represents a face by locally characterising the

contrast of the pixel values while keeping the spatial information of the pixels within the given area. In all the experiments only the top half of the face was involved, proving this representation suitable for face gender classification in situations where the bottom half of the face is not available due to occlusions possibly caused by clothing accessories. An experiment simulating inaccurate face detections showed that *Ranking Labels* characterisations are robust to errors in the location of the face.

The classification performance achieved by *Ranking Labels* was compared with those obtained when characterising the faces by means of other types of local features. Particularly, Local Binary Patterns (LBP) and Local Contrast Histograms (LCH). In this comparison, the experiments were presented under various scenarios. In the first two experiments, it was assumed that the face detection was sufficiently accurate and the classification was conducted taking into account the position of each patch. In the other two experiments, the classification was performed by patches independently of their position in the image which provided a certain tolerance to errors in the face detection. These experiments showed that LBPs and LCHs correctly address the problem with reasonably good detected faces. However, these face representations were less reliable in situations with inaccurate face detections, since there is an important loss of spatial information. In an inaccurate environment, *Ranking Labels* achieved classification accuracies of about 90%. Whereas the other face characterisations (LBPs, LCHs and their combinations) did not reach an 86% of correct classification rate.

Summarising, *Ranking Labels* have been proved a reliable face representation as it performs similarly in all the situations considered in this experimental study. Although, LBPs and LCHs successfully address the gender classification task, they were more dependent on the accuracy of the face detections.

CHAPTER 5

Classification based on Local Neighbourhoods

Most authors are inclined to choose global solutions when addressing gender classification problems from completely visible faces. Those approaches may seem more appropriate than local solutions since all facial information is available. In this chapter, we propose a new classification method based on local neighbourhoods and provide a statistical comparison of the proposed approach with some widely employed global solutions.

5.1 Motivation and Background

Faces can be described by global or local representations. In the literature we can find many works following either of those approaches. Intuitively, when the whole face is visible, holistic solutions seem to be more likely to achieve higher classification rates. This is based on the fact that global characterisations provide configural information (i.e. relations among face parts) as well as featural (i.e. characteristics of face parts), whereas local descriptors mostly provide featural information. However, this has only been tested using standard classification techniques. By standard classification, we mean that one face is represented by one feature vector and the classifier is trained with those face descriptions. Those single feature vectors can contain global or local features. In the local case, the features would be concatenated to form one feature vector [4, 58, 60]. This standard approach expects faces to be perfectly detected and aligned, having all facial features in the same position within all images. In other words, it is based on the assumption that regions at the same location have the same content in all face images. However, this is certainly an ideal scenario because faces do not always have the same exact proportions. For instance, the separation between the eyes usually varies from one face to another and the nose length

is not always the same. Therefore, a technique for aligning the faces should be applied in order to have as many face parts as possible in the same position within all images. These automatic alignment methods, as well as the face detectors, are likely to commit mistakes. With this in mind, the usual local approach does not seem to exploit all the potential of local features. Instead of classifying all regions at once (concatenating the local features into a single vector), we could classify each of them independently. For that, we could learn about their content considering only those regions extracted from the same position, or taking into account a wider area, that is, considering also neighbouring regions. These two alternatives were used in the experiments presented in Section 4.3. Particularly, the second option was taken to the extreme case, where all the patches in the image were considered as neighbours. The improvement in the classification accuracies reported in those previous experiments raised questions about whether a local classification based on a more restrictive definition of neighbourhood would perform better than global solutions. Considering a less extreme neighbourhood seems a more realistic approach, since the facial features tend to be in the same areas of the face, even in misaligned faces. Generally, the right eye will always be in the top right area of the face, the nose in the centre, and so on. In addition, if neighbourhoods cover a smaller area, some mistakes could be avoided. For example, right eyes cannot be confused with left eyes, if the patches containing them are not in the same neighbourhood.

To the best of our knowledge, the superiority of global versus local approaches has only been proved on single-database experiments [19, 70, 71], which are not representative of real world settings. More realistic scenarios can be simulated by using different databases for training and testing. When crossing databases, the acquisition conditions and demographic characteristics of the training images vary notably with respect to the test images. Therefore, the generalisation capabilities of the classifiers can be evaluated.

In this chapter, we propose a local classification technique based on neighbourhoods of regions. The knowledge of the position of a region in the image is utilised to create one classifier specialised in each region's neighbourhood. This method has been designed to compensate for the lack of configural information with some tolerance towards misaligned faces. Thus, even in situations where faces are fully available, our local classification can be competitive. In order to prove its suitability, we present an experimental study where the performances obtained by global approaches are compared with those achieved by the proposed local technique. This comparison is accomplished using three well-known classifiers (k-Nearest Neighbour, Linear Discriminant Analysis and Support Vector Machine) and three different types of features (grey levels, Principal Components Analysis and *Ranking Labels*). In addition to the common single-database experiments, more realistic conditions are simulated by cross-database experiments involving three different face databases. Supporting the discussion of the results, three statistical tests provide information about the existence of significantly relevant differences among performances.

5.2 Classification based on Neighbourhoods

The aim of this classification based on neighbourhoods is to use the spatial information of local characterisations to gain a certain level of tolerance towards misaligned faces, inaccuracies in the face detection and faces presenting different facial proportions. The idea is to learn about the appearance of neighbouring regions (or patches) as opposed to learning from only those patches that are located at the exact same position within the image.

An essential requirement of this method is that the face representation must be local, that is, the description must consider a set of local regions (patches). If this condition is fulfilled, the classification is carried out in two stages. First, each patch is individually classified as belonging to a male or female and, second, the local predictions are combined to obtain the final class label. The individual classification per patches is performed taking into account the neighbourhood to which the patch belongs. This is achieved by having local classifiers which specialise in some particular patches (those included in the same neighbourhood). Concretely, there is one classifier per patch which learns from the patches pertaining to the same neighbourhood as its associated patch. Then, the outputs of those local classifiers are combined by majority voting to predict the gender of the given face. Following, these concepts are defined more formally.

Let $B_{i,j}$ be the neighbourhood associated to position (i, j) in an image. For a given patch $p_{k,l}$, centred at position (k, l), $p_{k,l} \in B_{i,j}$ iff $|i - k| \leq T$ and $|j - l| \leq T$, where T defines the size of the neighbourhood. Now, let $C_{i,j}$ be the local classifier trained with $B^{tra}_{i,j}$, which denotes the set of patches within the neighbourhood $B_{i,j}$ from all of the training images. Given the S patches extracted from a test image, a class label for patch $p_{i,j}$ is predicted by $C_{i,j}$ $\forall i, j$, resulting in S predicted classes. Finally, the predicted gender of the face is obtained by a majority vote of the S local predictions.

As it was introduced in the previous chapter, the patches can be overlapped up to a maximum level of overlapping, that is, when there is a one-pixel shift from one patch to the next (in all directions). Independently of the overlapping degree of the patches, the classification process is exactly the same. The overlapping only affects the size of the area covered by the patches of the same neighbourhood. Figure 5.1 shows in blue all the patches belonging to the neighbourhood associated to the patch in white (with $T = 2$ and maximum overlapping). For classifying that white patch, the classifier will be trained with the coloured patches (including the white patch) extracted from the training images.

5.3 Experimental Study

5.3.1 Experimental Methodology

The experiments are designed following the general methodology presented in Chapter 2. Next, only the details which are specific to this study are indicated.

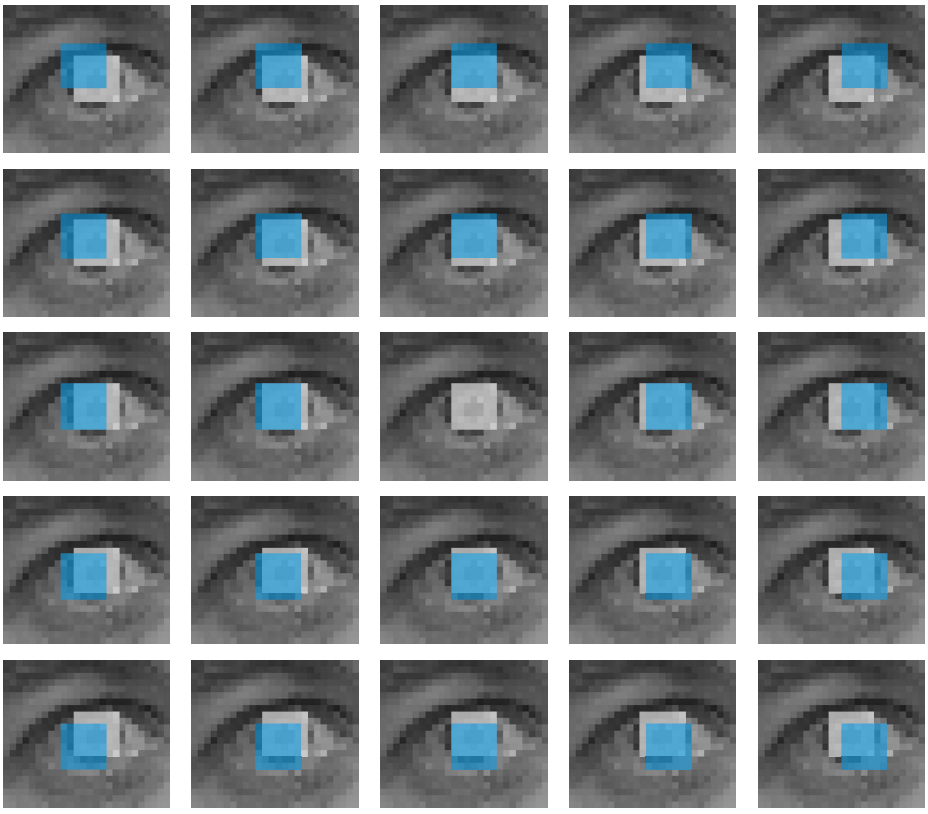


Figure 5.1: Patches belonging to the white patch's neighbourhood which spans $T = 2$ patches in all directions (the patches have maximum overlapping).

Step 1. Preprocessing

The faces are automatically detected by the Viola-Jones algorithm, equalised and scaled to a given size. It should be noted that no aligning techniques are applied, so in the end unaligned face images are classified. Therefore, the whole face area returned by the detector is used in the experiments.

Step 2. Feature Extraction

Given a preprocessed face image, both global and local approaches are taken to characterise the face. From the area of the image where the face is detected, one feature vector is extracted. In the local case, only featural information is supplied by describing overlapping patches of $L \times L$ pixels considered over the face image. We choose to utilise overlapping patches because they provide an exhaustive description of the face. Additionally, overlapping patches reported the best classification results in the experiments presented in

Section 4.3. A feature vector is extracted from each one of these patches, consequently a face is described by a set of feature vectors. Three different types of features are considered, in both global and local approaches:

- Grey levels,
- PCA, and
- *Ranking Labels*.

For information about extracting these features go to Section 2.3.

Step 3. Classification

The proposed local classification is compared to standard classifications. The standard (also referred to as global) versions of the classifiers employed are those explained in Section 2.4. Three classifiers are included in the experiments:

- k-Nearest Neighbour classifier (k-NN),
- Linear Discriminant Analysis (LDA), and
- Support Vector Machine (SVM).

For more details about these classifiers see Section 2.4.

Step 4. Performance Assessment

The measure utilised to evaluate the performance of the classification models is the classification accuracy (for more details see Section 2.5).

Step 5. Statistical Analysis

Due to the large number of experiments, a detailed comparison of the performances is difficult to provide. In order to ease the comparison task, several tests have been applied to show whether statistical differences exist among the performances of the classifiers. For further details about these tests see Section 2.6.

5.3.2 Face Image Dataset

The experiments involve non-occluded frontal faces from three well-known databases, FERET, PAL and AR (for specifics about them see Appendix A). From FERET database, 2,014 frontal face images are selected. Specifically, there are 1,173 male and 841 female faces corresponding to 787 different subjects (427 males and 360 females). From PAL database, all the images are used, that is, 575 face images. Concretely, there is one image per subject of a total of 224 males and 351 females. From AR database, 130 occlusion-free frontal

Global	Local
(1) 1NN-grey-G	(6) 1NN-grey-L
(2) 1NN-pca-G	(7) 1NN-pca-L
(3) LDA-pca-G	(8) 1NN-rank-L
(4) SVM-grey-G	(9) LDA-pca-L
(5) SVM-pca-G	

Table 5.1: Classification models considered in the experiments (classifier-feature-G/L).

face images with neutral expressions are chosen, which correspond to 74 males and 56 females. The ratio of males to females in each dataset is 1:0.7 for FERET, 0.6:1 for PAL and 1:0.8 for AR.

5.3.3 Experimental Setup

A number of experiments are designed to assess how robust global and local approaches are when training and test faces are acquired under different conditions. In order to evaluate the independence of each approach to the type of feature or classifier used, several experiments with different combinations of those factors are performed. From now on, the term *classification model* refers to a combination of:

- an approach (global or local),
- a type of feature (grey levels, PCA or *Ranking Labels*), and
- a classifier (k-NN, LDA or SVM).

A particular classification model is referred to as classifier-feature-G/L. A summary of these classification models is provided in Table 5.1. It should be noted that *Ranking Labels* are only defined as local features (for such definition see Chapter 4), therefore they cannot be considered in a global version. The SVM classifier is not used with a local approach because building a SVM per each neighbourhood has been proved an impractical solution. After conducting an empirical study, we concluded that local SVMs were computationally unaffordable due to the large amount of time required for training purposes. With regard to models based on LDA classifier, only PCA features are used. The reason is that this classifier is most commonly employed after applying PCA (further information in Section 2.4).

Each of those classification models is tested using all possible combinations for training and test of the three datasets indicated in Section 5.3.2. Consequently, 81 experiments (9 classification models $\times$ 9 training-test datasets) are performed. When the same database is used for training and testing, 5 repetitions of a 5-fold cross validation technique are implemented (that is, 25 runs of the experiment). The partition of the database is made by subjects, not by images. Therefore, one subject can only be in the training or test set,

Table 5.2: Number of global and local features involved in the experiments. In the local case, the number of features passed to each local classifier is given (there are a total of 1170 local classifiers). For local PCA, the average number of features per neighbourhood is provided.

	Grey levels	PCA	*Ranking Labels*
GLOBAL	1620	FERET: 236 PAL: 184 AR neutral: 49	–
LOCAL	49	FERET: 14 PAL: 17 AR neutral: 15	49

but never in both. In cross-database experiments, only one simulation is executed, training with one dataset and testing with the other.

In order to gain more insight about how certain degree of similarity of the face images affects the classification models, three statistical tests are applied to two groups of experiments. The first statistical study considers all the experiments which will give us information about which models are more suitable for the task. The second study only considers the experiments crossing databases which will provide further details about the performance of the models in more realistic scenarios.

Implementation Details

In this section, some implementation details needed for replicating the experiments are given. After detecting the face in the image, the face area is reduced to 45×36 pixels. When following a global approach, the features are extracted from the mentioned face area. In the local approach, the parameters involved are the size of the square patches, which is set to $L = 7$ and the number of patches that every neighbourhood spans in each direction, which is $T = 2$. As a result, each neighbourhood consists of 25 patches of size 7×7 pixels. The overlapping degree of the patches is maximum. From one patch to the next (in all directions), there is a one pixel shift.

For extracting the different types of features, the value of some parameters have to be decided. For PCA, we keep those components which account for 95% of the total variance of the training data. For *Ranking Labels*, the threshold that indicates the maximum difference to give two grey levels the same label is set to $G = 8$. For each type, the exact number of features passed to the global and local classifiers is given in Table 5.2.

With respect to the classifiers, the k-NN classifier searches only for the nearest neighbour, that is, $k = 1$. The LDA classifier is implemented using the numerical analysis library, ALGLIB [16]. The kernel of the SVM is a third degree polynomial. Particularly, we use the SVM implementation provided with LIBSVM [21].

Table 5.3: Classification accuracies (%) obtained in all experiments (marked in bold the best result per experiment).

		Global Classification					Local Classification			
		1-NN		LDA	SVM		1-NN			LDA
Training Dataset	Test Dataset	Grey Levels	PCA	PCA	Grey Levels	PCA	Grey Levels	PCA	Ranking Labels	PCA
FERET	FERET	85.31	85.57	91.86	93.66	92.83	92.35	91.29	**94.54**	85.07
	PAL	66.03	64.98	**71.25**	66.72	62.55	66.03	62.19	67.60	60.80
	AR Neutral	79.17	82.31	77.69	81.54	84.62	86.15	86.92	**90.77**	83.08
PAL	FERET	66.53	65.56	75.22	72.99	70.66	63.16	62.07	52.98	**77.11**
	PAL	77.42	77.35	82.72	85.23	**85.61**	83.73	83.52	80.17	73.69
	AR Neutral	81.25	82.31	89.23	**92.31**	91.54	90.00	90.00	86.15	87.69
AR Neutral	FERET	76.02	76.86	80.09	**80.83**	77.21	78.90	78.90	75.42	78.20
	PAL	73.35	72.30	71.43	75.09	70.38	74.39	73.17	**80.31**	65.51
	AR Neutral	83.99	82.46	87.54	90.42	**98.15**	88.92	89.08	95.54	86.31

5.3.4 Results and Discussion

The classification accuracies obtained in each one of the experiments are shown in Table 5.3. Looking at those results, the first impression is that the classification models using a global SVM or a local classifier obtain higher accuracies than the rest. In order to check whether these performance differences are statistically relevant or not, we applied three statistical tests. Two different statistical analyses are presented, the first one includes all the experiments, whereas the second one only includes cross-database experiments.

Study considering all the experiments

The results of this first study are shown in Tables 5.4(a-c).

Iman-Davenport's statistic (see Table 5.4(a)) finds significant differences among the performances of all classification models (that is, $F_F > F(8, 64)$) with a 95% confidence level, which is corroborated by the results of the other two tests.

Holm's method results with a 95% confidence level are presented in Table 5.4(b). All models above the double line performed significantly worse than the most significant model (marked in bold at the bottom of the table). These results indicate that the global models based on SVMs (both, with grey levels and PCA features) and LDA, together with local 1-NN classifiers using grey levels and *Ranking Labels* are statistically superior to the rest.

Wilcoxon's Signed Ranked test provides a pairwise comparison among all classification models which is summarised in Table 5.4(c). The symbol "•" indicates that the classification model in the row significantly outperforms the model in the column, and the

Table 5.4: Statistical tests applied to the classification accuracies obtained in all experiments and in only cross-database experiments. (a) Iman-Davenport's statistic. It is marked in bold if differences were detected. (b) Holm's results with a 95% significance level. Models above the double line performed significantly worse than the most significant model (marked in bold at the bottom). (c) Wilcoxon's Signed Ranked test summary. Above the main diagonal with a 90% significance level, and below it with a 95%. Symbol "•": model in row outperforms model in column, and "○": model in column outperforms model in row.

Statistical tests applied to all experiments

(a) Iman-Davenport's

$\mathbf{F_F = 3.48}$
$F(8, 64)_{0.95} = 2.09$

(b) Holm's

Model	P_{Holm}
1NN-pca-G	0.006313
1NN-grey-G	0.008742
LDA-pca-L	0.011675
1NN-pca-L	0.425958
LDA-pca-G	0.485341
1NN-rank-L	0.684693
1NN-grey-L	0.684693
SVM-pca-G	0.684693

SVM-grey-G

(c) Wilcoxon's

	1	2	3	4	5	6	7	8	9
1NN-grey-G (1)	-		○	○	○	○	○		
1NN-pca-G (2)		-	○	○	○	○	○		
LDA-pca-G (3)	•	•	-						
SVM-grey-G (4)	•	•		-		•	•		•
SVM-pca-G (5)	•				-				•
1NN-grey-L (6)	•	•				-			
1NN-pca-L (7)							-		
1NN-rank-L (8)								-	
LDA-pca-L (9)				○					-

Statistical tests applied to cross-database experiments

(d) Iman-Davenport's

$F_F = 1.12$
$F(8, 40)_{0.95} = 2.18$

(e) Holm's

Model	P_{Holm}
1NN-pca-G	0.163159
1NN-grey-G	0.164032
LDA-pca-L	0.346677
SVM-pca-G	0.700083
1NN-pca-L	0.700083
1NN-rank-L	0.700083
LDA-pca-G	0.700083
1NN-grey-L	0.700083

SVM-grey-G

(f) Wilcoxon's

	1	2	3	4	5	6	7	8	9
1NN-grey-G (1)	-		○						
1NN-pca-G (2)		-	○						
LDA-pca-G (3)			-						
SVM-grey-G (4)	•			-					
SVM-pca-G (5)					-				
1NN-grey-L (6)						-			
1NN-pca-L (7)							-		
1NN-rank-L (8)								-	
LDA-pca-L (9)									-

symbol "∘" indicates that the model in the column significantly surpasses the model in the row (above the main diagonal with a 90% confidence level, and below it with a 95%). Wilcoxon's results reveal that a global SVM model using grey levels outperforms the global 1-NN classifiers and all local models with the exception of 1-NN with *Ranking Labels*. This test also shows that a global 1-NN with grey level features achieves statistically worse performances than its local version and that most of the global models (except global 1-NN using PCA features).

Considering the findings of this first analysis, a straightforward conclusion would be that global methods are more suitable for dealing with a gender classification problem than local models. Particularly, a global SVM with simple grey level features appears to be the best choice. However, we would like to check whether these results hold if we apply the statistical tests only to the accuracies obtained in cross-database experiments.

Study considering only cross-database experiments

The results of the statistical tests omitting three experiments that were carried out using the same database for training and testing are shown in Tables 5.4(d-f).

Iman-Davenport's statistic (see Table 5.4(d)) does not find significant differences among classification models ($F_F \not> F(8, 40)$).

Holm's method (see Table 5.4(e)) only rejects global 1-NN with PCA, meaning that all the other models obtained statistically equal performances which were superior to the performance achieved by the mentioned model.

Wilcoxon's Signed Ranked test (see Table 5.4(f)) results supports the findings of Holm's test, since only a couple of statistical differences are found where global SVM with grey levels outperforms both global 1-NN models with a 90% confidence level. When increasing the confidence level to 95%, such SVM is statistically superior to only one of those models, global 1-NN with grey levels.

Summary of the Discussion

After the two statistical studies of the performances of all experiments and of a subset of them, an interesting fact was discovered: differences among the classification accuracies of the implemented models only exist when single-database experiments are taken into account. When training and test images do not share the same acquisition conditions nor the demography of subjects (that occurs in the cross-database experiments presented), no significant differences are found in the performances of the models. This leads to the conclusion that in real scenarios, where the characteristics of the training and test images are more likely to differ, due to the multiple uncontrollable factors, global solutions are as suitable as the proposed local approach to address gender classification tasks.

It is interesting to mention that the proposed classification based on local neighbourhoods outperformed the classification presented in the previous chapter where the neighbourhood considered included all the patches of the image.

5.4 Conclusions

In this chapter, we proposed a new classification method based on local neighbourhoods. In addition, we provided a comprehensive statistical study of the suitability of the proposed local classification for gender classification when compared to global approaches. Realistic conditions were simulated by cross-database experiments involving three face image databases with a wide range of ages and races, and different acquisition conditions. The comparison included three popular classifiers using three different types of features.

The main conclusion drawn from the results is that when addressing gender classification problems from neutral non-occluded faces, global methods and the proposed local approach achieve statistically equal accuracies. However, if we can ensure similar acquisition condition (i.e., similar to the experiments using the same database for training and testing), global features are generally more suitable for addressing the task.

As regards the classifiers and features, when the training and test images share the same characteristics, a global SVM is more likely to obtain the highest classification accuracies. Although the mentioned classifier was proved superior to many of the others, no statistical differences were found when compared to a local 1-NN using *Ranking Labels* or to a global LDA. In other cases, no significant differences were found among the three classifiers studied nor the different types of features considered.

The studies have provided statistical support to refute the generally accepted assumption that global techniques provide more useful information for discriminating between genders than local solutions. There are certain situations where the training and test images do not share the same characteristics and, in those cases, global solutions and the proposed classification based on local neighbourhoods perform in a similar manner.

CHAPTER 6

Gender Classification including Partially Occluded and Expressive Faces

Fairly often, the solutions presented for addressing gender classification problems are only tested using completely visible faces. That is a good practice, if those systems are going to be set in controlled environments. However, if we aim for a solution that can be employed in quite realistic scenarios, it is wise to check its performance using images of expressive and partially occluded faces. In this chapter, we analyse the performance of the solutions presented in previous chapters including the mentioned types of face images.

6.1 Motivation and Background

Most of the areas where automatic gender classification has interesting applications are usually set in real environments. In those scenarios, the accessories and clothes worn by the individuals are beyond our control. Furthermore, people do not normally show neutral faces, instead they express their feelings through facial expressions. These are the main reasons why automatic gender classification systems should be able to properly classify expressive and partially occluded faces. Many studies have been published proposing several methodologies for recognizing faces in the presence of occlusions [40, 64, 68], as opposed to the very few published studies on gender classification of occluded faces [61]. In that work, Toews and Arbel [61] propose a methodology for classifying visual traits using the Object Class Invariant (OCI) model. Faces are described by an OCI consisting of a segment line from the bottom of the nose to the forehead and a set of model features denoted by scale-invariant geometric and appearance image information. Using images from the FERET

database, the best classification rate is 88.10% obtained using a Bayesian classifier. In addition, the authors test their OCI model for classifying gender from simulated occluded faces. That is, images from the FERET database with a resolution of 256×384 pixels were artificially obscured by a black circle of different radii. With an occlusion of radius 40 pixels, the classification rate is 75%, however when the occluding radius goes up to 80 pixels, the classification rate drops to 60% which is roughly the percentage of male faces in the dataset.

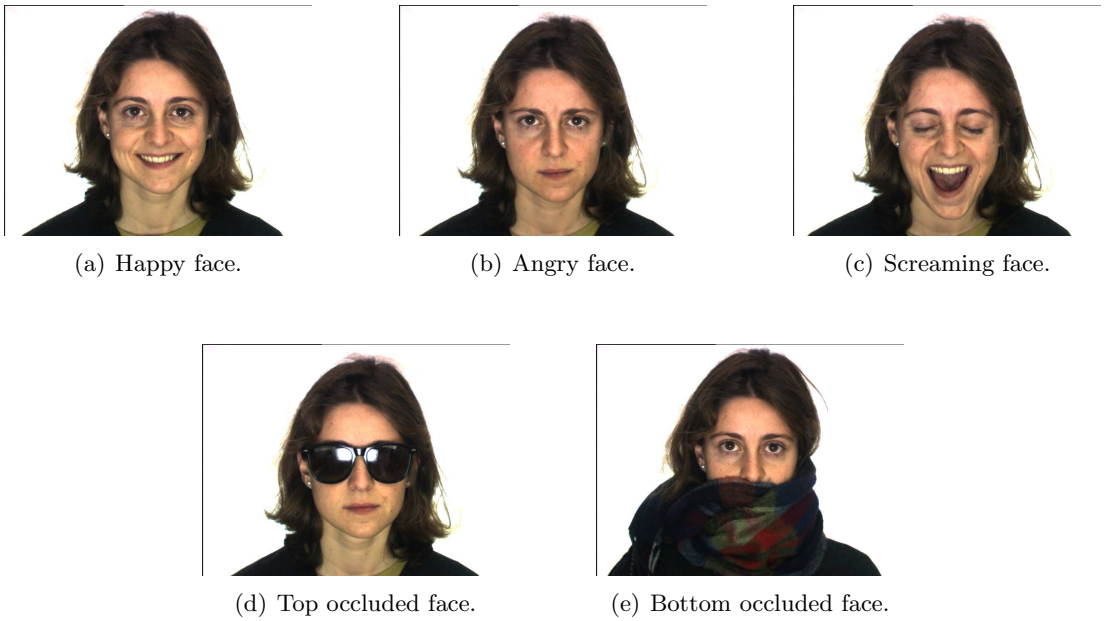
In the current literature, most of the automatic gender classification systems use the same face database for obtaining the training and test sets [35, 58, 49]. In those cases, the acquisition conditions of training and test images are practically the same which is far from a realistic scenario. Bekios-Calfa et al. [13] presented a single- and cross-database study on gender classification and proved that single-database experiments are optimistically biased. In cross-database experiments with a reasonable amount of training samples, a SVM with Radial Basis function kernel roughly achieved 80% of success. However, when there was less training data and a broad demography, all the compared classifiers achieved lower classification rates of around 70-75%. All three face databases used in that work contained non-occluded faces.

To the best of our knowledge, the problem of assessing the consequences of including expressive and occluded faces in the training and the evaluation of classifiers has not been extensively addressed in previous works. Therefore, the main aim of this chapter is to study the performance of the solutions presented in previous chapters in a more realistic scenario where there could be expressive and partially occluded faces. To that end, we present a comprehensive experimental study of gender classification techniques using face images showing different facial expressions and partially occluded faces. We compare global and local representation approaches, four types of features and three classifiers using two performance measures. In addition, the experiments are carried out on single databases and crossing databases to explore more realistic scenarios. Furthermore, the conclusions extracted from the experimental results are supported by three statistical tests applied to both performance measures.

6.2 Face Images with Distortions

In order to check the suitability of the proposed techniques for gender classification in more real conditions, we use images showing three different facial expressions and two types of occlusions. Apart from neutral and completely visible faces, the study includes the following types of images:

- Images of happy faces.
- Images of angry faces.
- Images of faces screaming.



(a) Happy face. (b) Angry face. (c) Screaming face.

(d) Top occluded face. (e) Bottom occluded face.

Figure 6.1: Example of expressive and partially occluded faces.

- Images of faces wearing sunglasses (top occlusion).
- Images of faces wearing a scarf (bottom occlusion).

An example of each type of expression and occlusion is shown in Figure 6.1. The figure presents the images as provided with the database, they have not been preprocessed. It should be noted that the expressive faces correspond to fake emotions, not real ones. Due to the fact that the individuals were asked to pretend to be happy, angry and screaming.

We consider partially occluded or expressive faces as faces with distortions, while faces without distortions would be those which are not occluded and show a neutral expression. Based on this definition of distortion, we analyse the behaviour of several classification models in different scenarios. Those scenarios vary depending on the presence or absence of distortions in the training and test images. There are nine scenarios each one consisting of several experiments which share the same characteristics. A summary of the nine groups is provided in Table 6.1.

Some of these nine scenarios represent classification problems with dataset shifts [56]. Dataset shift refers to those situations where the training and testing data do not follow the same distribution. This often occurs in real-world applications, that is probably why the machine learning community has shown a growing interest in this topic. In our case, dataset shifts occur when the types of distortion present in the training set do not appear in the test set, or vice versa. Hence, we could interpret the scenarios with respect to whether

Table 6.1: Nine different scenarios (G1-G9) are considered with respect to the involvement of distorted and non-distorted faces in the training and test sets.

		Training Set		
		With and Without Distortions	With Distortions	Without Distortions
Test Set	With and Without Distortions	**G1**	**G2**	**G3**
	With Distortions	**G4**	**G5**	**G6**
	Without Distortions	**G7**	**G8**	**G9**

there is dataset shift or not. We could say that the training set is representative of the test set in those scenarios where there is no dataset shift. That is the case of G1, G5 and G9. On the contrary, in the remaining cases, the training set can be considered unrepresentative of the test set due to the fact that both sets do not contain the same type of distortions (that is, there exists a dataset shift).

6.3 Experimental Study

6.3.1 Experimental Methodology

In this section, apart from giving the specifics of the experiments, we go into further details with respect to the groups introduced in the previous section.

Step 1. Preprocessing

The faces are automatically detected by Viola-Jones algorithm, equalised and resized. All these preprocessing techniques applied to the images are described in Section 2.2.

Step 2. Feature Extraction

Global as well as local representations are included in the present experimental comparison. For the local approach, the method for describing images based on patches introduced in Chapter 4 is employed. Specifically, the four types of features involved in the experiments are:

- Grey Levels
- Principal Components Analysis (PCA)
- Local Binary Patterns (LBP)
- Ranking Labels

For details about how to represent face images using the first three types of features see Section 2.3, and for information about Ranking Labels see Section 4.2.2. It is important to note that Ranking Label features were defined as a local descriptor, hence a global version of these features is not considered.

Step 3. Classification

This study also involves global and local classification methods. Particularly, the three classifiers included in the comparison are the following:

- k-Nearest Neighbour (k-NN)
- Support Vector Machine (SVM)
- Linear Discriminant Analysis (LDA)

The global versions of these classifiers were explained in depth in Section 2.4, whereas the local versions employ the classification based on neighbourhoods proposed in Chapter 5. It is worth mentioning that SVM is not considered with a local approach, due to the high computational cost of training the local SVMs.

Step 4. Performance Assessment

The performance of the classifiers is evaluated by means of two performance measures, accuracy and *D-prime*. Both measures were detailed in Section 2.5.

Step 5. Statistical Tests

In order to support the conclusions drawn from the experimental results, three statistical tests are applied over the two performance measures computed. Those tests are: Iman-Davenport's test, Holm's method and Wilcoxon's Signed Rank test (for further details about them see Section 2.6).

6.3.2 Face Image Dataset

Three face databases are used in the experiments, two of them containing only neutral faces (FERET and PAL databases), and a third database containing neutral, expressive, and realistically occluded faces (AR database). From FERET database, we use 2,014 frontal face images of 1,173 male and 841 female faces corresponding to 787 different subjects (427 males and 360 females). All the images included in PAL database are used, which correspond to one face image per subject of a total of 575 individuals (224 males and 351 females). From AR database, we use images corresponding to neutral, happy, angry and "screaming" faces, and faces presenting top occlusions caused by sunglasses and bottom occlusions caused by wearing a scarf. Concretely, we use images of 130 individuals (74 males and 56 females) per each facial expression, images of 129 individuals (74 males and

Table 6.2: Datasets involved in the experiments simulating scenarios with different distortions in the face images.

(a) Combinations of training and test datasets. A: Both, training and test without distortions. B: Training with distortions and test without them. C: Training without distortions and test with them. D: Both, training and test with distortions. Class balance ratios of male to female faces are shown below each training dataset.

		Training datasets				
		FERET 1:0.7	PAL 0.6:1	AR neutral 1:0.8	AR light distortions[1] 1:0.8	AR heavy distortions[2] 1:0.8
Test datasets	FERET	A	A	A	B	B
	PAL	A	A	A	B	B
	AR Neutral	A	A	A	B	B
	AR light distortions[1]	C	C	C	D	D
	AR heavy distortions[2]	C	C	C	D	D

[1]The dataset *AR light distortions* contains neutral and expressive faces.
[2]The dataset *AR heavy distortions* contains neutral, expressive and occluded faces.

(b) Dataset combinations included in each group of experiments from G1 to G9.

		Training datasets		
		With and Without Distortions	With Distortions	Without Distortions
Test datasets	With and Without Distortions	**G1**: A∪B∪C∪D	**G2**: B∪D	**G3**: A∪C
	With Distortions	**G4**: C∪D	**G5**: D	**G6**: C
	Without Distortions	**G7**: A∪B	**G8**: B	**G9**: A

55 females) wearing sunglasses and images of 125 individuals (72 males and 53 females) wearing a scarf.

All those face images are divided into several groups for having sets of images with or without distortions, and sets with both types of images. The exact class balance ratios of each set of images involved in the experiments are indicated in the next section.

6.3.3 Experimental Setup

As introduced in Section 6.2, face images of different characteristics (with respect to facial expression and occlusions) are included in the experiments. In addition, to recreate realistic conditions, those images are from different databases so they have different demography and acquisition conditions. All the combinations of training-test datasets are shown in Table 6.2(a). In that table, the letters A, B, C and D indicate if the dataset contains

Table 6.3: Classification models considered in the experiments (classifier-Feature-G/L).

Global	Local
(1) 1NN-grey-G	(8) 1NN-grey-L
(2) 1NN-pca-G	(9) 1NN-pca-L
(3) 1NN-lbp-G	(10) 1NN-lbp-L
(4) LDA-pca-G	(11) 1NN-rank-L
(5) SVM-grey-G	(12) LDA-pca-L
(6) SVM-pca-G	
(7) SVM-lbp-G	

(or not) distortions in the training/test sets. The class balance ratio for each dataset is also provided in the table. The face images from the AR database are divided into three datasets with an increasing level of difficulty: *AR neutral* contains only neutral faces, *AR light distortions* contains neutral and expressive faces, and *AR heavy distortions* contains the images of the previous dataset and also occluded faces.

The statistical study of the results is performed over the nine groups of experiments introduced in Section 6.2 which are built regarding some distortion criteria to assess specific experimental scenarios. The experiments included in each group are shown in Table 6.2(b).

As was indicated in Section 6.3.1, four types of features and three classifiers are used in the experiments. All possible combinations of classifiers and features, with three exceptions, are considered following both global and local approaches. A combination of classifier, feature and approach is referred to as *classification model*. Table 6.3 shows all the classification models. In order to name one particular model, we use the nomenclature classifier-feature-G/L. As can be seen on the table, the first exception is that the classifiers based on LDA are just combined with PCA features, because this classifier is commonly employed with those features (for further information see Section 2.4). The second exception relates to the local versions of SVM which are not included in the experiments due to the fact that the time needed to train all the local SVMs is computationally unaffordable. The third exception is based on the fact that the Ranking Label representation was defined for describing local regions, hence they are not considered with a global approach.

To assess classifier performances in single-database experiments, that is, experiments where the training and test sets are extracted from the same database, 5 repetitions of a 5-fold cross validation technique are executed (25 runs in total). The partitions needed for conducting these experiments are based on subjects instead of images. Therefore, images of the same individual could only be found in the training or the test set, but never in both. In cross-database experiments, only one simulation is performed, training with one database and testing with the other.

Table 6.4: Number of global and local features involved in the experiments. In the local case, the number of features passed to each local classifier is given (there are a total of 1170 local classifiers). For local PCA, the average number of features per neighbourhood is provided.

	Grey levels	LBP	PCA		Ranking Labels
GLOBAL	1620	1180	FERET: 236 PAL: 184 AR neutral: 49	AR light: 142 AR heavy: 165	–
LOCAL	49	59	FERET: 14 PAL: 17 AR neutral: 15	AR light: 16 AR heavy: 14	49

Implementation Details

In this section we give all the details that are necessary for replicating the experiments. The images are resized to 45×36 pixels and described using several types of features. Table 6.4 shows the number of features passed to the global and local classifiers in each case. The specifics regarding the parameters and its values for extracting each type of features are given next. In the case of global approaches the parameters are the following:

- Global grey level features consist of the pixel values. Therefore, the number of pixels (which is 1620) and the number of features coincide.
- Global LBP characterisations are based on $LBP^{u}_{8,2}$ sensitive to rotation, that means that each patch is represented by a histogram of 59 bins. The size of the patches is 9×9 pixels which has been chosen for being a reasonable size considering the resolution of the face images. For covering the whole image with patches of that size, a total of 20 patches with no overlapping are used. After concatenating the 59 bins extracted from the 20 patches, 1180 features are obtained.
- Global PCA features account for 95% of the variance of the training data. Hence, depending on the training data the number of features varies. In Table 6.4, the number of features resulting from each training dataset is indicated.

In local approaches, patches of 7×7 pixels are considered over the image with maximum overlapping (from one patch to the next there is a one-pixel shift). The local neighbourhoods span $T = 2$ positions in each direction from its center, resulting in neighbourhoods that cover 25 patches. Given the image size and the previous details, there is a total of 1170 patches per image. As a reminder, the classification based on local neighbourhoods considered one neighbourhood per patch and there is a local classifier which specialises in each neighbourhood. Following, we indicate the parameters of each type of local features:

- Local grey level features correspond to the pixel values within the patch, since the patch's size is 7×7, there are 49 grey level features.

- Local LBP features, as the global version, are based on $LBP^{u}_{8,2}$ sensitive to rotation. Hence, from each patch 59 features are extracted. Local LBP features were also extracted from patches larger than 7×7 pixels to test if the size of the patch influenced the classification task. We concluded that the performance was not strongly affected. Therefore, the size of patches is kept the same as for the other local features (that is, 7×7 pixels) to make possible a direct comparison with them.

- Local PCA features, as in the global case, account for 95% of the variance of the training data. As has been explained, each local classifier is trained with the patches belonging to the same neighbourhood. Hence, the training data depends on the neighbourhood being considered. In Table 6.4, we provide the average number of the PCA features selected per neighbourhood for each dataset.

- Local Ranking Label descriptions assign a label to each pixel within a patch. Hence, being the size of the patch 7×7, a total of 49 features are extracted.

As regards the classifiers, we use 1-NN, SVM with a third degree polynomial (the implementation provided with LIBSVM 3.0 [21]), and the LDA classifier is implemented using the numerical analysis library ALGLIB [16].

6.3.4 Results and Discussion

In this section, we present a wide comparison of the performance of all the classification models involved in the study. In order to provide a comprehensive analysis, we applied the statistical tests to the nine groups of experiments detailed in Table 6.2(b) for both performance measures, classification accuracy (ACC) and *D-prime* (the result of each experiment can be seen in Tables 6.7 and 6.8). Therefore, eighteen groups of experiments are considered for each statistical test.

Iman-Davenport's Statistic

According to Iman-Davenport's statistic (see Table 6.5), significant differences exist among the performance of all classification models using both measures. In other words, the value of the statistic (F_F) is always higher than the corresponding value of the F-distribution.

Holm's Method

Holm's method results for ACC and *D-prime* are shown in Figure 6.2 and Figure 6.3, respectively. Holm's null hypothesis assumes statistical equality to the control model (shown in bold at the bottom of each table). Those hypotheses associated to the models above the double line were rejected with a 95% significance level.

Taking into account Holm's results over both measures, there are various models which are always rejected. Those are the three global 1-NNs and the local LDA. This indicates that those models are not the most suitable for addressing gender classification problems.

Table 6.5: Iman-Davenport's statistic applied to *ACC* and *D-prime* values (marked in bold when statistical differences are found).

	ACC	*D-prime*	F-distribution
G1	$\mathbf{F_F = 20.17}$	$\mathbf{F_F = 43.44}$	$F(11, 264)_{0.95} = 1.83$
G2	$\mathbf{F_F = 22.41}$	$\mathbf{F_F = 26.95}$	$F(11, 99)_{0.95} = 1.87$
G3	$\mathbf{F_F = 7.75}$	$\mathbf{F_F = 23.92}$	$F(1, 154)_{0.95} = 1.85$
G4	$\mathbf{F_F = 15.76}$	$\mathbf{F_F = 32.11}$	$F(11, 99)_{0.95} = 1.87$
G5	$\mathbf{F_F = 70.41}$	$\mathbf{F_F = 26.02}$	$F(11, 33)_{0.95} = 2.09$
G6	$\mathbf{F_F = 6.75}$	$\mathbf{F_F = 37.37}$	$F(11, 55)_{0.95} = 1.97$
G7	$\mathbf{F_F = 8.68}$	$\mathbf{F_F = 19.76}$	$F(1, 154)_{0.95} = 1.85$
G8	$\mathbf{F_F = 8.92}$	$\mathbf{F_F = 11.88}$	$F(11, 55)_{0.95} = 1.97$
G9	$\mathbf{F_F = 3.88}$	$\mathbf{F_F = 11.20}$	$F(11, 88)_{0.95} = 1.90$

Besides, the hypothesis associated to the global version of LDA is only supported for groups G5 and G9 (in this last group just when considering accuracy results). In those two groups of experiments the training set is fully representative of the test set which leads us to think that the mentioned classification model (global LDA) is mainly appropriate in that particular situation.

If we focus the analysis on the classification accuracy results (Figure 6.2), global SVMs and local 1-NNs perform statistically better than the rest. However, if we consider the *D-prime* measure, global SVMs are not always among the models achieving statistically superior performances. That leaves local 1-NNs as the only models that are always among the statistically best for both measures. It is worth reminding at this point, that *D-prime* measure is less influenced by skewed classes than the classification accuracy.

Wilcoxon's Signed Rank Test

A summary of the results of Wilcoxon's signed rank test applied to the *ACC* and *D-prime* values of the different groups of experiments is depicted in Figure 6.4 and Figure 6.5, respectively. Remember that, in these figures, the symbol "•" indicates that the classification model in the row significantly outperforms the model in the column, and the symbol "◦" indicates that the classification model in the column significantly surpasses the model in the row. Above the main diagonal, the confidence level is 90% and, below it, it is 95%. In those tables the classification models are represented by a number from 1 to 12, see Table 6.3 for matching each number with the corresponding classification model.

Groups G1, G2, G3, G4, and G7: Images of both types (with and withouth distortions) were used either for training or for testing. Wilcoxon's results show that global SVMs and local 1-NNs statistically outperform the rest of the classification models. However,

Training Dataset

Test Dataset

With & Without Distortions

G1		G2		G3	
1NN-lbp-G	0	1NN-lbp-G	0.000001	1NN-lbp-G	0.000164
LDA-pca-L	0	LDA-pca-L	0.000005	LDA-pca-L	0.001616
1NN-pca-G	0.000003	1NN-grey-G	0.000383	1NN-pca-G	0.002398
1NN-grey-G	0.000008	1NN-pca-G	0.000506	1NN-lbp-L	0.003462
1NN-lbp-L	0.000344	LDA-pca-G	0.004043	1NN-grey-G	0.005822
LDA-pca-G	0.000526	SVM-lbp-G	0.021354	LDA-pca-G	0.054675
SVM-lbp-G	0.216919	1NN-lbp-L	0.054998	1NN-rank-L	2.16264
1NN-pca-L	0.957266	1NN-pca-L	0.689789	1NN-pca-L	2.16264
SVM-pca-G	1.668895	SVM-pca-G	1.370251	SVM-lbp-G	2.16264
SVM-grey-G	1.668895	SVM-grey-G	1.370251	SVM-pca-G	2.16264
1NN-rank-L	1.668895	1NN-grey-L	1.370251	SVM-grey-G	2.16264
1NN-grey-L		**1NN-rank-L**		**1NN-grey-L**	

With Distortions

G4		G5		G6	
1NN-lbp-G	0.000027	LDA-pca-L	0.000273	1NN-lbp-G	0.001847
LDA-pca-L	0.000027	1NN-lbp-G	0.001312	LDA-pca-G	0.003147
1NN-pca-G	0.002279	1NN-pca-G	0.012943	LDA-pca-L	0.005182
LDA-pca-G	0.002279	1NN-grey-G	0.016076	1NN-lbp-L	0.014346
1NN-lbp-L	0.004532	1NN-lbp-L	0.075512	1NN-pca-G	0.024319
1NN-grey-G	0.005441	LDA-pca-G	0.111618	1NN-grey-G	0.062434
SVM-lbp-G	0.768765	SVM-lbp-G	0.197366	SVM-grey-G	0.299516
SVM-grey-G	1.004995	1NN-pca-L	0.565304	1NN-rank-L	0.919065
1NN-pca-L	1.314634	1NN-grey-L	0.980399	SVM-lbp-G	1.010005
1NN-rank-L	1.314634	SVM-grey-G	1.247857	SVM-pca-G	1.010005
SVM-pca-G	1.314634	1NN-rank-L	1.247857	1NN-pca-L	1.010005
1NN-grey-L		**SVM-pca-G**		**1NN-grey-L**	

Without Distortions

G7		G8		G9	
1NN-lbp-G	0.00001	1NN-lbp-G	0.000242	1NN-lbp-G	0.005839
LDA-pca-L	0.00051	LDA-pca-L	0.005758	1NN-pca-G	0.010808
1NN-grey-G	0.000569	1NN-grey-G	0.014076	1NN-grey-G	0.013694
1NN-pca-G	0.000626	1NN-pca-G	0.024422	LDA-pca-L	0.021099
1NN-lbp-L	0.032011	LDA-pca-G	0.024422	1NN-lbp-L	0.022846
LDA-pca-G	0.084321	SVM-lbp-G	0.078391	1NN-pca-L	0.902289
SVM-lbp-G	0.341569	1NN-lbp-L	0.546573	LDA-pca-G	0.902289
1NN-pca-L	0.822161	SVM-pca-G	0.693942	1NN-rank-L	1.307198
SVM-pca-G	0.862834	1NN-pca-L	1.010005	1NN-grey-L	1.307198
1NN-grey-L	1.408222	SVM-grey-G	1.010005	SVM-lbp-G	1.307198
1NN-rank-L	1.408222	1NN-grey-L	1.010005	SVM-pca-G	1.307198
SVM-grey-G		**1NN-rank-L**		**SVM-grey-G**	

Figure 6.2: Holm's method results for measure accuracy (ACC) of all classification models compared with the most significant in each case (showed in bold at the bottom of each table) with a 95% significance level. All the models above the double line performed significantly worse than the model in bold. Refer to Table 6.2 for the description of the groups (G1 to G9) of experiments.

Training Dataset

Test Dataset	With & Without Distortions	With Distortions	Without Distortions
	G1	**G2**	**G3**
With & Without Distortions	1NN-lbp-G 0 1NN-grey-G 0 1NN-pca-G 0 LDA-pca-L 0 LDA-pca-G 0 SVM-lbp-G 0.000162 SVM-pca-G 0.000372 SVM-grey-G 0.013055 1NN-lbp-L 0.024322 1NN-pca-L 0.215604 1NN-grey-L 0.216633 **1NN-rank-L**	1NN-lbp-G 0.000001 LDA-pca-L 0.000006 1NN-grey-G 0.000293 1NN-pca-G 0.001587 LDA-pca-G 0.002855 SVM-lbp-G 0.006077 SVM-pca-G 0.219228 SVM-grey-G 1.111148 1NN-lbp-L 1.111148 1NN-pca-L 1.196182 1NN-rank-L 1.196182 **1NN-grey-L**	1NN-lbp-G 0 1NN-pca-G 0 1NN-grey-G 0 LDA-pca-G 0 LDA-pca-L 0.000003 SVM-pca-G 0.003139 SVM-grey-G 0.017978 SVM-lbp-G 0.024998 1NN-lbp-L 0.031659 1NN-pca-L 0.19965 1NN-grey-L 0.19965 **1NN-rank-L**
	G4	**G5**	**G6**
With Distortions	1NN-lbp-G 0 LDA-pca-G 0.000004 1NN-pca-G 0.000014 1NN-grey-G 0.000031 LDA-pca-L 0.000199 SVM-lbp-G 0.034705 SVM-pca-G 0.034705 SVM-grey-G 0.13992 1NN-lbp-L 0.753747 1NN-pca-L 0.753747 1NN-grey-L 0.753747 **1NN-rank-L**	LDA-pca-L 0.00314 1NN-lbp-G 0.004154 1NN-grey-G 0.063047 1NN-pca-G 0.063047 LDA-pca-G 0.130221 SVM-lbp-G 0.374686 SVM-pca-G 1.515966 1NN-pca-L 1.848297 1NN-lbp-L 1.848297 1NN-rank-L 1.848297 SVM-grey-G 1.848297 **1NN-grey-L**	1NN-lbp-G 0.000008 LDA-pca-G 0.000016 1NN-pca-G 0.000237 1NN-grey-G 0.000592 SVM-pca-G 0.027634 SVM-grey-G 0.027634 LDA-pca-L 0.041195 SVM-lbp-G 0.099899 1NN-lbp-L 0.643825 1NN-pca-L 0.643825 1NN-grey-L 0.643825 **1NN-rank-L**
	G7	**G8**	**G9**
Without Distortions	1NN-lbp-G 0 1NN-grey-G 0 LDA-pca-L 0 1NN-pca-G 0 LDA-pca-G 0.000105 SVM-lbp-G 0.009311 SVM-pca-G 0.01949 1NN-lbp-L 0.052375 SVM-grey-G 0.11365 1NN-pca-L 0.468117 1NN-grey-L 0.468117 **1NN-rank-L**	1NN-lbp-G 0.000202 LDA-pca-L 0.002306 1NN-grey-G 0.006947 SVM-lbp-G 0.027793 LDA-pca-G 0.031355 1NN-pca-G 0.031355 SVM-pca-G 0.249063 SVM-grey-G 0.512821 1NN-lbp-L 0.786995 1NN-pca-L 1.558611 1NN-grey-L 1.558611 **1NN-rank-L**	1NN-pca-G 0.000017 1NN-grey-G 0.000025 1NN-lbp-G 0.000092 LDA-pca-L 0.000092 LDA-pca-G 0.007566 1NN-lbp-L 0.132818 SVM-pca-G 0.168104 SVM-lbp-G 0.35677 SVM-grey-G 0.451145 1NN-grey-L 0.451145 1NN-pca-L 0.451145 **1NN-rank-L**

Figure 6.3: Holm's method results for measure *D-prime* of all classification models compared with the most significant in each case (showed in bold at the bottom of each table) with a 95% significance level. All the models above the double line performed significantly worse than the model in bold. Refer to Table 6.2 for the description of the groups (G1 to G9) of experiments.

if we focus on the results obtained with *D-prime*, most of the times local 1-NNs are significantly better than global SVMs, supporting the conclusion extracted from Holm's method. Specially, local 1-NN using Ranking Label features are statistically superior to all the classification models with only three exception in group G2 (local 1-NNs using grey levels and PCA features, and global SVMs using grey leves).

Group G5: Only distorted faces were used both for training and testing. Wilcoxon's test cannot find differences among the performances of the classification models due to insufficient data; to approximate a normal distribution, six or more experiments are needed [15] and this group only has four experiments (see Table 6.2). Note that, in this group of experiments, Holm's method finds reasonable statistical differences since having a small number of experiments can only cause the non rejection of false null hypotheses [25].

Group G6: Non-distorted faces were used for training and distorted faces for testing. Looking at the results obtained when considering the accuracy measure, three classification models stand out for significantly outperforming all global 1-NNs and also global LDA. Those three models are: the global SVM using grey levels and the local 1-NNs based on grey levels and PCA features. However, focusing on the *D-prime* results, all local 1-NNs are statistically superior to all global models. The local 1-NN using LBP features is the only one that just outperforms global 1-NNs and global LDA, but not global SVMs. It is worth highlighting, that the local 1-NN using Ranking Labels also outperforms the remaining local models which does not occur for any of the other local models.

Group G8: Distorted faces were used for training and non-distorted faces for testing. In this group, the global SVMs do not perform as well as they do with other experiment settings. All global SVMs only statistically outperform global 1-NN with LBP features with a 95% confidence level. However, local 1-NNs are found to be significantly superior to all global 1-NNs and some global SVMs (particularly, when taking into account *D-prime* results). This shows that in some circumstances when the training set is not representative of the test set, global SVMs might not be as suitable as local 1-NNs for dealing with gender classification problems.

Group G9: Only non-distorted faces were used both for training and testing. Wilcoxon's test findings differ depending on the measure employed. In *ACC* based tests, global SVM with grey levels statistically outperform most of the other models. However, in *D-prime* based tests, global SVMs and local 1-NNs show a statistical superiority to the rest of global models (with the exception of local 1-NNs using LBP features). The fact that local 1-NNs are among the best models only considering the *D-prime* measure suggests that local 1-NNs lead to more balanced performance rates between classes.

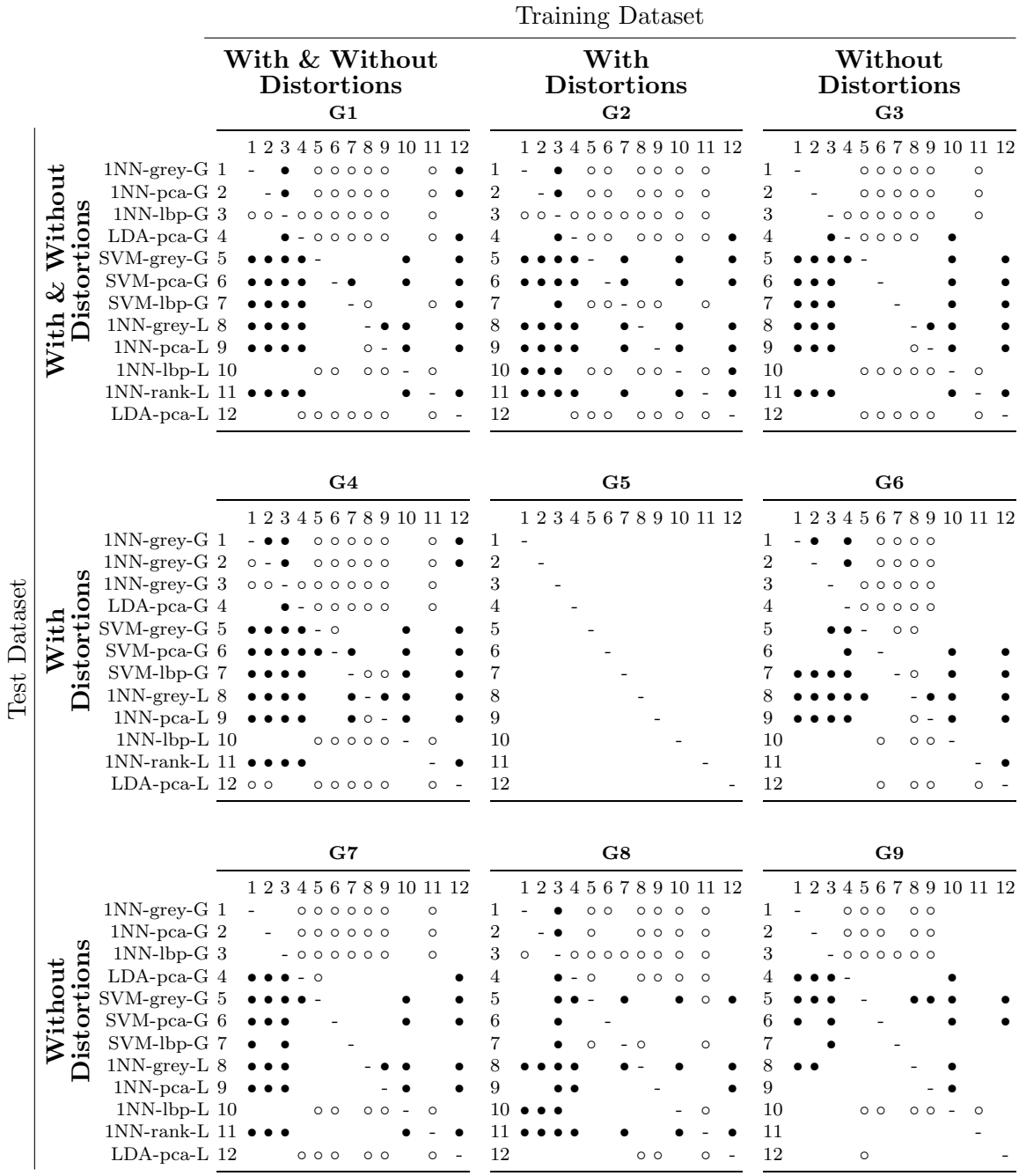


Figure 6.4: Summary of the Wilcoxon's Signed Rank test applied to the ACC values of all groups of experiments. The symbol "•" indicates that the classification model in the row significantly outperforms the model in the column, and the symbol "○" indicates that the model in the column significantly outperforms the model in the row (above the main diagonal with 90% confidence level, and below it with 95%). Refer to Table 6.3 for the description of the classification models numbered from 1 to 12 and to Table 6.2 for the description of the groups of experiments (G1 to G9). In group G5, Wilcoxon's cannot find differences due to insufficient data.

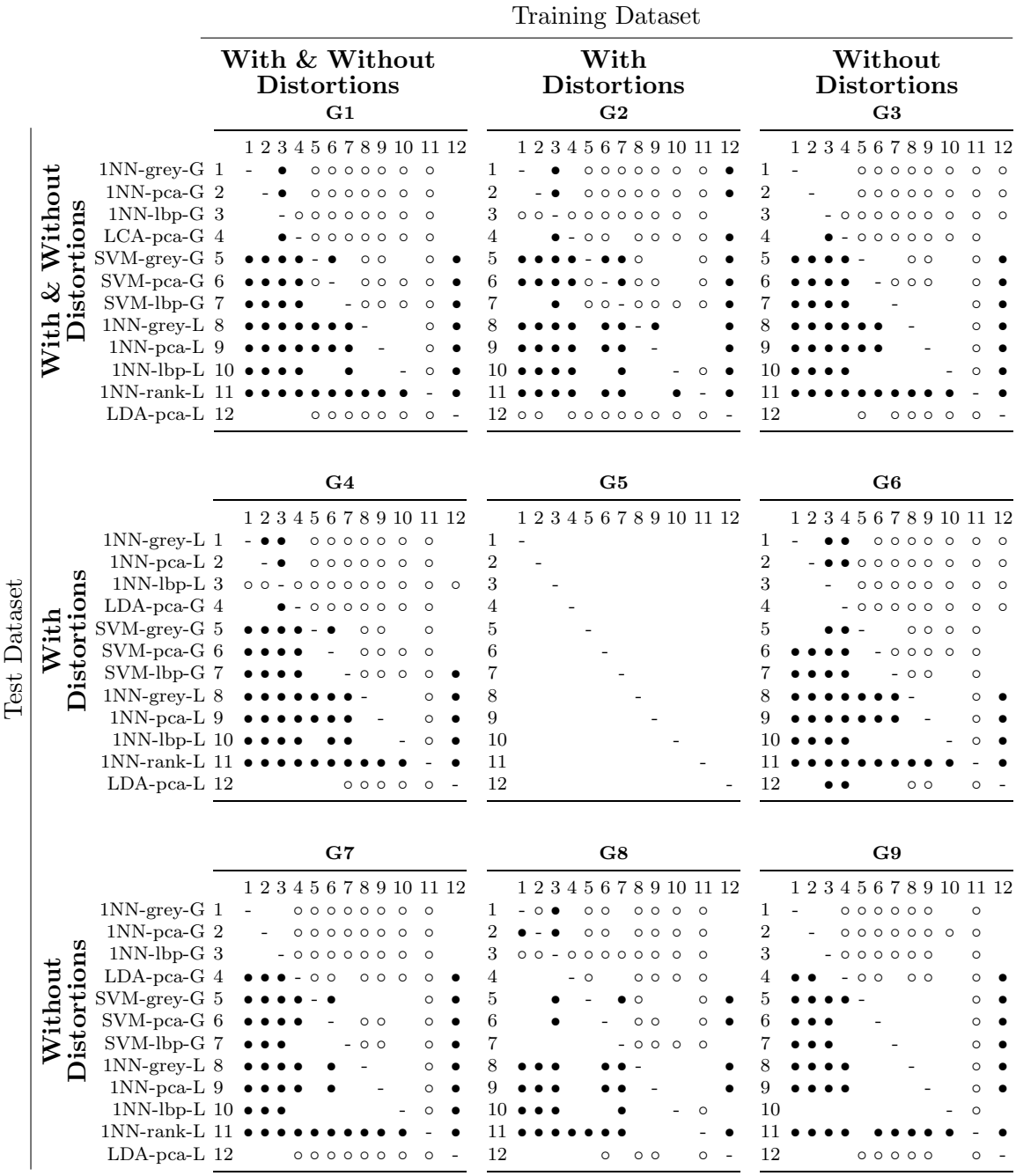


Figure 6.5: Summary of the Wilcoxon's Signed Rank test applied to the *D-prime* values of all groups of experiments. The symbol "•" indicates that the classification model in the row significantly outperforms the model in the column, and the symbol "○" indicates that the model in the column significantly outperforms the model in the row (above the main diagonal with 90% confidence level, and below it with 95%). Refer to Table 6.3 for the description of the classification models numbered from 1 to 12 and to Table 6.2 for the description of the groups of experiments (G1 to G9). In group G5, Wilcoxon's cannot find differences due to insufficient data.

Summary of the Discussion

In general, looking at the results, there are two differentiated sets of classification models, global SVMs together with the proposed local approach using 1-NNs, and the rest. The former set undoubtedly obtains better performances than the latter in most groups of experiments. In cases where the training set is unrepresentative of the test set (training faces present distortions and test faces do not or vice versa), local 1-NNs tend to be superior to the rest in statistical terms. However, when the training set is representative of the test set (the same type of faces are found in training and test sets), global SVMs and local 1-NNs behave similarly.

The statistical results showed that there is a tendency for local models to be superior to global solutions. This conclusion was strongly supported by *D-prime* based tests where local 1-NNs perform statistically better than most of the models. In particular, local 1-NN using Ranking Label features was superior to all global models with a 95% confidence level in most groups. The only exception occured in groups G2 and G9, where that model did not outperform global SVM with grey levels.

A clear advantage of one type of feature over the rest was detected, that is the case of the proposed Ranking Label features. Considering *D-prime* measure, this type of feature shows an obvious superiority over the other features. In terms of ACC values, these features are among the best in most groups of experiments.

Analysing the results (provided in Tables 6.7 and 6.8), the performances of the different models did not seem to be strongly affected by the datasets employed with the exception of FERET and PAL. When using those databases (in both combinations of training with one and testing with the other) lower performances were obtained. This is probably due to the different acquisition conditions of the face images in each database.

Comparison with other works

There is a considerable variability among the actual classification accuracies (see Table 6.7) which is due to the different levels of difficulty of each experiment. Considering only the best accuracy per experiment, the lowest and highest ACC values are 71.60% and 99.07%, respectively. However, this does not provide information about how good or bad the proposed solutions are with respect to what has been already published.

Although not all our results can be directly compared to those in the literature, Table 6.6 shows the accuracies obtained in similar published studies. As can be seen, our results are at the same level of the state of the art in face gender classification. It should be taken into account that the configurations of the experiments are different (for example, the partitions of the data into training and test sets probably differ).

Table 6.6: Comparison of different gender classification studies.

(a) Single-database experiments.

Database	Study	Classifier	Features	*ACC*
FERET	Our results	Local 1-NN	Ranking Labels	94.47
	Tapia et al. [60]	Feature selection (CMIFS) + SVM	LBP (various spatial scales)	97.53
	Moghaddam et al. [46]	SVM RBF kernel	PCA	96.62
	El-Din [27]	Bayesian Mixture of Experts	Haar-like + geometric	95.10
	Bekios-Calfa et al. [13]	SVM RBF kernel	Grey levels	93.95
	Alexandre [4]	SVM linear kernel	LBP	93.46
	Toews and Arbel [61]	Bayesian Classifier	Geometric and appearance	88.10
	Mäkinen et al. [39]	SVM RBF kernel	Grey levels	86.54
PAL	Our results	Global SVM	LBP	88.57
	Bekios-Calfa et al. [13]	SVM RBF kernel	Grey levels	89.81
AR Neutral	Our results	Global SVM	PCA	98.15
	Mozaffari et al. [47]	1-NN	LBP + DCT + geometric	96.00

(b) Cross-database experiments.

Databases (training/test)	Study	Classifier	Features	*ACC*
FERET/PAL	Our results	SVM	LBP	71.60
	Bekios-Calfa et al. [13]	LDA	PCA modified	71.50
PAL/FERET	Our results	LDA	PCA	77.11
	Bekios-Calfa et al. [13]	SVM RBF kernel	Grey levels	78.65

6.4 Conclusions

In this chapter, we presented a comprehensive experimental study on gender classification techniques using expressive, partially occluded and completely visible neutral faces. An extensive comparison of two approaches (global and local), four types of features (grey levels, PCA, LBP and Ranking Labels) and three classifiers (1-NN, SVM and LDA) was provided by means of three statistical tests applied to two performance measures (classification accuracy and *D-prime*).

The findings of these statistical tests indicated that global as well as our local approach can successfully solve gender classification problems when the training set is representative of the test set. However, in the case of training and test sets with different face distortions, the proposed local approach significantly outperformed global solutions according to the results of both performance measures. This superiority is due to the fact that the classification based on local neighbourhoods was designed to better deal with distortions in the images and an expressive/occluded face can be seen as a distorted face.

As regards the type of feature, models based on grey levels, PCA or Ranking Labels provided the best classification performances. In general, Ranking Label features showed to be the best choice. It should also be noted that LBP features, which are widely used in

facial analysis problems, were in many cases outperformed by models using other features. They are particularly not good when using a local approach which is probably caused by the complete lack of spatial information (it is a descriptor based on histograms).

Summarising, global SVMs together with local 1-NNs are the best models to address gender classification problems among those considered in this work. However, when the training set is unrepresentative of the test set, local 1-NNs surpass global solutions.

6.A Numerical Results

This sections shows the classification accuracies (Table 6.7) and *D-prime* values (Table 6.8) of each of the experiments carried out for the present study.

Table 6.7: Classification accuracies obtained in each of the experiments. Class balance ratios of males to females are shown below each training dataset. The best result for each training-test combination is marked in bold.

		Global							Local				
		1-NN			LDA	SVM			1-NN				LDA
Training Dataset	Test Dataset	Grey Levels	PCA	LBP	PCA	Grey Levels	PCA	LBP	Grey Levels	PCA	LBP	Ranking Labels	PCA
FERET 1:0.7	FERET	85.31	85.57	86.40	91.86	93.66	92.83	94.06	92.35	91.29	85.58	**94.47**	85.07
	PAL	66.03	64.98	58.19	71.25	66.72	62.55	**71.60**	66.03	62.19	43.03	67.77	60.80
	AR Neutral	79.17	82.31	75.37	77.69	81.54	84.62	84.69	86.15	86.92	61.20	**89.55**	83.08
	AR Light Dis.	82.79	82.60	70.21	78.39	81.07	82.60	83.80	86.99	86.62	64.71	**89.37**	79.73
	AR Heavy Dis.	76.06	74.90	67.43	72.84	74.00	76.71	78.74	83.66	83.14	63.12	**86.06**	76.32
PAL 0.6:1	FERET	66.53	65.56	71.49	75.22	72.99	70.66	69.62	63.16	62.07	56.50	52.88	**77.11**
	PAL	77.42	77.35	79.23	82.72	85.23	85.61	**88.57**	83.73	83.52	79.06	80.24	73.69
	AR Neutral	81.25	82.31	85.07	89.23	**92.31**	91.54	88.63	90.00	90.00	88.81	85.82	87.69
	AR Light Dis.	80.88	80.69	80.27	82.03	85.66	84.32	85.14	**86.99**	85.66	81.21	79.32	67.88
	AR Heavy Dis.	75.68	75.80	76.30	74.00	77.48	75.93	**79.73**	78.12	76.96	75.92	69.84	65.51
AR Neutral 1:0.8	FERET	76.02	76.86	65.23	80.09	**80.83**	77.21	65.29	78.90	78.90	69.86	74.96	78.20
	PAL	73.35	72.30	68.12	71.43	75.09	70.38	72.13	74.39	73.17	74.56	**76.92**	65.51
	AR Neutral	83.99	82.46	88.81	87.54	90.42	**98.15**	91.96	88.92	89.08	92.24	95.22	86.31
	AR Light Dis.	88.18	87.76	85.28	85.66	88.30	**94.65**	88.96	89.79	89.45	87.29	91.92	85.32
	AR Heavy Dis.	82.47	82.34	82.10	80.46	82.52	**92.66**	83.27	85.95	85.69	84.36	89.02	83.53
AR Light Distortions 1:0.8	FERET	72.59	72.94	69.20	76.56	77.66	75.22	74.52	80.59	**81.23**	75.72	77.56	77.41
	PAL	72.47	72.65	69.51	72.64	76.48	73.52	73.39	73.69	73.34	74.74	**81.36**	65.85
	AR Neutral	91.23	91.38	90.00	91.08	95.93	**96.92**	93.13	95.54	94.62	92.84	95.97	86.15
	AR Light Dis.	91.24	91.24	87.70	92.82	95.66	**99.07**	92.17	94.22	93.69	91.20	94.83	86.42
	AR Heavy Dis.	85.15	85.22	83.80	85.28	88.89	**92.79**	86.61	89.32	88.65	87.48	91.03	83.14
AR Heavy Distortions 1:0.8	FERET	71.50	72.10	68.65	73.58	75.77	75.22	69.22	82.17	**82.97**	74.53	77.81	77.56
	PAL	72.82	72.82	69.34	70.56	72.82	70.38	72.69	74.91	71.43	73.87	**81.18**	66.03
	AR Neutral	90.46	90.31	90.45	90.46	94.78	**98.46**	92.66	96.00	94.92	92.39	95.22	85.69
	AR Light Dis.	91.05	90.90	88.31	91.43	**95.73**	95.41	91.48	94.23	93.46	91.46	94.99	85.85
	AR Heavy Dis.	87.57	87.28	85.83	89.21	91.85	**98.06**	89.72	91.02	90.60	89.02	92.55	83.14

Table 6.8: *D-prime* values of each experiment. Class balance ratios of males to females are shown below each training dataset. The best result for each training-test combination is marked in bold.

		Global							Local				
		1-NN			LDA	SVM			1-NN				LDA
Training Dataset	Test Dataset	Grey Levels	PCA	LBP	PCA	Grey Levels	PCA	LBP	Grey Levels	PCA	LBP	Ranking Labels	PCA
FERET 1:0.7	FERET	2.08	2.11	2.23	2.78	3.04	2.92	3.12	3.13	3.18	2.69	**3.25**	2.16
	PAL	1.14	1.11	1.19	1.44	1.66	1.38	1.53	1.77	1.67	1.13	**1.90**	1.01
	AR Neutral	1.96	1.88	2.11	1.97	2.19	2.39	2.44	2.36	2.55	1.25	**2.99**	2.29
	AR Light Dis.	1.89	1.88	1.27	1.74	1.98	1.99	2.24	2.42	2.45	1.98	**2.96**	2.13
	AR Heavy Dis.	1.39	1.32	0.98	1.18	1.26	1.43	1.71	2.04	2.08	1.79	**2.29**	1.78
PAL 0.6:1	FERET	1.26	1.22	1.16	1.51	1.82	2.05	1.39	2.13	1.86	1.94	**2.17**	1.47
	PAL	1.46	1.46	1.74	1.86	2.05	2.08	2.37	2.21	2.19	2.27	**2.76**	1.19
	AR Neutral	1.89	1.83	2.44	2.39	2.89	2.74	2.88	2.55	2.58	**3.28**	3.10	2.35
	AR Light Dis.	1.76	1.74	1.73	1.82	2.12	2.05	2.32	2.42	2.36	3.07	**3.24**	1.98
	AR Heavy Dis.	1.41	1.42	1.46	1.28	1.51	1.48	1.92	2.03	2.01	2.58	**2.94**	1.68
AR Neutral 1:0.8	FERET	1.52	1.55	1.05	1.76	**1.77**	1.57	1.21	1.57	1.57	1.07	1.30	1.52
	PAL	1.07	1.12	0.83	1.06	1.33	1.08	1.29	1.27	1.22	1.58	**1.60**	0.84
	AR Neutral	1.98	1.86	2.46	2.34	2.63	2.46	2.84	2.52	2.53	3.31	**3.50**	2.37
	AR Light Dis.	2.37	2.34	2.09	2.18	2.49	2.47	2.46	2.66	2.64	2.63	**2.83**	2.22
	AR Heavy Dis.	1.88	1.88	1.82	1.74	1.95	1.89	1.93	2.30	2.30	2.50	**2.53**	2.02
AR Light Distortions 1:0.8	FERET	1.32	1.34	1.21	1.44	1.51	1.53	1.36	1.69	**1.75**	1.43	1.63	1.48
	PAL	1.13	1.15	0.92	1.39	1.40	1.19	1.19	1.24	1.24	1.67	**1.82**	0.83
	AR Neutral	2.82	2.99	2.57	2.71	3.44	3.21	3.01	3.59	3.50	3.42	**3.62**	2.29
	AR Light Dis.	2.74	2.76	2.33	2.84	**3.48**	3.31	2.84	3.35	3.28	3.08	3.25	2.36
	AR Heavy Dis.	2.18	2.19	1.96	2.08	2.49	2.47	2.21	2.76	2.72	**2.86**	2.74	1.96
AR Heavy Distortions 1:0.8	FERET	1.27	1.29	1.16	1.25	1.39	1.53	1.09	1.81	**1.89**	1.34	1.63	1.49
	PAL	1.15	1.16	0.91	1.15	1.13	0.98	1.15	1.36	1.17	1.54	**1.79**	0.89
	AR Neutral	2.66	2.71	2.64	2.64	3.21	3.06	2.91	**3.73**	3.59	3.38	3.36	2.31
	AR Light Dis.	2.71	2.69	2.39	2.72	**3.44**	3.17	2.74	3.37	3.26	3.05	3.29	2.29
	AR Heavy Dis.	2.29	2.26	2.14	2.46	2.77	2.68	2.53	2.77	2.73	2.89	**2.90**	1.99

CHAPTER 7

The Effect of Image Resolution on Face Gender Classification

Rationally, it would seem that the higher the resolution of the face images, the easier the gender classification task. In this chapter, we search for empirical evidence suggesting whether the mentioned assumption based on common sense is true. With that purpose, we statistically study if the classification results are affected by the resolution of the face images considering extremely low to fairly high image resolutions.

7.1 Motivation and Background

In the last two decades, automatic gender classification has received a broad research interest due to its wide range of applications. Studies on automatic face gender classification have been reported using many different learning algorithms. However, most of them focus on selecting the best classification technique and face representation using a fixed image resolution. Only a few works analyse the influence of the image size on the performance achieved by face gender classification systems. Tamura et al. [59] studied which of three resolutions (8×8, 16×16 and 32×32 pixels) reported better performances using Neural Networks, concluding that the best accuracy, which was 90%, was obtained with images of 32×32 pixels. However, using the lowest resolution the performance was only decreased by 3%. Moghaddam and Yang [46] investigated the performance of Support Vector Machines (SVMs) with Radial Basis Function and cubic polynomial kernels using two face image resolutions (21×12 and 84×48 pixels). They claimed that SVMs performed equally well (reaching about 95% accuracy) with both image sizes, reporting a difference of only

1%. Mäkinen and Raisamo [39] evaluated several classification models (Neutral Networks, SVMs and Adaboost) using three image resolutions (24×24, 36×36 and 48×48 pixels). They found that SVM with face images of 36×36 pixels achieved the best classification accuracy (which was, 86.54%). Although, that same classifier with 48×48 size images obtained similar results, images of 24×24 pixels provided not as good results. Additionally, they tested SVMs using grey level features and Local Binary Patterns, concluding that grey levels are more suitable for addressing gender classification problems. El-Din et al. [27] conducted a brief study for selecting the most suitable image size for their experiments. They tested two classifiers, SVM and Adaboost, with image resolutions starting from 8×8 to 40×40 with steps of 8 pixels. The resolution of 16×16 pixels was chosen as it provided the best results with a reasonable computational time. Those best accuracies were 92.91% with SVM and 93.71% with Adaboost, in both cases using FERET database. However, they claimed that increasing the image resolution can significantly degrade the classifiers performance (particularly, in the case of SVM using grey level features).

Although these works have addressed face gender classification problems comparing the performances obtained with different image resolutions, most of them were only carried out using a small set of image sizes. Besides, they were focused on finding the best image resolution for a specific combination of type of features and classifier. The experiments designed with that purpose always used images from the same databases for training and testing. This is very common practice in the field and it assumes that the characteristics of the training images coincide with those of the test images. This condition is rarely satisfied in real scenarios where many factors cannot be fully under control.

As far as we know, none of the published studies have explored the lower limit of meaningful resolutions of face images. The knowledge of a lower boundary could provide guidance on the design of vision systems when the circumstances do not make possible the acquisition of images of a reasonable size. In those situations, it is fundamental to know which is the smallest image resolution that carries discriminant information and, hence, it is useful for our purposes.

In this chapter, we present a detailed experimental study on the influence of the resolution of face images for automatic gender classification. The face images considered go from extremely low resolution images to the highest resolution available. For providing more realistic conditions, three different face databases are involved in both single- and cross-database experiments. With cross-database experiments, we simulate scenarios where the training and test images do not share the same characteristics, since face images from different databases have various acquisition conditions and diverse demography (age range, races, etc.). Additionally, the performances of two well-known and frequently utilised classifiers are employed, Support Vector Machine and k-Nearest Neighbour. In order to rigorously compare the classification performances, three statistical tests are applied over two performance measures.

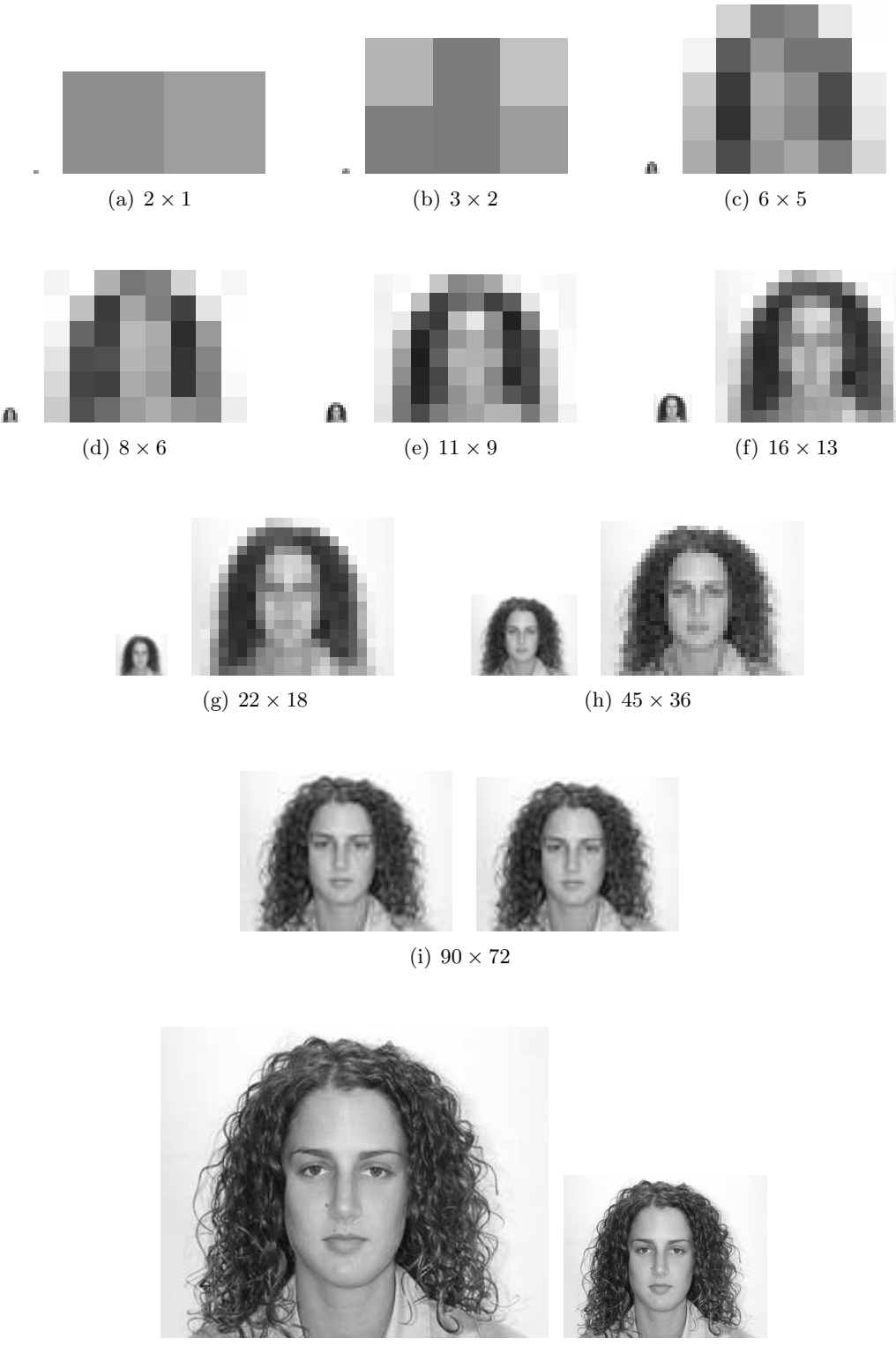
7.2 Image Resolutions

In most of the previous related works, no more than three image resolutions have been tested [59, 46, 39]. Moreover, in those studies where more image sizes (up to five) were compared [27], the aim was to find out which resolution was more suitable for a particular classification methodology. Our goal is to design an experimental study using a broad range of image resolutions, together with two different classifiers, to answer several questions:

- Which image sizes provide useful information to distinguish between genders?
- Which resolution provides enough discriminant information to achieve the best classification performance?
- Which is the smallest image size that improves the performance of a random classifier?

For this study, we consider ten different image sizes ranging from extremely low resolutions to the highest available resolution. Specifically, those ten image resolutions are: 2×1, 3×2, 6×5, 8×6, 11×9, 16×13, 22×18, 45×36, 90×72, and 329×264 pixels. For selecting the highest resolution, we consider the images with the lowest resolution among the highest image resolutions provided with the face databases. When using those images, the face detector returns areas of the same proportions but different sizes. Among those areas, the size that appears more often is 329×264 pixels. That is why we choose that particular highest resolution. The second highest resolution (90×72 pixels) is selected for making possible to appreciate the facial details while the computation time is reasonable for a realistic gender classification system. From that size on, the resolutions are chosen by dividing the width of the images by two. There are three exceptions which are the sizes 16×13, 8×6 and 2×1 pixels. The first two sizes are included in the study for gaining more insight around the resolutions where the classification performances are not stable. The lowest resolution is set to 2×1 pixels to keep the rectangular shape of the image. Figure 7.1 depicts an example of each image resolution. For each resolution two images are shown, the images on the left are the actual resized images while the images on the right are all scaled to the same size to better see the details.

In addition to the previous resolutions, the expected performance of a random classifier based on the *a priori* probabilities of classes in the training set is compared to those obtained with the different image sizes. From such comparative analysis, we draw conclusions about which resolution provides enough information to be more useful than a random classifier. This would be of valuable help in real systems for knowing whether an automatically captured face image could provide any useful discriminant information.



(a) 2×1 (b) 3×2 (c) 6×5

(d) 8×6 (e) 11×9 (f) 16×13

(g) 22×18 (h) 45×36

(i) 90×72

(j) 329×264

Figure 7.1: Example of each of the image resolutions involved in the study. For each resolution, the image on the left is shown at the actual image size (in the case of (j) the image is shown at half its size), while the image on the right is scaled to a common size to better see the details.

7.3 Experimental Study

7.3.1 Experimental Methodology

The experiments follow the general methodology presented in Chapter 2. Below, the specific details of each step are described.

Step 1. Preprocessing

The faces are detected using Viola-Jones algorithm and, afterwards, the area containing the face is scaled down to each one of the resolutions involved in the experiments. For the resizing process, a three-lobed Lanczos windowed sinc function [62] is utilised which keeps the original image aspect ratio. It is worth noting that the image sizes considered are rectangular due to the fact that the area of the image resulting from the face detector has that shape. Only one of the image sizes does not maintain the same aspect ratio, that is the case of 2×1 pixel images. The rest of the tasks applied to the images in this step are those introduced in Section 2.2.

Step 2. Feature Extraction

After rescaling, each face image is described by the grey level values of its pixels. An explanation of how to represent images with grey level features is provided in Section 2.3. These features are chosen for having been proved suitable for describing faces with gender classification purposes [39]. We also verified this fact as was shown in Section 5.3.4.

Step 3. Classification

The classifiers used in the experiments are the k-Nearest Neighbour and Support Vector Machine (SVM), both of them are very popular in the field. For further details about these models see Section 2.4.

Step 4. Performance Assessment

The performance of the classifiers is evaluated using two measures: classification accuracy (ACC) and geometric mean ($G\text{-}mean$), both described in Section 2.5.

Step 5. Statistical Analysis

This study consists in a considerable number of experiments, therefore in order to provide a rigorous comparison, three statistical tests are applied over the performances of the classification models. The tests are Iman-Davenport's statistic, Holm's method and Wilcoxon's Signed Rank test. All of them are explained in depth in Section 2.6.

7.3.2 Face Image Dataset

In the experiments, three different databases are used, FERET, PAL and AR (for a detailed description see Appendix A. From FERET database, we use 2,014 frontal view images of 1,173 male and 841 female faces corresponding to 787 different subjects (427 males and 360 females). The class imbalance ratio is 1:0.8. From PAL database all images are used, that is, 575 face images (225 males and 350 females). The class imbalance ratio is 0.6:1. From AR database, we use images of neutral faces (showing no facial expressions) of 130 subjects, resulting in 74 males and 56 females. The class imbalance ratio is 1:0.8.

7.3.3 Experimental Setup

For evaluating the influence of the resolution of the images when addressing face gender classification problems, we have designed single-database and cross-database experiments using the three databases described in Section 7.3.2. Specifically, all possible combinations of these three databases have been used for training and test, that is, three single-database experiments and six cross-database ones. As two different classifiers are evaluated (k-NN and SVM) per each of the ten image resolutions (see Section 7.2), a total of 180 experiments are carried out (9 training-test dataset combinations $\times$ 2 classifiers $\times$ 10 image resolutions).

In single-database experiments, the classification performances are estimated by two repetitions of a 5-fold cross-validation technique (that is, 10 runs of each experiment), where all face images of the same person are included in the same subset to avoid contamination effects between training and test partitions. These partitions keep the same ratio between female and male faces as the original database. In cross-database experiments, only one simulation is performed, training with one database and testing with the other.

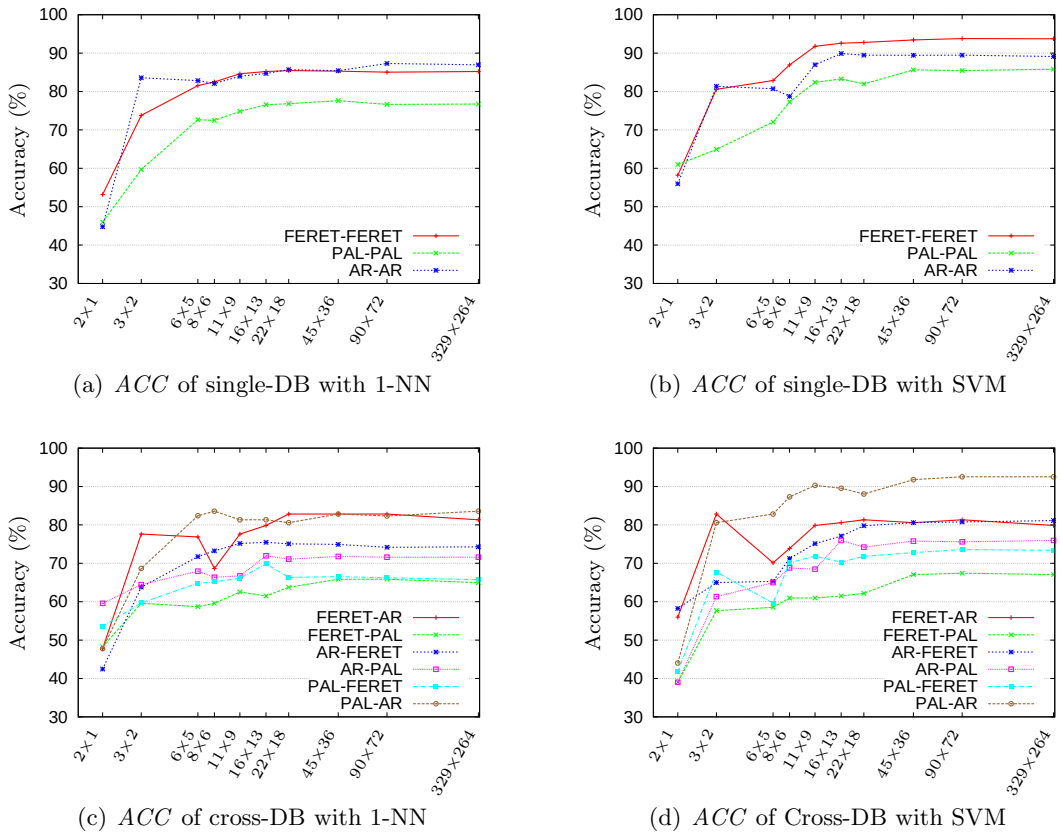
Implementation Details

With respect to the classifiers, the k-NN classifier searches for the nearest neighbour, that is, k = 1. The kernel of the SVM is a third degree polynomial, particularly we use the SVM implementation provided with LIBSVM [21].

7.3.4 Results and Discussion

Due to the large number of experiments, it is difficult to directly draw conclusions from the numerical results (which are presented in Section 7.A). Hence, we study these results graphically. The classification accuracies achieved in all experiments are depicted in Figure 7.2 while the geometric mean values are presented in Figure 7.3. Each figure shows four graphs, where (a) and (b) depict the performances over single-database experiments whereas, (c) and (d) show the performances over cross-database experiments. In both figures, (a) and (c) depict the results of 1-NN classifiers, and (c) and (d) show the results of SVM classifiers. In the graphs, the X axis presents the total number of pixels using a logarithmic scale and the Y axis corresponds to the value of the performance measure. From

a first analysis of these graphs, it can be seen that, for both measures, the performances achieved by SVMs are always higher than those obtained with 1-NN classifiers. Besides, the resolution at which the performances reach a stable state (where increasing the image size does not significantly improve the classification accuracy) is a bit higher for SVM than for 1-NN. Additionally, in cross-database experiments, the stability of performances is achieved with bigger image sizes than in single-database experiments. It seems reasonable that for dealing with more challenging classification problems (where training and test faces come from varied demographic populations, have different illumination conditions, etc.) more information is needed to distinguish between genders.



(a) *ACC* of single-DB with 1-NN

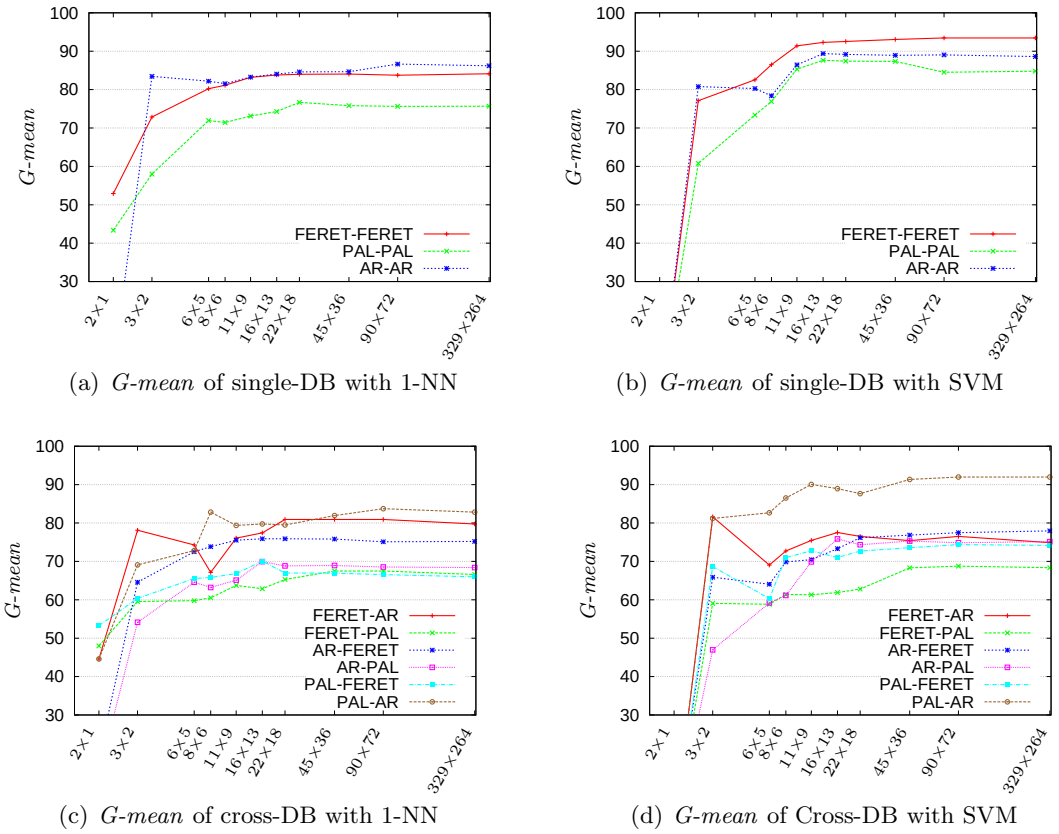
(b) *ACC* of single-DB with SVM

(c) *ACC* of cross-DB with 1-NN

(d) *ACC* of Cross-DB with SVM

Figure 7.2: Classification accuracies obtained in all experiments, separated in four plots depending on the classifier and the combination of training-test datasets.

Next, we discuss the results of the statistical tests (see Section 7.3.1) applied over the results of the experiments considering both measures. These tests have been applied to all experiments and also to four subgroups of them. This allow us to take a closer look to specific experiments to see if changes in the image resolution affect more to a



(a) *G-mean* of single-DB with 1-NN

(b) *G-mean* of single-DB with SVM

(c) *G-mean* of cross-DB with 1-NN

(d) *G-mean* of Cross-DB with SVM

Figure 7.3: Geometric mean values obtained in all experiments, separated in four plots depending on the classifier and the combination of training-test datasets.

particular classifier or to certain combinations of training-test datasets. Particularly, those four subgroups of experiments are: 1) experiments with 1-NN classifiers, 2) experiments with SVMs, 3) single-database experiments using both classifiers and 4) cross-database experiments using both classifiers. In the results of the statistical tests showed in this section, the performance of a random classifier based on the *a priory* probabilities of each class (considering only the training set) is labelled as "Baseline".

The Iman-Davenport's statistics over both ACC and *G-mean* values are shown in Table 7.1, respectively. If the value of the statistic (F_F) is larger than the corresponding value of the F-distribution, it means that there are significant differences among the performances included in the study. Looking at those results, we could see that this test strongly supports the existence of statistical differences among the classification results achieved using the image resolutions considered in all five groups of experiments. In order to gain some insight into these differences, we applied Holm's method.

Table 7.1: Iman-Davenport's statistic (F_F) applied to the results obtained in different groups of experiments (statistic marked in bold if differences were detected).

	ACC	*G-mean*	F-distribution
All experiments	$\mathbf{F_F = 67.13}$	$\mathbf{F_F = 69.71}$	$F(10, 170)_{0.95} = 1.89$
Only 1-NN exp.	$\mathbf{F_F = 28.51}$	$\mathbf{F_F = 36.23}$	$F(10, 80)_{0.95} = 1.95$
Only SVM exp.	$\mathbf{F_F = 42.57}$	$\mathbf{F_F = 33.07}$	$F(10, 80)_{0.95} = 1.95$
Single-DB exp.	$\mathbf{F_F = 40.46}$	$\mathbf{F_F = 33.37}$	$F(10, 50)_{0.95} = 2.03$
Cross-DB exp.	$\mathbf{F_F = 34.95}$	$\mathbf{F_F = 39.38}$	$F(10, 110)_{0.95} = 1.92$

Table 7.2: Holm's method applied to the results of all experiments. The image sizes above the double line achieved statistically worse performances than those below it with a significance level $\alpha = 0.95$.

(a) *ACC* values

resolution	P_{Holm}
Baseline	0
2x1	0
3x2	0.000013
6x5	0.000038
8x6	0.000431
11x9	0.044863
16x13	1.119829
22x18	1.119829
329x264	1.603227
90x72	1.603227
45x36	

(b) *G-mean* values

resolution	P_{Holm}
2x1	0
Baseline	0
6x5	0.000019
3x2	0.000034
8x6	0.00059
11x9	0.079307
16x13	2.054317
22x18	2.054317
329x264	2.054317
90x72	2.054317
45x36	

Holm's method applied over the *ACC* values with a confidence level $\alpha = 0.95$ (see Table 7.2(a)) revealed that a set of five image resolutions are always selected for achieving the best classification rates without statistical differences among them. These resolutions are the five highest ones: 16×13, 22×18, 45×36, 90×72 and 329×264. For most subgroups of experiments (see Table 7.3(a-d)), with the exception of the group using a SVM classifier, Holm's selected 45×36 pixels as the most discriminant image size. These results are consistent with the ones obtained when applying Holm's method to *G-mean* values which are shown in Table 7.2(b) and Table 7.3(e-h). It can be seen that the classifications based on the same five resolutions along with an extra one (11×9 pixels) provide statistically better results than the rest. Regarding the most significant resolution, it varies depending on the group of experiments. When using a 1-NN classifier (Table 7.3(e)), images of 45×36 pixels provide the best performances. For solutions based on a SVM classifier (Table 7.3(f)), the 90×72 pixel resolution is considered the most significant and the same happens when training and test images come from different databases (Table 7.3(h)). However, when only images from one database are used (Table 7.3(g)), the size 22×18 pixels is the most

Table 7.3: Holm's method applied to the results of the four subgroups of experiments. The image resolutions above the double line achieved statistically worse performances than those below it with a significance level $\alpha = 0.95$.

ACC values

(a) Only 1-NN exp.

resolution	P_{Holm}
2x1	0.000002
Baseline	0.000002
3x2	0.001148
6x5	0.01095
8x6	0.015149
11x9	0.214128
16x13	1.497439
90x72	1.497439
329x264	1.497439
22x18	1.497439
45x36	

(b) Only SVM exp.

resolution	P_{Holm}
Baseline	0
2x1	0.000001
6x5	0.000552
3x2	0.001536
8x6	0.005706
11x9	0.086389
22x18	0.542372
16x13	0.542372
45x36	0.954579
329x264	0.954579
90x72	

(c) Single-DB exp.

resolution	P_{Holm}
Baseline	0.000294
2x1	0.000561
3x2	0.024665
6x5	0.0488
8x6	0.0488
11x9	0.490904
16x13	2.055564
22x18	2.382008
329x264	2.382008
90x72	2.382008
45x36	

(d) Cross-DB exp.

resolution	P_{Holm}
2x1	0
Baseline	0
3x2	0.001229
6x5	0.001752
8x6	0.018809
11x9	0.21127
22x18	1.125831
16x13	1.166656
329x264	1.563628
90x72	1.563628
45x36	

G-mean values

(e) Only 1-NN exp.

resolution	P_{Holm}
2x1	0.000001
Baseline	0.000006
3x2	0.001756
6x5	0.004527
8x6	0.015149
11x9	0.165031
16x13	1.575075
329x264	1.575075
90x72	1.575075
22x18	1.575075
45x36	

(f) Only SVM exp.

resolution	P_{Holm}
2x1	0.000001
Baseline	0.000025
6x5	0.001324
3x2	0.007553
8x6	0.017025
11x9	0.378111
22x18	1.497439
16x13	1.709015
45x36	1.709015
329x264	1.709015
90x72	

(g) Single-DB exp.

resolution	P_{Holm}
2x1	0.0002
Baseline	0.000808
3x2	0.032602
6x5	0.0488
8x6	0.0488
11x9	0.490904
16x13	2.653686
90x72	2.653686
329x264	2.653686
45x36	2.653686
22x18	

(h) Cross-DB exp.

resolution	P_{Holm}
2x1	0
Baseline	0.000001
6x5	0.000389
3x2	0.001217
8x6	0.013891
11x9	0.244499
22x18	1.624194
16x13	1.624194
329x264	1.624194
45x36	1.624194
90x72	

Table 7.4: Summary of the Wilcoxon's Signed Rank test applied to the results of all experiments. "•"= the resolution in the row improves that of the column, and "○"= the resolution in the column improves that of the row. Above main diagonal, significance level $\alpha = 0.90$, below main diagonal $\alpha = 0.95$.

(a) *ACC* Values

	1	2	3	4	5	6	7	8	9	10	11
Baseline (1)	-		○	○	○	○	○	○	○	○	○
2x1 (2)		-	○	○	○	○	○	○	○	○	○
3x2 (3)	•	•	-	○	○	○	○	○	○	○	○
6x5 (4)	•	•		-	○	○	○	○	○	○	○
8x6 (5)	•	•	•	•	-	○	○	○	○	○	○
11x9 (6)	•	•	•	•	•	-	○	○	○	○	○
16x13 (7)	•	•	•	•	•	•	-		○	○	○
22x18 (8)	•	•	•	•	•	•		-	○	○	○
45x36 (9)	•	•	•	•	•	•	•	•	-		
90x72 (10)	•	•	•	•	•	•	•	•		-	
329x264 (11)	•	•	•	•	•	•		•			-

(b) *G-mean* Values

	1	2	3	4	5	6	7	8	9	10	11
Baseline (1)	-	•	○	○	○	○	○	○	○	○	○
2x1 (2)	○	-	○	○	○	○	○	○	○	○	○
3x2 (3)	•	•	-		○	○	○	○	○	○	○
6x5 (4)	•	•		-	○	○	○	○	○	○	○
8x6 (5)	•	•	•	•	-	○	○	○	○	○	○
11x9 (6)	•	•	•	•	•	-	○	○	○	○	○
16x13 (7)	•	•	•	•	•	•	-		○		
22x18 (8)	•	•	•	•	•	•		-	○		
45x36 (9)	•	•	•	•	•	•			-		
90x72 (10)	•	•	•	•	•	•				-	•
329x264 (11)	•	•	•	•	•	•					-

significant. These findings lead to the conclusions that SVM classifiers (which are based on a global model) take advantage of higher resolutions while simpler classifiers, such as the 1-NN which is based on local comparisons, do not. Moreover, when addressing more complex problems (for example, those presented with cross-database experiments), an increase in the information provided by moderately high resolution images is useful for better dealing with gender classification problems.

Holm's results show that the gender classification performance reaches certain stability when using medium and high image resolutions. However, this performance is significantly, and always negatively, affected when using images smaller than 11×9 pixels.

To deepen our understanding of the differences among image resolutions, a pairwise comparison is provided by applying Wilcoxon's Signed Rank test. The results of Wilcoxon's test are summarised in the form of tables. In such tables, the symbol "•" indicates that the resolution in the row achieves significantly better results than the one in the column, and the symbol "○" indicates that the resolution in the column obtains a significant improvement over the one in the row. The level of significance for those differences marked above the main diagonal is $\alpha = 0.90$ and below it, $\alpha = 0.95$.

Wilcoxon's results over the *ACC* and *G-mean* values of all experiments are shown in Table 7.4. Focusing on the *ACC* measure (Table 7.4(a)), the higher the resolution, the statistically better the results (up to 45×36 pixels). However, with the *G-mean* measure (Table 7.4(b)), the performances achieved with resolutions of 16×13 pixels and higher show no statistical differences. According to Wilcoxon's test findings with respect to the random classifier (labelled "Baseline" on the tables), face images of size 3×2 pixels and higher provide a certain amount of useful information for distinguishing between genders. Although, extremely low-resolution face images, such as 3×2 pixels, may provide some

Table 7.5: Summary of the Wilcoxon's Signed Rank test applied to the results of the four subgroups of experiments. "•"= the resolution in the row improves that of the column. "○"= the resolution in the column improves that of the row. Above the main diagonal, the significance level is $\alpha = 0.90$, below it, it is $\alpha = 0.95$.

ACC values

(a) Only 1-NN exp.

	1	2	3	4	5	6	7	8	9	10	11
Baseline (1)	-	○	○	○	○	○	○	○	○	○	○
2x1 (2)		-	○	○	○	○	○	○	○	○	○
3x2 (3)	•	•	-	○		○	○	○	○	○	○
6x5 (4)	•	•		-		○	○	○	○	○	○
8x6 (5)	•	•			-	○	○	○	○	○	○
11x9 (6)	•	•	•			-	○	○	○	○	○
16x13 (7)	•	•	•	•	•		-				
22x18 (8)	•	•	•	•	•	•		-			
45x36 (9)	•	•	•	•	•	•			-		
90x72 (10)	•	•	•	•	•	•				-	
329x264 (11)	•	•	•	•	•	•					-

(b) Only SVM exp.

	1	2	3	4	5	6	7	8	9	10	11
1	-	○	○	○	○	○	○	○	○	○	○
2		-	○	○	○	○	○	○	○	○	○
3	•	•	-		○	○	○	○	○	○	○
4	•	•		-	○	○	○	○	○	○	○
5	•	•		•	-	○	○	○	○	○	○
6	•	•	•	•	•	-		○	○	○	○
7	•	•	•	•	•		-		○	○	○
8	•	•	•	•	•			-	○	○	○
9	•	•	•	•	•	•	•	•	-	○	
10	•	•	•	•	•	•	•	•	•	-	
11	•	•	•	•	•	•	•	•			-

(c) Single-DB exp.

	1	2	3	4	5	6	7	8	9	10	11
1	-	○	○	○	○	○	○	○	○	○	○
2		-	○	○	○	○	○	○	○	○	○
3	•	•	-		○	○	○	○	○	○	○
4	•	•		-	○	○	○	○	○	○	○
5	•	•			-	○	○	○	○	○	○
6	•	•	•	•	•	-	○	○	○	○	○
7	•	•	•	•	•	•	-	○			
8	•	•	•	•	•			-			
9	•	•	•	•	•	•			-		
10	•	•	•	•	•	•				-	
11	•	•	•	•	•	•					-

(d) Cross-DB exp.

	1	2	3	4	5	6	7	8	9	10	11
1	-	○	○	○	○	○	○	○	○	○	○
2		-	○	○	○	○	○	○	○	○	○
3	•	•	-		○	○	○	○	○	○	○
4	•	•		-	○	○	○	○	○	○	○
5	•	•			-	○	○	○	○	○	○
6	•	•	•	•	•	-	○	○	○	○	○
7	•	•	•	•	•	•	-			○	
8	•	•	•	•	•	•		-	○	○	
9	•	•	•	•	•	•	•	•	-		
10	•	•	•	•	•	•	•	•		-	
11	•	•	•	•	•	•					-

G-mean values

(e) Only 1-NN exp.

	1	2	3	4	5	6	7	8	9	10	11
Baseline (1)	-	•	○	○	○	○	○	○	○	○	○
2x1 (2)	○	-	○	○	○	○	○	○	○	○	○
3x2 (3)	•	•	-	○		○	○	○	○	○	○
6x5 (4)	•	•		-		○	○	○	○	○	○
8x6 (5)	•	•			-	○	○	○	○	○	○
11x9 (6)	•	•	•	•		-	○	○	○	○	○
16x13 (7)	•	•	•	•	•		-				
22x18 (8)	•	•	•	•	•	•		-			
45x36 (9)	•	•	•	•	•	•			-		
90x72 (10)	•	•	•	•	•	•				-	•
329x264 (11)	•	•	•	•	•	•					-

(f) Only SVM exp.

	1	2	3	4	5	6	7	8	9	10	11
1	-	•	○	○	○	○	○	○	○	○	○
2	○	-	○	○	○	○	○	○	○	○	○
3	•	•	-		○	○	○	○	○	○	○
4	•	•		-	○	○	○	○	○	○	○
5	•	•		•	-	○	○	○	○	○	○
6	•	•	•	•	•	-	○	○	○	○	○
7	•	•	•	•	•		-				
8	•	•	•	•	•			-			
9	•	•	•	•	•	•			-		
10	•	•	•	•	•	•				-	
11	•	•	•	•	•	•					-

(g) Single-DB exp.

	1	2	3	4	5	6	7	8	9	10	11
1	-	•	○	○	○	○	○	○	○	○	○
2		-	○	○	○	○	○	○	○	○	○
3	•	•	-		○	○	○	○	○	○	○
4	•	•		-	○	○	○	○	○	○	○
5	•	•			-	○	○	○	○	○	○
6	•	•	•	•	•	-					
7	•	•	•	•	•	•	-				
8	•	•	•	•	•	•		-			
9	•	•	•	•	•	•			-		
10	•	•	•	•	•					-	
11	•	•	•	•	•						-

(h) Cross-DB exp.

	1	2	3	4	5	6	7	8	9	10	11
1	-	•	○	○	○	○	○	○	○	○	○
2	○	-	○	○	○	○	○	○	○	○	○
3	•	•	-		○	○	○	○	○	○	○
4	•	•		-	○	○	○	○	○	○	○
5	•	•			-	○	○	○	○	○	○
6	•	•	•	•	•	-	○	○	○	○	○
7	•	•	•	•	•		-				
8	•	•	•	•	•	•		-	○	○	
9	•	•	•	•	•	•			-		
10	•	•	•	•	•	•				-	•
11	•	•	•	•	•	•					-

useful information, this information is clearly insufficient when compared to moderate image sizes, such as 11×9 pixels.

The results obtained applying Wilcoxon's test on the subgroups of experiments are shown in Table 7.5. In all cases, higher resolutions appear to be more useful for achieving statistically better performances than lower ones. This occurs up to images of 16×13 pixels. From that resolution on, an increase in the size of the image does not result in statistically superior classification rates. Focusing on Wilcoxon's results over ACC values (Table 7.5(a-d)), we find an exception to this behaviour: the experiments using a SVM classifier (Table 7.5(b)). SVM solutions using higher resolutions obtain statistically better performances than using smaller sizes.

7.4 Conclusions

In this chapter, we presented a detailed statistical study on the influence of image resolution for addressing automatic face gender classification problems. In our experimental study, two classifiers (SVM and 1-NN), three face image databases and ten image resolutions were tested. Single-database and cross-database experiments simulated different degrees of difficulty in the classification task. The classification performances were assessed with two measures (classification accuracy and geometric mean) and then analysed by means of three statistical tests.

After comprehensively comparing the performances of the experiments, the questions raised at the beginning of the chapter can now be answered.

Which image sizes provide useful information to distinguish between genders? In general, a moderate image size of 45×36 pixels provides enough information to discriminate between genders with the statistically best performances for both ACC and *G-mean* measures. Those performances remain consistent independently of the classifier and the training-test datasets used.

Which resolution provides enough discriminant information to achieve the best classification performance? In order to obtain the statistically best accuracies an image size between 22×18 and 90×72 pixels is recommended. For selecting a resolution within this range the classifier to be used should be taken into consideration. SVMs take advantage of higher resolutions than 1-NN classifiers, although the performances achieved by 1-NN classifiers were always lower than those obtained by SVMs. From a reasonable image size, both classifiers showed robustness with respect to scale and degree of facial details.

Which is the smallest image size that improves the performance of a random classifier? From the results, we can conclude that an extremely low resolution image of 3×2 pixels is enough to provide some discriminant information, since the classification results using images of that size were significantly better than a random classifier.

Summarising, in situations where different resolution face images are available, moderately sized faces from 22×18 to 90×72 pixels are optimal for gender classification tasks.

Besides, the performance of the classifiers was robust towards changes in the image resolution (using medium to the highest tested sizes). Only when the image resolution was 8×6 pixels or smaller, the classification results were significantly affected.

7.A Numerical Results

This sections shows the classification accuracies (Tables 7.6 and 7.7) and geometric means (Tables 7.8 and 7.9) of each one of the 180 experiments carried out for the present study. The columns labelled as "Baseline" in the tables show the *ACC* and *G-mean* values of a random classifier based on *a priori* probabilities.

Table 7.6: Classification accuracies achieved by 1-NN classifiers over all training-test datasets using ten image resolutions. The result of a random classifier is labelled as "Baseline". Marked in bold the best result per experiment.

			Image Resolutions									
Training Dataset	Test Dataset	Baseline	2×1	3×2	6×5	8×6	11×9	16×13	22×18	45×36	90×72	329×264
FERET	FERET	51.28	53.17	73.78	81.55	82.49	84.63	85.20	**85.45**	85.30	85.03	85.20
	PAL	48.24	48.25	59.58	58.71	59.58	62.54	61.48	63.76	**65.85**	**65.85**	64.98
	AR	51.12	47.76	77.61	76.86	68.65	77.61	79.85	**82.83**	**82.83**	**82.83**	81.34
PAL	FERET	48.24	53.57	59.68	64.84	65.14	66.13	**69.86**	66.33	66.53	66.23	65.73
	PAL	52.42	45.99	59.68	72.64	72.47	74.82	76.56	76.82	**77.61**	76.65	76.74
	AR	48.56	47.76	68.65	82.38	**83.58**	81.34	81.34	80.59	82.83	82.32	**83.58**
AR	FERET	51.12	42.45	63.80	71.74	73.23	75.17	**75.47**	75.07	74.92	74.18	74.28
	PAL	48.56	59.58	64.45	67.94	66.37	66.72	**71.95**	71.08	71.77	71.60	71.60
	AR	50.98	44.77	83.58	82.83	82.08	83.95	84.70	85.70	85.44	**87.31**	86.94

Table 7.7: Classification accuracies achieved by SVMs over all training-test datasets using ten image resolutions. The result of a random classifier is labelled as "Baseline". Marked in bold the best result per experiment.

			Image Resolutions									
Training Dataset	Test Dataset	Baseline	2×1	3×2	6×5	8×6	11×9	16×13	22×18	45×36	90×72	329×264
FERET	FERET	51.28	58.21	80.56	82.87	86.92	91.77	92.57	92.79	93.44	**93.79**	93.74
	PAL	48.24	39.02	57.66	58.53	60.97	60.97	61.49	62.19	67.07	**67.42**	67.07
	AR	51.12	55.97	**82.83**	70.14	73.88	79.85	80.59	81.34	80.59	81.34	79.85
PAL	FERET	48.24	41.75	67.72	59.63	70.15	71.84	70.30	71.84	72.79	**73.58**	73.38
	PAL	52.42	60.98	64.91	72.04	77.27	82.41	83.28	81.97	85.63	85.45	**85.80**
	AR	48.56	44.02	80.59	82.83	87.31	90.29	89.55	88.05	91.79	**92.53**	**92.53**
AR	FERET	51.12	58.24	64.99	65.29	71.25	75.12	77.11	79.79	80.58	80.78	**81.13**
	PAL	48.56	39.02	61.32	64.98	68.81	68.46	**75.95**	74.21	75.78	75.60	**75.95**
	AR	50.98	55.96	81.35	80.71	78.73	86.97	**89.89**	89.47	89.44	89.48	89.11

Table 7.8: *G-mean* values obtained by 1-NN classifiers over all training-test datasets using ten image resolutions. The result of a random classifier is labelled as "Baseline". Marked in bold the best result per experiment.

			Image Resolutions									
Training Dataset	Test Dataset	Baseline	2×1	3×2	6×5	8×6	11×9	16×13	22×18	45×36	90×72	329×264
FERET	FERET	49.34	52.92	72.87	80.23	81.13	83.19	83.80	84.00	84.04	83.75	**84.12**
	PAL	48.14	48.00	59.54	59.77	60.52	63.73	62.88	65.29	**67.50**	67.49	66.60
	AR	49.51	44.59	78.11	74.29	67.15	76.06	77.42	**80.93**	**80.93**	**80.93**	79.75
PAL	FERET	48.95	53.25	60.35	65.55	65.82	66.83	**70.09**	66.97	66.94	66.57	65.98
	PAL	48.77	43.38	57.99	71.94	71.42	73.13	74.28	**76.67**	75.82	75.64	75.66
	AR	49.53	44.59	69.09	72.81	82.83	79.38	79.75	79.48	81.93	**83.73**	82.83
AR	FERET	49.36	19.22	64.57	72.50	73.84	75.52	**75.90**	75.89	75.82	75.11	75.21
	PAL	48.77	14.66	54.15	64.58	63.21	65.12	**69.94**	68.82	68.93	68.54	68.40
	AR	49.59	8.16	83.44	82.20	81.55	83.27	84.04	84.61	84.67	**86.64**	86.19

Table 7.9: *G-mean* values obtained by SVMs over all training-test datasets using ten image resolutions. The result of a random classifier is labelled as "Baseline". Marked in bold the best result per experiment.

			Image Resolutions									
Training Dataset	Test Dataset	Baseline	2×1	3×2	6×5	8×6	11×9	16×13	22×18	45×36	90×72	329×264
FERET	FERET	49.34	0.00	77.10	82.56	86.49	91.39	92.29	92.53	93.06	**93.46**	93.45
	PAL	48.14	0.00	59.07	58.83	61.38	61.31	61.89	62.81	68.32	**68.76**	68.36
	AR	49.51	0.00	81.62	69.04	72.72	75.46	**77.53**	76.51	75.40	76.51	74.89
PAL	FERET	48.95	0.00	68.65	60.29	71.00	72.77	71.02	72.69	73.56	**74.38**	74.17
	PAL	48.77	0.00	60.78	73.37	76.89	85.31	**87.63**	87.42	87.34	84.52	84.80
	AR	49.53	0.00	81.18	82.67	86.51	90.05	88.94	87.66	91.34	**91.98**	**91.98**
AR	FERET	49.36	0.00	65.89	64.05	69.86	70.45	73.32	76.18	76.85	77.49	**77.95**
	PAL	48.77	0.00	46.96	59.20	61.14	69.87	**75.86**	74.35	75.29	74.87	75.15
	AR	49.59	0.00	80.80	80.27	78.39	86.44	**89.37**	89.17	88.93	89.05	88.61

CHAPTER 8

Conclusions and Future Work

EVERY step of the automatic gender classification process has been revised and several improvements have been proposed to address the problem under more realistic circumstances. In this chapter, we revisit the proposed improvements as well as the conclusions drawn from the various empirical studies performed. We then indicate some lines of work that remain open and other interesting extensions of our work.

8.1 Conclusions

The primary goal of the thesis was to improve face gender classification in fairly realistic scenarios. Under those conditions, many problems can arise due to the lack of control over the individuals and their surroundings. The main issues we have focussed on are related to the distortions present on the faces caused by expressing emotions or wearing pieces of clothing. In order to tackle the mentioned problems among others, we have presented new approaches for addressing several of the tasks involved in the classification process together with empirical proof of the suitability of the proposed methods. Additionally, we have experimentally studied how different factors (such as image size or inaccurate face detections) affect the gender classification rates.

Our first attempt on face gender classification was focused on understanding what parts of the face contain more discriminant information. We studied the role of different face parts by comparing the performances of several classifiers trained with individual parts (the eyes, the nose, the mouth and the chin). In the comparison, we also included classifiers based on holistic parts of the face (particularly, the internal face, the external face and the full face). As a result, the individual parts were proved to provide enough discriminant information to

deal with gender classification, although holistic parts were always superior. From all the individual parts involved, the eyes were the most robust part since they always achieved one of the highest classification accuracies in experiments over two face image databases. The fact that classifiers based on just one individual part were capable of reasonably solve the classification problem raised the question of whether different face parts provided complementary information. Hence, we evaluated that complementarity using various ensembles of classifiers where each base learner was trained with a different part. The experimental comparison showed that ensembles based on three face parts reached similar classification rates to those based on five parts. This data indicated that gender classification problems could be satisfactorily resolved even when the face is not completely visible.

The next step we took was towards handling situations where the demography of the individuals is diverse, the faces can present local distortions and the detection of the faces may be inaccurate. On one hand, we designed new local features (Ranking Labels) to characterise face images. Those features separately describe regions of the images by means of local contrast values whereas spatial information is maintained. The suitability of Ranking Labels to address gender classification problems was confirmed by an experimental comparison with other widely used methods (Grey Levels, PCA, Local Binary Patterns and Local Contrast Histograms). On the other hand, we introduced a new classification approach based on local neighbourhoods. It consists of an ensemble method designed to be used in combination with a local face description. Thus, each member of the ensemble specialises in a specific area of the face. The use of local neighbourhoods provides a certain level of flexibility in case of unaligned or misaligned faces and inaccurate face detections. To test the adequacy of the proposed method when compared to other highly effective techniques, we employed a complete set of single-database and cross-database experiments (including only neutral non-occluded faces). The latter type of experiment simulates scenarios with considerable variability of the demography. The results showed that the proposed classification technique in conjunction with Ranking Label features is an approach as suitable as global solutions for addressing gender classification. In addition, several tests provided statistical evidence to refute the assumption that global approaches achieve better performances than local procedures. The situations where global classifications outperformed the proposed local approach were reduced to those in which training and test images shared the same characteristics. However, this is unlikely to occur in real environments since we have no control whatsoever.

Following, we analysed the performance of gender classification techniques when are presented with more challenging face images. Now, the image set contained neutral, expressive and partially occluded faces. These face images were combined for different training and testing purposes in single and cross-database experiments. The experiments were addressed by the proposed classification method as well as various extensively employed global approaches. A comprehensive statistical comparison of the results produced several interesting findings. If the training set is representative of the test set, global as well as local approaches successfully deal with gender classification. However, in situation where the training/test set contains expressive or occluded images and the other one does not, the

performance of local approaches clearly surpasses those obtained by global models. This empirical data also suggested that Ranking Label features provide more discriminant information than the other descriptions tested (which were based on grey levels, PCA, and LBP). As regards the classifiers, whereas global SVMs are among the best models only in situations where the training set is representative of the test set, local 1-NNs achieve good classification rates in all cases.

Aside from the main objective, we also performed a detailed experimental study to gain insight into the influence of the image resolution in face gender classification tasks. We went beyond previously published works with respect to the number of different image sizes involved and the diversity of classifiers. Empirical data indicated that a moderate image size between 22×18 and 90×72 pixels is optimal for addressing gender classification. Besides, the experimental results statistically supported that an image as small as 3×2 pixels provides discriminant information to distinguish between males and females. This knowledge can be crucial for those systems where it is not always possible to acquire reasonably sized images.

Summarising, we have revised all the steps in the gender classification process from face images, proposing improvements in each one of them. These improvements can be all applied or just a selection of them can be independently employed.

8.2 Future Work

In the thesis, we have presented substantial improvements applicable to face gender classification systems. However, the effectiveness of the systems can be limited in certain real scenarios. Therefore, there is still some work to do until the problem can be considered fully resolved. Various interesting research lines towards reaching that objective are described next, as well as some open lines of work that we commenced during the thesis but that have not yet been completed.

Open Lines of Work of the Thesis

Following, we list several lines of work which remain under development after the conclusion of the thesis.

- We are conducting an online study with the aim of comparing the performance of humans and machines in face gender classification tasks. It consists of a large set of preprocessed face images that the participants have to label as belonging to a male or female person. The set of images contains neutral and expressive faces as well as faces wearing sunglasses and scarves. In order to replicate as much as possible the same conditions of automatic systems, the images are showed with the same reduced size. After collecting a reasonable amount of data, we plan to statistically analyse which personal traits make difficult the task for humans and for automatic systems. In addition, we have asked the participants to indicate their age, race and gender in

order to study the influence of those factors when recognising the gender of faces of the same (or different) age group, race and gender.

- A functional prototype which shows, in real time, the gender of the people appearing on camera is being implemented. It consists of a camera capturing images in real time, and as soon as a face is detected, it is passed to the classification system. The gender prediction is shown on the screen represented by a square of a specific colour surrounding the person's face. This prototype is almost fully functional, it just needs to pass an exhaustive debugging process.

Future Research Lines

Several research lines that could improve the currently available automatic face gender classification systems are described next. We focus on some extension that could be easily applied to our work.

- We paid attention to databases with fairly controlled illumination conditions and only frontal faces. However, a gender classification system which is reliable in all kinds of situations should be able to cope with extremely different illumination conditions, such as intensive light coming from one side. In addition, it should provide accurate gender classifications of faces at different angles. Hence, experiments crossing databases containing those types of face images are an interesting extension of our work.

- Intuitively, the facial features which show the gender of a person seem to vary with the age of the person. An empirical study of face gender classification separating the individuals depending on their age range would shed light on this topic.

- According to psychological studies, humans suffer from *cross-race identification bias* effects [43]. It refers to the tendency of having more difficulty to identify people of different races to our own. Consequently, it seems that the characteristics for identifying people may differ depending on their race. It would be interesting to study whether automatic systems show that behaviour too. This could be done by comparing the performances of a generic classifier trained with faces from different races and specific classifiers trained with faces of a particular race.

- The proposed methods have not been exclusively designed for addressing gender classification problems. Therefore, they could be applied to other classification problems. The most direct applications could be facial expression, age or race recognition.

- As mentioned in the introduction, many authors propose to employ face recognition systems to solve gender classification problems. Hence, we think that the field could gain from a conclusive empirical analysis of the solutions for addressing both problems.

APPENDIX A

Face Image Databases

EACH of the face image databases involved in the experiments of the thesis is publicly available to the scientific community and researchers have been utilising them for face analysis tasks since their publication. Specifically, we have used four databases which are the AR database, the Facial Recognition Technology database (FERET), the Productive Aging Laboratory (PAL) and the extended M2VTS database (XM2VTS). In this appendix, we give general information about each database and the particular characteristics of the face images provided with them.

A.1 AR

The AR database [41] was created in the Computer Vision Centre at the Universitat Autònoma de Barcelona in 1998. It consists of over 4,000 colour images of 768×576 pixels corresponding to 136 people's faces (76 males and 60 females). Consequently, the ratio of males to females is 1:0.8. The images feature frontal views of faces with different illumination conditions, facial expressions, and occlusions. Although the illumination conditions were controlled, in order to achieve different lightning conditions an intense source of light was placed to the right, left and both sides of the subject. As regards the facial expression, each person was asked to pose with a happy face, an angry face and screaming. These emotions were faked, hence the face images show facial expression which cannot be considered completely real. In the case of occlusions, this were produced by real clothing accessories being worn by the subjects. Particularly, sunglasses and a scarf. As a result, there are images with partial occlusions on the top half of the faces and others on the bottom half.

Each person participated in two sessions (separated by 2 weeks) where the same set of images was taken. Hence, the database has two samples per person of each of the different images. No restrictions with respect to clothes, accessories (glasses, jewellery, etc.) or hairstyle were imposed to participants. No information is provided with the database regarding the age and race of the individuals. However, after majority sampling the database it can be said that most individuals are young Caucasian adults.

In the thesis, we employ only images of the first session. In particular, 130 images of each expression (neutral, happy, angry and screaming) corresponding to 74 males and 56 females, 129 images of individuals wearing sunglasses (74 males and 55 females) and 125 images of individuals wearing a scarf (72 males and 53 females). Not all the 136 images of each type were used due to errors in the face detection process. Figure A.1 shows all the images involved in the experiments of the thesis taken from the same subject.

(a) Neutral face.

(b) Happy face.

(c) Angry face.

(d) Screaming face.

(e) Wearing sunglasses.

(f) Wearing scarf.

Figure A.1: Example of the images of the same provided with AR database involved in the experiments of the thesis.

A.2 FERET

The Facial Recognition Technology Database (FERET) [55] was a project of the U.S Department of Defence. The images were collected in 15 sessions between August 1993 and July 1996 from 1,199 individuals. For some individuals, over two years had elapsed between their first and last sittings, with some subjects being photographed multiple times. This time lapse enables researchers to study the robustness of their algorithms with respect to changes in the subjects' appearance that occur over a year. The illumination conditions were controlled in every session, and there are individuals from different races and age groups. The database provides colour images of size 512×768 pixels which feature different poses. The angle of the head with respect to the camera varies in each pose going from frontal to profile faces. All the different poses considered are detailed in Table A.1. The database does not contain images from all poses of all subjects, in the table is indicated the number of images and subjects of each pose.

In the thesis, we only use frontal images from the regular and alternative image sets. Figure A.2 shows some examples of images of those sets acquired in different sessions.

Table A.1: Total number of people in FERET database for each of the images taken (the angle of the head for each pose is indicated in degrees).

Description	Pose	Number of Images	Number of Subjects
Regular frontal image	0	1762	1010
Alternative frontal image	0	1518	1009
Quarter left	−22.5	763	508
Quarter right	+22.5	763	508
Half left	−67.5	1246	904
Half right	+67.5	1298	939
Profile left	−90	1318	974
Profile right	+90	1342	980

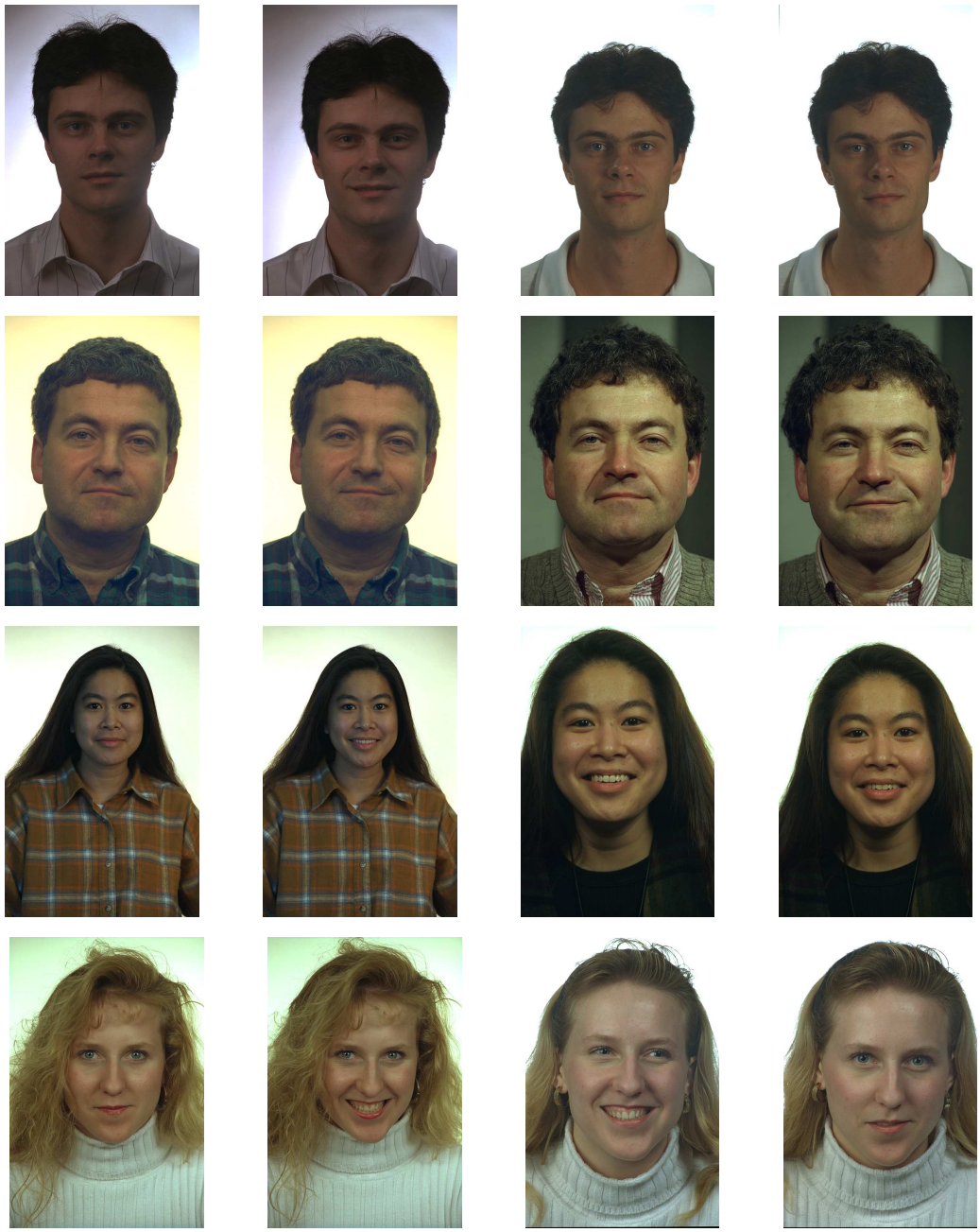


Figure A.2: Example of the regular and alternative frontal images from different sessions of several individuals provided with FERET database. The first two columns correspond to images of one session and the last two columns to images of another sessions.

A.3 PAL

The Productive Aging Laboratory (PAL) [45] face database was created at the University of Michigan in 2004. The database contains colour images of 575 individuals (219 males and 356 females) ranging from ages 18 to 93 and various races. One to three pictures were taken from each person. From all participants a frontal neutral face image was acquired. Table A.2 details the number of frontal images broken down by age group, race and gender. Additionally, some individuals were asked to show a happy face and to pose for a profile picture. The images were acquired in different locations such as college students unions, shopping centres and senior citizen festivals. In those places, the illumination conditions could not be controlled. Hence, it was decided that all images were taken under natural lightning. The images are provided with a resolution of 640×480 pixels. Some examples of face images from different age groups, races and gender are depicted in Figure A.3.

It is worth highlighting that this database has an unusually large variation in the demographic characteristics of the participants. It specially includes a high number of mature adults. In addition, it differs from most of the databases available in the field in the ratio of male to female faces (which is 0.6:1), containing more females than males.

Table A.2: Total number of frontal images in PAL database broken down by age group, race and gender.

Race	Age	Males	Females
African-American	18-29	14	29
	30-49	7	9
	50-69	3	12
	70-93	2	13
Caucasian	18-29	62	65
	30-49	22	38
	50-69	23	81
	70-93	46	97
Other	18-29	38	11
	30-49	0	0
	50-69	2	1
	70-93	0	0

(a) Images of faces from people of different races whose ages range from 18 to 29.

(b) Images of faces from people of different races whose ages range from 30 to 49.

(c) Images of faces from people of different races whose ages range from 50 to 69.

(d) Images of faces from people of different races whose ages range from 70 to 93.

Figure A.3: Example of the images provided with PAL database. The first two columns of each row are females and the last two males.

A.4 XM2VTS

The extended M2VTS Database (XM2VTS) [44] is a large multi-modal database collected in the Centre for Vision, Speech and Signal Processing at the University of Surrey in 1999. It consists of several datasets containing diverse types of data, such as audio, images and videos. The description is focused on the dataset of face image which is named "CDS001+CDS006". This dataset contains frontal views of faces corresponding to 295 subjects. The images, which have a resolution of 720×576 pixels, were extracted from a high quality digital video. The recording was repeated in four sessions within a four-month time frame. Concretely, two images were taken at the beginning of a head rotation shot and in the middle of it. As a result, eight images per person were taken considering that two images were taken in each of the four sessions, which makes a total of 2,360 images. Figure A.4 shows some examples of male and female faces from the four different sessions.

Figure A.4: Example of the images provided with XM2VTS database. The first two rows show female faces and the two last male faces. Each column depicts an image taken in a different session.

Bibliography

[1] OpenCV, Open Source Computer Vision Library.

[2] Ahonen, T., Hadid, A., and Pietikäinen, M. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence 28*, 12 (2006), 2037–2041.

[3] Alcalá-Fernández, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., and Herrera, F. KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic and Soft Computing 17*, 2-3 (2011), 255–287.

[4] Alexandre, L. A. Gender recognition: A multiscale decision fusion approach. *Pattern Recognition Letters 31*, 11 (2010), 1422–1427.

[5] Andreu, Y., García-Sevilla, P., and Mollineda, R. A. Dealing with inaccurate face detection for automatic gender recognition with partially occluded faces. In 14^{th} *Iberoamerican Congress on Pattern Recognition* (2009), vol. 5856 of *Lecture Notes in Computer Science*, pp. 749–757.

[6] Andreu, Y., García-Sevilla, P., and Mollineda, R. A. Gender classification from neutral and expressive. In 6^{th} *International Conference on Machine Vision* (2013), vol. 9067 of *SPIE Proceedings*, pp. 906723–26.

[7] Andreu, Y., García-Sevilla, P., and Mollineda, R. A. Face gender classification: A statistical study when neutral and distorted faces are combined for training and testing purposes. *Image and Vision Computing 32*, 1 (2014), 27–36.

[8] Andreu, Y., López-Centelles, J., Mollineda, R. A., and García-Sevilla, P. Analysis of the effect of image resolution on automatic face gender classification. In 22^{nd} *International Conference on Pattern Recognition* (2014).

[9] Andreu, Y., and Mollineda, R. A. On the complementarity of face parts for gender recognition. In 13^{th} *Iberoamerican Congress on Pattern Recognition* (2008), vol. 5197 of *Lecture Notes in Computer Science*, pp. 252–260.

[10] Andreu, Y., and Mollineda, R. A. The role of face parts in gender recognition. In *International Conference on Image Analysis and Recognition* (2008), vol. 5112 of *Lecture Notes in Computer Science*, pp. 945–954.

[11] Andreu, Y., Mollineda, R. A., and Garcia-Sevilla, P. Gender recognition from a partial view of the face using local feature vectors. In 4^{th} *Iberian Conference on Pattern Recognition and Image Analysis* (2009), vol. 5524 of *Lecture Notes in Computer Science*, pp. 481–488.

[12] Andreu, Y., Mollineda, R. A., and García-Sevilla, P. Assessing the effect of crossing databases on global and local approaches for face gender classification. In 15^{th} *International Conference on Computer Analysis of Images and Patterns* (2013), vol. 8047 of *Lecture Notes in Computer Science*, pp. 204–211.

[13] Bekios-Calfa, J., Buenaposada, J. M., and Baumela, L. Revisiting Linear Discriminant Techniques in Gender Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence 33*, 4 (2011), 858–864.

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

[15] Bellera, C. A., Julien, M., and Hanley, J. A. Normal approximations to the distributions of the wilcoxon statistics: Accurate to what n? graphical insights. *Journal of Statistics Education 18*, 2 (2010), 1–17.

[16] Bochkanov, S. ALGLIB, numerical analysis library.

[17] Bruce, V., Burton, A., Hanna, E., Healey, P., Mason, O., Coombes, A., Fright, R., and Linney, A. Sex discrimination: how do we tell the difference between male and female faces. *Perception 22*, 2 (1993), 131–152.

[18] Brunelli, R., and Poggio, T. HyberBF Networks for Gender Classification. In *Proceedings of the DARPA Image Understanding Workshop* (1992), pp. 311–314.

[19] Buchala, S., Davey, N., Frank, R. J., Loomes, M., and Gale, T. M. The role of global and feature based information in gender classification of faces: A comparison of human performance and computational models. *International Journal of Neural Systems 15*, 01n02 (2005), 121–128.

[20] Burton, A., Bruce, V., and Dench, N. What's the difference between men and women? evidence from facial measurement. *Perception 22*, 2 (1993), 153–176.

[21] Chang, C.-C., and Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology 2*, 3 (2011), 27:1–27:27.

[22] Chu, W.-S., Huang, C.-R., and Chen, C.-S. Identifying gender from unaligned facial images by set classification. In *Pattern Recognition (ICPR), 20th International Conference on* (2010), pp. 2636–2639.

[23] Cortes, C., and Vapnik, V. Support-vector networks. *Machine Learning 20* (1995), 273–297.

[24] Cottrell, G. W., and Metcalfe, J. EMPATH: Face, emotion, and gender recognition using holons. *Advances in Neural Information Processing Systems 3 (NIPS)* (1990), 564–571.

[25] Derrac, J., García, S., Molina, D., and Herrera, F. A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation 1*, 1 (2011), 3–18.

[26] Duin, R., Juszczak, P., Paclik, P., Pekalska, E., de Ridder, D., and D.M. Tax, J. *PRTools4, A Matlab Toolbox for Pattern Recognition*, 4.0 ed. Delft University of Technology, The Netherlands, 2004.

[27] El-Din, Y. S., Moustafa, M. N., and Mahdi, H. Gender classification using mixture of experts from low resolution facial images. *Multiple Classifier Systems, LNCS 7872* (2013), 49–60.

[28] Fasel, B., and Luettin, J. Automatic facial expression analysis: a survey. *Pattern Recognition 36*, 1 (2003), 259 – 275.

[29] Fisher, R. The statistical utilization of multiple measurements. *Annals of Eugenics 8* (1938), 376–386.

[30] Golomb, B. A., Lawrence, D. T., and Sejnowski, T. J. SEXNET: A neural network identifies sex from human faces. *Advances in Neural Information Processing Systems 3 (NIPS)* (1990), 572–579.

[31] Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6* (1979), 65–70.

[32] Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology 24*, 7 (1933), 498–520.

[33] Huang, C., Ai, H., Li, Y., and Lao, S. High-performance rotation invariant multiview face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence 29*, 4 (2007), 671–86.

[34] Iman, R., and Davenport, J. Approximations of the critical region of the friedman statistic. *Communications in Statistics 9* (1980), 571–595.

[35] Jabid, T., Kabir, M. H., and Chae, O. Gender classification using local directional pattern (ldp). In *Pattern Recognition (ICPR), 20th International Conference on* (2010), pp. 216–2165.

[36] Kawano, T., Kato, K., and Yamamoto, K. A comparison of the gender differentiation capability between facial parts. In *Pattern Recognition (ICPR), 17th International Conference on* (2004), vol. 1, pp. 350–353.

[37] Kumari, S., Sa, P. K., and Majhi, B. Gender classification by principal component analysis and support vector machine. *Proc Int Conf Communication, Computing and Security - ICCCS* (2011), 339–342.

[38] Kuncheva, L. I. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley Press, 2004.

[39] Mäkinen, E., and Raisamo, R. Evaluation of gender classification methods with automatically detected and aligned faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence 30*, 3 (2008), 541–547.

[40] Martinez, A. M. Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. *IEEE Transactions on Pattern Analysis and Machine Intelligence 24*, 6 (2002), 748–763.

[41] Martinez, A. M., and Benavente, R. The AR face database. Tech. rep., CVC Technical Report 24, June 1998.

[42] Mayo, M., and Zhang, E. Improving face gender classification by adding deliberately misaligned faces to the training data. In *Image and Vision Computing New Zealand (IVCNZ), 23rd International Conference on* (Nov 2008), pp. 1–5.

[43] Meissner, C. a., and Brigham, J. C. Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law 7*, 1 (2001), 3–35.

[44] Messer, K., Matas, J., Kittler, J., Lüttin, J., and Maitre, G. XM2VTSDB: The extended M2VTS database. In *Audio and Video-based Biometric Person Authentication, AVBPA'99* (1999), pp. 72–77.

[45] Minear, M., and Park, D. C. A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments & Computers 36*, 4 (2004), 630–633.

[46] Moghaddam, B., and Yang, M.-H. Learning gender with support faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence 24*, 5 (2002), 707–711.

[47] Mozaffari, S., Behravan, H., and Akbari, R. Gender Classification Using Single Frontal Image Per Person: Combination of Appearance and Geometric Based Features. In *Pattern Recognition (ICPR), 20th International Conference on* (2010), pp. 1192–1195.

[48] Naini, F. B., Moss, J. P., and Gill, D. S. The enigma of facial beauty: Esthetics, proportions, deformity, and controversy. *American Journal of Orthodontics and Dentofacial Orthopedics 130*, 3 (2006), 277–282.

[49] Nazir, M., Sciences, E., and Arabia, S. Multi-view gender classification using hybrid transformed features. *International Journal of Multimedia and Ubiquitous Engineering 7*, 2 (2012), 515–520.

[50] Oh, H., Lee, K., and Lee, S. Occlusion invariant face recognition using selective local non-negative matrix factorization basis images. *Image and Vision Computing 26*, 11 (2008), 1515–1523.

[51] Ojala, T., Pietikainen, M., and Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence 24*, 7 (2002), 971–987.

[52] Ojala, T., Pietikäinen, M., and Mäenpää, T. Gray scale and rotation invariant texture classification with local binary patterns. *European Conference on Computer Vision 1842* (2000), 404–420.

[53] O'Toole, A., Deffenbacher, K., Valentin, D., McKee, K., Huff, D., and Abdi, H. The perception of face gender: The role of stimulus structure in recognition and classification. *Memory & Cognition 26*, 1 (1998), 146–160.

[54] Parzen, E. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics 33*, 3, 1065–1076.

[55] Phillips, H., Moon, P., and Rizvi, S. The FERET evaluation methodology for face recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*, 10 (2000).

[56] Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset shift in machine learning*. The MIT Press, 2009.

[57] Rajagopalan, A. N., Rao, K., and Kumar, Y. Face recognition using multiple facial features. *Pattern Recognition Letters 28*, 3 (2007), 335–341.

[58] Shan, C. Learning local binary patterns for gender classification on real-world face images. *Pattern Recognition Letters 33*, 4 (2012), 431–437.

[59] Tamura, S., Kawai, H., and Mitsumoto, H. Male/female identification from 8 x 6 very low resolution face images by neural network. *Pattern Recognition 29*, 2 (1996), 331–335.

[60] Tapia, J. E., and Perez, C. A. Gender Classification based on Fusion of Different Spatial Scale Features Selected by Mutual Information from Histogram of LBP, Intensity and Shape. *IEEE Transactions on Information Forensics and Security 8*, 3 (2013), 488–498.

[61] Toews, M., and Arbel, T. Detection, Localization, and Sex Classification of Faces from Arbitrary Viewpoints and under Occlusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence 31*, 9 (2009), 1567–1581.

[62] Turkowski, K. Filters for common resampling tasks. *Graphics Gems I* (1990), 147–165.

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

[64] Wang, H., and Wu, X. Eigenblock approach for face recognition. *International Journal of Computational Intelligence Research 3*, 1 (2007), 72–77.

[65] Wang, H. L., Wang, H. L., Ye, M., and Yau, W.-Y. Real-time gender recognition with unaligned face images. In *Industrial Electronics and Applications (ICIEA), 5th IEEE Conference on* (2010), pp. 376–380.

[66] Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bulletin 1*, 6 (1945), 80–83.

[67] Wiskott, L., Fellous, J.-M., Kruger, N., and von der Malsburg, C. Face Recognition and Gender Determination. In *Proceedings of International Workshop on Automatic Face and Gesture Recognition* (1995), vol. 97, pp. 92–97.

[68] Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., and Ma, Y. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 31*, 2 (2009), 210–227.

[69] Xiao, R., Zhao, Q., Zhang, D., and Shi, P. Facial expression recognition on multiple manifolds. *Pattern Recognition 44*, 1 (2011), 107 – 116.

[70] Xu, Z., Lu, L., and Shi, P. A Hybrid Approach to Gender Classification from Face Images. In *Pattern Recognition (ICPR), 19th International Conference on* (2008), pp. 2–5.

[71] Yang, W., Chen, C., Ricanek, K., and Sun, C. Gender Classification via Global-Local Features Fusion. vol. 7098 of *Lecture Notes in Computer Science*. 2011, pp. 214–220.

[72] Zhao, W., and Chellappa, R. *Face Processing: Advanced Modeling and Methods.* Academic Press, 2006.