SOBRE LA DESCRIPCIÓN VISUAL AUTOMÁTICA DEL SER HUMANO PARA APLICACIONES DE INTERACCIÓN HOMBRE-MÁQUINA

Tesis Doctoral presentada por
DAVID S. FREIRE OBREGÓN

Bajo la supervisión del Doctor
Modesto F. Castrillón-Santana

Universidad de Las Palmas de Gran Canaria

Instituto Universitario en Sistemas Inteligentes y Aplicaciones Numéricas en Ingeniería

Programa de doctorado: Sistemas Inteligentes y Aplicaciones Numéricas en Ingeniería

Las Palmas de Gran Canaria - Septiembre de 2014

Agradecimientos

Hacer una tesis doctoral no es un trabajo, ni siquiera es un estudio sobre una temática, se trata de un proceso que transcurre en una etapa de nuestras vidas. Y como sucede en todos los procesos que ocurren a lo largo de nuestra vida, es una experiencia en la que aprendemos de otras personas y de nosotros mismos. No hace mucho, una persona me recordó una célebre cita de Newton; "si he logrado ver más lejos, ha sido porque he subido a hombros de gigantes", y me dí cuenta que "mis gigantes" me han mantenido en pie todo este tiempo, y a ellos quiero dedicar estos agradecimientos. Gigantes que me han levantado cuando me he caído, gigantes que me han impulsado cuando he dejado de girar y gigantes que me han enseñado el camino en los momentos más sombríos.

En primer lugar quiero agradecer a mi familia, a mis padres Bueno y Berta, y por supuesto a mi querida hermana Cristina. Ellos me han hecho la persona que soy hoy, han sido el pilar de mi vida y, todo lo que soy, se lo debo a ellos. Gracias a Cecilia, que llegó a tiempo para darme el suspiro de ánimo necesario, para superarme y todo ello con una sonrisa, siempre paciente, siempre aportando ideas, pero sobre todo aportando su infinito amor. También agradezco a mi cuñado Jorge por mandarme ánimos cada vez que ha tenido la oportunidad y con el que he compartido telemáticamente éxitos y frustraciones.

Sin embargo, este documento no existiría sin una persona que ha sabido llevarme de forma excelente en lo profesional y, sobre todo, me ha ayudado en lo personal. Me ha enseñado que nada se consigue en el mundo de la investigación sin honestidad, rigor y esfuerzo. Siempre me ha esperado con su puerta abierta, aportando nuevas ideas y nunca encontraré forma alguna de agradecerle que se haya volcado tanto conmigo. Por ello, quiero agradecer especialmente a mi director de tesis, el Dr. Modesto Castrillón por haber entendido el perfil de persona que soy, por su perseverancia, por su paciencia y, sobre todo, por su buen hacer.

No quisiera olvidar a nadie, especial mención al Dr. Luis Antón, a su consejo y a su ingenio, sin lugar a dudas, su creatividad no conoce fronteras y su aportación tiene un valor incalculable en esta tesis. También ofrezco mi gratitud al profesor Thomas B. Moeslund, quien ha desarrollado una nueva dimensión de lo que significa ser paciente desde que trabaja conmigo. Las aportaciones del profesor Quoc Vuong tienen también una valiosa contribución en esta tesis doctoral, le agradezco sus clases impartidas así como su consejo a lo largo de la

investigación realizada en el ámbito de las expresiones faciales.

Quiero agradecer a todos esos compañeros de la Universidad, del Departamento y del SIANI que alguna que otra vez han sabido recordarme que tenía una tesis pendiente de hacer y que me han aportado las vivencias vitales para ayudarme a madurar personalmente.

Por último deseo hacer mención de diversas entidades que han confiado en nosotros a la hora de financiar o colaborar en parte de los trabajos desarrollados en este documento y sus posibles ampliaciones futuras, concretamente a la Fundación Canaria Universitaria de Las Palmas (a través de las becas otorgadas dentro del marco del Programa INNOVA Canarias 2020) y al Ministerio de Ciencia e Innovación (Ref. TIN2008-06068) del Gobierno de España.

Resumen

Desde hace varias décadas, las máquinas forman parte de la vida cotidiana de las personas. Levantarse de la cama gracias a un despertador digital, encender la cafetera mientras consultamos nuestro correo electrónico a través de una PDA o lavarnos los dientes con un cepillo de dientes eléctrico escenifican la importancia que ha adquirido el papel que juegan las máquinas en nuestro entorno. Sin embargo, aún existe margen para mejorar la calidad de la comunicación entre humanos y máquinas. En los ejemplos anteriormente expuestos, sólo es posible el servicio de las máquinas cuando estas son esclavas desde una perspectiva comunicativa; sin jugar un papel proactivo ante las necesidades que pudiera tener un usuario. Por otro lado, la comunicación visual entre los seres humanos proporciona a la visión por computador la inspiración necesaria para abordar la problemática de la mejora de las interfaces perceptuales de usuario a través de modelos bioinspirados. Se trata de ayudar a que las máquinas comprendan mejor el entorno en el que se desenvuelven y sean capaces de ofrecer al ser humano asistencia en función de sus necesidades y no en la mera función de sus demandas. Esta tesis doctoral aborda el problema de la comunicación desde dos vertientes; la emocional y la contextual. En primer lugar se ha estudiado el origen de las emociones y se han articulado los mecanismos necesarios para inducir estas emociones en las personas. Una vez que se ha comprendido el origen de las emociones, se han estudiado diversas técnicas que permiten ampliar el estado del arte acerca de la detección de una de las expresiones faciales más universales; la sonrisa. Para ello, se han desarrollado diversos enfoques estáticos basados en la extracción de descriptores novedosos, además de un enfoque dinámico para estudiar la coherencia temporal en la detección de sonrisas. La segunda parte de este trabajo aborda la detección de componentes que ayuden a dilucidar características demográficas de los sujetos capturados. Para ello, se ha desarrollado un novedoso algoritmo basado en tríxeles (procedente del término anglosajón *"Triangular Superpixels"*) que permite generar una malla de tríxeles. Esta malla es interpretada como un grafo y segmentada a través de un algoritmo de flujo máximo para obtener una imagen segmentada. El algoritmo de corte empleado es similar al que se usa en el GrabCut, pero con una serie de ventajas muy provechosas. Por último, se aprovecha la información extraída de los tríxeles para determinar el sexo de una persona o detectar componentes contextuales de la indumentaria de la misma.

Índice general

Agradecimientos I

Resumen III

1 Introducción 1
- 1.1 La comunicación . . . 2
  - 1.1.1 La comunicación y los computadores . . . 3
- 1.2 Retos actuales en la interacción hombre-máquina . . . 5
- 1.3 Contribuciones de la memoria de tesis . . . 9
- 1.4 Estructura de la memoria de tesis . . . 10

2 Estado del arte 13
- 2.1 El rostro humano . . . 13
  - 2.1.1 Estructura facial . . . 15
  - 2.1.2 Interacción humana . . . 17
- 2.2 Bases de datos . . . 19
  - 2.2.1 Bases de datos existentes . . . 20
- 2.3 Reconocimiento de expresiones faciales . . . 29
  - 2.3.1 Tipos de expresiones faciales . . . 31
  - 2.3.2 Trabajos relevantes en el área . . . 35
- 2.4 Reconocimiento de características demográficas . . . 40
  - 2.4.1 Trabajos relevantes en el área . . . 42

3 Del origen a la detección de la expresión facial 53
- 3.1 Generación de expresiones faciales . . . 54
  - 3.1.1 Configuración experimental de Vedana . . . 55
  - 3.1.2 Base de datos Vedana . . . 65
- 3.2 Detección de expresiones faciales . . . 72
  - 3.2.1 Enfoques estáticos . . . 75
  - 3.2.2 Un enfoque dinámico . . . 102
  - 3.2.3 Discusión . . . 110

4 Extracción de descriptores demográficos fusionando información facial y de contexto 115
4.1 Antecedentes . . . 116
4.1.1 Superpíxeles . . . 116
4.1.2 GrabCut . . . 118
4.2 Creación de la malla de tríxeles . . . 121
4.2.1 Agrupamiento DMUM . . . 122
4.2.2 Topología de la malla . . . 124
4.2.3 Algoritmo de generación de la malla . . . 126
4.3 Segmentación de la malla . . . 131
4.3.1 Inicialización de la malla . . . 133
4.3.2 Algoritmo de maximización de flujo . . . 138
4.4 Análisis del rendimiento . . . 142
4.4.1 Indice de calidad (Qindex) . . . 150
4.4.2 Otro caso de estudio: Supertríxeles . . . 151
4.5 Evaluación del rendimiento de extracción de ropa y aplicaciones . 153
4.5.1 Configuración del experimento . . . 155
4.5.2 Modelo geométrico adoptado . . . 155
4.5.3 Representación de los datos . . . 157
4.5.4 Resultados . . . 158
4.5.5 Detección de componentes específicas en la indumentaria 161

5 Conclusiones 165
5.1 Principales aportaciones . . . 166
5.2 Ponencias y publicaciones . . . 169
5.3 Trabajo futuro . . . 170

Bibliografía 173

Índice de figuras

1.1 Evolución de la comunicación entre personas vs. evolución de la comunicación hombre-máquina. 5

2.1 Músculos de la mímica. 16
2.2 Escena con diferentes gestos y expresiones faciales 18
2.3 La base de datos FERET 22
2.4 Base de datos MIT/CMU 24
2.5 Paradigmas de inducción en la psicofisiología del estudio de las emociones 26
2.6 Base de datos DaFEx 28
2.7 Las seis expresiones faciales universales de las emociones 31
2.8 El sistema de codificación de acción facial 32
2.9 Movimiento facial desde las expresiones hasta el reconocimiento . 35
2.10 Enfoque geométrico para la detección de expresiones faciales . . . 37
2.11 Aplicación de filtros de Gabor para la detección de expresiones faciales 38
2.12 Complejidad de detección de sexo a través de los rasgos faciales . 41
2.13 Distancias de Fellous 44
2.14 Obtención tridimensional de datos faciales 45
2.15 La base de datos CAS-PEAL 46
2.16 Histogramas de gradientes orientados 49
2.17 Obtención de las características capilares según Yacoob 51
2.18 Mecanismo de extracción de propiedades del pelo 52

3.1 Esquema básico del estudio de grabación 56
3.2 Vista externa del estudio de grabación 56
3.3 Vista interna del estudio de grabación 57
3.4 Configuración para el proceso de captación de expresiones faciales 60
3.5 Configuración para el experimento de captación del torso 61
3.6 Percepción directa vs. percepción indirecta 62
3.7 Imágenes por resonancia magnética funcional 63
3.8 Composición de una imagen RISE 65
3.9 Participantes grabados desde las distintas cámaras 66
3.10 Distribución de edades en Vedana 67

3.11 Participantes de Vedana con gafas . . . 68
3.12 Sujetos de Vedana con distinto corte de pelo . . . 69
3.13 Grabaciones con capuchas en Vedana . . . 69
3.14 Diversidad de Vedana en cuanto al vello facial . . . 70
3.15 Imágenes del experimento de captación del torso . . . 70
3.16 Sonrisa auténtica vs. sonrisa fingida . . . 73
3.17 Versión básica del LBP . . . 75
3.18 Características del procesado LBP . . . 76
3.19 Versión simplificada del LBP . . . 77
3.20 Descomposición PCA de un rostro humano . . . 78
3.21 Clasificador lineal . . . 80
3.22 Problemas del clasificador lineal . . . 81
3.23 Ejemplo de transformación no lineal . . . 81
3.24 Aplicación de algoritmo k-NN . . . 83
3.25 Captación de imágenes faciales en archivos audiovisuales . . . 84
3.26 Ejemplos de imágenes normalizadas . . . 85
3.27 Comparativa de distintos enfoques para la detección de sonrisas . 93
3.28 Aplicación del detector de sonrisas combinado sobre Vedana . . . 94
3.29 Valor promedio de la sonrisa sobre Vedana . . . 95
3.30 Auto-similitud en imágenes . . . 96
3.31 Creación de la firma digital . . . 98
3.32 Descriptor de firma digital para la detección de sonrisas . . . 99
3.33 Exploración de posibles configuraciones de la firma digital para la detección de sonrisas . . . 101
3.34 Ciclo de vida de una expresión facial . . . 102
3.35 Derivadas temporales sobre acciones faciales . . . 104
3.36 Esquema del clasificador dinámico . . . 105
3.37 Distintas intensidades de emociones en la base de datos DaFEx . 106
3.38 Gráficas del clasificador temporal sobre secuencias de vídeo . . . 109

4.1 SLIC Superpixels . . . 117
4.2 Fundamentos del Graph Cut . . . 119
4.3 Relación entre el mínimo coste de corte y el flujo máximo . . . 120
4.4 Fases para crear TriToM . . . 122
4.5 Cálculo de DMUM . . . 123
4.6 Triangulación de Delaunay . . . 125
4.7 Diagrama de bloques del proceso para la segmentación de tríxeles 132
4.8 Selección geométrica del trimap . . . 134
4.9 Asociación entre tríxeles y gaussianas . . . 135
4.10 Distribución topológica de TriToM . . . 136
4.11 Arboles de búsqueda . . . 138
4.12 Aplicación de GrabCut . . . 143
4.13 Simplificación de procesado haciendo uso de tríxeles . . . 145
4.14 Comparación temporal entre píxeles y tríxeles . . . 146
4.15 Comparación temporal de la ejecución de las fases del algoritmo 147
4.16 Definición del trimap para extraer la ropa . . . 150

4.17 Distribución de la precisión según el Qindex 151
4.18 Segmentación de TriToM en base al algoritmo de superpíxels . . 152
4.19 Distribución de píxeles en función del kernel de los tríxeles 153
4.20 Generación de TriToM para segmentar piel y ropa 154
4.21 Imágenes piel y ropa segmentadas con TriToM 156
4.22 Problemática a la hora de seleccionar el trimap 156
4.23 Aplicación de LBP para el reconocimiento del sexo 157
4.24 Aplicación de HOG para el reconocimiento del sexo 158
4.25 Etiquetado manual de la colección de datos ClothesDB 161

Índice de tablas

2.1 Bases de Datos para el reconocimiento facial 21
2.2 Colecciones para la detección facial 27
2.3 Bases de Datos para el reconocimiento facial 30
2.4 Métodos para la detección de sexo centrados únicamente en el rostro humano 47
2.5 Métodos para la detección de sexo centrados únicamente en la ropa de los sujetos 48

3.1 Especificaciones técnicas del equipo de grabación 58
3.2 Categorías de las secuencias de vídeo 64
3.3 Esquema de elementos contextuales de Vedana 67
3.4 Comparación de Vedana con bases de datos existentes 71
3.5 Tasa de error con el clasificador k-NN considerando partes del rostro. 87
3.6 Tasa de fallos con el clasificador k-NN considerando todo el rostro. 88
3.7 Tasa de fallos con el clasificador SVM considerando partes del rostro. 89
3.8 Tasa de fallos con el clasificador SVM considerando todo el rostro 90
3.9 Tasa de fallos en la detección de sonrisas de la firma digital . . . 100
3.10 Características faciales temporales que son medibles 103
3.11 Tasa de fallos obtenida por cada enfoque sobre la base de datos DaFEx en función de la intensidad de la expresión 108
3.12 Tasa de fallos obtenida por el nuevo enfoque sobre la base de datos DaFEx en función de la intensidad de la expresión 110
3.13 Comparación de métodos para la detección de sonrisas 111

4.1 Creación de tríxeles 130
4.2 Asociación de T-enlace con tríxeles 137
4.3 Comparación entre el uso de píxeles o tríxeles 144
4.4 Análisis del proceso de segmentación de la base de datos Vedana 148
4.5 Comparación de resultados entre GrabCut y TriToM 149
4.6 Reconocimiento de sexo en la base de datos ClothesDB. 159
4.7 Segmentación de piel y ropa para el reconocimiento del sexo . . . 160
4.8 Comparativa para la detección de tipos de cuellos 162

4.9 Comparativa para la detección de corbatas 162

Capítulo 1

Introducción y motivación

A medida que la tecnología ha avanzado, diversas transformaciones sociales se han plasmado en la historia del ser humano. En este sentido, la tecnología digital ha sido el motor que ha impulsado profundos cambios sociales, económicos y políticos a lo largo de las últimas décadas de la humanidad. El voto electrónico, el comercio electrónico o el auge de las redes sociales, certifican la presencia de la componente tecnológica en los últimos avances que ha vivido la sociedad en cada una de las perspectivas que la definen (política, economía, social,..). Es más, la propia tecnología digital tiene un papel cada vez más activo en la vida cultural, por ejemplo, en la música, el cine o el arte digital. Toda esta evolución digital tiene una consecuencia directa en la mentalidad de las personas. Así, el cambio del soporte de la información del papel a la pantalla (de computadores, portátiles, *smartphones* o *tablets*), ha condicionado la forma en la que el ser humano accede a la información. En otras palabras, ha permitido ampliar, no sólo la cantidad de consumidores de información, sino que también ha propulsado la diversidad de contenidos para adaptarlos a las preferencias de estos consumidores.

La comunicación hombre-máquina no ha quedado exenta de este proceso temporal; por ejemplo, desde los primeros computadores cuya interfaz se gestionaba a través del uso de tarjetas perforadas hasta las actuales que hacen uso del teclado o el ratón. Sin embargo, este proceso no se ha estancado, sigue en continua evolución. En este sentido, podría afirmarse que estamos en la "Era de la interacción" y, por tanto, se está desarrollando la tecnología acorde a dicha era. Precisamente, la comunicación está íntimamente ligada al proceso de interacción social a través de la transmisión de información entre dos entidades (personas, máquinas, animales, etc.). Y se expresa a través de un nexo que está basado en los elementos esenciales de la comunicación humana; los sentidos. Así, en el campo de la interacción entre las personas y las máquinas, Abascal prevé dos tendencias fundamentales [1]; por un lado una tendencia evolutiva en la que se pretende avanzar en los sistemas de interacción actuales mejorando su usabilidad, su metodología, su diseño y, sobre todo, tratando de adaptarlas

al entorno industrial. Por otro lado, la tendencia más revolucionaria que trata de crear una nueva generación de interfaces más inteligente, móviles pero, sobre todo, menos visible al usuario.

Resulta especialmente interesante el segundo enfoque en el que se predice que el computador dejará de existir en un periodo de tiempo relativamente corto. Técnicamente no desaparecerá, sino que estará integrado en otros dispositivos de uso cotidiano y, por tanto, su presencia como entidad independiente se verá difuminada entre estos dispositivos. El propio Abascal indica que este cambio implica la desaparición de la interfaz de usuario "explícita" en beneficio de las interfaces "implícitas". Por otro lado, Weiser, el padre de la computación ubicua, indicó que estamos llegando a una fase de postmodernismo en el que el diseño de interfaces sigue dos líneas; la línea visible y atractiva que pretende convencer a los usuarios de la necesidad del uso del computador, así como de seducirlos. Y la línea invisible, que pretende hacer del computador algo tan incorporado, adaptable y natural que se pueda usar sin siquiera pensar en él [222].

Por tanto, estamos en un escenario en el que la comunicación entre humanos y máquinas nace de la propia adaptación de los canales que el ser humano utiliza para transmitir mensajes. En otras palabras, las nuevas interfaces implícitas, definidas por Abascal, deben ser más inteligentes y ser capaces de interactuar de forma proactiva y dinámica con otros elementos, ya sean humanos o no. De ello se trata esta tesis doctoral, de aportar capacidades que faciliten la comunicación entre hombres y máquinas.

1.1. La comunicación

La comunicación viene definida, desde una perspectiva formal, como el proceso de transmisión de información entre dos o más entidades (humanos, máquinas, etc.) a través de un canal en el que se envía una información codificada. Así pues, se identifica un emisor encargado de emitir el mensaje y, al menos, un receptor encargado de recibirlo. Por otro lado, el canal de comunicación es el medio a través del cual se transmite el mensaje entre emisor y receptor, y la codificación se define como el conjunto de símbolos y reglas utilizados para enviar dicha información.

Históricamente, la primera comunicación que se conoce es la que hace referencia a los signos y las señales. Resulta erróneo afirmar la existencia de símbolos a este nivel, los símbolos conllevan un significado semántico no desarrollado en esta primera etapa. Realmente, lo que ocurre en esta etapa es que los seres humanos establecen canales muy básicos para transmitirse hechos y, la forma de transmitirlos, es a través de respuestas instintivas [135]. Podría afirmarse que estamos en un escenario de acción-reacción en el que ser humano experimenta una sensación que le permite reaccionar de forma inmediata como respuesta a lo que ha percibido. Se hace uso del término "sensación" porque no todas las personas responden de igual manera ante los eventos que perciben. Por tanto, el comportamiento comunicativo a este nivel es mínimo, siendo predominante la, ya mencionada, respuesta instintiva [149].

Estas respuestas instintivas se regían, en primer lugar por gestos y, en un segundo momento a través de los sonidos emitidos haciendo uso de las cuerdas vocales. Con el paso del tiempo y, sobre todo, gracias a la evolución del cerebro se empezaron a vislumbrar los primeros atisbos de comunicación oral. Este fenómeno fue crucial, ya que la voz facilita la transmisión de conocimientos y de información a las generaciones sucesivas. Comenzaban así las sociedades donde la comunicación oral permitía la pervivencia de la información a través de las distintas generaciones [149]. En paralelo con el desarrollo vocal comenzó el arte rupestre. Así, de la misma forma en la que los niños comienzan a dibujar antes de tener formación alguna sobre figuras geométricas, el homo sapiens hizo su primer intento en transmitir la información a través del tiempo dibujando un esbozo de aquello con lo que convivía a diario. Fue la primera evidencia de que estos seres comenzaban a desarrollar la conciencia sobre el concepto del tiempo.

Posteriormente llega la escritura, que es divisible en tres categorías: logográfica (pictoramas e ideogramas), silábica y alfabética. Las tres categorías se pueden encontrar, normalmente, en cualquier sistema de escritura [177]. Desde la escritura cuneiforme de los sumerios a la escritura actual, pasando por los jeroglíficos egipcios, este arte ha experimentado un notable avance en su forma y en su fondo. El alfabeto, se desarrolló como consecuencia de la extensión de la escritura por los distintos puntos de la geografía. En ese momento de la historia también se desarrollaron eventos que hoy perviven, por ejemplo, los griegos desarrollaron las 5 vocales actuales para adaptar el nuevo alfabeto a su lengua, dando origen al primer alfabeto escrito de izquierda a derecha.

1.1.1. La comunicación y los computadores

El desarrollo de los computadores suele estar ligado a cinco generaciones [33]. En cada generación la máquina se ha vuelto más sofisticada, más eficiente, más pequeña y más barata. Sin lugar a dudas, uno de los factores esenciales que ha permitido la comercialización de estas máquinas y su integración en la vida cotidiana del ser humano ha sido la mejora de los canales de comunicación. Curiosamente, la evolución de los computadores trajo consigo la mejora de los mecanismos de comunicación con las mismas. Así, la primera generación de computadores (1938-1958) eran máquinas que ocupaban habitaciones enteras, usaban tubos de vacío para procesar la información y la forma de comunicación con ellas era a través del cableado de un tablero, o bien, a través de tarjetas perforadas. Este tipo de tarjetas eran unas láminas de cartulina conteniendo la información en forma de perforaciones siguiendo un código binario. Se trataba de la manera existente en aquella época para introducir el programa a ejecutar en la máquina, de lo que se deduce que no existía interacción alguna, en tiempo real, entre máquina y humano. Por otro lado, la forma de comunicarse no sólo fue externa, la comunicación interna con el computador fue desarrollándose y en la segunda generación (1938-1958) nuevos lenguajes de programación como COBOL y FORTRAN emergieron siendo comercialmente accesibles.

Las generaciones sucesivas permitieron la aparición del teclado y, posteriormente, del ratón. El teclado surgió con el declive de las tarjetas perforadas.

Con el teclado y la pantalla, nació la posibilidad de establecer una comunicación escrita en tiempo real, donde el canal de comunicación para la interacción entre el humano (emisor) y la máquina (receptor), lo constituyó el teclado y el canal de comunicación entre la máquina (emisor) y el humano (receptor) corrió por parte de la pantalla. Precisamente, durante la tercera generación de computadores (1964-1970), se comenzaron a utilizar los circuitos integrados, lo que permitió abaratar los costes, al tiempo que se aumentaba la capacidad de procesamiento y se reducía el tamaño de las máquinas. La capacidad de estas máquinas aumentaba de acuerdo con la demanda creciente de prestaciones y, en la cuarta generación (1971-1983), se mejoró el diseño de estas máquinas a través del desarrollo de los microprocesadores. El objetivo fue aprovechar el desarrollo de los circuitos VLSI (*Very Large Integration Ciruit*) para ofrecer computadores multitarea. Con este tipo de computadores, la interfaz de comunicación entre humanos y máquinas mejoró, dotando al ser humano de una percepción de realismo a la hora de comunicarse con la máquina en tiempo real. El problema de esta época radica en que estos equipos informáticos no tienen la capacidad de "decidir" por sí mismos, siguen siendo totalmente dependientes de las instrucciones dadas por los usuarios.

La antesala a la época actual lo pone la quinta generación (1984-1999) en la que IBM propone su computador comercial y revoluciona por completo el sector informático. Este salto evolutivo supuso llevar la informática a cada hogar y, por tanto, el surgimiento de desarrolladores en cada lugar, que pueden ser definidos en otras palabras como generadores potenciales de nuevas demandas comunicativas con la máquina y precursores de mejoras en los protocolos de comunicación con la máquina de la época. El software se optimiza y las principales empresas informáticas empiezan a competir por la capacidad de comunicarse con la computadora, en un lenguaje más cotidiano y no a través de códigos o lenguajes de programación. En esta época se desarrollan supercomputadores de la misma forma que disciplinas como la inteligencia artificial, la robótica o las redes de comunicaciones. Todas estas disciplinas sufren un gran avance debido a la accesibilidad de la información a nivel global. Precisamente, la sexta generación, que llega hasta nuestros tiempos, es la que ha recogido el avance tan vertiginoso de esos campos y ello ha quedado plasmado desde los ordenadores de mesa, hasta los pórtatiles, los smartphones y los dispositivos tipo *Google Glasses*. Destacan como factores clave de esta época el procesamiento vectorial y las redes de computadores. Se mejoran los dispositivos que usan pantalla táctil como las tablets. La pantalla táctil es una forma intuitiva que integra las mismas funcionalidades que el ratón y teclado en los dispositivos móviles de alta generación. Si bien es un invento que data del año 1971, no es hasta la época actual en la que ha tenido un mayor auge. Además, los smartphones de última generación incluyen reconocedores de voz muy robustos que permiten a los usuarios comunicarse con la máquina a través de la voz. ¿Hacia dónde nos encaminamos? El desarrollo de aplicaciones de última generación apunta a un camino bien definido marcado por la adaptación más creciente de la máquina hacia los canales de comunicación que utiliza el ser humano en su día a día.

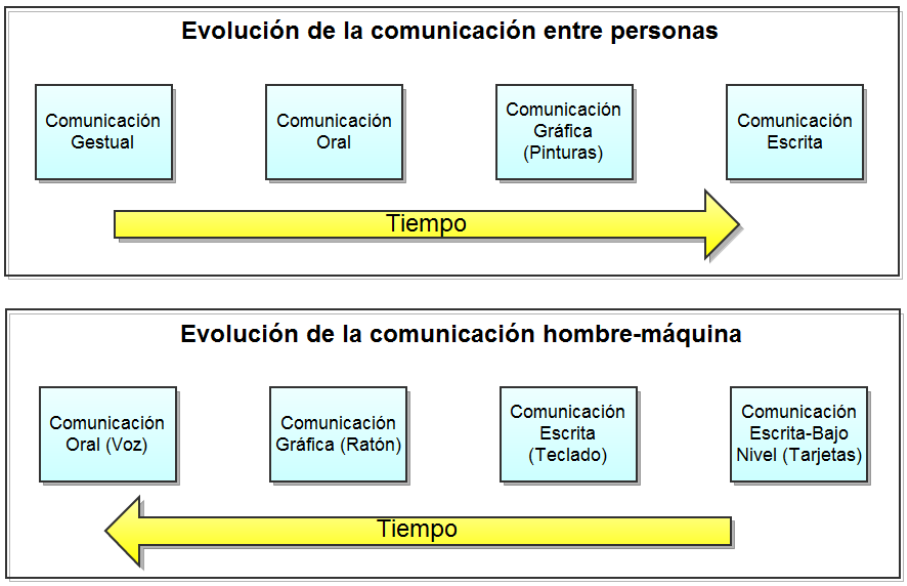


Figura 1.1: Evolución de la comunicación entre personas vs. evolución de la comunicación hombre-máquina.

1.2. Retos actuales en la interacción hombre-máquina

En la Figura 1.1 puede observarse una regresión natural de la propia evolución comunicativa del ser humano entre sí a la hora de comunicarse con las máquinas. Como se ha mencionado anteriormente, el ser humano comenzó a comunicarse de forma instintiva a través de gestos, para luego desarrollar la voz, la comunicación gráfica y, finalmente, la escritura. Con los equipos informáticos el fenómeno ha sido el inverso; empezando con las tarjetas perforadas (considerada comunicación escrita de bajo nivel por dos motivos: se comunican siguiendo un alfabeto (código binario) y, aunque sea perforada para que pueda ser leída por un sensor, es una información que plasma el ser humano sobre una superficie en forma de lámina). Tras las tarjetas perforadas llegó el teclado (comunicación escrita), el ratón (comunicación gráfica), hasta llegar a los potentes reconocedores de voz actuales (comunicación oral). El uso del ratón es considerado como una comunicación gráfica y no gestual debido a que, de la misma forma que se hace con la pintura, el transmisor final de la información es un instrumento (en este caso, el ratón) y no la mera acción humana a través de sus gestos. Por ello, si bien falta la comunicación gestual en ese esquema dentro de la comunicación hombre-máquina, a continuación veremos que es una tecnología con la que ya se está trabajando pero que no ha sido incluida porque su uso no está aún extendido.

Por tanto, tal y como se ha comentado anteriormente, la interacción entre los humanos y las máquinas puede ir enfocada hacia dos vertientes diferenciadas en el camino pero unidas en el fin. La primera vertiente es una evolución de los sistemas de interacción actuales, ser capaces de mejorar su usabilidad y adaptabilidad a través del desarrollo de mejores metodologías y herramientas. La segunda vertiente está encaminada hacia una revolución de nuevas interfaces más inteligentes, móviles y transparentes a los usuarios. Se trata de integrar las máquinas dentro de la vida cotidiana de las personas de la forma menos invasiva posible para su interacción pero con todos los beneficios de su uso [1]. Precisamente, el término usabilidad cobra una especial relevancia en lo referente a la interfaz dentro de esta vertiente pues actualmente se tiene un conocimiento amplio desde la perspectiva del usuario, lo que permitirá aplicar el desarrollo de nuevas ideas que faciliten el manejo de los dispositivos. Según la Figura 1.1, podemos apreciar un hecho bastante interesante en la parte referente a la evolución de la comunicación hombre-máquina; el salto cualitativo sucede cuando se introducen nuevos dispositivos capaces de captar una nueva dimensión de la condición humana. Esto es, para que exista un avance significativo en la forma de interacción entre personas y máquinas, es muy importante introducir nuevos estilos de interacción, incluyendo nuevos dispositivos de entrada/salida. Hace años la previsión iba enfocada hacia el desarrollo de la realidad virtual, sin embargo ahora, debido a la explosión de los dispositivos móviles y smartphones esta tendencia se ha orientado más hacia la computación móvil, computación afectiva, procesado del lenguaje natural, etc.

Volviendo a la propuesta de Abascal, la idea fundamental es que los equipos informáticos de mesa estén incorporados en dispositivos cotidianos ofreciendo funcionalidades pero cuya interfaz sea "invisible" al ser humano. Con esto, desaparecería la interfaz tal y como la conocemos hoy en día, pero permanecería de forma implícita enfocada a los servicios para los que ha sido desarrollada y estableciendo comunicaciones con otras interfaces de otros dispositivos con el fin de poder acometer su propósito, todo ello de forma invisible al usuario. Por otro lado, anteriormente se comentaba que la historia de los ordenadores venía evolucionando desde los computadores de tiempo compartido (un computador para muchas personas) hasta los ordenadores personales (un computador para cada persona). Si seguimos esta tendencia y unimos el desarrollo de las redes ubicuas, el futuro más cercano apunta a que, cada usuario pueda tener acceso a múltiples computadores en tiempo real. La idea fundamental es romper con el concepto de computador de escritorio permitiendo una computación dependiente del contexto en todo momento. Este hecho daría lugar a la posibilidad del desarrollo de entorno activos con el usuario (algo parecido a lo que ocurre en el campo de la domótica) en los que la interacción hombre-máquina se produce entre ambas entidades y entre los distintos ordenadores, todo de forma no invasiva, para favorecer la comodidad y satisfacer las necesidades de los humanos. En este caso, por ejemplo, podríamos hacer referencia a un hogar, un vehículo o, incluso, una tienda.

En segundo lugar, si bien los smartphones suponen un salto cualitativo en los dispositivos móviles, nuevos productos emergentes serán la confirmación de

una nueva tendencia de dispositivos que "vestirán" a los humanos en las próximas décadas. Se trata de desarrollar nuevos dispositivos que incorporen toda aquella tecnología que le resulta útil al ser humano en sus tareas diarias (agenda, realidad aumentada, acceso a comunicaciones, etc). La miniaturización de la tecnología, la alta duración de las baterías, la enorme cantidad de servicios ubicuos y el coste asequible de este tipo de dispositivos son los factores que hacen viable esta tecnología. La forma de interacción puede ser definida de múltiples formas; por voz, orientada a través del movimiento de los ojos o, incluso, de forma gestual. La salida de la información que ofrecen estos dispositivos sería a través de una pequeña pantalla o, bien, a través de unos auriculares. La idea es preservar la confidencialidad de la comunicación entre el humano y su máquina pero también garantizar su seguridad permitiendo al humano ser consciente de todo lo que sucede a su alrededor. La tecnología de estos dispositivos incluirá todas las funcionalidades que existen en los smartphones actuales: GPS, comunicaciones, videocámara, etc.

La inteligencia ambiental plantea un reto bastante interesante dentro del campo de la interfaz entre humanos y máquinas. Se define como un entorno sensible que reconoce y responde a la presencia de los seres humanos de forma invisible y anticipándose a sus necesidades. De hecho, son dispositivos embebidos en los objetos cotidianos y están conectados entre sí. La idea básica es la de dotar a las personas de todo aquello que necesitan en su vida diaria de forma natural haciendo uso de la información y la inteligencia que subyace en las redes que conectan estos dispositivos [32]. Precisamente, este tipo de tecnología está basada en los siguientes fundamentos; está embebida, entiende el contexto, está personalizada, se adapta y se anticipa a las necesidades. Por ejemplo, imagínese que usted tiene en su agenda una reunión de trabajo programada para el martes por la mañana. Se levanta el martes, pasa por delante del espejo con cámara incorporada y el sistema se percata de que tiene un poco de barba. Posteriormente, le muestra a través de su espejo un aviso para recordarle que debe afeitarse. Mientras lo hace, el sistema calcula la ruta más eficiente a su reunión y se lo hace saber, además, en función del clima, le asesora sobre qué indumentaria debe ponerse. En definitiva, se trata de analizar la actividad (ir a la reunión) y su estado actual (aspecto, tono de voz, etc.), prever posibles acciones (afeitarse, tomar medicamentos) o acciones seguras (vestirse, desayunar) y calcular todas las posibles condiciones (tráfico, clima,...) que faciliten el proceso en la toma de decisiones. Esta tecnología, aunque asombrosa desde un punto de vista funcional, no deja de tener ciertos críticos. Así, Wright alerta sobre la previsible falta de privacidad en tanto en cuanto una tercera empresa tenga control sobre la instalación domótica del hogar [227]. Sin embargo también precisa que esta privacidad puede estar sujeta a acuerdos restrictivos que la salvaguarden de las personas.

Otro aspecto muy interesante en el desarrollo de nuevas interfaces nace de la necesidad de explorar nuevas formas de interacción. En otras palabras, no ceñirse a lo que normalmente interpretamos, sino intentar dar un paso más allá en la comprensión del comportamiento de los humanos. En este sentido han surgido iniciativas bastante interesantes. Por ejemplo, en el campo de la compu-

tación afectiva, muchos trabajos recogen las emociones de sujetos a partir de las expresiones faciales [165, 168, 207] o bien, a partir de los distintos parámetros que ofrece la voz [186, 218]. Realmente, casi todos estos trabajos acerca del rostro, o de la voz, se fundamentan en que estos elementos son los que más utiliza el ser humano para identificar las emociones, de ahí a que se exploren enfoques adaptados a estos canales comunicativos con una mayor intensidad. Sin embargo, actualmente, también se están dando pasos en otras direcciones como, por ejemplo, la exploración del estado emocional de las personas a partir de su lenguaje corporal [119] o bien, a través de la respuesta cerebral ante estímulos o circunstancias externas [43]. En el primer caso, se pretenden recoger las emociones a través del lenguaje corporal con el objetivo de dotar a los robots con estos gestos tan característicos de los humanos. En otras palabras, se trata de humanizar a los robots a través de movimientos expresivos y, para ello, antes es necesario extraer el comportamiento expresivo que subyace en el lenguaje corporal de las personas [119]. No tiene porqué tratarse únicamente de humanizar robots bípedos, también pueden ser robots de otra índole los que muestren este tipo de actitudes, como por ejemplo el robot Aibo de Sony [92]. En esta línea, existen trabajos recientes en la literatura que recogen la forma en la que se interpretan las poses corporales así como su movimiento, y se discuten los principales retos que afectan a este tipo de información incluyendo diferencias entre individuos, el impacto cultural o el reconocimiento multimodal [121]. Por otro lado, otros investigadores también trabajan en el desarrollo de interfaces orientadas a detectar las emociones desde las respuestas de las ondas cerebrales, lo que se ha denominado como aBCI (del inglés, *"affective brain-computer interfaces"*) [19]. Se trata de hacer uso de la información neuronal ante la respuesta a un determinado estímulo para detectar el estado afectivo del sujeto. Según Bos, las respuestas emocionales ante eventos significativos pueden, ya sea en solitario o combinadas con otras fuentes de información afectiva, proporcionar la información necesaria para reconocer el estado emocional de una persona de forma continua, así como clarificar el contexto en el que ocurren las emociones y ayudar a guiar las interfaces hombre-máquina a que sean susceptibles de las emociones de sus usuarios (por ejemplo, la aparición de un asistente ante un estado de frustración durante el manejo de un determinado dispositivo) [19].

Por todo lo comentado anteriormente, se desprende que la interacción entre humanos y máquinas se estudia desde una triple vertiente [18]; 1) el hardware y software que afecta a la interacción, 2) la percepción de los usuarios frente a la percepción de la máquina y 3) las tareas que desempeña el sistema para adaptarse a las necesidades del usuario. Sin embargo, lo más importante es lo que afirma Booth cuando sugiere que los diseños deben estar dirigidos al usuario y no a la máquina [18]. Todas las tendencias anteriormente comentadas, enfocan sus esfuerzos considerando al usuario como el centro de toda actividad. En el escenario que nos encontraremos en los próximos años resulta vital el desarrollo e integración de nuevas formas de interacción que favorezcan la distribución de servicios a los humanos y que tengan un carácter lo menos invasivo posible, para lo cual es importante dotar a las máquinas con ciertas habilidades que tienen las personas; ser capaces de describir las personas con las que interactúa en

términos no sólo de identidad, sino también de emociones, estados de ánimo, identificar características demográficas de las personas (sexo, edad, etc.), así como identificar la vestimenta que utilizan. El fin de todo este proceso debe ser facilitar a la máquina toda la información necesaria para que pueda tomar una decisión sobre una determinada circunstancia y, de esta forma, proporcionar un servicio adecuado en función de dicha decisión.

1.3. Contribuciones de la memoria de tesis

Esta tesis doctoral está enfocada hacia la comunicación entre hombres y máquinas, más específicamente, en estudiar distintas posibilidades para dotar a las máquinas de la información necesaria para facilitar la toma de decisiones por parte de la máquina. Con este fin, se abordan dos líneas maestras que responden a los objetivos específicos:

En la primera parte de esta tesis doctoral se trabaja con la computación afectiva desde la detección de una emoción en particular hasta proponer un esquema para inducir emociones en las personas. El objetivo nace de la necesidad de entender cómo se producen las emociones y, para ello, la mejor manera de conseguirlo es estudiando la inducción de estas emociones en los humanos, así como hacer un estudio para la detección de alguna de ellas a través del uso de técnicas utilizadas con asiduidad en el mundo de la visión por computador. En esta parte del trabajo se ha generado una base de datos propia en la que se recogen las expresiones faciales de sujetos a través de un proceso de inducción visual y un sistema de grabación múltiple que incluye desde una webcam hasta una cámara de grabación de alta velocidad que permitiría el estudio y el análisis de las microexpresiones generadas por los sujetos. Posteriormente, siguiendo la línea argumental de la computación afectiva, se realiza el análisis de técnicas recientes en el mundo de la visión por computador para la detección de un tipo de expresión facial; la sonrisa. Para ello se utilizan técnicas basadas en los patrones binarios locales o el análisis de componentes principales. Además se ha diseñado un sistema de autosimilitud basado en la comparación de las distintas regiones de un área de interés para la detección de expresiones faciales que se ha denominado "firma digital". Por último, en esta primera parte, se ofrece un interesante enfoque acerca de la detección dinámica de las sonrisas considerando, además de las características anteriormente mencionadas, la respuesta temporal de los clasificadores.

La segunda parte de esta tesis doctoral está relacionada con la elaboración de una técnica de segmentación que contribuye a aportar un avance significativo en el análisis de componentes estructurales. Se trata de una técnica alternativa a los, tan difundidos, *superpíxeles* pero con un conjunto de variaciones que permiten la adaptabilidad automática de la técnica (sin la necesidad de especificar parámetros externos), lo cual es interesante para el funcionamiento autónomo de la máquina. Para ello se divide la imagen en un conjunto de regiones que permiten, a través de la aplicación de un algoritmo de flujo máximo, discernir las regiones pertenecientes a los distintos objetos de la escena. La principal ven-

taja que aporta la nueva propuesta sobre los superpíxeles es que se propone un grafo conexo en el que la información de los elementos del mismo, así como los vecinos de dichos elementos, aportan una información esencial y de forma rápida durante el proceso de segmentación. La técnica propuesta para la segmentación se aplica dentro del escenario interacción hombre-máquina y, previa detección del rostro, se aplica un proceso de segmentación de la imagen. Posteriormente, se extrae la información descriptiva de los sujetos a partir de estas regiones segmentadas con el fin de combinar la información visual con la contextual. Por último, se trabaja con la extracción de información demográfica a partir de los componentes estructurales anteriormente segmentados. De esto trata esta parte del trabajo; dotar a un sistema de la posibilidad de, a partir de una imagen en cualquier escenario posible, ser capaz de extraer características del aspecto e indumentaria del sujeto con el fin de obtener información demográfica, en nuestro caso nos centraremos, tanto en el sexo de las personas como en el uso de otros descriptores para poder extraer información de su indumentaria y, de esta forma, poder obtener alguna conclusión relevante acerca del individuo.

1.4. Estructura de la memoria de tesis

A razón de lo comentado anteriormente sobre las distintas líneas de actuación, esta tesis doctoral sigue el esquema, por capítulos, que se describe a continuación:

- **Capítulo 2:** El estado del arte pretende introducir al lector en los aspectos más relevantes de esta tesis doctoral incluyendo una revisión de la literatura existente acerca del rostro humano y de las técnicas descriptivas más relevantes a la hora de contextualizar al individuo. El rostro humano pero, sobre todo, su estructura y sus expresiones son los elementos centrales en la primera parte de la tesis. El motivo es que es necesario entender la estructura facial para conocer las distintas emociones que pueden extraerse del rostro, del mismo modo que es interesante conocer la forma en la que interaccionan los distintos músculos faciales para poder comprender la génesis de las expresiones faciales. Después del estudio del rostro, el lector se adentrará en la revisión de las bases de datos dentro del área de las expresiones faciales. Esta parte del trabajo resulta de especial relevancia porque se incide en la complejidad que entraña provocar una emoción que se exteriorice a través de una expresión facial. La mayoría de las bases de datos existentes dentro de esta temática están orientadas a recoger la mera simulación de las expresiones por parte de sujetos en lugar de recoger expresiones genuinas para un procesado óptimo. Una vez que se ha entendido cómo se generan las expresiones faciales, en el siguiente punto del trabajo se enumeran aquellas que son consideradas como básicas en la literatura, además de hacer un estudio exhaustivo sobre los trabajos más relevantes en el área del reconocimiento de las expresiones faciales. Por último, dentro del capítulo se hace un análisis de los trabajos más relevantes

a la hora de detectar una característica de biometría blanda, en concreto el sexo de las personas, en base a dos criterios bien diferenciados; por un lado el reconocimiento demográfico basado únicamente en el rostro y, por otro lado, se hace un estudio similar pero, en esta ocasión, considerando también aquellos elementos de contexto que rodean a los seres humanos, como puede ser la ropa o el pelo.

- **Capítulo 3:** El tercer capítulo de esta tesis doctoral está orientado, en exclusiva, a las expresiones faciales desde una doble vertiente que va desde una propuesta para la inducción de expresiones faciales hasta la detección de las mismas. En este apartado se presenta Vedana, que es una base de datos formada a partir de expresiones faciales de un número limitado de sujetos y, todo ello, a través de un método de inducción emocional mediante la proyección de material multimedia. La segunda parte de este capítulo recoge un estudio extenso sobre técnicas utilizadas para la detección de sonrisas.

- **Capítulo 4:** Se propone el desarrollo de un algoritmo genérico que usa convergencia de *clusters* para la identificación de regiones en la escena. Está demostrado que el procesado de regiones mayores que los píxeles permite una correcta discriminación entre los distintos objetos que pueden componer una escena. En este capítulo se ha hecho uso de un algoritmo para la detección de mínimos locales con el fin de generar una malla de triángulos a lo largo de toda la escena. Cada uno de estos triángulos conforma la nueva unidad elemental de información dentro de la escena y, de forma similar a otras técnicas ya conocidas como los superpíxeles, se trabaja para la convergencia de regiones que ofrecen características de información similares. Un aspecto muy interesante acerca de este enfoque de segmentación es que la malla de triángulos ofrece la posibilidad de hacer dos aproximaciones bien diferenciadas a la hora de unir las distintas regiones que conforman la escena; por un lado se hace un estudio de las características internas de los triángulos (color, textura, etc.) y, por otro lado, la malla de triángulos guarda características geométricas que permiten su uso en forma de grafo con el objetivo de identificar los grupos de triángulos que pertenecen a distintas regiones en la imagen. En la segunda parte del cuarto capítulo se ofrece un conjunto de aplicaciones de la técnica desarrollada. Por ello, en esta parte del capítulo se esboza un algoritmo que se divide en tres partes fundamentales para poder extraer al sujeto de un entorno no controlado; en primer lugar se utiliza un enfoque geométrico adaptativo para poder determinar la potencial zona de interés con el fin de marcar un conjunto de datos que servirá como entrada a la segunda parte del algoritmo. En la segunda parte, se aplica el modelo de triángulos descrito anteriormente con el fin de extraer la figura de la escena y así poder discriminar aquellas partes de la imagen en las que estamos interesados. Precisamente, estas partes son utilizadas a la hora de determinar el sexo de los sujetos en combinación con la información facial..

- **Capítulo 5:** Ofrece las conclusiones sobre el trabajo realizado a lo largo de esta tesis doctoral además de añadir pinceladas sobre posibilidades del trabajo futuro.

Capítulo 2

Estado del arte

En las últimas dos décadas, los periféricos más utilizados para interactuar con los computadores han sido el ratón y el teclado. Estos dos dispositivos han tenido un éxito bastante notable, en tanto en cuanto, la sociedad digital se ha desarrollado bajo su periodo hegemónico que, incluso, llega hasta nuestros días. Sin embargo, resulta visible que, tanto el ratón como el teclado, tienen un conjunto de limitaciones comunicativas. Por ejemplo, no permiten la comunicación de los usuarios haciendo uso de un lenguaje natural, es decir, mediante canales verbales y no verbales, mediante el habla o mediante la vista.

La potencia de los computadores actuales hace posible procesar tanto imágenes como datos relativos a la voz. Además, si a este factor se le une el asequible precio de los sensores para los equipos informáticos, surge un escenario lo suficientemente interesante que permite considerar la interacción de estas máquinas a través de elementos visuales, como por ejemplo, el rostro humano. De esta forma, en este escenario, el mayor reto se centra en cómo crear interfaces perceptuales que permitan su uso en múltiples tipos de aplicaciones. Así, el objetivo de este capítulo está centrado en mostrar el estado del arte en algunos de los campos más atractivos de las interfaces perceptuales; desde la detección de expresiones humanas, el análisis de las mismas, así como el reconocimiento de características demográficas, en concreto el sexo, haciendo uso de técnicas de biometría adaptadas al escenario de la biometría blanda.

2.1. El rostro humano

El rostro humano es capaz de proporcionar una gran variedad de señales que pueden ser detectadas e interpretadas desde una perspectiva social. Por ejemplo, un rostro puede indicar si la persona es hombre o mujer, joven o anciana, si está feliz o triste, si la estamos aburriendo o si siente curiosidad por lo que le estamos contando, si se siente atraída por nosotros o si, por el contrario, se siente repelida ante nuestra presencia.

Muchas especies animales poseen características distintivas en sus rostros,

estas características pretenden jugar un rol desde un punto de vista social o bien, desde un punto de vista sexual [26]; como los cuernos de los ciervos o rinocerontes. Tanto los humanos como algunas especias de primates pueden poseer algunas características bastantes distintivas. Por ejemplo, la ausencia de pelo en los rostros de los mandriles ha dado como resultado una extraordinaria evolución del rostro del animal mostrando una inherente coloración facial. Según [26], esto se debe a que el rostro del mandril recrea la forma de sus genitales, quedando la región roja como la zona del pene y el área azul como el escroto. La autora de esta obra va un paso más allá, y afirma que la característica del rostro de los mandriles sirve como elemento amenazador a otros machos de la camada.

En este sentido, el ser humano no está exento de esta evolución y presenta singularidades en su rostro; el rostro de un hombre tiene vello mientras que el de una mujer carece de él salvo en las cejas y el cabello. El pelo tiene como objetivo evitar la pérdida de calor del cuerpo, tanto en animales como en humanos. Sin embargo, las cejas siempre se han asociado con un mecanismo de defensa para evitar que el sudor penetrara en los ojos. Ekman apunta que las cejas tienen, además, un papel muy importante en la forma de mostrar las expresiones faciales, de hecho enfatiza dichas expresiones, y por ello tiene un papel muy activo durante el proceso comunicativo [73].

Por otro lado está la relevancia que tiene el rostro cuando se trata de distinguir el sexo de las personas. Realmente, el ser humano es bastante preciso distinguiendo el sexo de las personas, simplemente, en base a su rostro. En un experimento ocultando el pelo, con hombres afeitados y personas sin artificios de maquillaje, el 95 % de las personas que hicieron la prueba acertaron si el sujeto mostrado era un hombre o una mujer [26]. Descubrir las componentes faciales que permiten discernir el sexo de una persona es un trabajo que, en primer lugar, hace necesario encontrar las variables a analizar. De esta forma, Burton hizo un experimento en el que, a partir del rostro de 91 hombres jóvenes y 88 mujeres jóvenes, analizó un conjunto de medidas para encontrar las singularidades que permitan especificar el sexo (por ejemplo, la longitud de la nariz, el ancho de los ojos, la distancia entre componentes faciales, etc.) [31]. Burton encontró que se puede detectar el sexo de una persona sólo considerando 16 medidas faciales. Algunas de estas medidas son consideradas como simples, y directamente visibles, por ejemplo, el espesor de las cejas es mayor en los hombres, o la distancia entre las cejas y los ojos, mayor en la mujer (sobre todo si la mujer se depila la ceja). Otras medidas son más complejas, como por ejemplo medidas tridimensionales que se desprenden de la protuberancia de la nariz. Precisamente Enlow, justifica las diferencias de formas entre la nariz de un hombre y una mujer debido a que, al ser el hombre más grande, necesita consumir más oxígeno. De ahí a que las fosas nasales masculinas estén más dilatadas y la nariz sea más ancha [74].

Lo cierto es que el rostro humano tiene una importancia notable tanto desde una perspectiva biológica como social. Ya desde sus primeras impresiones en la vida de las personas, los neonatos, desarrollan una tremenda fijación por los patrones faciales de forma natural, tendiendo a imitarlos [26]. Ello, unido con el paso del tiempo, convierte a las personas en auténticos expertos en la percepción

facial. Además, el rostro sirve a las personas como canal de comunicación cuando se tiene una conversación pero también para expresar el estado de ánimo a través de las expresiones faciales. Se trata, por tanto, en un elemento que proporciona mucha información acerca de una persona, de ahí a que sea el órgano más estudiado de la fisionomía humana y merece un análisis más profundo en la siguiente sección.

2.1.1. Estructura facial

La estructura facial está divida en los huesos de la cara y los músculos que la componen. Los huesos que forman la cabeza son parte del denominado esqueleto axial. Las funciones que tienen los huesos de la cara son, al margen de mantener la morfología facial, proteger los sistemas respiratorio y digestivo de la cabeza, además de proveer al humano de una herramienta móvil tan útil como la mandíbula. Por el contrario, los músculos son aquellos que dotan de elasticidad al rostro humano además de permitir distintas acciones que tienen lugar en el rostro como, por ejemplo articular la mandíbula a la hora de masticar los alimentos. Precisamente, las expresiones faciales vienen determinadas por éstos músculos faciales, por lo que resulta interesante enumerar cuáles son los principales músculos que conforman la cara y la relevancia que tienen en la formación de las distintas expresiones faciales. En la Figura 2.1 puede apreciarse el esquema facial así como los principales músculos que componen el rostro de una persona. Todos estos músculos son los considerados como músculos de la mímica o de la expresión facial, y se activan gracias a la interacción del nervio facial lo que permite al ser humano realizar gestos, tanto de forma voluntaria como de forma involuntaria. De hecho, según ciertos estudios como el de Suárez [114], la propia manifestación del músculo permite determinar si la contracción se acompaña de la voluntad del sujeto (de forma fingida) o si se produce de manera involuntaria (con naturalidad).

Volviendo a la Figura 2.1 pueden apreciarse los 15 músculos principales responsables de la mímica facial. Los músculos analizados de forma independiente no permiten mostrar su verdadera relevancia en el espectro de las expresiones faciales, pues como se desarrolla en la sección 2.1.2, la mayor parte de las expresiones faciales tiene lugar cuando estos músculos se expresan de forma simultánea formando distintas configuraciones faciales. Sin embargo, resulta interesante, al menos, nombrar algunos y apreciar la importancia de su acción en la formación de las expresiones faciales. El primer músculo, el occipitofrontal (1), situado en la frente, es el encargado de elevar la piel de las cejas, mientras que el músculo superciliar (2) permite expresar un gesto de pensamiento forzado a través de su contracción. El músculo orbicular (3) de los párpados tiene como función fundamental orientar las lágrimas hacia los puntos lagrimales, mientras que cuando se contrae permite expresar un gesto de preocupación. Por otro lado, el músculo transverso (5) de la nariz da forma a esa región manteniendo la nariz arriba y facilitando, a través de su contracción, un semblante de aspecto alegre y divertido. Como su propio nombre indica, el músculo dilatador de la nariz (6) permite dilatar las aperturas nasales. El músculo canino (8) es el encargado de

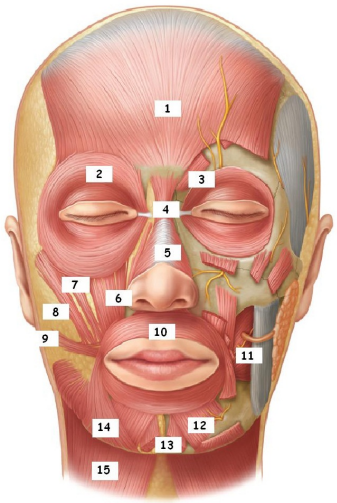


Figura 2.1: Músculos de la mímica: 1) occipitofrontal, 2) superciliar, 3)orbicular de los párpados, 4) piramidal, 5) transverso de la nariz, 6) dilatador de la nariz, 7) mirtiforme, 8) canino, 9) risorio de Santorini, 10) orbicular de los labios, 11) buccinador, 12) cuadrado de la barba, 13) borla del mentón, 14) triangular de los labios, 15) cutáneo del cuello. Fuente de la imagen [211].

atraer hacia arriba y adentro la comisura de los labios, de forma similar a como hacen los perros cuando se irritan. Al contrario que el canino, el músculo risorio de Santorini (9) retrae la comisura labial. Cuando los dos músculos homólogos de ambos lados de la cara se contraen juntos, aumentan el diámetro transversal de la boca, disposición que caracteriza la sonrisa, y de ahí el nombre del músculo. En los labios se ubica el músculo orbicular de los labios (10) o músculo "besador" cuya función principal es producir el cierre de los labios y ayudar al vaciado del vestíbulo bucal. Un músculo que tiene una importancia crucial a la hora de definir la morfología facial es el músculo buccionador (11). De hecho, su principal función es la de dar forma al rostro, agrandar la hendidura bucal. Tira hacia atrás la comisura labial aumentando el diámetro transversal de la boca e interviene en la acción de silbar, soplar, succionar o tocar instrumentos de viento. Un músculo que tiene acción directa sobre la boca es el triangular de los labios (14), que permite bajar o deprimir las comisuras labiales.

Una vez se han presentado los distintos músculos faciales y, sobre todo, de haber esbozado la acción que tiene la contracción o relajación de algunos de ellos, conviene estudiar el rostro como instrumento de interacción en la sección 2.1.2.

2.1.2. Interacción humana

La cabeza humana es un órgano central y complejo que aglutina la mayor parte de nuestros sentidos. Anteriormente, se ha descrito la fisonomía de la cara humana además de indicar la funcionalidad de los distintos elementos que componen los músculos de la mímica. Al margen de las expresiones faciales, el rostro dota a las personas de apariencia y de identidad, por lo que también juega un papel relevante a la hora de determinar quiénes son las personas que rodean a un sujeto en un determinado escenario. En el nivel más elevado de su taxonomía, los movimientos faciales pueden ser clasificados como rígidos o no rígidos. Los movimientos rígidos de la cabeza incluyen rotarla sobre el eje vertical, agitarla, asentir y girarla. Son movimientos en los que, no existe uso de la musculatura facial, sólo interaccionan los músculos del cuello cambiando la perspectiva de la cara respecto a un observador externo. También es importante matizar que cada uno de estos movimientos está relacionado con señales o connotaciones habituales en las relaciones sociales; asentir con la cabeza está relacionado con el hecho de afirmar o aceptar, mientras que agitarla hacia atrás y hacia delante está relacionado con el hecho de estar en desacuerdo o, un caso más común, girar la cabeza hacia una persona o en sentido contrario a una persona puede significar iniciar o romper una comunicación con dicha persona respectivamente. Por su parte, los movimientos no rígidos de la cara tienen lugar durante el transcurso de una conversación y se combinan con cambios en el rostro, tanto en las expresiones faciales como en la mirada. Este tipo de movimientos resulta interesante ya que implica la posibilidad de generar un conjunto variable de imágenes acerca de una persona y que puede distorsionar muchas "características identificativas" de un rostro; por ejemplo, a través de la modificación de la distancia relativa entre los elementos estructurales que forman parte del rostro. Así, la diferencia entre un rostro que sonríe y otro que muestra una expresión de sorpresa puede ser bastante considerable; la relativa posición de las cejas cambia de forma notable para cada una de esas expresiones faciales en relación, por ejemplo, a la posición de la boca o de otros rasgos faciales identificativos del individuo. Por otro lado, la expresión facial a lo largo de un discurso, el tono durante el discurso o los movimientos con las miradas también ofrecen un mensaje social. Resulta extraño que una persona hable manteniendo una pose estática o una expresión facial neutra, lo normal es que contextualice lo que está contando a través de una respuesta física. En este sentido, los cambios en las miradas pueden denotar aburrimiento o interés y la expresión facial puede proveer una cantidad ilimitada sobre el estado interno de una persona en lo concerniente a un determinado tema. La unión de todos estos movimientos, rígidos y no rígidos, ofrecen al ser humano la habilidad de crear miles de expresiones faciales distintas. Además, la funcionalidad del comportamiento humano se ve reflejada en varias facetas de la vida diaria de un individuo entre las que destacan:

- Exposición durante un discurso. Un discurso no está exento de expresiones faciales, por lo que no es necesaria una interacción activa por parte de dos sujetos. Por ejemplo, un monólogo puede denotar, por sí sólo, este tipo de comportamiento. Una situación que puede apreciarse durante cualquier

Figura 2.2: En esta imagen pueden apreciarse diferentes comportamientos que se ven reflejados en las expresiones faciales y en los gestos de los protagonistas de la escena. Por un lado, en los círculos rojos pueden apreciarse gestos simbólicos y emblemáticos de shock, en el círculo verde un gesto deíctico para llamar la atención del cuerpo médico y el círculo azul muestra una respuesta facial a través de una expresión de dolor de alta intensidad. Imagen obtenida de [9].

discurso es que, normalmente, las personas suelen elevar sus cejas cuando son curiosas y, por el contrario, suelen bajar las cejas cuando bajan el tono de la voz.

- Regulación de la conversación. Ayuda a gestionar los intervalos de habla dentro de una conversación. Un sujeto proporciona pistas a otras personas cuando ha finalizado su turno de hablar y desea ceder el testigo a otra persona, o no, en función de la expresión facial y de la voz.

- Cognición. Cuando una persona se concentra, la cara no pasa inadvertida. El ser humano, en muchas ocasiones, suele fruncir el ceño cuando está concentrado o bien, cuando se queda perplejo ante cualquier evento. Además, una persona suele apretar los labios cuando se haya inmersa en medio de un proceso mental.

- Hablando y comiendo. Las personas suelen hacer uso de los músculos que rodean la región de la boca cuando están hablando y alimentándose. Además esta región sirve para articular ciertos sonidos durante una conversación.

- Señalando emoción. Las emociones y los estados de ánimo tienen cabida a través de las expresiones faciales cuando se combinan distintos músculos de la mímica para mostrar dicho estado.

- Regulación expresiva. Las expresiones faciales no son discretas en su forma, están sujetas a las intensidades con las que son producidas. La regulación expresiva es el proceso que sucede cuando el sujeto controla los músculos para regular la intensidad de una expresión.

- Gestos emblemáticos. Se trata de un conjunto de gestos con un significado propio dentro de un contexto sociocultural. Este tipo de acciones tienen un significado simbólico verbal que puede ser respaldado con palabras. Por ejemplo, la mirada dudosa que se produce al elevar el labio superior haciendo uso del inferior para empujarlo hacia arriba. De hecho, los gestos pueden tener distintos propósitos, una taxonomía elaborada por [146] sugiere la siguiente clasificación:

  - Gestos simbólicos. Son gestos que, dentro de un entorno cultural, tienen un significado único. El gesto de "OK" es uno de este tipo de gestos más difundido dentro del contexto sociocultural de la sociedad occidental. Los gestos de la lengua de signos también forman parte de esta categoría (véase la Figura 2.2).
  - Gestos deícticos. Este tipo de gestos son bastante usuales dentro las interfaces hombre-máquina existentes. Se caracterizan porque suelen apuntar. Pueden ser usados para dirigir la atención de un público hacia un evento específico o bien, hacia objetos de la escena. También suelen ser utilizados en un contexto para orientar a un grupo de personas.
  - Gestos icónicos. Están dirigidos a especificar la información sobre el tamaño, la forma o la orientación del objeto de la discusión. Un ejemplo evidente de este tipo de gestos sucede cuando un sujeto comenta "el pájaro volaba de esta forma" mientras mueve las manos simulando el vuelo de un pájaro.
  - Gestos pantomímicos. Estos gestos son usados de forma común cuando un sujeto muestra el uso a través del movimiento de una determinada herramienta u objeto invisible en sus manos. Por ejemplo, cuando el sujeto especifica "moví la piedra a la izquierda" mientras hace la acción de mover una piedra pesada con las dos manos en orientación a su izquierda, es lo que se denomina gesto pantomímico.
  - Gestos rítmicos. Hace referencia a situaciones en las que el sujeto realiza movimientos en acompañamiento de un ritmo. Por ejemplo, mover las manos arriba y abajo con el ritmo del discurso o acompañar el ritmo del diálogo.
  - Gestos cohesivos. En este caso, no se habla de un gesto discreto sin más, realmente se trata de la combinación en el tiempo de gestos icónicos, pantomímicos y deícticos que son usados para unir partes del discurso separadas en el tiempo pero relacionadas desde un punto de vista temático.

2.2. Bases de datos

En los últimos años, el análisis facial automático se ha convertido en uno de los temas más activos de la visión por computador. Sin embargo, a pesar del

auge de esta línea de investigación, existen distintos estudios que agradecen la suma de nuevas aportaciones con datos que sean accesibles desde una perspectiva realista con el marco descriptor de estas investigaciones. En otras palabras, nuevas bases de datos proporcionan conjuntos de imágenes que podrían ser utilizadas como base de pruebas en los avances del campo de investigación. El análisis del rostro humano es sensible desde muchas perspectivas (identidad, pose facial, iluminación, expresión facial, edad, oclusión y vello facial). Esto es consecuencia de la estructura tridimensional del rostro además de la no rigidez del mismo. Tantas variables aumentan la complejidad sobre la dispersión de los datos necesarios para analizarlas. Por otro lado, el desarrollo de algoritmos robustos frente a todas estas variables requiere de un vasto conjunto de imágenes con el tamaño suficiente que recoja todas las posibles variaciones de la características que se desee analizar. Es más, las colecciones de datos normalmente utilizadas en el área deben proveer la posibilidad de evaluar comparativas entre algoritmos. Aunque la disponibilidad pública de la base de datos facial es importante para lograr avances en el campo, conseguir una colección con una alta calidad de datos resulta una tarea compleja. En esta sección se proporciona una extensa revisión de algunas bases de datos públicas para el análisis automático de rostros, centrándonos en los tres aspectos más frecuentes: el reconocimiento facial, la detección facial y el análisis de expresiones faciales.

2.2.0.1. Tipología de bases de datos

Como se ha comentado anteriormente, las bases de datos relativas al análisis facial se centran en los tres aspectos que son estudiados con más frecuencia. En esta sección se muestra un compendio de las bases de datos faciales más relevantes clasificadas dentro de cada uno de esos aspectos. Sin embargo, para simplificar esta sección, no se van a discutir todas las bases de datos al mismo nivel de detalle porque no todas son relevantes para el trabajo realizado, o bien, porque no resultan significativas para su análisis en profundidad. Por ejemplo, en el ámbito de las expresiones faciales, la *Belfast Naturalistic Database* [65] y la *PICS* [10] son bases de datos que sesgan el conjunto de emociones ya que solamente contemplan una única emoción; por otro lado, la *Yale Face Database* está dividida en dos colecciones bien diferenciadas pero contiene solamente tres expresiones emocionales (triste, dormido y gritando) de las cuales dos de ellas (dormido y gritando) no son consideradas como verdaderas expresiones faciales. La mayoría de las colecciones están disponibles de forma gratuita para la comunidad científica, sin embargo, existen algunas bases de datos como la CMU (*Carnegie-Mellon University Facial Expression Database*) que son colecciones restringidas únicamente a estudiantes de visión por computador.

2.2.1. Bases de datos existentes

Colecciones para el reconocimiento facial

El reconocimiento facial se ha beneficiado de forma notable a lo largo de

Colección	Año	Sujetos	Pose	Exp. Facial	Disponibilidad	Website
AR	1998	126	1	4	Online	[52]
BANCA	2004	208	1	1	Comercial	[53]
CAS-PEAL	2004	1040	21	6	Bajo demanda	[79]
CMU Hyper	2002	54	1	1	Bajo demanda	[61]
CMU PIE	2000	68	13	3	Bajo demanda	[111]
Equinox IR	2004	91	1	3	Comercial	[48]
FERET	2003	1199	9-20	2	Bajo demanda	[54]
Harvard RL	2000	200	3	1	Online	[49]
NIST MID	2000	1573	2	1	Comercial	[50]
ORL	1994	40	1	++	Online	[51]
Sheffield	1998	20	++	++	Online	[56]
XM2VTS	1999	295	++	++	Comercial	[152]
Yale	1997	15	1	6	Online	[57]
Yale B	2000	10	9	1	Online	[11]

Tabla 2.1: Descripción de las condiciones de grabación para todas las bases de datos discutidas en relación al reconocimiento facial. Los casos en los que las condiciones no pudieron ser verificadas con marcadas como "++".

los últimos años del gran número de bases de datos que han emergido para su estudio. La mayoría de estas colecciones ha sido creada bajo condiciones controladas con el fin de poder facilitar el estudio de los parámetros específicos del problema que supone el reconocimiento facial. Estos parámetros pueden incluir diversos tipos de variables como la pose, la condición lumínica, la expresión, el fondo de la escena, la calidad de la cámara, la inclusión de elementos que ocluyen total o parcialmente el rostro, la edad y el sexo. A pesar de la existencia de un gran número de aplicaciones para la tecnología del reconocimiento facial en las que un investigador puede controlar los parámetros para la adquisición de imágenes, también existen otras aplicaciones en las que el investigador tiene poco o ningún control sobre dichos parámetros. La tabla 2.1 ofrece una revisión de las principales bases de datos discutidas en este apartado.

Por mencionar algunas, la base de datos FERET [169] ofrece un conjunto de imágenes obtenidas desde distintas perspectivas frontales y desde los perfiles laterales (véase la Figura 2.3). Sólo considera la parte superior del torso de los sujetos (cabeza y cuello en su mayoría) por delante de un fondo uniforme. Además, la base de datos FERET ha sido utilizada con asiduidad para evaluar las fortalezas y debilidades de los distintos algoritmos [169]. Por otro lado, la base de datos de los Laboratorios AT&T de Cambridge (más conocida como la base de datos Olivetti (ORL)) consiste en 10 imágenes distintas de cuarenta sujetos [182]. En este caso, las imágenes fueron obtenidas en distintos momentos, variando la luz del entorno, la expresión facial y los detalles faciales (incluyen elementos como gafas).

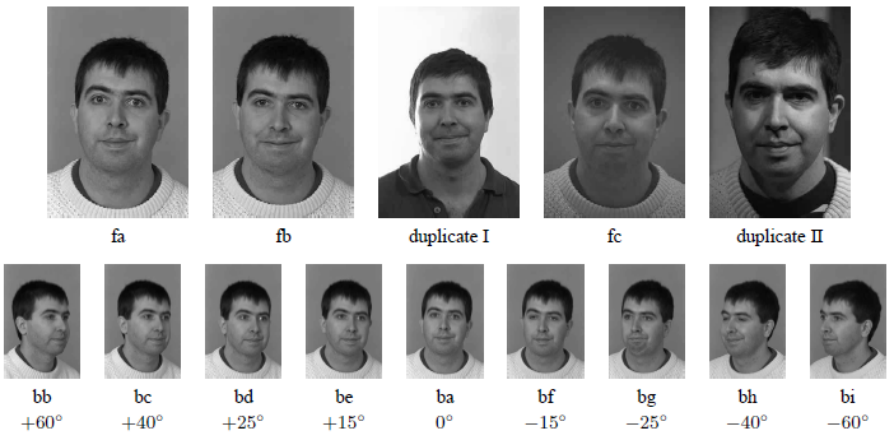


Figura 2.3: Imagen frontal y variaciones de la pose usadas en las evaluaciones de FERET [169]. La imagen *duplicate I* fue tomada dentro del mismo año que la imagen **fa** mientras que la imagen *duplicate II* y la **fa** distan en un año de diferencia. También es apreciable que las poses varían desde un ángulo de +60 grados (**bb**) hacia la imagen frontal (**ba**) y continua hasta -60 grados (**bi**).

La base de datos de Harvard consta de imágenes recortadas de caras desde una perspectiva frontal y tomadas desde un variable entorno lumínico [107]. Esta base de datos fue utilizada por Hallinan para un estudio sobre el reconocimiento facial bajo un entorno de condiciones lumínicas variable. Por su parte, con un total de 15 individuos, la base de datos de Yale contiene 10 imágenes frontales por persona, cada una con una expresión facial diferente, con y sin gafas, y bajo diferentes condiciones lumínicas [17]. Por otro lado, la base de datos multimodal M2VTS nace como proyecto europeo ACTS y fue desarrollada para experimentos de control de acceso haciendo uso de entradas multimodales [172]. Precisamente, esta base de datos en concreto contiene imágenes de 37 individuos y, años más tarde, fue reemplazada por su versión extendida (*Extended M2TVTS - XM2VTS*) [147]. La nueva versión ofrece una colección para el desarrollo de aplicaciones con el fin de verificar usuarios haciendo uso de una base multimodal (rostro y voz) y contiene 295 sujetos que fueron grabados en cuatro sesiones a lo largo de cuatro meses. Durante cada sesión se grabó al usuario mientras rotaba la cabeza además de registrar seis fragmentos de voz (los sujetos leen tres frases en dos ocasiones). La base de datos AR ofrece más de tres mil imágenes a color de 126 personas (70 hombres y 56 mujeres) desde una perspectiva frontal [144]. Esta base de datos fue diseñada para experimentos de reconocimiento facial bajo un conjunto de variaciones notables: diversas expresiones faciales, cambios en la iluminación y el uso de elementos de oclusión como gafas de sol, bufanda, etc. Resulta interesante añadir que las expresiones faciales que se consideran en este experimento son; neutral, sonriente, enfado y grito. Precisamente, con el grito, la intención de los creadores es distorsionar de

forma notable la estructura del rostro. Por último, mencionar que la iluminación de las imágenes capturadas es variable para cada subconjunto; fuente de luz a la izquierda del sujeto, fuente de luz a la derecha del sujeto y fuente de luz a ambos lados del sujeto. Además, las imágenes fueron tomadas en un intervalo de dos semanas y en un entorno con fondo uniforme.

Colecciones para la detección de caras

En las últimas décadas, la cara ha atraído mucho el interés de los científicos en el campo, sin embargo, no es un campo novedoso pues el rostro humano se ha estudiado a lo largo de más de treinta años por psicofísicos, neurocientíficos e ingenieros. Normalmente, el primer paso de toda aplicación que desee hacer procesamiento facial es detectar la posición de los rostros. No obstante, la detección de los rostros de una escena puede ser una ardua tarea debido a múltiples situaciones que pueden sufrir las caras que se desean detectar; el escalado, las rotaciones, la orientación o la pose (frontal o de perfil). Además, la expresión facial, la oclusión de parte de la cara o las condiciones lumínicas pueden tergiversar elementos identificativos de los rostros de una escena. Según Yang [234], la detección facial se define como, dada una imagen cualquiera, la capacidad de determinar la existencia o no de cualquier rostro en la escena y, si hay alguna cara presente, devolver la localización y la extensión de dicho rostro. Entre los retos asociados a la problemática de la detección facial destacan los que se enumeran a continuación:

- Pose. Las imágenes de una cara pueden variar debido a la posición relativa de la misma con respecto a la cámara (frontal, 45 grados, perfil, perspectiva superior o bien perspectiva inferior) y, debido a ello, algunos elementos característicos del rostro (por ejemplo, los ojos o la nariz) pueden llegar a quedar parcial o totalmente ocluidas.

- Presencia o ausencia de componentes estructurales. Características faciales como la barba, el bigote o las gafas pueden estar, o no, presentes y existe una gran variedad entre estos mismos componentes; tamaño, color o el tamaño.

- Expresión facial. La apariencia de las caras se ve directamente afectada por la expresión facial que tenga el sujeto, pues modifica su morfología.

- Oclusión. Las caras pueden estar parcialmente tapadas por otros objetos. En una imagen de un grupo de personas, algunos rostros pueden tapar de forma parcial a otros.

Para poder comparar el rendimiento de dos algoritmos de detección facial es necesario proveer este tipo de conjunto de datos. Los datos de prueba deben proporcionar una representación de los datos existentes en el mundo real, esto es, conteniendo imágenes de rostros en varias orientaciones con un fondo complejo. De acuerdo con Yang, la elecciones más populares son las mencionadas

Figura 2.4: Ejemplos de imágenes del conjunto de prueba de la base de datos MIT/CMU para la detección facial frontal [197]. Algunas imágenes contienen dibujos hechos a mano de caras. La mayoría de las imágenes contienen más de una cara y el tamaño de la cara varía de forma significativa.

anteriormente, las bases de datos FERET, MIT, ORL, Hardvard y AR respectivamente [234] (véase la tabla 2.1). Estas bases de datos son diseñadas con el objetivo de medir el rendimiento de los algoritmos para el reconocimiento de rostros y, por tanto, cada imagen contiene únicamente un individuo. Es por ello que dichas colecciones de datos son más adecuadas para entrenar que para hacer pruebas. La razón tácita de comparar clasificadores sobre conjuntos de prueba radica en que estos conjuntos de datos representan problemas que el sistema puede afrontar en el mundo real y, la mejora en el rendimiento sobre este conjunto de algoritmos puede ser trasladada a una mejora en el rendimiento de las tareas del mundo real. Para este fin, los investigadores han ido recopilando una gran colección de datos sobre una gran variedad de posibles escenarios. Entre las colecciones para la detección facial, dos conjuntos de datos de carácter público han emergido de forma estándar para la evaluación de los conjuntos de prueba: el conjunto de prueba MIT/CMU para la detección de rostros frontales [197] y el conjunto de datos de prueba II CMU para la detección de rostros frontales y no frontales [185]. El conjunto de prueba MIT/CMU (ver Figura 2.4) incluye 180 imágenes organizadas en dos conjuntos que incluyen; personas y patrones faciales.

Colecciones para el análisis de expresiones faciales

De acuerdo con los métodos para generar emociones en las personas, las expresiones faciales pueden ser categorizadas según tres tipos: inducidas, fingidas o naturales. Sin lugar a dudas, las naturales suponen el conjunto de datos más fehaciente posible a la hora de poder evaluar sistemas afectivos en situaciones del mundo real, sin embargo existen diversas consideraciones adicionales sobre este tipo de datos. Conseguir un conjunto de datos tan real supone enfrentarse a diversos inconvenientes, por ejemplo, problemas con los derechos de imagen, el uso de técnicas sofisticadas para conseguir grabar este tipo de colecciones o la verificación de las emociones obtenidas no es siempre fiable. Todos estos

factores constituyen un conjunto de barreras bastante complejo de superar a la hora de adquirir este tipo de datos. Por otro lado, la técnica para obtener emociones fingidas está basada en grabar a un conjunto de personas que actúan frente a la cámara, y estos, pueden ser actores profesionales [14] o no profesionales [141, 183]. Esta es una solución relativamente rápida ante las dificultades, previamente comentadas, que supone completar una colección de imágenes con expresiones faciales naturales. Sin embargo, entre las expresiones faciales naturales y las fingidas (o simuladas) existen varias técnicas para inducir emociones en las personas a través de distintos métodos; escuchar música emotiva, observar imágenes de películas emocionantes o bien jugar a juegos especialmente diseñados para provocar estas emociones. No obstante, dichos datos son difíciles de etiquetar debido a que no todos los estímulos generan la mismas respuestas en las personas; una imagen puede provocar asco en algunos sujetos, mientras que la misma imagen podría despertar miedo en otros sujetos. Debido a ello, la primera de todas las cuestiones trata acerca de determinar si los estados emocionales que están directamente relacionados con patrones fisiológicos específicos descuidan, de alguna manera, que los estados emocionales no son estáticos, sino que son dinámicos y que muchas veces dependen de las experiencias vitales del sujeto [24]. Esto es, desde una perspectiva fisiológica, que la mayoría de los índices de actividad fisiológica periférica pueden variar en función de la cantidad y el tipo de participación somática además del acompañamiento de soporte metabólico que se produce ante tal demanda. En otras palabras, cuando una persona decide hacer ejercicio físico, correr o prepararse para hacerlo puede producir una configuración diferente de la actividad fisiológica que sentarse y observar. De hecho, la actividad en un solo sistema (por ejemplo, el cardiovascular) está condicionada, en cierta medida, por la actividad en otro sistema (por ejemplo, el somático). Lacey [127] comentó la gran variedad de contextos en los cuales una emoción puede ser generada en un laboratorio:

> *"La amenaza para el organismo puede ser real o imaginaria, presente o posterior o anticipada, social o física, expresiva o totalmente inaccesible a la expresión en el momento. En todas estas situaciones se producen muchos cambios fisiológicos. . . . De hecho, en individuos predispuestos, cambios somáticos dolorosos e incluso peligrosos - tales como dolor de cabeza, dolor de espalda. . . producción de la sangre, bilis y exceso de ácido clorhídrico en el estómago. . . pueden ser generados por la mera discusión de material conflictivo y psicológicamente amenazante."*

Lacey también fue consciente de lo que él mismo denominó *"situational stereotypy"*. Esto significa que la naturaleza y la dirección de los cambios que ocurren desde una perspectiva psicológica dependen, en gran medida, en el contexto en el que tiene lugar el experimento. En la literatura existente sobre los estados emocionales se describe con gran interés cómo se generan las emociones en función del contexto. Por ejemplo, las inferencias acerca de la fisiología del miedo se realizan, a menudo, mediante la comparación de los datos de contextos

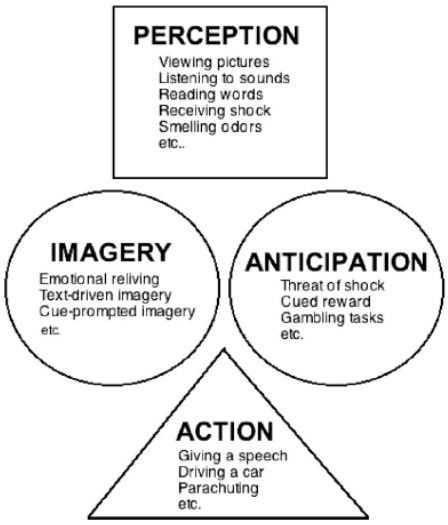


Figura 2.5: Los paradigmas más comunes a la hora de inducir emociones en la psicofisiología pueden estar clasificadas como la participación de la percepción, la imaginación, la anticipación, o la acción [127].

tan diversos como oír ruidos fuertes, anticipar un shock, imaginar un intruso en la casa, mirar una foto de una pierna amputada, ver una película de miedo, dar un discurso en público, poner la mano del sujeto en agua fría, o escuchar un grito de angustia. Por el contrario, las respuestas a estímulos como el de recibir dinero, escuchar música alegre, mirar una foto de cachorros, ver una película erótica, imaginar un día en la playa, recibir buenas calificaciones, pensar en ganar la lotería, o imaginarse las vacaciones son la base de contextos que sugieren un estado emocional feliz. Los diversos procesos sensoriales, cognitivos y motores provocados por estos procedimientos de inducción pueden incitar muy diferentes perfiles fisiológicos, con independencia de la modulación por la emoción. Por otra parte, la naturaleza y la dirección de la fuente generadora del afecto también pueden depender del contexto inducción específica. Tomados en conjunto, la comprensión de la psicofisiología de la emoción dependerá de especificar claramente el contexto de la inducción emocional en el laboratorio.

En el laboratorio, los contextos que se utilizan habitualmente para inducir reacciones afectivas se organizan en aquellos canales para buscar la emoción del individuo; anticipación, imaginación o acción (véase la Figura 2.5). Aunque es evidente que no son mutuamente excluyentes, estos dominios de inducción representan contextos similares con respecto a las respuestas psicofisiológicas. Por ejemplo, las reacciones fisiológicas de la percepción difieren cuantitativa y cualitativamente de los de la imaginación. Por otra parte, tanto dentro como a través de los contextos de inducción, los parámetros específicos pueden tener fuertes efectos sobre el grado y la naturaleza de compromiso emocional - es

Colección	Año	Sujetos	Exp. Facial	Disponibilidad	Website
JAFFE	1999	126	6 Emociones Básicas & Neutral	Online	[55]
Cohn-Kanade	2000	100	6 Emociones Básicas & Neutral	Bajo demanda	[47]
DaFEx	2004	8	6 Emociones Básicas & Neutral 3 Niveles de intensidad	Bajo demanda	[155]
MMI	2005	61	6 Emociones Básicas	Bajo demanda	[78]
U. Texas	2005	284	++	Bajo demanda	[156]

Tabla 2.2: Revisión de las condiciones de grabación para todas las bases de datos para la detección facial discutidas en esta sección. Los casos en los que las condiciones no pudieron ser verificadas con marcadas como "++". Este caso es posible porque la medida que se pretende obtener sea continua en el tiempo o bien porque la condición no se tuvo en consideración a la hora de grabar las imágenes.

decir, la facilidad con la que se activan los sistemas neurales que median en las distintas respuestas emocionales - y el patrón resultante de cambio fisiológico que se puede asociar con el compromiso afectivo.

Es importante destacar que, hasta hace bien poco, no existían colecciones que recogieran estímulos dinámicos. Sin embargo, a día de hoy, DaFEX (Base de datos de las expresiones humanas), la base de datos Cohn-Kanade, MMI (*"M&M Intiative Database Face"* [165]) y la base de datos de la Universidad de Texas son las colecciones de expresiones faciales básicas dinámicas más utilizadas por los investigadores (véase la tabla 2.2). La mayoría de las colecciones se han creado solicitando a actores que mostraran a través de sus rostros ciertas configuraciones de la unidad de acción (en inglés *Facial Action Coding System (FACS)*, más información en la sección 2.3.1), ya sea después de una pequeña orientación para conseguir la configuración deseada o bien después de un entrenamiento más extenso por parte del actor. En la colección JAFFE (*Japanese Female Facial Expression Database*), los actores mostraron emociones de forma libre [142]. El escenario para la adquisición de emociones fue configurado de tal forma que los actores eran capaces de monitorizar sus rostros además de tomar fotografías de los mismos.

En la colección de datos DaFEx [14], actores profesionales escuchaban pequeñas historias que incitaban a producir emociones mientras los actores tenían que mostrar cierta empatía que se veía reflejada en sus rostros. Realmente, la única colección que contiene expresiones faciales espontáneas es la de la Universidad

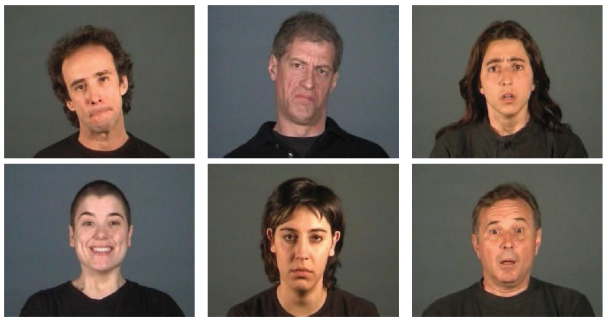


Figura 2.6: Seis emociones básicas presentadas por seis individuos distintos; imagen obtenida de la base de datos DaFEx [14]. Las emociones recogidas en dicha colección son: enfado, asco, miedo, felicidad, tristeza y sorpresa.

de Texas [161], donde los sujetos fueron grabados mientras observaban películas que les evocaban emociones. Es importante matizar que ciertas emociones son francamente complicadas de evocar, posiblemente debido a las restricciones que se derivan del entorno de presentación de contenidos para inducir tales emociones, y por ello, es tan complicado obtenerlas de forma natural. La base de datos de Cohn-Kanade [118] está disponible de forma libre desde la Universidad de Carnegie Mellon (CMU). Contiene secuencias de imágenes con expresiones faciales de hombres y mujeres de diferentes orígenes étnicos. Los sujetos simularon series de 23 poses faciales que incluían unidades de acción facial individuales, FACS. De hecho, las colecciones de expresiones faciales deben ser evaluadas desde una doble perspectiva; con la codificación FACS y a través de la evaluación individual del comportamiento facial del sujeto, aunque la segunda perspectiva podría ser considerada más interesante pues confirma que el material recogido es realmente percibido como estaba previsto en un primer momento. Las colecciones CMU y MMI han sido codificadas sólo con FACS mientras que, por otro lado, la colección de la Universidad de Texas no incluye ninguna evaluación sobre las expresiones faciales recogidas.

Las colecciones de datos anteriormente comentadas proporcionan un marco adecuado para la localización y el reconocimiento de rostros o de emociones por parte de las máquinas. En otras palabras, suponen un nexo comunicativo necesario para poder entrenar a las máquinas con el fin de ayudarlas a interpretar qué es lo que están viendo en el mundo real. En las secciones 2.3 y 2.4 se ofrece una revisión del estado del arte acerca de cómo los investigadores han hecho que las máquinas reconozcan tanto las emociones humanas como el sexo de los sujetos en base a los descriptores que se extraen de dichas colecciones de datos faciales.

Colecciones para el análisis de elementos contextuales

El último grupo de bases de datos que se analizan en esta sección está re-

lacionado con el análisis de elementos contextuales; ropa, pelo, accesorios, etc. Así, los conjuntos de datos presentados contienen imágenes de sujetos que permiten la extracción de estos elementos que pueden servir para definir aspectos relevantes del sujeto que los porta. Es importante señalar que no todas las bases de datos son exclusivas; por ejemplo la base de datos VOC PASCAL contiene imágenes muy variadas, no sólo de personas, sin embargo es posible extraer un amplio conjunto de imágenes en las que aparecen personas para un posterior análisis. Es por ello que en la tabla 2.3 sólo se consideran aquellas imágenes que son de interés para el propósito del análisis de elementos contextuales.

VIPeR [160] es una base de datos en las que los peatones son capturados desde múltiples perspectivas y presentan variaciones lumínicas. Se trata de imágenes de 128×48 pixels capturadas a lo largo de varios meses por estudiantes. Por el contrario, la base de datos de MIT-CBCL [100] ofrece una imagen mucho más cercana de los sujetos. En esta ocasión no se recoge el tronco inferior de las personas que son grabadas, tan sólo se recoge el tronco superior del sujeto así como un mayor nivel de detalle en el peinado así como los accesorios que pudiera portar. A pesar de sólo registrar a diez sujetos, se toman múltiples instantáneas de los mismos hasta llegar a un total de 200 imágenes por sujeto (imágenes frontales, imágenes tomadas desde múltiples ángulos, variando la iluminación de la escena, etc.). La colección de imágenes PASCAL VOC [77] contiene 11530 imágenes en las que están anotados hasta 27450 objetos. En esta ocasión el número de imágenes consideradas para el análisis de elementos contextuales ha sido reducido a 912 imágenes en las que se han llegado a detectar después de aplicar el detector ENCARA2 [37] hasta 1514 sujetos.

La base de datos Clothes [39] muestra un conjunto de 1858 imágenes con un sujeto por imagen. Aunque la varianza de los poses es pequeña, esta colección de imágenes está encaminada a la extracción y el análisis de la indumentaria de los participantes registrados. Por último, *The Images of Groups* [95] agrupa distintas colecciones de imágenes clasificadas en tres grandes bloques; bodas, grupos e imágenes familiares. Esta colección permite el análisis de los sujetos en distintos contextos sociales.

2.3. Reconocimiento de expresiones faciales

El reconocimiento de las personas es una tarea que conlleva una complejidad bastante elevada debido a la enorme diversidad de características que cohabitan en los rostros. Se define como reconocimiento interpersonal, no a la mera identificación de una persona (reconocimiento facial), sino más bien a la identificación de la persona además de la evaluación inicial sobre el lenguaje expresivo de dicha persona. En cuanto a la parte que concierne al proceso de identificación, el problema es menor, rara vez dos sujetos son iguales, y de serlo, llaman poderosamente la atención de las personas que los observan (por ejemplo, en el caso de gemelos). La identificación puede considerarse como una capacidad sujeta a las características físicas de las personas y, por regla general, supone poca complejidad al ojo humano. Sin embargo, establecer una línea emocional previa al

Colección	Año	Sujetos	Website
VIPeR	2007	632	[160]
MIT-CBCL	2004	10	[100]
PASCAL VOC	2012	1514	[77]
Clothes	2012	1858	[39]
The Images of Groups	2009	4739	[95]

Tabla 2.3: Descripción de las condiciones de grabación para todas las bases de datos discutidas en relación al reconocimiento facial. Los casos en los que las condiciones no pudieron ser verificadas con marcadas como "++". Este caso es posible porque la medida que se pretende obtener sea continua en el tiempo o bien porque la condición no se tuvo en consideración a la hora de grabar las imágenes.

inicio de la comunicación comprende una mayor dificultad en el reconocimiento de una persona. De hecho, se puede afirmar que cuando una persona visualiza a otra, su cerebro está procesando, al menos, dos tareas de forma concurrente; la primera tarea trata sobre identificar a la persona que tiene enfrente y, en segundo lugar, trata de reconocer su contenido emocional [214]. Si el sujeto es una persona ya conocida por el observador, entonces el reconocimiento del sujeto (primera tarea anteriormente descrita) resulta instantáneo y el cerebro apenas realiza esfuerzo alguno. Sin embargo, tal y como se argumentó en la sección 2.1.2, el reconocimiento facial implica una interacción interpersonal dinámica y no verbal (por parte de ambas personas, el observado y el observador), por lo que existen factores sociales relevantes a la hora de reconocer a una persona.

Otro factor de especial relevancia acerca de las emociones es que, a pesar de poderlas catalogar de forma universal [70], no existe un paradigma de universalidad de cara al propio reconocimiento de las expresiones. Por ejemplo, lo que para una persona puede resultar una sonrisa, para otra persona puede denotar un estado de nerviosismo. Estamos pues, ante un problema en el que la interpretación del observador se hace indispensable cuando no se está familiarizado con la persona observada. No conocer personalmente a alguien supone un importante matiz que implica que la red neuronal del observador no ha sido entrenada de forma que pueda reconocer al sujeto observado de forma eficiente. Cuando dos personas se conocen, este fenómeno queda diluido y la comunicación no verbal tiene un importante componente comunicativo en la relación de estas dos personas. En la sección 2.3.1, se muestra que existen patrones interculturales de conducta facial que son fácilmente reconocibles y rara vez tienden a la confusión pero, no es posible hacer una generalización axiomática de la interpretación de las emociones, y de ahí la enorme complejidad que éstas pueden entrañar. En realidad no todas las máquinas deben prestar atención a las emociones o tener habilidades emocionales; algunas máquinas no tienen la necesidad de tener este tipo de habilidades (una tostadora, una cafetera, etc.), no obstante, existen situaciones en las que la interacción entre máquinas y personas (HCI, del

Figura 2.7: Las seis expresiones faciales universales de las emociones más la expresión neutral. Las expresiones faciales son bastante significativas en la comunicación, la interacción y la relación entre las personas, ya que revelan información sobre el estado emocional, la actividad cognitiva, la personalidad, las intenciones y el estado psicológico de una persona. Esta información es muy difícil de enmascarar [70].

inglés *"Human Computer Interaction"*) podría mejorarse al tener las máquinas habilidades que se adaptan a sus usuarios. La computación afectiva es la ciencia que estudia esta respuesta de la inclusión emocional como una nueva forma de proveer información a las máquinas. En el capítulo 1, se estableció una línea evolutiva sobre los mecanismos de comunicación, por un lado, entre las personas y, por el otro lado, entre las personas y las máquinas. La computación afectiva supone, a través del uso de los canales de comunicación existentes, proveer de una dimensión más a la información que las máquinas pueden captar. Realmente, no estamos ante una situación extraordinaria que requiera un esfuerzo adicional por parte de las personas, más bien todo lo contrario. Los humanos suelen expresar emociones cuando están afrontando cualquier situación de la vida cotidiana. Por ejemplo, al leer la etiqueta de un envase de plástico una persona puede llegar a fruncir el ceño y ello indicará que está en un estado de confusión. Este proceso no es distinto cuando la persona se sienta frente a una pantalla de un computador, los patrones emocionales son inherentes a la condición humana y su manifestación sucede de forma inconsciente en base a lo que pueda estar experimentando dicha persona en ese preciso instante.

2.3.1. Tipos de expresiones faciales

Las emociones han evolucionado como un sistema de respuesta rápida y coordinada que permite a los seres humanos responder rápida y eficazmente a los acontecimientos que afectan a su bienestar [46, 70, 129]. Por su parte, las expre-

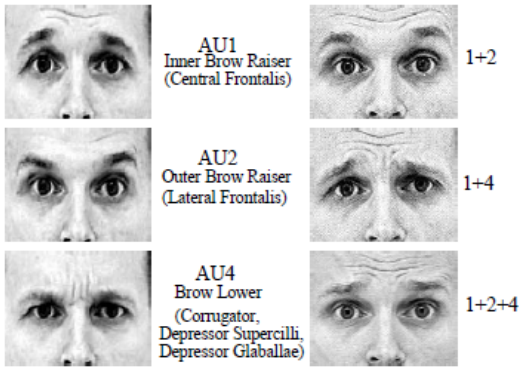


Figura 2.8: El sistema de codificación de acción facial clasifica las expresiones faciales en acciones individuales de las distintas componentes faciales. En esta Figura se puede apreciar cómo las tres regiones individuales de las cejas se activan y cómo se combinan. Cuando un sujeto muestra miedo suelen presentar las acciones 1+2 (parte superior derecha de la imagen), mientras que una acción de miedo espontánea activa las acciones 1+2+4.

siones faciales son parte de ese sistema de respuesta; la musculatura facial tiene la posibilidad de realizar más de 40 acciones independientes que pueden ocurrir de forma concurrente, lo que acarrea un número extremadamente grande de posibles configuraciones expresivas. Sin embargo, a pesar de este enorme potencial de configuraciones faciales, existe una fuerte evidencia de que un pequeño número de configuraciones faciales específicas y discretas que son universalmente reconocidas cuando se generan emociones [68]. Ekman observó que los miembros de una cultura aislada podrían identificar con un alto grado de fiabilidad las expresiones faciales en las fotografías de personas de culturas con las que aún no estaban familiarizados. También podían atribuir las expresiones faciales a las descripciones de situaciones. En esta prueba, se llegó a la conclusión de que las expresiones asociadas con algunas emociones son básicas o biológicamente universales a todos los seres humanos [72]. La lista de las emociones básicas propuestas por Ekman la componen; ira, asco, miedo, alegría, tristeza y sorpresa (ver Figura 2.7). Sin embargo, en la década de 1990, Ekman amplió su lista de emociones básicas [69], incluyendo una gama de emociones positivas y negativas. En esta nueva gama no todas las emociones están codificadas de acuerdo a la musculatura facial. Las emociones recién incluidas son: la diversión, el desprecio, la alegría, la vergüenza, el entusiasmo, la culpa, el orgullo (por un logro), alivio, satisfacción, placer y la vergüenza.

El sistema de codificación de acción facial (del anglosajón *The facial action coding system* (FACS)) de Ekman y Friesen [71] es el principal método para medir el movimiento facial en la disciplina que estudia el comportamiento facial. Se trata de un sistema que actualmente se realiza sin la ayuda de la visión por computador, sino que lo realiza un especialista. En FACS, las codificadores de

humanos descomponen las expresiones faciales en unidades de acción (UA) que corresponden, de forma muy aproximada, a los movimientos musculares independientes en la cara (véase la sección 2.1.1). Así pues, se concluye que Ekman y Friesen describieron 46 movimientos faciales independientes, o "acciones faciales" (véase la Figura 2.8). Estas acciones faciales guardan una estrecha relación con las distintas configuraciones de las expresiones faciales. Más de 7.000 posibles combinaciones de tales movimientos se han observado en el comportamiento espontáneo de una persona. Las ventajas que ofrece el FACS incluyen:

- Objetividad. No se aplican las etiquetas de las unidades de acción de forma interpretativa a las expresiones sino que, más bien, se realiza a partir de una descripción de los cambios físicos en la cara. Es decir, se descompone la expresión facial en unidades de acción más pequeñas a las que corresponde, de forma directa, la musculatura facial. Esto permite el estudio de nuevas relaciones entre el movimiento facial y el estado interno de las personas, tales como las señales de estrés o fatiga en el rostro.

- Integridad de los datos. Los códigos FACS para todos los movimientos faciales son observados y validados por psicólogos especialistas en el comportamiento humano.

- Validación robusta de los datos con respecto a la realidad. Existen bastantes datos sobre las relaciones entre los parámetros de movimiento FACS y los estados emocionales subyacentes. La codificación automática de la acción facial es eficaz para el desarrollo de herramientas de interacción persona-máquina. Además, la codificación de la animación facial a bajo nivel tiene un enorme impacto en la ciencia del comportamiento proporcionando la medida del comportamiento objetivo más accesible.

Sin embargo, existen ciertas cuestiones relativas al propio origen de las FACS que merecen la pena comentar. De hecho, se conocen dos enfoques teóricos acerca de la universalidad de las expresiones faciales [145]. El primero de ellos sugiere que las expresiones universales son generadas en un contexto cultural de continuo aprendizaje. Este punto de vista afirma que las personas se instruyen a través del aprendizaje por observación, modelado y refuerzo, con el fin de asociar las configuraciones faciales con los mismos estados emocionales. Las expresiones faciales de las emociones son, por tanto, universales porque las mismas expresiones son observadas y modeladas a lo largo de todo el mundo como respuesta a los mismos estímulos. El segundo enfoque apunta a que las expresiones universales proceden de un sistema emocional de respuesta que ha evolucionado. Este enfoque afirma que las distintas configuraciones faciales están codificadas genéticamente para todos los seres humanos y son parte de un sistema de respuesta más amplio que incluye cambios cognitivos, fisiológicos, y fenomenológicos. Por ejemplo, generar ira en un individuo consistiría en conseguir una serie de respuestas fisiológicas que incita a un individuo a pelear (con el consiguiente aumento del ritmo cardíaco y la frecuencia de la respiración), mantener la atención mental

alerta ante posibles adversarios, y generar expresiones amenazantes que permiten la mayor atención en el entorno, dejando al descubierto otras expresiones subyacentes en el rostro como enseñar los dientes (preparados para morder), y la percepción de la amenaza por otros. De acuerdo con este punto de vista, este sistema de respuesta coordinada se produce a partir de una fuente residente de base biológica que requiere poco o ningún aprendizaje y las expresiones faciales, por tanto, son universales, ya que son un producto de nuestra historia evolutiva.

Expresiones y reconocimiento

Mención especial requiere una temática interdisciplinar entre las expresiones faciales y el reconocimiento facial. Ciertas investigaciones apuntan a la existencia de una correlación directa entre las expresiones faciales y el reconocimiento facial, es decir, la dinámica del movimiento facial puede encauzar la identificación del sujeto [192]. Esta interesante tesitura plantea que, no sólo el rostro "sólido" puede ayudar a su reconocimiento, sino que también su propia dinámica lo hace. Evidentemente, si ya resulta complejo analizar las expresiones faciales, la afirmación sobre que una expresión contribuya a la identificación de un sujeto no está exenta de cierta complejidad. En primer lugar, no se trata de una posibilidad aplicable a cualquier circunstancia, tan sólo ocurre en algunos escenarios. Por ejemplo, el movimiento rígido (como el que ocurre cuando se gira la cámara alrededor de una cabeza inmóvil) no dificulta el reconocimiento con respecto a caras identificadas anteriormente [162, 184], pero esta ventaja se desvanece si sucede el mismo movimiento rígido de un sujeto durante el proceso de aprendizaje y asimilación de dicho sujeto por parte del observador. Por el contrario, en el movimiento no rígido (por ejemplo, cuando los individuos exhiben expresiones faciales o movimientos a través del habla) este enfoque juega un papel más relevante.

En los experimentos que hacen uso de sutiles formas morfológicas a través de caricaturas que simulan el movimiento facial de un sujeto a través de marcadores en el rostro, mostró que el movimiento facial no rígido de una persona desconocida (para el observador) aplicado a la caricatura puede llegar a dificultar que un observador identifique de forma correcta la persona a la que se está simulando (véase la Figura 2.9) [122]. Por el contrario, los experimentos de simulación de expresiones a partir de rostros conocidos (es decir, caras con las que el observador está altamente familiarizado) demostraron, de nuevo, una facilidad en el reconocimiento con señales dinámicas a través de movimientos expresivos o del habla. Este experimento no fue aplicable a los movimientos rígidos porque las caricaturas no poseen esa funcionalidad, además, debe tenerse en cuenta que los movimientos rígidos son muy similares entre sí independientemente de la persona que lo realiza. La familiaridad y la identificación fue más relevante en aquellas caras cuyo movimiento fue juzgado como "distintivo", es decir, en expresiones o poses faciales muy típicas de un sujeto. También debe considerarse que la familiaridad viene dada a partir de una secuencia natural de imágenes en movimiento, y no sólo a través de múltiples perspectivas disponibles de un mismo rostro. Además resulta interesante resaltar que la correcta identificación

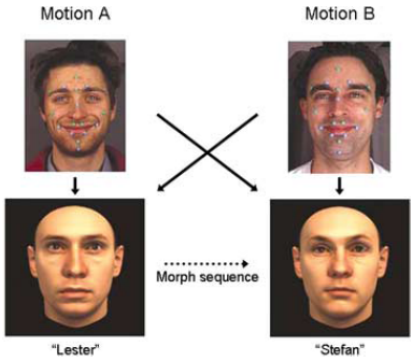


Figura 2.9: Experimento en el que el movimiento facial a partir de expresiones, o a través del habla, se identifica a través de las formas de "Lester" y "Stefan". La idea radica en que, a partir de una caricatura que simula el movimiento facial de una persona, se puede orientar la personalidad de un sujeto hacia la forma de Lester o hacia la forma de Stefan en función de dichos movimientos [122].

de los rostros del experimento morfológico empeoró cuando las simulaciones eran presentadas en orden aleatorio (con respecto a la persona simulada) sobre las caricaturas [122]. Estos resultados sugieren que el movimiento de la cara es mucho más que una sucesión de puntos de vista en el sistema humano de reconocimiento de rostros. Las señales dinámicas que proporcionan los movimientos expresivos proporcionan información sobre aspectos de las estructura facial de exceden los beneficios de tener múltiples imágenes estáticas de un sujeto.

2.3.2. Trabajos relevantes en el área

Desde hace relativamente bien poco se sabe que las emociones juegan un papel bastante significativo en la toma de decisiones por parte de los humanos [45, 170]. Por ejemplo, el contenido de un correo electrónico puede generar que una persona se sienta confusa y, en consecuencia, frunza el ceño además de inclinarse ligeramente hacia la pantalla. La capacidad para poder mostrar y, sobre todo, para poder interpretar las emociones resulta esencial en la interacción hombre-máquina. Existen autores como Picard que justifican la necesidad de desarrollar esta nueva dimensión comunicativa con las máquinas, no sólo por la posibilidad del desarrollo de nuevas formas de interacción interpretativa o gestual, sino que además puede ayudar a reducir, desde una perspectiva psicosocial, la frustración de los usuarios [171]. En este contexto, el rostro humano se ha convertido en el eje central de este tipo de investigaciones. De hecho, el análisis del comportamiento facial se encuentra actualmente como un tema de investigación relevante dentro de la comunidad de la visión por computador. De hecho, ya existen empresas que proporcionan servicios derivados de estas investigaciones [3]. Dentro de la literatura pueden encontrarse tres enfoques que

permiten abordar la temática desde tres puntos de vista bien diferenciados; por un lado se identifica el reconocimiento de expresiones en base a las características geométricas (localización o tamaño) de los rasgos faciales (por ejemplo, distancia entre puntos) [21, 29, 40], el segundo enfoque está basado en la apariencia del propio rostro (a través del uso de gradientes, filtros de Gabor, etc.) [13, 112, 221] y, por último, el tercer enfoque sugiere una perspectiva mixta de los dos enfoques anteriores [58, 204].

Recientemente, Valstar [212, 213] ha demostrado que los métodos basados en características geométricas proporcionan un rendimiento similar, incluso mejor, que aquellos métodos basados en la apariencia del rostro cuando se trata de encontrar una correlación entre las distintas unidades de acción del FACS y la expresión facial correspondiente. Sin embargo, los métodos basados en enfoques geométricos, por norma general, necesitan una detección y un seguimiento de las características faciales de forma estable y precisa, lo cual es difícil de conseguir en muchas ocasiones debido a la enorme plasticidad del rostro humano, a los cambios de iluminación, etc. En los métodos basados en la apariencia, filtros de imágenes (como los filtros de Gabor) son aplicados a lo largo de todo el rostro o bien en regiones específicas de la cara con vistas a poder analizar los cambios en la apariencia del rostro de los individuos. Debido a que, con el uso de este tipo de filtros, el rendimiento obtenido es bastante alto, no resulta extraño encontrar en la literatura una gran variedad de métodos basados en la apariencia que hacen uso de los filtros de Gabor como método de representación de la información [63, 142, 204, 240]. Sin embargo, este tipo de métodos tiene una desventaja en forma de tiempo de ejecución o en cuanto al consumo de recursos de memoria del computador, ya que obtener los coeficientes a nivel de múltiples escalas y orientaciones utilizando un banco de filtros de Gabor es bastante costoso. Por otro lado, es necesario destacar una limitación que suele mermar la eficiencia de bastantes métodos de reconocimiento facial que hacen uso de datos provenientes de colecciones de imágenes las cuales ofrecen rostros frontales con una gran resolución y una gran calidad del entorno lumínico [204]. Esta limitación viene dada porque, en un entorno real, las imágenes tienen una iluminación bastante irregular y una resolución mucho más baja, lo cual hace que reconocer las expresiones faciales a distancia y con estas características sea bastante complejo en un entorno no controlado. De hecho, algunos autores han enfocado sus investigaciones siguiendo esta línea de trabajo, en el análisis de expresiones faciales con imágenes de baja resolución, estudiando así los efectos que tienen las distintas resoluciones sobre el estudio del problema en cuestión [204, 205]. Otros autores como Buenaposada, en cambio, han obtenido notables resultados en el análisis de expresiones faciales aplicando métodos basados en la apariencia considerando variaciones lumínicas [28]. Para ello, se ha analizado el rostro a través de los cambios que presenta cuando la iluminación de la escena es variable.

Más allá de la distinción entre los distintos tipos de métodos para la correcta clasificación de expresiones faciales, lo cierto es que este campo ha tenido un largo recorrido en la última década. En realidad, haciendo un barrido de la literatura sobre la existencia de las distintas técnicas utilizadas para clasificar las

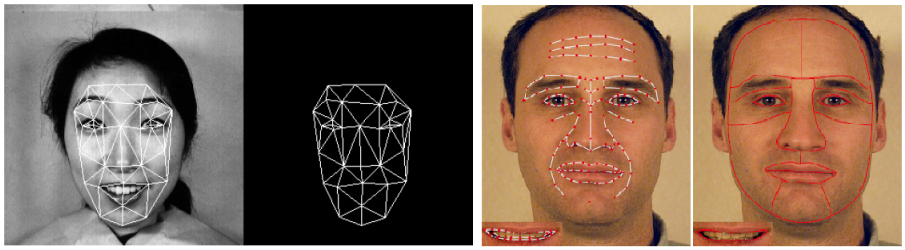


Figura 2.10: Enfoque geométrico para la detección de expresiones faciales. La selección de puntos relevantes en la región facial permite determinar, a través de las distintas distancias entre puntos, la posible expresión facial que el usuario esté mostrando [141, 239].

expresiones faciales, es posible encontrar una gran variedad de técnicas que hacen uso de redes neuronales [164, 204, 240], máquinas de soporte vectorial (SVM) [13], redes bayesianas [40] y clasificadores basados en reglas [165, 166, 167]. Por su parte, Lyon hace uso del análisis de componentes principales (*Principal Component Analysis (PCA)*, véase sección 3.2.1.2) de las características faciales obtenidas a partir de las imágenes de entrenamiento. Además, este autor hace uso del análisis de discriminantes lineales (*Linear Discriminant Analysis (LDA)*), con el fin de discriminar los vectores de características faciales y realizar una clasificación de la expresión facial de las imágenes de pruebas haciendo uso de estos vectores.

Por su parte, otros autores como Cohen, realizaron comparaciones entre distintos clasificadores bayesianos y llegaron a la conclusión de que los clasificadores gaussianos (*Gaussian Tree-Augmented-Naive (TAN)*) obtenían los mejores resultados [40] entre todos los clasificadores analizados. Bartlett realizó una comparación de distintas técnicas que incluían AdaBoost, SVM y LDA para el reconocimiento de expresiones faciales, obteniendo los mejores resultados se seleccionó un subconjunto de filtros de Gabor a través del AdaBoost para después entrenar el clasificador SVM con las salidas de los filtros seleccionados [13]. Pantic afrontó la temática a través de un razonamiento basado en reglas con el objetivo de reconocer las distintas unidades de acción (AU) y las combinaciones que estas ofrecían [167]. Sin embargo, en la sección 2.3.1 ya se estableció que las expresiones faciales no son estáticas, sino que tienen una importante componente temporal, en otras palabras, una persona no refleja una expresión facial a través de su rostro en un determinado instante sino que, más bien, la expresión facial es una evolución natural de los músculos faciales. Muchos autores se percataron de este hecho y, con el fin de explorar el comportamiento temporal de las expresiones faciales se aplicaron distintas técnicas de reconocimiento facial sobre secuencias de imágenes. Los primeros intentos en la temática vinieron de la mano del seguimiento y el reconocimiento de expresiones faciales a través del tiempo haciendo uso de técnicas de flujo óptico. En estos trabajos se analizó el comportamiento de una serie de puntos distribuidos a lo largo de la superficie

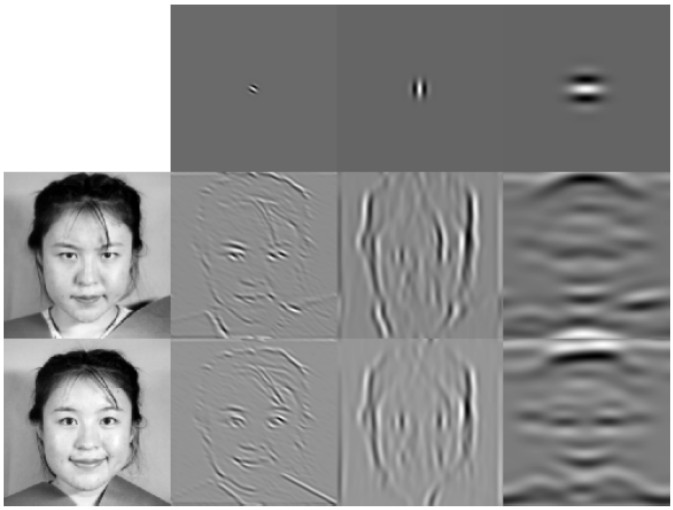


Figura 2.11: Respuestas antes dos expresiones faciales después de aplicar tres configuraciones posibles de filtros de Gabor [141].

facial [76], o bien, a lo largo de regiones en concreto como lo son ojos y boca [232]. Por su parte, Tian presentó un enfoque basado en una red neuronal para reconocer las distintas unidades de acción en secuencias de imágenes [206]. En el ámbito temporal también se ha hecho uso de modelos ocultos de Markov (*Hidden Markov Models (HMM)*) para modelar el comportamiento facial a partir de secuencias de imágenes [40, 236]. En su trabajo, Cohen propuso un clasificador HMM de múltiples niveles, ello permitió clasificar expresiones faciales en secuencias de vídeo además de segmentar secuencias duraderas de vídeo en función de las expresiones detectadas sin necesidad de hacer uso de métodos heurísticos para la segmentación de expresiones [40].

Kaliouby propuso un sistema que infiere el estado mental de una persona a partir de secuencias de vídeo en las que se aprecia una notable riqueza de movimientos no rígidos (expresiones faciales) así como de movimientos rígidos (gestos con la cabeza). Para ello hicieron uso de un clasificador DBN (siglas procedentes de *Deep Belief Network*) multinivel que permitió modelar un conjunto de estados mentales asociándolos a un conjunto de imágenes de movimientos rígidos y no rígidos [117]. Una DBN es una red de creencias profundas la cual es, básicamente, una red neuronal de tipo probabilístico con un conjunto de capas ocultas. Zhang y Ji continuaron explorando esta técnica en el uso de las expresiones faciales a través del uso de DBNs para el modelado y el entendimiento de la evolución temporal en la formación de las expresiones faciales [238]. Por último, autores como Lee y Elgammal han desarrollado un modelo que puede reconocer diferentes apariencias faciales para distintas personas y para distintas expresiones, la idea fundamental de este proyecto es permitir el reconocimiento simultáneo de cara y de expresiones faciales [130]. Dornaika abordó el problema

desde otra perspectiva, a través del análisis de las características faciales más relevantes en un entorno dinámico [64]. Para ello, hace uso de una técnica denominada "modelos de apariencia en tiempo real" (*Online Appearance Models*) que, aplicada sobre un conjunto de fotogramas, permite obtener hasta 54 características relevantes. A través de esta técnica se obtuvo notables resultados (97 % de precisión) con SVM. Además, entre las medidas más relevantes para dicho clasificador, destacan principalmente los extremos de la comisura labial así como la parte externa de las cejas.

La mayoría de los autores citados anteriormente resalta en sus trabajos la enorme subjetividad existente a la hora de poder reconocer expresiones faciales no sólo a través de las máquinas, incluso entre los propios humanos. Por otra parte, el rango de interacción entre humanos y máquinas es bastante reducido, no existe una comunicación activa por parte del usuario hacia la máquina, lo que dificulta la contextualización de la emoción en un determinado instante. Además, el rol adoptado por el sujeto que muestra la expresión es dependiente del ente (máquina o humano) a la que se la muestra dicha expresión. Es decir, si bien existen ciertas expresiones faciales que suelen ocurrir frente a la pantalla de forma similar a la que sucede frente a un humano (por ejemplo, una sonrisa), otro tipo de expresiones (por ejemplo, una expresión de sorpresa o de enfado) es raro que se muestren de la misma forma natural a la que serían mostradas frente a otro humano.

En un contexto más reducido es posible estudiar la detección de expresiones faciales para las que no existe ambigüedad interpretativa; la detección de sonrisas. Sin duda, la sonrisa humana constituye una configuración facial más fácilmente reconocible por un computador con una alta precisión y robustez debido al contraste que refleja la apariencia de una persona cuando sonríe en relación a cuando lo hace. Existen hasta siete músculos que intervienen en el desarrollo de la sonrisa, alguno de los cuales ya ha sido nombrado en la sección 2.1.1, pues forma parte de los músculos principales de la mímica. Sin embargo, existen otros músculos más secundarios en la mímica, pero que juegan un papel relevante en la formación de la sonrisa. Los siete músculos son el cigomático mayor, el cigomático menor, el risorio, el buccinador, el elevador del labio superior, el elevador del ángulo de la boca y, por último, el orbicular del ojo. Aunque la modulación de estos músculos puede generar una amplia variedad de sonrisas, en la práctica, la sonrisa prototípica sólo debe implicar el cigomático mayor [15]. Además, se trata de una expresión facial bastante útil y muy significativa, pues permite denotar alegría, diversión e, incluso, aprobación (y también la falta de cualquiera de estos estados) [73, 106].

A diferencia del reconocimiento de expresiones faciales generales, la investigación en la detección de sonrisas había producido, hasta hace bien poco, mucha menos literatura que otros enfoques para el reconocimiento de expresiones. Sin embargo, en los últimos años este tipo de trabajos han sido proyectados con vistas a ser enfocados para el desarrollo de dos aplicaciones fundamentales: por un lado la "sonrisa disparador" de las cámaras digitales que permite obtener una imagen de forma automática cuando las personas de la escena sonríen a la lente y, por otro lado, detectores sociales para detectar sonrisas en humanos durante

su vida cotidiana [148]. Para ambas aplicaciones el sistema detector de sonrisas necesita operar con un rango amplio de condiciones lumínicas, localización geográfica, etnia, sexo o edad, además de lidiar con el hardware del dispositivo de adquisición de las imágenes (resolución de la cámara, zoom, etc.). En el trabajo de Ito se hizo uso de los bordes labiales y un perceptrón [113]. Ello denota que la parte del labio es la región más importante pues envuelve el movimiento de los, previamente mencionados, músculos cigomáticos para elevar los extremos de la boca. Sin embargo, los bordes de los labios pueden resultar características insuficientes a la hora de determinar si una persona sonríe o no. Así, Kowalik hizo uso del seguimiento de una serie de puntos ubicados alrededor de la boca con el fin de poder entrenar un clasificador neuronal [125]. Por su parte, Déniz propuso una técnica para la interacción en tiempo real [62]. Este trabajo sugiere que ojos, nariz y boca proporcionan la información suficiente para decidir si una persona está sonriendo. Mientras, el trabajo de Shinohara [190] ofrece un sistema de autocorrelación de orden superior para la detección de sonrisas, llegando a alcanzar tasas de acierto muy elevadas en entornos controlados. Más recientemente, las conclusiones del trabajo de Whitehill sostienen que existe una gran diferencia entre las pruebas típicas de la literatura y los resultados que se obtienen en situaciones más reales [223]. Este autor afirma que el conjunto de entrenamiento entraña una importancia casi tan relevante como el método de detección que se desee aplicar, sobre todo en términos de variabilidad y tamaño, afirmando que lo normal es que se manejen miles de imágenes durante la fase de entrenamiento. Recientemente, han sido publicados diversos trabajos que hacen uso de bases de datos públicas para la detección de sonrisas [110, 189, 233]. Estos trabajos, así como dos sistemas comerciales, son discutidos en la sección 3.2.3.

2.4. Reconocimiento de características demográficas

Como se ha comentado en la sección 1.1, los seres humanos son criaturas sociales que establecen la comunicación a través de los distintos canales existentes para ello. Evidentemente, estos canales permiten al ser humano basar buena parte de su existencia a través del desarrollo de relaciones significativas con las personas que le rodean y, por ende, la supervivencia en la sociedad actual depende de esa habilidad para comunicarse de forma eficaz con este grupo de personas. Sin embargo, para sobrevivir en este contexto, es tan necesario tener un dominio para interpretar las emociones de otras personas como tener la habilidad de mostrar las expresiones apropiadas. Las expresiones faciales proporcionan la evidencia de la existencia de emociones en el ser humano y son considerados la "señal social" más importante, de hecho, el rostro no sólo revela las emociones sino que también supone la base para el reconocimiento personal. Por otro lado, los rostros humanos en sí también son un estímulo visual igual de relevante que las emociones. El sexo de los interlocutores es realmente importante en las

Figura 2.12: La detección de sexo a partir de los rasgos faciales no siempre es tarea sencilla para las máquinas. En múltiples ocasiones, la no contextualización del rostro puede dar lugar a "parecidos razonables" que dificulten la tarea. Imagen obtenida de [203].

propias interacciones humanas ya que es complicado omitir dicha componente sexual durante el proceso de interacción. En otras palabras, la transmisión de un mismo mensaje entre dos personas puede sufrir diversos ajustes en cuanto a connotaciones en función del sexo de las personas que interactúan. Por ello, es importante determinar el sexo de una persona a la hora de dar respuesta a un cierto compendio de poses y expresiones durante la interacción. La investigación acerca de la determinación automática de sexo data de los años sesenta y se sitúa en un escenario en el que se comenzó a extender el estudio de rasgos distintivos de los seres humanos como la raza o, las ya mencionadas, expresiones faciales.

La realidad es que la identificación de los atributos demográficos de los seres humanos tales como edad, sexo y origen étnico ha tenido una atención especial en el campo de investigación de la biometría blanda a lo largo de los últimos años. Dichos atributos pueden desempeñar un papel particularmente relevante en tareas como: interacción hombre-máquina, estudios demográficos o la publicidad dirigida. Precisamente, la publicidad dirigida, ha tenido un eco mayor en el foro de investigadores debido a la enorme aplicabilidad así como a los beneficios económicos directos que se desprenden de su aplicación. Por ejemplo, una de las aplicaciones más interesantes que puede ofrecer el campo de la publicidad es el que consiste en recopilar toda la información estadística sobre el número de personas que se detienen frente a un escaparate así como el tiempo que dura esta parada. Durante ese periodo de tiempo, tanto la edad como el sexo, pueden proveer una información vital para que los agentes de marketing estudien el impacto de un producto comercial. Se trata de una alternativa a los usuales cuestionarios que suelen resultar molestos para las personas que son encuestadas en base a sus gustos y esta alternativa es potencialmente exitosa porque no es invasiva. El reconocimiento de sexo es un problema de clasificación de dos clases: hombre y mujer. Sin embargo, aunque a priori pueda parecer un problema sencillo, existe un compendio de factores que dificultan la tarea. Para empezar resulta interesante observar la Figura 2.12 donde queda patente la complejidad

que se le puede presentar a una máquina (incluso a un humano) a la hora de distinguir el sexo de las personas que se muestran en dicha imagen si sólo se considerara un marco facial cerrado (sin considerar factores contextuales como pelo o ropa). Aunque el método más extendido en la literatura para abordar este problema por parte de las máquinas es el que está basado en las características del rostro, en dicha imagen puede constatarse que no siempre el rostro ofrece una respuesta categórica, por lo que también ha sido necesario el desarrollo de estudios que basan la cuestión de sexo en otros elementos contextuales que rodean a las personas.

Como se podrá apreciar en la sección 2.4.1.1, una gran mayoría de los trabajos de la literatura a la hora de clasificar el sexo se basan en la extracción total o parcial de características faciales. Además, de forma análoga a los tipos de técnicas para la detección de expresiones faciales descritas en dicha sección, la extracción de este tipo de características también puede dividirse en métodos basados en la apariencia o en métodos basados en la geometría. Los primeros procesan el propio contenido de la imagen a través de filtros mientras que los basados en características geométricas lo hacen a través de distintas medidas que pueden proporcionar las distintas localizaciones de zonas concretas dentro de la imagen. Además, en cuanto a la forma de procesar la información, es posible que no se procese toda la imagen y sea posible identificar el sexo en función, únicamente, del procesado de partes concretas del rostro. Todos estos enfoques dejan de relieve que el rostro, por sí sólo, supone una de las herramientas biométricas más relevantes a la hora de identificar a un ser humano. Los rostros son innatos y presentan diferencias entre sí de manera regular. Sin embargo, en este tipo de aplicaciones de detección de sexo existe cierta carencia de sistemas automatizados que presenten resultados robustos dentro de cualquier contexto; la pose, la iluminación, elementos de oclusión, etc., pueden dificultar la correcta identificación del sexo de una persona. Es por ello que la diversidad de los seres humanos les permite elevar el estándar de reconocimiento de sexo a través de elementos contextuales como por ejemplo la ropa o el pelo. Por este motivo, en la sección 2.4.1.2 también se analizan trabajos en la literatura considerando esta integración de elementos complementarios al rostro a la hora de determinar el sexo de las personas.

2.4.1. Trabajos relevantes en el área

2.4.1.1. Reconocimiento del sexo basado en componentes estructurales faciales

Desde un punto de vista biológico, el rostro humano presenta una dicotomía estructural acentuada si lo que se pretende considerar es el sexo. Diversos estudios avalan esta teoría y recorren la problemática desde la estructura facial ósea [99] hasta las medidas estructurales entre componentes faciales [31]. Wild demostró que el reconocimiento de sexo en las personas está considerado como un proceso cognitivo que es eficiente y muy rápido. Para ello, recortaron imágenes de personas con el objetivo de eliminar todas las claves culturales que

aparecían en la escena (peinado, ropa, accesorios, etc.), dejando a la vista sólo el rostro humano. La idea del experimento era examinar la eficacia con la que las personas adultas y los niños eran capaces de identificar el sexo de la persona que estaban visualizando. Los resultados fueron concluyentes, el 100 % de los adultos encuestados acertaron todas las imágenes propuestas de 255×240 píxeles, mientras que los niños de siete años de edad sólo alcanzaron el 80 % en el desarrollo de la misma tarea [225]. Estos resultados sugieren que, si bien los rasgos biológicos de la anatomía facial suponen muestra suficiente para un reconocimiento de sexo óptimo, no es menos cierto que el proceso de identificación de sexo es evolutivo y, por tanto, debe ser entrenado desde temprana edad. Los rostros de hombres y mujeres suelen diferir tanto en la forma como en la textura y, precisamente, estas características son tenidas en cuenta por el ojo humano a la hora de diferenciar entre ambos sexos. Desde un punto de vista frontal del rostro, la textura es más relevante que la forma a la hora de clasificar el sexo, mientras que desde un punto de vista lateral, o de perfil, la forma adquiere una especial relevancia en el proceso [25, 108].

Por otro lado, aunque en el estudio de Wild [225] quedó patente que los adultos eran capaces de acertar el sexo de las personas que les era mostrado, resulta interesante señalar que el estudio se realizó en unas condiciones óptimas en la adquisición de las imágenes de los sujetos. En un entorno lumínico variable el rostro no supone un objeto fácilmente discriminante para el sexo debido a que las sombras, la profundidad o la forma en el rostro influyen de forma notable en la percepción que se puede tener del mismo. Por ejemplo, supongamos una naranja y una sandía; ambas son frutas y pueden ser categorizadas como comida a pesar de que tienen aspectos diferentes, mientras que si consideramos una naranja y una bola de billar del mismo color, entonces ambas son categorizadas por un ser humano de forma diferente a pesar de tener el mismo aspecto en términos de forma, color y tamaño. Cuando las condiciones en las que el humano percibe la imagen se ven distorsionadas por elementos externos lo que ocurre es que se pierde la facultad de poder adquirir ciertas características; si bajamos la intensidad de la luz que enfoca la bola de billar y la naranja o si alejamos al observador de la escena, se perderá la capacidad de poder adquirir información básica (en este caso, la textura ya no sería apreciable de forma óptima) para la correcta clasificación de clases. Otra interesante cuestión a la hora de poder distinguir entre hombres y mujeres radica en el propio proceso de toma de decisión que ocurre en la mente del ser humano. Aunque, normalmente, las personas son capaces de distinguir entre el sexo de las personas con relativa facilidad, no son capaces de explicar en qué se basan para hacer semejante juicio. Las personas presentan deficiencias cuando se le pregunta por qué catalogan a otra como hombre o mujer basándose sólo en el aspecto facial (más allá de posibles elementos estructurales como barba o bigote), es decir, no son capaces de explicar la definición que hace que una cara sea hombre o mujer. Se trata pues, de una decisión que las personas toman sin apenas conciencia de ello y lo hacen de una forma eficiente con pocos recursos a su alcance [173].

Estas cuestiones hacen que la clasificación de sexo en base al rostro de una persona suponga un reto en el campo de la visión por computador. Dentro de

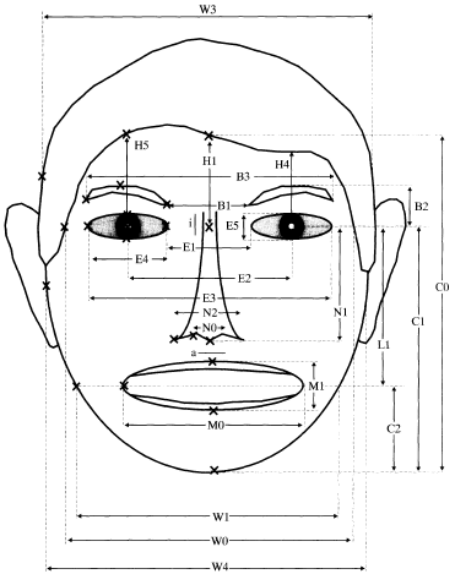
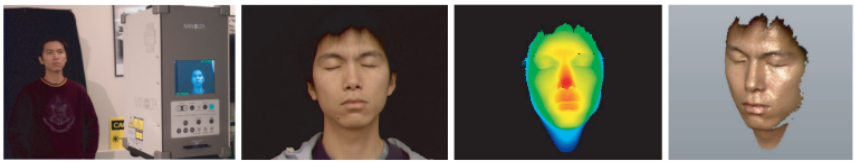


Figura 2.13: Las 22 distancias basadas en los puntos más relevantes de la estructura facial según Fellous [80].

la literatura, la detección de sexo ha atraído mucho la atención de los investigadores. Entre los trabajos del área, a la hora de considerar únicamente la cara, pueden distinguirse los, previamente mencionados, métodos geométricos y los métodos basados en la apariencia. Ya luego, dentro de cada tipo, también puede considerarse el estudio del rostro por completo o bien la selección de regiones específicas con el fin de poder determinar el sexo de una persona con mayor grado de precisión.

En los métodos basados en componentes geométricos destacan aquellos que se basan en el reconocimiento de las regiones que señalan las características más significativas del rostro como los ojos, la nariz o la boca. Brunelli hizo uso de una máscara facial la cual le permitió obtener 18 distancias entre puntos con el fin de poder entrenar su clasificador de detección de sexo [27]. Sin embargo, Fellous aumentó la complejidad de la máscara llegando a extraer hasta 40 puntos faciales que le permitieron obtener 22 distancias normalizadas que fueron utilizadas como un vector de características (véase Figura 2.13) [80].

Los métodos basados en la apariencia, en cambio, trabajan a partir de un conjunto de transformaciones efectuadas sobre los píxeles del rostro. Este conjunto de transformaciones puede ser del rostro integral o bien de forma local a través de patrones. Entre los trabajos más destacados para este tipo de clasificación se pueden enumerar algunos de especial relevancia; Sun [196] propuso un análisis de componente principales (PCA) para representar las características de las imágenes en un espacio con pocas dimensiones. Se combinaron esta representación con algoritmos genéticos con el fin de determinar qué características



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

Figura 2.14: Proceso de selección de características faciales a través del escaneado tridimensional del rostro de las personas [140].

eran más distintivas a la hora de seleccionar el sexo de una persona. Con dichas características entrenó cuatro tipos de clasificadores obteniendo resultados para cada uno de ellos; un clasificador Bayesiano (77.62 % de tasa de acierto), una red neuronal (82.3 %), un clasificador basado en LDA (85.8 %) y, por último, un clasificador SVM (91.1 %). Para este experimento se hizo uso de una base de datos propia con 400 imágenes frontales y normalizadas de rostros. Por su parte, Gutta hizo uso de la base de datos FERET para clasificar el sexo a partir de un clasificador híbrido de rostros humanos [103]. Su idea consistió en un conjunto de funciones de base radial (RBF) combinadas con árboles de decisión (DTs) y, según su estudio, consiguió una tasa de aciertos del 100 % con dicha base de datos.

Saatci y Town también propusieron un sistema híbrido entre reconocimiento de expresiones faciales y reconocimiento de sexo [179]. En primer lugar, el rostro era modelado a través de la aplicación de un *Active Appearance Model (AAM)* para poder, de esta forma, extraer las características más relevantes. Se hizo uso de clasificadores SVM de base radial y polinomial. Lo realmente interesante de este trabajo es que aprovechó la información del sexo del individuo para poder reconocer la expresión facial mostrada. Precisamente, en su trabajo sobre detección de sexo, Bekios-Calfa resalta que, siempre que exista un buen conjunto de datos de entrenamiento, el uso de un clasificador SVM con kernel de base radial proporcionará los mejores resultados siempre que se posean los recursos computacionales adecuados [16]. En caso de que estos recursos sean escasos, enfoques lineales o los basados en *boosting* también proporcionan buenos resultados.

Nakano centró su atención en la información proporcionada en los bordes de los rostros y filtró la información a través de una red neuronal para clasificar el sexo. Este trabajo hizo uso de histogramas de densidad proporcionados por los bordes de la imagen, los cuales servían de entrada a la, ya mencionada, red neuronal [153]. Lu exploró el rango y la intensidad de la información de los rostros humanos para la identificación de la etnia y del sexo haciendo uso de clasificadores SVM [140]. En este trabajo, en primer lugar, se discrimina sobre si un rostro es asiático, o no lo es, con el fin de explotar una relación entre la forma tridimensional del rostro analizado y su origen étnico. Los datos sobre los que trabaja esta investigación son modelos tridimensionales en los que el investigador seleccionó, de forma manual, las áreas relevantes para el procesado

Figura 2.15: La base de datos CAS-PEAL [79] suele ser usada en el reconocimiento de sexo. Dicha base de datos ofrece sujetos de ambos sexos en un entorno controlado y, las adquisiciones de datos se producen con semanas de diferencia para garantizar la diversidad de aspecto de los sujetos registrados.

de la imagen (véase Figura 2.14). Para ello, el investigador introduce una serie de puntos relevantes y luego se aplica una rejilla que sirve para recortar el área de la cara y así generar los vectores de características. Los resultados ofrecieron una tasa de aciertos del 91 % haciendo uso de la base de datos de la Universidad de Notre Dame (UND) y la colección de la Universidad Estatal de Michigan (MSU).

En la tabla 2.4 se ofrece una revisión sobre otros enfoques utilizados, todos ellos basados en la apariencia. En dicha tabla, también se ofrecen las técnicas utilizadas para la extracción de características, el tipo de clasificador empleado y la base de datos utilizada para el entrenamiento y la prueba del método.

Técnicas empleadas	**Clasificador**	**Datos para Entrenar**	**Datos para Pruebas**	**Tasa de Aciertos %**	**Ref.**
PCA	SVM + fusion temporal	Secuencias de vídeo	Imágenes locales	98.5	[38]
LBP, histograma	SVM polinomial	CAS-PEAL [79]	CAS-PEAL	94.96	[134]
Local Gabor Binary Mapping Pattern (LGBMP), histograma	SVM-RBF	CAS-PEAL	CAS-PEAL	94	[230]
Patch-based	Bayesiano	Imágenes de Internet	Imágenes de Internet	89	[4]
PCA	SVM-RBF	CAS-PEAL	CAS-PEAL	95.33	[139]
SIFT	Bayesiano	FERET [54]	Secuencias de vídeo	90	[60]
Comparación píxeles	Adaboost	FERET	Validación Cruzada	94.3	[12]
SIFT, Filtros de Gabor	Adaboost	Desconocido	Desconocido	97	[219]
LBP, bordes	SVM-lineal	FERET	FERET	99.07	[6]
Local Gabor Binary Pattern (LGBP)	SVM-RBF	CAS-PEAL	CAS-PEAL	95	[229]
LGBP-LDA	SVM-lineal	FERET y CAS-PEAL	FERET y CAS-PEAL	99.1	[241]
LBP, HOG	SVM-lineal	GROUPS [95]	GROUPS	89.7	[36]
LBP, PCA	SVM-lineal	Imágenes de Internet	Imágenes de Internet	87.2	[35]

Tabla 2.4: Métodos para la detección de sexo centrados únicamente en el rostro humano. Como puede apreciarse, los resultados ofrecidos son bastante buenos y, la mayoría de los experimentos, hacen uso de las bases de datos FERET (véase la Figura 2.3) o la de CAS-PEAL (véase la Figura 2.15). Las imágenes de ambas bases de datos están capturadas bajo condiciones lumínicas variables pero en todas ellas se cuida la pose, es decir, no existe oclusión, imágenes con múltiples rostros o rostros rotados o inclinados. En algunas entradas de la tabla no ha sido posible detectar la fuente de las imágenes empleadas.

Técnicas empleadas	**Clasificador**	**Datos para Entrenar**	**Datos para Pruebas**	**Tasa de Aciertos %**	**Ref.**
HOG	Adaboost	MIT-CBCL [100]	Validación Cruzada	75	[34]
HOG, histogramas de color	SVM-lineal	VIPeR [160]	Validación Cruzada	80.6	[41]
Descriptor de texturas locales	SVM-lineal	MIT-CBCL [100]	Validación Cruzada	80.6	[102]
PCA	SVM-RBF	Atributos de personas	Atributos de personas	82.4	[20]

Tabla 2.5: Métodos para la detección de sexo centrados únicamente en la ropa de los sujetos analizados.

2.4.1.2. Reconocimiento del sexo considerando elementos de contexto (pelo-ropa)

En cuanto al análisis de sexo basado en elementos contextuales, en la literatura se consideran dos opciones para definir el contexto de una persona; el pelo y la ropa. Ambas características permiten especificar, de forma aproximada, el sexo de una persona en virtud de la diversidad existente entre hombres y mujeres a la hora de definir estas características; por ejemplo, los hombres pueden ser calvos, las mujeres pueden llevar el pelo suelto, además es más probable que los hombres lleven corbata a que las mujeres la lleven y, por otro lado, es más probable que las mujeres vistan con falda o luzcan escote a que lo hagan los hombres. Existen estos clichés culturales que, si bien no son universales, pueden ser considerados como base para la clasificación del sexo [20]. Por otro lado, es posible que estos clichés culturales no siempre resulten eficientes ya que, normalmente, personas del mismo sexo suelen hacer uso de estilos dispares a la hora de vestirse [34]. Si a eso se le añade la posibilidad de que hombres y mujeres puedan llevar ropa similar, o cortes de pelo parecidos, entonces es necesario que el clasificador desarrollado se muestre robusto ante tales circunstancias y ante la posibilidad de posibles variaciones; pose o la presencia de elementos de oclusión (por ejemplo, un bolso).

Dentro del análisis del cuerpo como elemento contextual para definir el sexo, en la tabla 2.5 se pueden encontrar unos enfoques bastante prometedores. El trabajo de Cao ofrece un interesante estudio en el que se utilizó un conjunto de imágenes centradas y con la altura normalizada para, posteriormente, dividir la imagen en varias partes procesables [34]. El descriptor utilizado en este trabajo fueron los histogramas de gradientes orientados (*Histograms of Oriented Gradients (HOG)*) el cual había sido desarrollado por Dalal [44] con otro fin; la detección de seres humanos en imágenes (véase Figura 2.16). Una de las

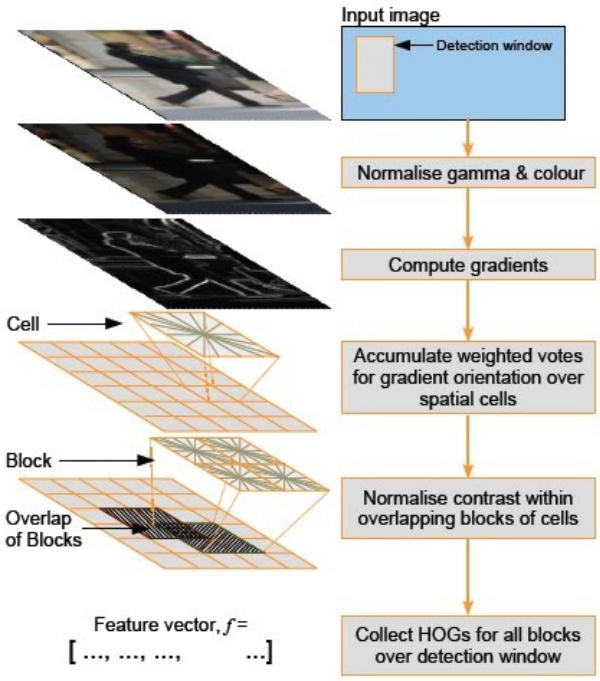


Figura 2.16: Para obtener las características de HOG, en primer lugar es necesario obtener una imagen del gradiente a través de la aplicación de una máscara unidimensional con las forma [-1, 0, 1] de forma vertical y horizontal. Después, dicha imagen es dividida en celdas y, para cada celda se obtiene un histograma de gradientes orientados. A su vez, las celdas son agrupadas en bloques que ofrecen un contraste normalizado y entre estos bloques suele haber solapes de celdas. El histograma obtenido de un bloque es concatenado para poder formar vectores de características de la información contenida en la imagen [44].

principales ventajas que ofrece el HOG es que es capaz de capturar, de forma eficiente, la información de cada parte de la imagen dividida haciendo uso del gradiente, el cual, es invariante frente a translaciones y rotaciones [44].

De hecho, después del trabajo de Cao, Collins ofreció una variación del HOG para este mismo estudio; el *PixelHOG* hace uso de características de densidad de los HOG obtenidos de una imagen de bordes. Además, la información del color también se tuvo en cuenta a través de los histogramas obtenidos de los valores de tonalidad y saturación de los píxeles que conforman la imagen [41]. Por su parte, Guo [102] desarrolla un modelo de características de inspiración biológica (*Biologically Inspired Features (BIF)*) que está basado en la aplicación de filtros de Gabor haciendo uso de doce escalas distintas y ocho orientaciones por escala.

Bourdev propone una serie de parches denominados *"poselet"* para deter-

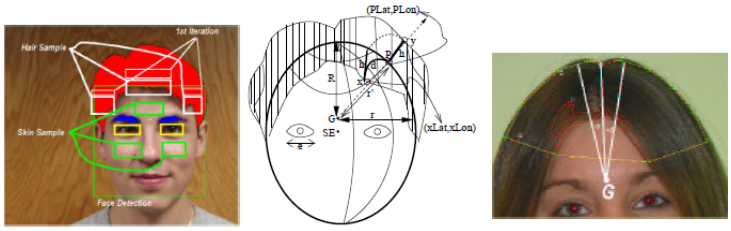
minar atributos en entornos no controlados [20]. Las características proporcionadas por los HOG, los histogramas de color y características de la piel son determinantes a la hora de definir los poselets. De esta forma, para entrenar el clasificador se utiliza como parámetro de entrada los propios poselets. Este enfoque es robusto frente a cambios de pose y oclusión pero requiere una fase de entrenamiento con datos que contienen una información muy detallada a través de anotaciones manuales del cuerpo humano.

De estos experimentos resulta interesante remarcar que se experimentó, únicamente, con las bases de datos del MIT [100] y la de VIPeR [160] las cuales sólo contienen imágenes de personas en posición vertical (no hay sujetos ni acostados ni en posiciones que permitan la rotación parcial del cuerpo con respecto a la cabeza). Existen otras bases de datos que suponen un verdadero reto para este tipo de estudios pero que aún no han sido aplicadas para la temática, como por ejemplo la PASCAL [77] que contiene imágenes de personas con una mayor variedad de poses. La base de datos del MIT sólo proporciona imágenes frontales y de espalda de personas, mientras que la VIPeR añade la vista lateral a las dos poses que ofrece la base de datos del MIT.

Por último, el problema puede ser abordado con el fin de contribuir a la mejora de un sistema biométrico en función de características como la distancia entre el dispositivo de adquisición y el sujeto. Recientemente, Fierrez obtuvo notables resultados a través del uso de las "etiquetas blandas" (*soft labels*), que eran clasificadas en tres categorias con el objetivo de analizar distintos aspectos de una persona [208]. La primera categoría es el cuerpo e identifica datos relativos al aspecto físico de la persona (longitud brazos o de piernas, proporciones, caderas, etc.). La segunda categoría hace referencia al aspecto global de la persona y permite determinar datos acerca del sexo, la raza o la edad. Por último, la categoría de la cabeza define aspectos como color de piel, color del cabello, grosor del cuello, etc.

Con respecto al uso del pelo para determinar el sexo de las personas, la complejidad para detectar la región capilar supone en sí una dificultad intrínseca de la propia investigación. Ha de tenerse en cuenta múltiples factores que pueden dificultar la obtención de la región del pelo de una persona; en primer lugar, la región puede ser variable en cuanto a posición debido a la pose y, en segundo lugar, la enorme diversidad de los tipos de cabello puede complicar su identificación si el fondo de la imagen no está normalizado. La pose cobra especial relevancia debido a múltiples razones; las imágenes tomadas desde un ángulo superior al sujeto o desde un ángulo inferior pueden aumentar o disminuir la región de pelo respectivamente. Además, el pelo puede ser no uniforme combinando colores o bien alternando regiones de pelo y de calvicie. Es por ello que la diversidad de los tipos de pelo en cuanto a color, forma o textura pueden dificultar su identificación en entornos no controlados. Por todo lo comentado anteriormente se puede deducir que el cabello, pese a ser propio del ser humano y menos variable que la ropa, no deja de ser una característica muy compleja de la apariencia humana. Por este motivo, el cabello ha sido descartado en múltiples ocasiones a la hora de considerar el sexo de una persona [4, 12, 103].

En la literatura existen dos trabajos especialmente relevantes en el trata-



(a) (b) (c)

Figura 2.17: Obtención de las características capilares según Yacoob [231]. En primer lugar se identifica el rostro y se recogen datos relativos a una posición segura de la piel y del pelo además de aislar las regiones de interés. Posteriormente se recorren dichas regiones de forma que se identifica la piel y el pelo del sujeto. Por último se calcula, en base a la posición relativa de distintos puntos, las características geométricas y de apariencia del cabello.

miento del cabello a través de técnicas de segmentación. Según diversos autores, actualmente existen dos algoritmos acerca de la representación de características del cabello que han tenido una notable aceptación [116]. En primer lugar, Yacoob desarrolló un modelo de medición de la apariencia del cabello. En este enfoque se combina la geometría y la apariencia con el objeto de reunir un conjunto de características como color, volumen, área, simetría, longitud o textura, con el objeto de definir los descriptores que se ajustan de forma eficiente a las características del pelo del sujeto [231]. La salida de este modelo genera un vector de características del pelo (véase la Figura 2.17 para apreciar la obtención de los datos sobre el pelo). Por otro lado, Lapedriza propone un modelo basado en conjunto de bloques que se corresponden con fragmentos de imágenes representativas de zonas del pelo [128]. La idea básica es que el conjunto de bloques se superponen sobre una imagen vacía para reconstruir la imagen inicial y para ello se van colocando de forma que cubren aquellas coordenadas de la imagen original que les son más parecidas al bloque que se coloca en dicha región, reconstruyendo la zona "parecida" a través de estos bloques y detectando así, la región capilar del sujeto.

En lo referente a trabajos interesantes sobre la detección de sexo haciendo uso del pelo de una persona, si bien no existen investigaciones que hagan un uso exclusivo del cabello, si existen trabajos que combinan pelo, cara o ropa. Además, trabajos como el de Bing Li [131], ofrecen una estadística bastante notable sobre qué partes del ser humano son relevantes para la detección de sexo y en qué medida. En su trabajo, Li genera hasta siete clasificadores distintos (para pelo, frente, ojos, nariz, boca, barbilla y parte superior del pecho (incluyendo cuello) respectivamente) que, posteriormente, serán combinados a la hora de tomar una decisión sobre si el sujeto es hombre o mujer (véase el proceso en la Figura 2.18a). La bases de datos que utilizó fueron la FERET [54], la AR [52] y una base de datos propia en la que las imágenes tienen un contexto bien definido con rostros frontales y donde los sujetos están ubicados

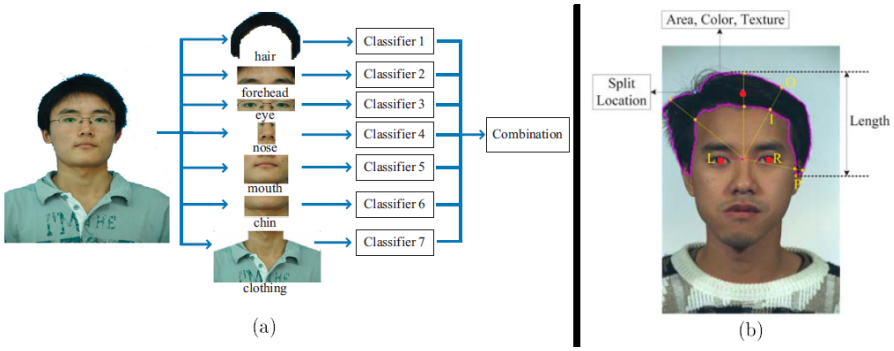


Figura 2.18: En la Figura (a) puede observarse la combinación de los distintos clasificadores propuesto por Bing Li [131] a partir de la extracción y el procesado, por separado, de cada una de las partes. Al final, se aplican distintas técnicas de fusión de clasificadores para determinar la decisión del sexo del sujeto. Por su parte, en la Figura (b) puede apreciarse la extracción de las características del pelo a partir de la ubicación de 3 puntos, uno en cada ojo y otro en la región central dónde exista cabello [116]. Además, se pueden observar las propiedades que considera el autor para formar el vector de características del pelo (longitud, area, colo, textura y punto de la raíz capilar).

sobre un fondo normalizado. Cada uno de los clasificadores es una máquina SVM y el autor experimenta con los distintos núcleos que ofrece el SVM para determinar el resultado óptimo. Para FERET, los mejores resultados obtenidos se consiguen con el núcleo de base radial. Considerando únicamente el pelo, el porcentaje de acierto es del 80.6 %. Combinando la información de los siete clasificadores antes mencionados mediante una técnica de lógica difusa, la tasa de aciertos se eleva hasta el 93.4 %. Otro trabajo relevante en el área que utiliza de forma explícita el uso del cabello en combinación con la información facial es el ofrecido por Ji [116]. Haciendo uso de las bases de datos FERET [54], AR [52] y CAS-PEAL [79], propone un enfoque en el que, a través de unos puntos introducidos de forma manual por el investigador, se estiman las posiciones del pelo y la piel sobre un fondo normalizado (véase Figura 2.18b). Posteriormente, el autor genera dos clasificadores, una con la información del pelo y otra con la información de la cara. A diferencia del trabajo de Li [131], Ji trata la cara como un conjunto inseparable. Los resultados de Ji, procesando exclusivamente el pelo, sobre el conjunto de datos FERET es de una tasa de acierto del 84.16 % que, en combinación con el clasificador facial (92.67 % de acierto), la tasa de aciertos se eleva hasta el 94.13 %. La fusión de los clasificadores se hizo a través de un enfoque basado en lógica difusa. Este autor expuso tres conclusiones fundamentales en su trabajo; que el pelo juega un papel importante cuando se trata de clasificar el sexo de una persona, que el rostro es más importante que el pelo a la hora de realizar esta tarea y que la combinación de ambas técnicas es mejor que los resultados individuales de cada una de ellas.

Capítulo 3

Del origen a la detección de la expresión facial

Como se ha descrito en el capítulo 2, las emociones suelen emerger de forma natural entre las personas que comparten una interacción social a través de los distintos canales de comunicación existentes. Estos canales (voz, gestos, etc.) vienen determinados por ciertos rasgos inherentes al propio canal, por ejemplo, en el caso de la voz destacan el tono, la frecuencia o el timbre. Por contra, en el caso de los gestos destacan el tipo de expresión facial o el comportamiento del lenguaje corporal. Si bien es posible que las personas no pretendan mostrar sus verdaderos sentimientos (o sensaciones) mientras están comunicándose, lo cierto es que las expresiones faciales afloran de forma inconsciente y dejan entrever lo que se está sintiendo en un momento determinado. Un claro ejemplo del descontrol que tienen los humanos con respecto a este tipo de expresiones lo conforman las partidas de póker profesionales; en dichas partidas queda permitido que los participantes puedan cubrir su rostro de forma parcial (gafas), su pelo, sus orejas. El motivo es que la respuesta del organismo humano ante una determinada circunstancia no responde de forma directa a la voluntad de las personas y en el póker se admite este hecho de forma natural, a través de la adaptación de su reglamentación.

En este capítulo se pretende abordar la temática de las expresiones faciales desde su génesis hasta su detección. Para ello, la primera parte del capítulo propone la creación de una base de datos de expresiones faciales denominada Vedana. Dicha base de datos es una colección de vídeos en las que se recogen las expresiones faciales de distintos sujetos a través de un estímulo audiovisual. Precisamente, este estímulo es un mecanismo de inducción que facilita la captación de una expresión facial genuina. Por otro lado, en la segunda parte del capítulo, se expone un compendio de técnicas de la literatura reciente, además de nuevas propuestas, cuyo propósito es detectar una de las expresiones más universales que existen: la sonrisa. En la sección 2.3.2 ya se argumentó acerca de las cualidades inherentes que tiene la sonrisa humana que, además, se trata

de una expresión en la que existe poco debate interpretativo sobre el motivo de su manifestación. La sonrisa humana se ha convertido en un tema recurrente a la hora de desarrollar ciertas aplicaciones informáticas orientadas a recoger el estado de ánimo de grupos de personas. Como ejemplo ilustrativo, en los últimos años, el MIT ha desarrollado un sistema que permite la detección de sonrisas entre las cámaras ubicadas en sus distintas facultades [148], lo que les ha permitido desarrollar lo que se ha denominado *"mood meter"* y generar un "mapa de la felicidad" del campus.

3.1. Generación de expresiones faciales

Determinar la génesis de las expresiones faciales requiere del análisis de las emociones que experimentan las personas ante un estímulo. Este estímulo puede ser externo, como respuesta ante una circunstancia externa al sujeto, o bien puede ser interno, como resultado de una reflexión del sujeto ante un cúmulo de posibles circunstancias que puede haber, o no, ocurrido. Fisiológicamente, las emociones están ubicadas en el sistema límbico. Este sistema es la porción del cerebro situada inmediatamente debajo de la corteza cerebral y comprende regiones importantes de otras zonas cerebrales como el hipotálamo, el tálamo, el hipocampo o la amígdala cerebral [154].

Diversos autores han afirmado que lo que se conoce comúnmente como "emoción" no está relacionado con un proceso cerebral aislado e independiente, sino que es el resultado de la interacción de múltiples mecanismos cerebrales que pueden diferir en función del tipo de emoción que experimente el sujeto [98]. De esta forma, es importante considerar que los componentes conscientes de las emociones, sentimientos como la alegría o el miedo, no son, desde una perspectiva cognitiva, diferentes procesos a lo que podría suponer para el cerebro resolver una ecuación o la percepción de que el documento que usted está leyendo es una tesis doctoral.

Por tanto, generar expresiones faciales es un proceso que se debe desarrollar a través de la intermediación de una persona. Esta afirmación resulta esencial por dos motivos; el primero es que, precisamente esta persona intermediaria, es la que debe exteriorizar la expresión facial y, el segundo motivo es que se debe activar el sistema límbico con el objetivo de que la persona en cuestión muestre alguna expresión facial. Por ejemplo, una canción puede generar nostalgia en unas personas y alegría en otras, todo depende de las características personales y de las experiencias vitales que haya experimentado el sujeto que la escucha. Memoria, aprendizaje y emociones están relacionados en la activación de este tipo de procesos [67]. La explicación fisiológica es que la amígdala cerebral, que juega un papel bastante relevante en el procesamiento de las emociones, guarda una relación esencial con otros procesos cognitivos. Fulton y Jacobsen demostraron que los pacientes que tienen esta región del cerebro lesionada no eran capaces de reconocer una expresión facial y tendían a perder habilidades sociales con las personas que les rodeaban [94]. Además, estos investigadores afirman que la amígdala también participa de forma activa en el aprendizaje y

la memoria, y lo confirmaron a través de un experimento; colocaron dos cuencos frente a varios chimpancés, unos con la amígdala dañada y otros que la tenían intacta. Uno de los cuencos tenía comida mientras que el otro estaba vacío. Posteriormente, taparon los cuencos (sin cambiarlos de lugar) y les pidieron que eligieran uno. Los chimpancés que tenían la amígdala intacta seleccionaron aquellos que tenían comida mientras que los que la tenían dañada basaron su elección de forma aleatoria. Fulton y Jacobsen llegaron a la conclusión de que no eran capaces de elegir de forma adecuada porque la comida no había producido ninguna excitación en estos chimpancés, por lo que afectó a su aprendizaje y a su memoria [115].

Provocar una expresión facial en una persona es una tarea compleja. En la sección 2.2.1 se explica, desde una perspectiva psicológica, el conjunto de procesos que puede experimentar una persona a la hora de generar las emociones como respuesta a una circunstancia determinada. Además, en dicha sección se especificó que sólo existen tres formas posibles para poder obtener una expresión facial de una persona; la primera ocurre a través de una actuación o de forma fingida, la segunda es de forma natural (sin que el sujeto sea consciente de que está siendo grabado) y la tercera se produce a través de un mecanismo de inducción (mediante la experimentación con las emociones del sujeto a través de la exposición de contenidos multimedia). En dicha sección se explicaron las ventajas y desventajas de cada enfoque y quedó de manifiesto que la inducción de emociones permite captar de forma verosímil la esencia de las expresiones al mismo tiempo que permite salvaguardar la privacidad de las personas, situación que no sucede en los contextos de captura de expresiones de forma natural. Sin embargo, conseguir transmitir de forma eficaz un contenido que permita aflorar expresiones faciales no sólo depende del contenido, sino que también depende de la preparación de un entorno adecuado para dicha tarea. Vedana es una base de datos que ha sido desarrollada cuidando los tres aspectos fundamentales de la inducción: entorno, contenido multimedia y dispositivos de adquisición de información.

3.1.1. Configuración experimental de Vedana

La base de datos Vedana ha sido desarrollada con el objetivo de disponer de un conjunto de datos que permita evaluar y comparar distintas técnicas para la detección de expresiones faciales. Si se consideran las características fundamentales de un proceso de inducción de expresiones faciales, se llega a la conclusión de que la preparación del experimento entraña un elemento crítico dentro de esta parte de la investigación. De hecho, el entorno debe aislar a los sujetos para que puedan centrarse, al máximo, en los materiales que están visualizando y, además, el proceso debe estar envuelto en un clima de tranquilidad con el fin de permitir que las emociones fluyan de la forma más natural posible. Además se busca provocar, no sólo un conjunto de expresiones faciales a través de distintos vídeos, sino también integrar diversidad de datos en los que cohabitan diversas poses o distintas expresiones, además de permitir la presencia de accesorios.

Para el entorno se habilitó un espacio en el sótano del SIANI [174] en el que

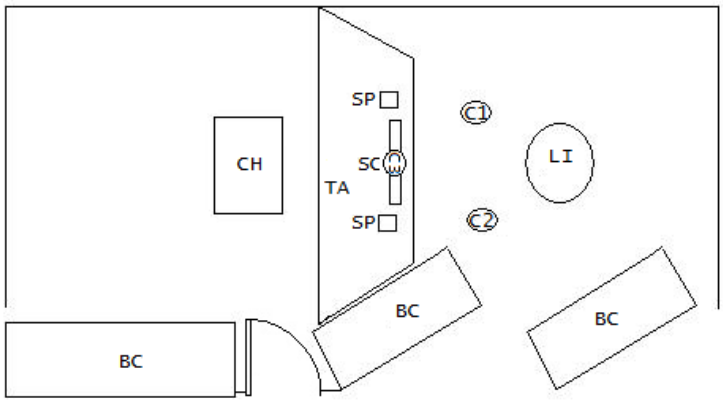


Figura 3.1: Esquema básico del estudio de grabación. La nomenclatura se resume en un foco de luz con difusor (LI), tres cámaras (C1, C2 y C3), una silla (CH), una mesa (TA), dos altavoces (SP), una pantalla (SC) y tres armarios (BC).

Figura 3.2: Vista externa del estudio de grabación. Como puede apreciarse, existe una entrada principal para los participantes que está cubierta con una cortina blanca. Además, existe una segunda entrada que permite la manipulación de las cámaras no controladas por el equipo informático de forma automática.

Figura 3.3: Vista interna del estudio de grabación. A la derecha puede apreciarse una vista frontal (lo que el participante visualiza durante el experimento) del estudio, mientras que en la izquierda se muestra una vista trasera (qué es lo que hay detrás del participante) del mismo.

se preparó un pequeño habitáculo para poder realizar la captura. En la Figura 3.1 puede apreciarse la distribución de elementos a lo largo de todo este pequeño estudio de grabación. Dicho estudio tiene unas medidas de 4,33 metros de largo por 1,99 metros de ancho y 1,97 metros de altura. Además, se ha configurado el equipo de grabación para poder capturar las expresiones faciales desde distintos puntos de vista. Tres cámaras fueron utilizadas para grabar, de forma simultánea, la reacción de los usuarios mientras estos visualizaban contenido multimedia a través de la pantalla. Los ángulos entre la cámara frontal (C3) y las cámaras laterales C1 y C2 son 50 y 41 grados respectivamente. Resulta interesante remarcar que la organización del habitáculo se desarrolló de acuerdo con posibles circunstancias que se podían presentar durante el proceso de adquisición de imágenes. Por ejemplo, no existió un mecanismo para la activación de las tres cámaras de forma simultánea, pues sólo una de ellas, la C3, estaba conectada al equipo informático. Por este motivo, en la Figura 3.2 puede apreciarse que se habilitaron dos entradas al habitáculo; la entrada principal para el acceso de los usuarios al habitáculo, mientras que una segunda entrada permitió la manipulación directa de las cámaras C1 y C2 al comienzo de cada sesión. En cuanto a la configuración de la iluminación, el habitáculo estaba cubierto por armarios y una cortina que permitían concentrar la luz dentro del entorno además de aislar a los usuarios ante cualquier posible distracción. Por último, con el objetivo de que el fondo de la imagen permaneciera de forma uniforme, se forró todo el estudio de grabación con cartulinas blancas (véase la Figura 3.3).

Un aspecto relevante del proceso de adquisición radica en las características del equipo utilizado durante el proceso de grabación. Como se ha comentado anteriormente, durante el proceso de grabación se hizo uso de tres cámaras digitales identificadas como C1, C2 y C3. Las cámaras C1 y C2 estaban sustentadas por sendos trípodes de aluminio para estabilizar el proceso de captura de imágenes, mientras que la cámara C3 estaba integrada en el monitor. Sobre las características de cada cámara resulta interesante que, independientemente de

Esp. Técnicas	Camera C1	Camera C2	Camera C3
Imagen			
Caract. de vídeo			
Resolución	480x360	1920x1080	640x480
Tasa de bits	7129 bps	3419 bps	1905 bps
Frecuencia	210 fps	30 fps	30 fps
Aspect Relation	4:3	16:9	4:3
Compresión de vídeo	IBM Motion JPEG	H.264/AVC	MPEG4
Caract. de audio			
Tasa de bits	-	448 bps	-
Canales	-	5	-
Frecuencia	-	48000 Hz	-
Formato	-	AC3	-

Tabla 3.1: Especificaciones técnicas del equipo de grabación. Los casos marcados como "-" hacen referencia a condiciones no existentes.

la posición que ocupaban en el plano, fueron configuradas de forma diferente. La cámara Sony Handycam HDR-XR520VE (etiquetada como cámara C2 en la Figura 3.1) realizó capturas de 30 fotogramas por segundo de 24 bits de color con una resolución de 1920x1080. Por su parte, la cámara etiquetada como C3, la Logitech Webcam C250, también grabó a 30 fotogramas por segundo de 24 bits de color pero con una resolución de 640x480. Por último, la Casio Exilim EXFH20 (etiquetada como la cámara C1), capturó 210 fotogramas por segundo de 24 bits de color, con una resolución de 480x360, en otras palabras, ésta última cámara capturó imágenes siete veces más rápido que las otras dos. Con esta última configuración se pretende dar respuesta a la necesidad de capturar microexpresiones faciales. La única desventaja de usar esta configuración en la cámara C1 radica en el aumento del tamaño del fichero resultante; si la duración de un vídeo, considerando 30 fotogramas por segundo, de una sesión es de 7 minutos y 36 segundos para las cámaras C2 y C3, la cámara C1 genera un fichero con una duración de 53 minutos y 53 segundos.

El segundo aspecto relevante es determinar las condiciones del entorno de grabación, así como la predisposición de los dispositivos de adquisición durante los mismos. Normalmente, en las bases de datos faciales descritas en la sección 2.2.1 se toman imágenes considerando diversos aspectos como iluminación, expresión, edad, etc. En el caso de los vídeos, muchos de estos aspectos se multiplican debido al factor temporal que, de forma inherente, incluye el propio

vídeo. Por un lado, en el modelo de grabación propuesto existen diferentes aspectos del proceso de grabación que condicionan la calidad del resultado del proceso de adquisición de imágenes: movimiento, pose, resolución de la imagen, tasa de captura y variaciones temporales. Tanto el movimiento como la pose fueron dejadas a libre disposición de los sujetos. Con el fin de poder adquirir a distintas resoluciones así como a distinto tasa de captura a lo largo de las distintas sesiones de grabación, se hizo uso de las tres cámaras anteriormente mencionadas (C1, C2 y C3). Las variación temporal se tradujo en la adquisición de dos sesiones por sujeto, con una diferencia de días entre la primera y la segunda sesión de cada sujeto. La iluminación es un factor estático durante todo el proceso de adquisición: una lámpara (LI) de grabación con un difusor. Así, considerando esta configuración de elementos, el proceso de adquisición de expresiones faciales se resume de la siguiente forma; se le presenta al sujeto, que está sentado frente a las cámaras, un contenido multimedia (pequeños vídeos) a través de una pantalla de 20 pulgadas. Estos vídeos de corta duración pretenden generar estados emocionales en los sujetos y que éstos estados se traduzcan en diversas expresiones faciales.

Además, con el fin de simular condiciones del mundo real, cada usuario pudo hacer uso de los accesorios que estimase oportunos. La mayoría de los accesorios proporcionados por los sujetos se resumen en gafas (de sol o de vista, con distintas características como el tipo de cristal, de montura, etc.) y en gorros (siendo las capuchas lo más utilizado). Por otro lado, todo el proceso de grabación se extendió a lo largo de 4 semanas, completando tres turnos de grabación por semana.

Volviendo al tema del tipo de grabación realizada en cada sesión, resulta indispensable explicar en qué consistió. El primer proceso, que ya ha sido explicado anteriormente, se basa en la grabación de expresiones faciales, mientras que un segundo proceso grabó la parte superior del cuerpo de los sujetos mientras elevaban y bajaban los brazos de forma síncrona. Este segundo proceso pretende aprovechar la oportunidad que otorga la situación para enriquecer la base de datos con otro tipo de datos útiles para la visión por computador, como por ejemplo, la detección de ropa. A continuación se describe de forma precisa el procedimiento seguido para cada uno de estos experimentos.

En primer lugar, se pidió a todos los participantes que firmaran un acuerdo para la cesión de derechos de imagen al grupo de investigación. Para el proceso de adquisición de expresiones faciales se siguió el siguiente protocolo:

1. Antes de comenzar la sesión de grabación, la silla se ajusta de forma que el sujeto queda situado horizontalmente alineado con la cámara C3 a una distancia de 80 centímetros (véase la Figura 3.4). Con fines estadísticos, el sujeto rellena, a través del software de gestión de contenidos, una encuesta en la que se recogen datos acerca de su aspecto así como de su edad, su sexo o forma de contacto.

2. Con el objetivo de sincronizar las tres cámaras y, dado que las cámaras C1 y C3 no tienen línea de audio, se solicita a cada participante que asienta

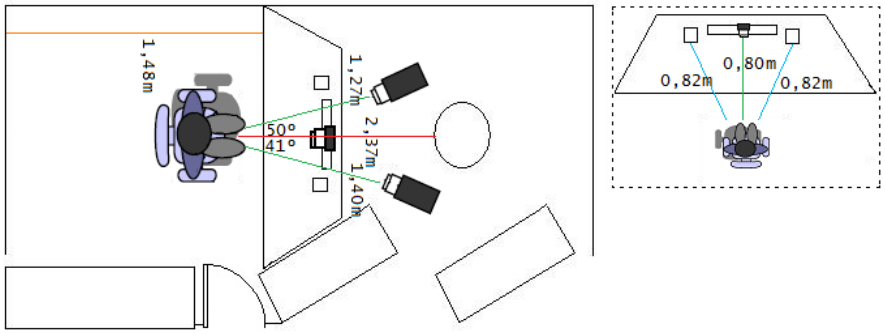


Figura 3.4: Configuración para el proceso de captación de expresiones faciales. Cada sujeto debe visualizar una serie de 17 vídeos de forma continua en las que, cada secuencia de vídeo, tiene una duración media de 20 segundos. Como puede apreciarse, la distancia entre el sujeto y el monitor es de 80 centímetros y resulta destacable la distancia entre el sujeto y la pared trasera. Guardar esta distancia permite asegurar la no existencia de sombras durante el proceso de grabación.

con la cabeza inmediatamente después de escuchar el pitido generado por el software.

3. Este pitido ocurre con posterioridad a que el técnico haya activado las cámaras C1 y C2, y pasados tres segundos de que el participante haya seleccionado la casilla de comienzo de la prueba.

4. La única restricción que tuvieron los participantes durante la duración de la prueba fue la prohibición de taparse el rostro mientras visualizaban los materiales multimedia. El estímulo para la generación de las expresiones faciales lo compuso una serie de 17 cortas secuencias de vídeo que formaban parte de una base de datos mucho mayor (temáticamente clasificada según el contenido de cada secuencia de vídeo, véase la tabla 3.2) y donde la secuencia mostrada era seleccionada de forma aleatoria por parte del software de gestión de contenidos. Además, antes del inicio de cada vídeo, se muestra una imagen alterada a través de una serie de transformadas de Fourier que se correspondían con el último fotograma del vídeo previamente reproducido. El objetivo de mostrar estas imágenes, denominadas RISE ("*Random Image Structure Evolution*"), durante tres segundos, entre las distintas secuencias de vídeo, no es otro que eliminar los efectos a bajo nivel de los vídeos en la retina [180]. Más adelante se describe con mayor precisión en qué consisten este tipo de imágenes intercaladas entre la exposición de estímulos. Los vídeos mostrados tienen una duración máxima de unos 20 segundos y la prueba dura en torno a 8 minutos, mostrados en una pantalla de 20 pulgadas bajo una resolución de 1024x768 píxeles.

Por todo lo comentado anteriormente se desprende que el objetivo era que

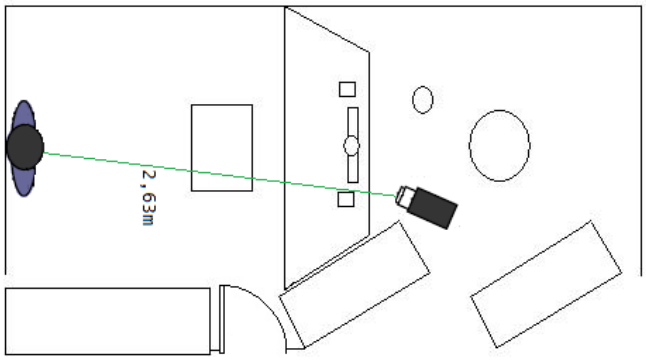
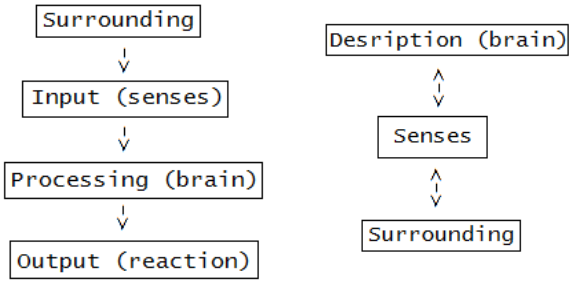


Figura 3.5: Configuración para el experimento de captación del torso. El experimento sólo fue grabado desde la cámara C2 y la duración aproximada de este experimento es de 8 segundos.

los participantes del experimento, al visualizar las distintas secuencias de vídeo, tuvieran la misma percepción que si estuvieran viendo el televisor en sus casas. Así, sería más probable poder captar la reacción natural de sus expresiones faciales y, sobre todo, hacerlo de la forma más precisa posible. Sin embargo, también se realizó un segundo experimento para la cual sólo se grabó desde un único ángulo a través de la cámara C2 (véase la Figura 3.5). Este experimento tuvo lugar después de la adquisición de expresiones faciales con el fin de no desconcentrar al sujeto antes del primer experimento tratando de minimizar la interacción con el técnico de laboratorio antes de la primera prueba en la que la concentración del sujeto es fundamental. El procedimiento para el segundo experimento es bastante sencillo, en primer lugar, se solicitó al participante que se levantara de la silla y se pusiera de frente a las cámaras con la espalda pegada a la pared y los brazos relajados, dejando caer ambas manos a la cintura. En ese momento comenzaba el proceso de grabación y el participante debía elevar los brazos de forma progresiva hasta alcanzar un ángulo de 180 grados para luego volver a bajar los brazos, al mismo ritmo, hasta el punto de partida.

El conjunto de datos elegido para que los usuarios visualicen durante la prueba de grabación de expresiones faciales es la tercera piedra angular en la creación de esta base de datos. Como ya se ha comentado anteriormente, esta base de datos está englobada dentro de las bases de datos de expresiones de tipo inducido y, por esta razón, la cuestión del conjunto de datos que se debe mostrar a los participantes es de vital importancia. El objetivo es hacer llegar la información al cerebro y activar las regiones adecuadas para poder establecer una correlación entre el sentimiento del participante y su expresión facial. Quizás, la primera cuestión de todas sea determinar por qué se ha seleccionado la secuencia



(a) Percepción directa (b) Percepción indirecta

Figura 3.6: Esquemas de la percepción [75].

de vídeo como el estímulo ideal para este experimento. La percepción se define como un proceso activo de adquisición, interpretación, selección y organización de la información sensorial como respuesta a un fenómeno específico (estímulo) en forma de sonido, evento, temperatura o gusto [75]. En resumen, la percepción es una búsqueda dinámica de la mejor interpretación posible ante algo que experimenta un sujeto. Precisamente, la ciencia cognitiva es un estudio interdisciplinario del esquema cerebral desde los campos más relevantes incluyendo la psicología, la filosofía, la neurociencia, la lingüística, antropología, ciencias de la computación, la biología y la física. De hecho, esta ciencia define dos posibles tipos de percepción; directa o indirecta [75]. En la percepción indirecta (*"top down perception"*) de toda la información que llega al ojo humano, el 90 % se pierde antes de alcanzar el cerebro. En este tipo de percepción, queda de manifiesto que el conocimiento previo y la experiencia del sujeto juegan un papel relevante. De hecho, Ernst afirmó que la percepción humana es una relación dinámica entre la descripción que llega al cerebro, los sentidos y el entorno del sujeto (véase Figura 3.6b). Este enfoque insta a que, para desarrollar una buena percepción de la información sensorial, es necesario combinar esta información con la experiencia del sujeto y, además, indica que la formación de una interpretación incorrecta da lugar a errores de percepción (ilusiones ópticas). Por otro lado, en la percepción directa (*"bottom up perception"*), no existe necesidad de interpretar la información de entrada pues ésta información que recibe el sujeto es suficiente como para que pueda interactuar de forma directa con el entorno. En este caso, la información sensorial se analiza en una dirección, comenzando por el análisis de los datos sensoriales de entrada, desde los más simples hasta las tareas más complejas (véase Figura 3.6a). De hecho, en la percepción directa, se proporciona suficiente información como para ayudar a la percepción humana.

Supóngase un ejemplo, una imagen de un vehículo, con todo lujo de detalles, y este vehículo está ubicado en una tarima con un fondo totalmente oscuro y, una luz, lo ilumina desde la parte superior de la escena. En base a lo comentado anteriormente sobre los tipos de percepción, la persona lo observa y determina que es un vehículo en función de la información recibida a través de los ojos y de

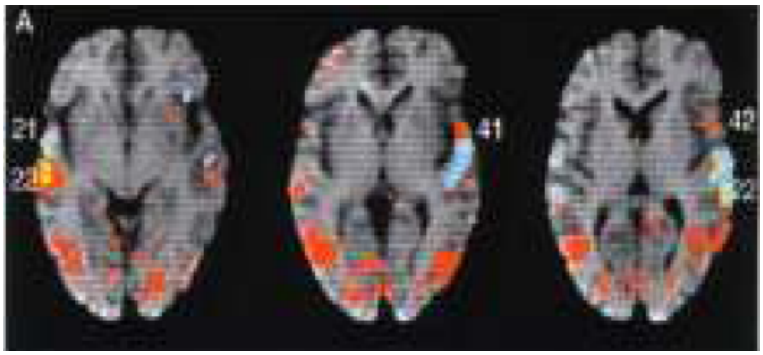


Figura 3.7: A través del uso de imágenes por resonancia magnética funcional (IRMf) los científicos fueron capaces de determinar cómo reaccionaba el cerebro humano ante la aplicación de distintos tipos de estímulos como auditivos, visuales o audiovisuales [126]. En esta imagen se pueden apreciar tres imágenes cerebrales con tres colores bien diferenciados; el color rojo se corresponde con la reacción cerebral de una persona que visualiza los labios de otra persona pero sin sonido, el color celeste se corresponde con que la persona sólo es capaz de escuchar al interlocutor pero no lo puede ver y, por último, el color naranja es la reacción que se produce cuando la entrada de datos en el cerebro es de tipo audiovisual.

la experiencia del observador (recuérdese que el vehículo está muy logrado, con todo lujo de detalles). Sin embargo, ¿no podría ser una maqueta de un vehículo hábilmente enfocado con una cámara para dar un efecto de realismo? Efectivamente, el observador hace una interpretación pero no tiene toda la información a su alcance para discernir si lo que ve es real o una ilusión óptica. Si se cambia la escena y, en lugar de una imagen con un fondo oscuro, el observador es trasladado a un espacio abierto en el que puede apreciar a una familia entrando en el vehículo, escucha a los niños, etc. entonces se produce una percepción directa, pues se recibe suficiente información como para que el observador interactúe con el entorno y pueda entender que lo que observa no es una maqueta.

Lo mismo ocurre con los vídeos y el entorno que se prepara para que los sujetos generen emociones a través de sus percepciones y que, estas emociones, se traduzcan en expresiones faciales. La selección de los materiales debe completar la percepción del individuo y, si no la hace directa, debe permitir al participante involucrarse de forma emocional con lo que está observando. Las imágenes estáticas limitan la percepción sensorial, haciéndola unimodal (sólo a través de la vista). Para una percepción directa es necesario que el cerebro sea capaz de extraer todas las características inherentes al objeto observado; por ejemplo, el cerebro requiere de la textura, la sombra e, incluso, del movimiento de un objeto para poder determinar su tamaño [126]. Lo que ocurre, es que el cerebro hace uso de todas las posibles fuentes de información comentadas para cubrir con la información de unas, la carencia de otras. Obviamente, el movimiento no es

Categoría de Video	Núm. de muestras	Palabras clave
Animales - Positivo	12	mascotas
Animales - Negativo	30	maltrato, cucarachas
Catástrofes	17	vuelo, tornados
Gente - Bebés	7	felices
Gente - Hombres	8	hablando
Gente - Mujeres	14	hablando, bailando
Políticos - Positivo	4	divertido
Políticos - Negativo	15	mentiras
Sexo - Bizarro	14	músculos, transexual
Sexo Mujer - Heterosexual	24	sexy
Sexo Mujer - Homosexual	22	lesbianas
Sexo Hombre - Heterosexual	11	cuerpos
Sexo Hombre - Homosexual	12	gay
Deportes - Positivo	3	F1, fútbol
Deportes - Negativo	8	lesiones
Naturaleza	22	plantas
Gente - Bostezos	7	bostezo, dormir

Tabla 3.2: Categorías de las secuencias de vídeo.

apreciable a través de imágenes estáticas y de ahí la limitación de la información de entrada. En la Figura 3.7 puede apreciarse cómo se traduce la riqueza en la información del canal de entrada a través de las distintas respuestas del cerebro, lo que justifica el uso de secuencias de vídeo como elemento inductor de emociones.

En cuanto a las propias secuencias de vídeo seleccionadas, se resumen en un total de 230 secuencias descargadas de Internet y editadas antes del experimento. Esta edición se basó fundamentalmente en la eliminación de información redundante así como en asegurar la temporización de la prueba en los límites justos para optimizar la respuesta de los participantes. Estas 230 secuencias de vídeos se clasificaron en las 17 categorías descritas en la tabla 3.2 y, aunque, la selección del vídeo a mostrar era aleatoria dentro de cada categoría, al menos se mostraba un vídeo de cada categoría a los participantes. Por tanto, cada participante visualizó un vídeo perteneciente a cada una de estas categorías siguiendo el mismo orden expuesto en la tabla 3.2. Este orden tiene un fin bastante concreto; controlar que las posibles emociones experimentadas por los participantes tuvieran una evolución gradual entre la reproducción de cada uno de los vídeos. Por ejemplo, la categoría naturaleza son vídeos tranquilos que pretenden recoger una expresión neutral del sujeto, mientras que la categoría gente-botezos, que viene inmediatamente después de la naturaleza, persigue llevar al participante a un estado lo suficientemente tranquilo como para generar un bostezo, o bien, como para generar una reacción de contención del bostezo.

Por último, siguiendo la sugerencia del Dr. Quoc Vuong (psicólogo del Insti-

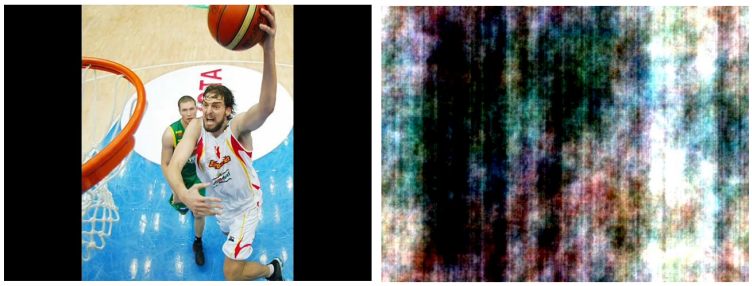


Figura 3.8: Un ejemplo de una imagen RISE generada a través de la degradación de la matriz de fase de la imagen original. Aunque esta técnica altera la estructura espacial de la imagen, propiedades importantes a bajo nivel de la imagen original, como el espectro de frecuencia espacial o la iluminación global de la imagen, son perfectamente preservadas.

tuto de Neurociencia de la Universidad de Newcastle) y, antes de la reproducción de cada vídeo, se muestra durante 3 segundos una imagen RISE del último fotograma del vídeo previamente mostrado. Estas imágenes RISE son usadas en investigaciones de alta percepción visual [181]. Son creadas de una manera en la que se controla de forma estricta un conjunto de características de la imagen a bajo nivel, tales como la iluminación y el espectro de frecuencia, reduciendo de este modo los factores que alteran el análisis de los procesos visuales a alto nivel. Lo que se persigue con el uso de este tipo de imágenes es eliminar cualquier adaptación de bajo nivel en la retina y en la memoria. Para generar este tipo de imágenes es necesario realizar un análisis del espectro de frecuencias de la imagen objetivo, en este caso el último fotograma del vídeo previo, y manipular la estructura espacial de la imagen sin alterar su espectro de potencia original. Esto puede llevarse a cabo en el dominio de Fourier mediante la alteración del denominado espectro de fase, al tiempo que se conserva la amplitud del espectro. De hecho, se ha demostrado que gran parte de la información que especifica la estructura natural de una imagen se encuentra en el espectro global [159], por lo que manipular el espectro de fase tiene el efecto de degradar la estructura espacial de una imagen en la retina del participante.

3.1.2. Base de datos Vedana

La base de datos Vedana está dividida en dos partes según los dos procesos de grabación previamente descritos. El conjunto principal de datos lo componen 345 secuencias de vídeo de expresiones faciales de 63 sujetos, mientras que el segundo conjunto de datos lo componen 126 secuencias de vídeo con el mismo número de sujetos. Además, con el fin de proporcionar un marco en el que se facilite la extracción de fondo, se proporcionan un conjunto de secuencias (una por cámara) en las que no existe sujeto en la escena. En total, la base de datos está formada por 474 secuencias de vídeo. Resulta importante resaltar que, de

Figura 3.9: Distintas perspectivas de las tres cámaras durante el proceso de grabación. En base a las características de la propia imagen capturada, resulta fácilmente apreciable que las propiedades del proceso de grabación son distintas para cada cámara. La imagen de la izquierda se corresponde con la cámara Sony, la central con la Webcam Logitech y, por último, la imagen de la derecha se corresponde con el resultado obtenido con la Casio. De esta forma, se consideran tres ángulos, lo cual puede servir de base para la evaluación de la robustez de algoritmos de reconocimiento facial combinados con otros algoritmos de estimación de pose del sujeto.

la cámara de alta velocidad (la cámara identificada como C1) no se realizaron grabaciones de las dos sesiones de cada sujeto, sino que se restringió a una sesión en algunos casos. El motivo radica en limitaciones debido la capacidad de almacenamiento disponible, por lo que, en lugar de realizar una grabación por usuario y sesión (126 grabaciones), para esta cámara, se limitó a 93 el número total de grabaciones.

De los 63 sujetos que forman parte de la base de datos, todos acudieron a las dos sesiones. El 27 % de ellos son mujeres mientras que el 73 % restante lo conforman varones. El participante más joven tiene 18 años mientras que el más longevo tiene 57 años, dejando la media de edad en $27,5$ años (véase la Figura 3.10). Por otro lado, en algunas bases de datos relevantes como FERET [54] la captura de sujetos con edad variable determina, en buena medida, la capacidad de proporcionar una base sólida de pruebas para la realización de ciertas tareas relativas a la visión por computador (por ejemplo, estimación de edad o identificación de sujetos). Esto supone un valor añadido pues muchas de las bases de datos presentadas en la sección 2.2.1 sólo proporcionan imágenes o secuencias de una sola sesión con los participantes implicados en el experimento, y esto ocurre debido a que suele ser complicado mantener el contacto con los sujetos voluntarios. En Vedana, los 63 sujetos han sido grabados en dos ocasiones con varios días entre cada sesión, normalmente se pretendió que este periodo fuera de, al menos, una semana.

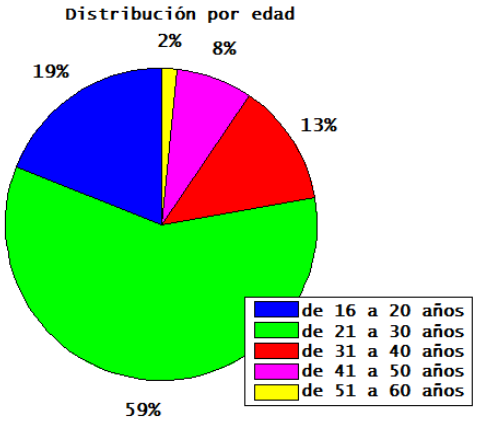


Figura 3.10: Distribución de edades en Vedana.

Elemento	**Categoría**	**Mujer**	**Hombre**
Gafas	Nada	22	56
Gafas	De vista	10	30
Gafas	De sol	4	4
Pelo	Rapado	0	2
Pelo	Corto	2	70
Pelo	Medio	8	11
Pelo	Largo	13	1
Pelo	Coleta	12	3
Pelo	Capucha	1	3
Vello facial	Nada	36	27
Vello facial	Barba	0	51
Vello facial	Perilla	0	11
Vello facial	Bigote	0	1
Nivel de vello facial	Nada	36	27
Nivel de vello facial	Incipiente	0	26
Nivel de vello facial	Diario	0	28
Nivel de vello facial	Semanal	0	9

Tabla 3.3: Esquema de elementos contextuales de la base de datos.

Figura 3.11: En estas imágenes se pueden apreciar distintos participantes que llevan gafas puestas. Resulta notable la enorme diversidad de los tipos de gafas proporcionadas por los participantes; con montura oscura, con poca montura, blancas y finas, etc.

Otro aspecto interesante es que Vedana presenta secuencias de vídeo con elementos oclusivos del rostro tales como vello facial, capuchas o gafas. En la tabla 3.3 pueden apreciarse un resumen de los accesorios usados por los participantes. En dicha tabla pueden observarse cuatro columnas; la primera columna se corresponde con los elementos que permiten variar el aspecto de los participantes, la segunda columna hace referencia a la categoría de dicho elemento, la tercera y la cuarta se corresponden con el número de usuarios, mujer y hombre respectivamente, que coinciden con dicha categoría. Debe tenerse en cuenta que estas cifras no son datos por cada usuario, sino que son datos por cada sesión, pues un usuario puede aparecer sin gafas en una sesión y con gafas en la siguiente. De esta forma, observando los elementos aparecidos en la tabla, se puede apreciar que las gafas, si bien tienen una menor aportación en las que son de sol, las de vista proporcionan un interesante marco de estudio en la literatura de este tipo de bases de datos. Por otro lado, las mujeres ofrecen mayor variedad de tipo de cabello mientras que los hombre lo hacen cuando se trata el vello facial. Muchos usuarios masculinos variaron este aspecto de su imagen entre las distintas sesiones a las que acudieron.

Figura 3.12: La tipología del cabello es uno de los puntos de mayor variabilidad en Vedana.

Figura 3.13: Aquellos participantes que llevaban puesto un jersey con capucha fueron grabados, para ambos experimentos, con la capucha en una de las sesiones y, sin la capucha, en la siguiente sesión.

Figura 3.14: El vello facial es un elemento distintivo en los participantes masculinos. La morfología que presenta este tipo de vello puede ser de lo más variado.

Figura 3.15: Imágenes del experimento de captación del torso.

	Cohn-Kanade	JAFFE	AR	MMI	**Vedana**
Referencia	[118]	[142]	[144]	[167]	[87]
Imágenes	No	219	4000	740	No
Vídeos	2105	No	No	848	475
Emociones	Sí	Sí	4	Sí	Sí
Múltiples exp. AU	Sí	Sí	4	Sí	Sí
Codificación AU	Sí	No	No	Sí	No
Núm. sujetos	210	10	126	19	63
Edad sujetos	18-50	?	?	19-62	18-60
Sexo	Ambos	Mujeres	Ambos	Ambos	Ambos
Iluminación	Única	?	Variable	única	única
Vello fac., gafas	No	No	Sí	Sí	Sí
Descargable	No	Sí	No	Sí	Pronto
Accesible	No	No	No	Sí	Sí

Tabla 3.4: Comparación de Vedana con bases de datos existentes.

Análisis comparativo

Las secuencias de vídeo grabadas para Vedana difieren de las bases de datos existentes en diversos aspectos. En primer lugar, las expresiones faciales capturadas no han sido verificadas ni identificadas dentro de las expresiones primarias definidas por Ekman y Friesen [71] de manera formal, es decir, no se ha contratado a un experto para que las clasificara. En segundo lugar, la mayor parte de las expresiones obtenidas son más sutiles que las que han sido expuestas en otras colecciones de datos; la expresión se genera en un rango temporal de forma gradual, no de forma discreta en un instante determinado. Además, debido a que estos estímulos son dinámicos, los movimientos de la cabeza y de los ojos acompañan a estas expresiones. Por último, las secuencias ofrecen más de una expresión facial ofreciendo un "puzle de expresiones"; por ejemplo, un rostro empieza con una sonrisa debido a un vídeo gracioso y acaba generando una expresión de asco en el vídeo posterior. De esta forma, combinando la información obtenida en las secuencia, las tramas de tiempo y la información concerniente al vídeo que está visualizando el sujeto, puede determinarse qué estímulos provocan qué reacciones, en otras palabras, qué vídeos generan qué expresiones faciales.

En la tabla 3.4 se puede apreciar un resumen comparativo con otras bases de datos disponibles en la literatura. De todas ellas, la base de datos de expresiones faciales Cohn-Kanade [47] proporciona el conjunto de datos más completo y más utilizado en la investigación acerca de la automatización de expresiones faciales. Sin embargo, los dos principales inconvenientes que presenta este conjunto de datos son que; en primer lugar, sólo se muestra la expresión facial en su máxima intensidad, lo que hace imposible analizar el proceso de formación de la expresión así como su posterior disipación del rostro del individuo y, en segundo lugar, la

mayoría de las grabaciones contienen la fecha y hora de la grabación en la región de la barbilla de los sujetos, lo que complica de forma notable la apariencia del sujeto pues la barbilla apenas es visible. Por otro lado, de todas las bases de datos mostradas en la tabla 3.4, ninguna de ellas contiene imágenes con todos los posibles códigos de activación (AU), ninguna contiene imágenes de perfil ni tampoco combinan imágenes estáticas con secuencias de vídeo. Tampoco, los metadatos (etiquetas) asociados a cada objeto de las distintas colecciones permiten identificar el segmento temporal (comienzo, esplendor, finalización) de la expresión facial mostrada.

Vedana es una base de datos que actualmente está parcialmente anotada. Ello es así porque el presupuesto que se nos proporcionó para realizar una anotación FACS de todos los fotogramas de las secuencias de vídeo de la cámara C3 (la webcam) estaba fuera de nuestro presupuesto; 25200 dólares americanos. Por ello, se optó por una solución intermedia. Dos becarios anotaron de forma manual un total de 1524 fotogramas (12 por sesión de grabación). Esta anotación consitió en determinar las zonas de la imagen pertenecientes a las zonas de la piel, el pelo, la ropa, el fondo o las gafas de sol.

Por último, el objetivo será dar disponibilidad bajo petición. Realmente, el volumen de los datos compuestos por secuencias de vídeo, complica notablemente la posibilidad de diseminar los datos. Por ello, resulta necesario proporcionar mecanismos de descarga parcial de los datos.

3.2. Detección de expresiones faciales

En la última década ha habido un enorme desarrollo de los sistemas informáticos que facilitan la interacción entre usuarios y máquinas y, en ello, la computación afectiva ha tenido una trascendencia de vital importancia. La capacidad de hacer dispositivos más "amigables" y "adaptables" pretende desarrollar una nueva generación de dispositivos dentro del mercado tecnológico. Un ejemplo notable de este tipo de "revolución adaptable" es el de un conjunto de pacientes con una lesión física severa que les impide hacer uso de un ratón o de un teclado. Existen trabajos orientados a dar respuesta a este tipo de circunstancias a través de la aplicación de distintas técnicas. Por ejemplo, actualmente existen dispositivos de seguimiento en el que, el usuario, es capaz de mover un ratón con su mirada y hacer clic sobre una letra a través de su parpadeo. Esta forma de comunicación, robusta y eficaz, facilita las tareas en la vida cotidiana de estas personas [110].

Sin embargo, el mercado exige algo más que adaptabilidad física, exige una adaptabilidad en términos de usabilidad o en la forma de entender la información semántica del contexto de la misma forma que la entienden los humanos. En el año 2007, Sony lanzó su cámara Cyber-shot DSC T200 con la funcionalidad de realizar fotografías al detectar rostros sonrientes [199]. El objetivo es que la cámara dispara una foto cuando detecta un número mínimo usuarios de la escena sonriendo. La idea pretendía ofrecer una simplificación del proceso de captación de fotografías; pasar de decir "sonrían" a que la cámara sacara la

Figura 3.16: En la primera fila puede apreciarse imágenes de sonrisas auténticas, mientras que en la fila inferior se exponen imágenes con sonrisas fingidas [136].

foto automáticamente cuando, las caras detectadas en la escena, sonreían. Este dispositivo generó cierta controversia entre los consumidores pues, si bien la cámara funcionaba para sonrisas notables, no era menos cierto que perdía mucha precisión cuando no captaba sonrisas muy marcadas o bien, si los usuarios, posaban con sonrisas sin abrir la boca [110]. Además, otro motivo de queja de esta falta de precisión surgió cuando los usuarios detectaron que el dispositivo disparaba fotografías ante la mera aparición de dientes sin que, necesariamente, existiera la presencia de una sonrisa. Por último, la cámara no era capaz de detectar el rostro de forma precisa cuando éste portaba gafas de sol y, por tanto, no se reconocía sonrisa alguna [226].

En cualquier caso, estas aplicaciones suponen los primeros pasos que se están dando en el mercado tecnológico con el fin de adaptar la computación afectiva en los dispositivos que los seres humanos utilizan de forma regular. Se desprende de todo ello que la máquina no espera un mecanismo de acción directa por parte del ser humano, sino que tiene un papel más proactivo con el usuario; lo observa, lo analiza y responde a la interacción no verbal. Precisamente, eso es lo que define las expresiones faciales; un mecanismo primario de comunicación no verbal que facilita el contexto de la información que se transmite. Sin embargo, se trata de un proceso no carente de cierta organización; para que la cámara pueda detectar la sonrisa es necesario que, previamente, haya detectado un rostro. Por ello, cuando se pretende analizar toda expresión facial, más allá de definir el tipo de descriptor a utilizar para definir si el usuario está expresando alegría o tristeza, es necesario detectar el rostro del individuo, extraer las características faciales y clasificar al sujeto en el rango de las expresiones. En la sección 2.3.2 se ha hecho un extenso estudio sobre el conjunto de técnicas más utilizadas en la literatura a la hora de detectar expresiones faciales y también, al final de dicha sección, se ha ofrecido un esbozo de los músculos más relevantes a la hora de formar esta singular expresión, la sonrisa.

Precisamente, la sonrisa, es una de las expresiones faciales que ocurre con más frecuencia en la vida cotidiana de los humanos y por ello requiere una es-

pecial atención. Las personas sonríen para ser agradables entre sí y, por regla general, es una expresión común que refleja placer aunque también puede ser concebida desde un sentimiento contrario, como la ira. En este sentido, dicha expresión suele suceder de forma involuntaria en un estado de ansiedad. Ekman demostró que la sonrisa es una reacción normal, y natural, a ciertos estímulos y sucede de forma independiente, sea cuál sea la cultura en la que sucede [163]. Entre los tipos de sonrisas existentes, la sonrisa común o "sonrisa de Duchenne" es la sonrisa más estudiada del mundo e involucra el movimiento de los músculos cigomáticos mayor y menor cerca de la región bucal y el músculo orbicular cerca de la región ocular. La "sonrisa de Duchenne", nombrada así por el investigador Guillaume Duchenne, es una sonrisa de corte simétrico que es producida como una respuesta involuntaria a una emoción genuina y es lo que se denomina una "sonrisa genuina" [163]. Dentro de la detección de este tipo de expresión, existen ciertos trabajos encaminados a detectar estrictamente esta sonrisa de Duchenne y, para ello, se basan en la codificación FACS explicada en la sección 2.3.1. Ciertos autores identifican este tipo de sonrisas como aquellas en las que las unidades de acción AU6 y AU12 suceden, de forma simultánea, en un determinado momento. En la Figura 3.16 puede apreciarse un conjunto de sujetos clasificados como sonrisas genuinas o sonrisas falsas según Liu [136]. Como puede apreciarse, las genuinas son aquellas en las que existen cambios en las, ya mencionadas, regiones ocular (músculo orbicular) y bucal (músculo cigomático). Ekman sostiene, en todo momento, que la presencia de la subjetividad de una persona sobre las emociones que percibe condiciona el proceso de toma de decisión acerca de la expresión facial que debe mostrar [163]. Este mismo autor también sugiere que la relatividad de la emoción percibida es extrapolable a la reacción que puedan experimentar los sujetos y, aún existiendo la posibilidad de una generalización a la hora de codificar la expresión, siempre existe la posibilidad de que este análisis no sea extrapolable a todas las personas. En otras palabras, si bien la sonrisa existe en todas las culturas, pueden existir matices entre la forma que se tiene para sonreír dentro de cada cultura y dependiendo de este contexto es posible que, en la Figura 3.16, puedan existir sonrisas genuinas marcadas como falsas.

Es por ello que, en esta sección del capítulo, no se pretende dilucidar la calidad de la sonrisa en base a la codificación FACS, sino que lo que se persigue es la mejora de los mecanismos existentes de detección de sonrisas utilizando nuevos enfoques que permitan extraer vectores de características novedosos. Para ello, se hacen uso de técnicas, como el PCA o LBP y sus variantes, para extraer las características necesarias que permitan una correcta clasificación de la expresión facial. Estos enfoques son descritos en las secciones 3.2.1.1 y 3.2.1.2 respectivamente. En la sección 3.2.1.3 se muestran los enfoques para la clasificación de datos que han sido considerados en los estudios expuestos en las secciones 3.2.1.4 y 3.2.1. En la sección 3.2.1.6 se ofrece un enfoque de autosimilitud para la detección de sonrisas haciendo uso de un descriptor, denominado *firma digital*, que ha sido usado tanto para la detección de sonrisas como para la segmentación de patrones en la imagen. Por último, en la sección 3.2.2 se ofrece un novedoso enfoque para la detección de la sonrisa en entornos dinámicos en

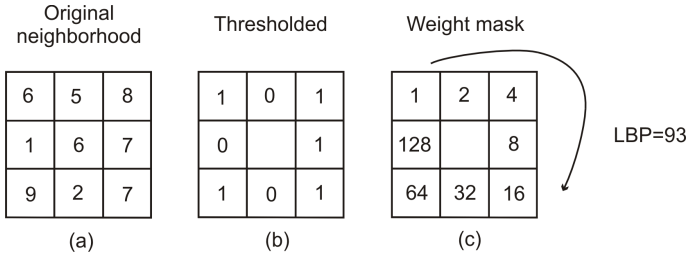


Figura 3.17: Versión básica del LBP [157].

el que se considera la evolución temporal de dicha expresión facial.

3.2.1. Enfoques estáticos

Esta sección está encaminada al estudio de técnicas para la detección de la sonrisa sobre imágenes estáticas. Para ello, se pretende cubrir tres objetivos fundamentales; en primer lugar, se trata de encontrar el conjunto de soluciones adecuado para obtener un clasificador fiable, en segundo lugar, se pretende estudiar su combinación con el fin de lograr mejorar las tasas obtenidas de manera individual y, por último, se exponen los resultados depués de la aplicación de la firma digital como descriptor para la detección de sonrisas. Por tanto, las pruebas se han dividido en tres categorías: estudio de enfoques para la detección de sonrisas, combinación de enfoques para la detección de sonrisas y, por último, estudio basado en la autosimilitud de las sonrisas. Sin embargo, antes de entrar a explicar el trabajo desarrollado es necesario establecer el marco de la investigación a través de las distintas técnicas que son utilizadas (secciones 3.2.1.1 y 3.2.1.2) así como los métodos de clasificación que han sido considerados (sección 3.2.1.3).

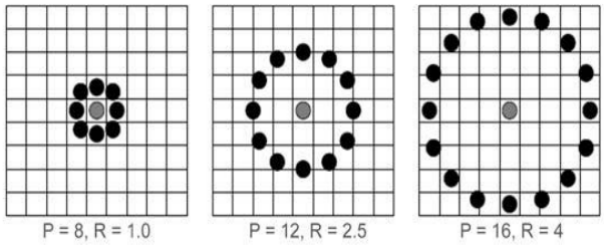
3.2.1.1. Patrones binarios locales

Los patrones binarios locales (*Local Binary Patterns (LBP)*) permiten obtener un descriptor que ha sido ampliamente utilizado en problemas de clasificación de imágenes. Si bien, en un principio, fue presentado por Ojala para la clasificación de texturas, su popularidad creció al destacar por la invariancia que presentan ante los cambios de iluminación así como el bajo coste de cómputo, lo que les ha permitido ser aplicados con relativo éxito en la resolución de problemas de clasificación en los que se necesita cierta robustez ante los cambios lumínicos en la imagen a procesar [201].

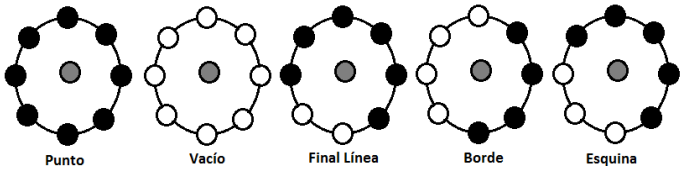
La creación de esta tipología de patrones es relativamente sencilla. Dada una imagen en escala de grises, se selecciona un píxel y se aplica el operador LBP sobre dicho píxel con el fin de obtener un valor. Para ello, se realiza una operación de umbral de forma circular a los píxeles que rodean al píxel central seleccionado y, posteriormente, se construye un patrón binario con el resultado de cada una de esas operaciones de umbral. Ojala consideró, en su versión básica, una ventana

de 3x3 y construyó el patrón binario local en base al píxel central (g_c) y sus ocho vecinos (g_p) (véase la Figura 3.17). Sin embargo, la propia definición de LBP le permite fácilmente adaptarse a otros radios (R) considerando un número P de vecinos [158]:

$$LBP_{P,R}(x_c, y_c) = \sum_{k=0}^{P-1} s(g_p - g_c) \cdot 2^k \ , \ s(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{3.1}$$



(a) Adaptación del número de vecinos (P) a considerar para el procesado LBP en función del radio (R) seleccionado.



(b) Comportamiento LBP de acuerdo con las texturas locales identificadas.

Figura 3.18: Características del procesado LBP [158].

En la Figura 3.18a puede apreciarse la adaptación del número de vecinos a considerar en función del radio de la vecindad a considerar. Por otro lado, una interesante propiedad del LBP es que, al ser un patrón calculado de forma circular, resulta invariante frente a rotaciones. Los experimentos de Ojala sugieren que sólo un subconjunto particular de patrones binarios locales está presente en la mayoría de los píxeles analizados en imágenes comunes. Este subconjunto de patrones son conocidos como patrones uniformes y se caracterizan por el hecho de que contienen, como máximo, dos transiciones de bit de 0 a 1 o viceversa. En otras palabras, un patrón uniforme es aquel que no tiene transiciones entre 0 y 1 (por ejemplo, 00000000 y 11111111) o bien aquel que tiene, como máximo, dos transiciones (por ejemplo, 00011100 o 100000011). De hecho, teniendo en cuenta que se trata de un patrón circular, es imposible la existencia de una única transición. Para patrones con dos transiciones existen $X \cdot (X-1)$ combinaciones posibles. Considerar, únicamente, patrones uniformes tiene dos importantes ventajas; la primera es que ahorra memoria teniendo en cuenta que los patrones no uniformes consumen 2^X posibles combinaciones y, la segunda ventaja, es que los patrones LBP uniformes (ULBP) sólo detectan importantes texturas

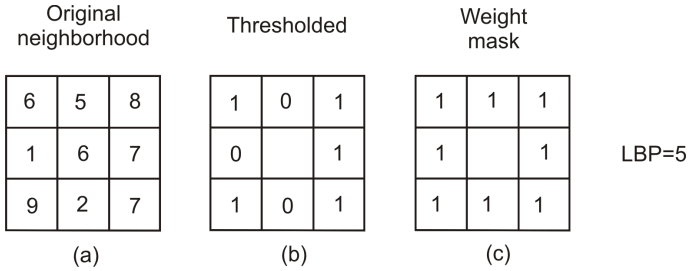


Figura 3.19: Versión simplificada del LBP [202].

locales (puntos, líneas, vértices, esquinas) tal y como se muestra en la Figura 3.18b. En los experimentos realizados por Ojala se demostró que los patrones uniformes conformaban cerca del 90 % de todos los patrones obtenidos en una imagen cuando se considera una vecindad de 3x3 [158].

Más recientemente se ha confirmado la validez de LBP a la hora de describir la apariencia facial. De hecho, una vez que el patrón LBP de la imagen es obtenido, la mayoría de los autores aplican una representación de datos basada en histogramas [178]. Por ejemplo, Ahonen propone una división de la imagen a través de un conjunto de rejillas para obtener un conjunto de histogramas que, concatenizados, le permiten obtener el descriptor deseado [5]. Sin embargo, tal y como apuntan los últimos trabajos que han hecho uso de estos patrones, el enfoque basado en histogramas LBP pierde la relativa localización de la información [178, 202], sobretodo si no se ha dividido la imagen en una rejilla, separando los patrones por localización [5]. Por este motivo se observa que el uso de LBP como método de preprocesado no favorece ciertas estrategias a la hora de clasificar pues tiene un efecto de enfatizar bordes y ruido. Con el fin de reducir este ruido, Qian Tao propuso una modificación de la ecuación 3.1 [202]. En lugar de aplicar un peso único a cada vecino, los pesos son los mismos de forma que se obtiene un patrón LBP simplificado (*Simplified Local Binary Patterns (SLBP)*), véase la Figura 3.19.

$$LBP_{P,R}(x_c, y_c) = \sum_{k=0}^{P-1} s(g_p - g_c) \ , \ s(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{3.2}$$

Según sus autores, este nuevo enfoque muestra algunos beneficios aplicados a la verificación facial debido al hecho de que, simplificando los pesos, la imagen es más robusta frente a los cambios de iluminación (proporcionando un máximo de nueve posibles valores por píxel). De esta forma, el número de patrones locales se reduce de forma drástica y la imagen se mueve en un espacio de representación más compacto.

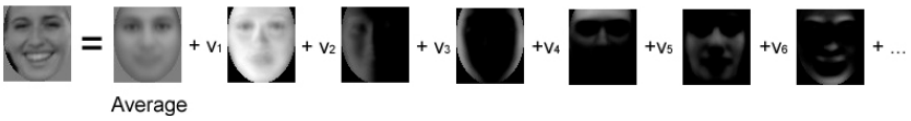


Figura 3.20: Descomposición PCA de un rostro humano.

3.2.1.2. Análisis de componentes principales

El análisis de componentes principales (*"Principal Components Analysis" (PCA)*) es una técnica estadística muy útil que ha tenido una aplicación notable en los últimos años en el área del reconocimiento facial así como en la compresión de imágenes [120]. De hecho, se trata de una técnica bastante común para encontrar patrones en datos que ofrecen múltiples dimensiones y expresar estos datos de forma que se resalten, tanto las similitudes existentes entre los datos, como las diferencias que presenten entre sí. Desde una perspectiva gráfica, ofrece un marco incomparable debido a que los datos multidimensionales son muy complicados de representar gráficamente. Otra notable ventaja que ofrece este enfoque es que permite empaquetar los datos resaltando sus diferencias o similitudes; una vez que se encuentran estos patrones, es posible reducir el número de dimensiones sin perder demasiada información durante el proceso. Entre los objetivos para los que se hace uso del PCA destacan cuatro de manera fundamental:

1. Extraer la información más relevante de un conjunto de datos.
2. Comprimir el tamaño de un conjunto de datos manteniendo, únicamente, la información más relevante.
3. Simplificar la descripción de un conjunto de datos.
4. Analizar la estructura de la observaciones y de las variables.

Es por ello que, para conseguir estos objetivos, el PCA procesa variables denominadas "componentes principales" y que son obtenidas a partir de combinaciones lineales de las variables originales de la imagen. Así, la primera componente principal es aquella que es necesaria para obtener la mayor varianza posible, la segunda componente es calculada teniendo en cuenta que debe ser ortogonal con respecto a la primera componente y así sucesivamente. Los valores que se desprenden de estas nuevas variables son denominados factores o variables latentes (*"factor scores"*) y pueden ser interpretados de forma geométrica como proyecciones. Por ejemplo, los rostros humanos pueden ofrecer un alto grado de dimensionalidad en este contexto. La descomposición a través del PCA permite reducir esta alta dimensionalidad sin perder, tal y como se ha especificado anteriormente, una cantidad significativa de información. Para ello, se realiza un análisis de covarianza entre los distintos factores. Así, dada una imagen objetivo, el PCA permite proyectarla en el espacio (véase Figura 3.20) y

la apariencia de los distintos individuos es representada en una dimensionalidad menor a través de la media de las variables latentes v_i [210].

3.2.1.3. Clasificadores: SVM vs. k-NN

Dentro del contexto de las emociones, se pueden definir múltiples enfoques a la hora de seleccionar el clasificador apropiado para etiquetar la expresión facial que se está proyectando; usar un clasificador multiclase, un clasificador binario, varios clasificadores binarios, etc. Todo depende de la naturaleza de los datos que se deseen clasificar. Por ejemplo, en el trabajo de Bing Li [131], que se puede observar en la Figura 2.18a, se optó por el uso de siete clasificadores SVM con salidas probabilísticas [228]. Al final del proceso, se hizo uso de cinco estrategias diferentes para combinar los resultados de cada uno de los clasificadores y, así, poder tomar una decisión sobre si el sujeto presentado era hombre o mujer.

Algo similar ocurre en el caso del estudio de las emociones pero, en lugar de determinar la probabilidad de que el pelo de una persona sea de hombre o de mujer [131], se estudia la probabilidad sobre si una determinada emoción es expresada o no. Por tanto, en esta sección, se enfoca la resolución del problema de clasificación desde una perspectiva binaria. Un problema de clasificación binaria se define de forma matemática de la siguiente forma; sea un conjunto de datos χ, un espacio de etiquetas γ y una distribución Δ sobre χ. Además, dada la secuencia $\Theta = \{x_i, y_i\}_{i=1}^{m}$ de muestras etiquetadas entonces cada elemento $x_i \in \chi$ es independiente y está distribuido según la distribución de Δ, donde cada $y_i \in \gamma$ es asignado en función de una regla estocástica. En el caso de la clasificación de tipo binaria, se restringe el valor de $\gamma = \{-1, +1\}$.

Por otro lado, durante el proceso de clasificación, se define la regla de clasificación a través de una hipótesis. Esta hipótesis es una función $\eta : \chi \rightarrow \gamma$ que asigna una etiqueta a cada elemento del espacio de entrada. Considerando el problema de clasificación binaria, la expresión queda reducida en $\eta : \chi \rightarrow [-1, +1]$ donde el signo de $\eta(x)$ es interpretado como la predicción de la etiqueta definida para la instancia x y la magnitud $|\eta(x)|$ indica el grado de confianza de dicha predicción.

Sin embargo, esta predicción puede sufrir un error de generalización R y un error empírico R_{emp}, que son definidos como:

$$R(\eta) = P_{(x,y)\sim\Delta}\{signo(\eta(x) \neq y\} \tag{3.3}$$

$$R_{emp}(\eta, \Theta, D) = \sum_{i=1}^{m} D(i) I_{[sgn(\eta(x)\neq y]} \tag{3.4}$$

Donde $D \in \mathbb{R}^m$ es una distribución discreta sobre el conjunto de muestras etiquetadas e I es la función indicadora [66].

Por su parte, el problema que plantea una clasificación lineal viene dado por un conjunto de n datos de entrenamiento de la forma especificada en la ecuación 3.5.

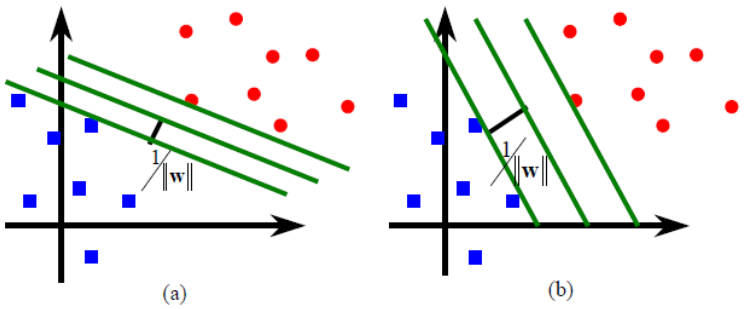


Figura 3.21: Problema lineal que puede separarse con un posible hiperplano separador, en la primera imagen con un hiperplano separador de bajo margen y, en la segunda imagen, con un hiperplano separador con alto margen [66].

$$\mathcal{D} = \{(x_i, y_i)/x_i \in \mathbb{R}^m, y_i \in \{-1, +1\}\}_{i=1}^{n} \tag{3.5}$$

Donde y_i es -1 o $+1$ e indica la clase a la que pertenece el dato x_i. El objetivo es, con los datos proporcionados, encontrar un clasificador que maximice el margen en el hiperplano que divide los puntos etiquetados como $y_i = -1$ de los que son etiquetados como $y_i = +1$. Cualquier hiperplano puede ser escrito como el conjunto de puntos x que satisfagan $w \cdot x - b = 0$, donde se expresa el producto escalar a través del vector normal w al hiperplano. El parámetro $\frac{b}{||w||}$ determina el desplazamiento del hiperplano desde el origen a lo largo del vector normal w. Así, tal y como puede observarse en la Figura 3.21, el objetivo no es sólo encontrar un clasificador que separe ambas clases, pero sí uno que maximice la frontera de separación entre ambas clases n función de las características proporcionadas por los datos de entrada χ. De esta forma, ante la posibilidad de introducir nuevos datos en el clasificador, la salida que éste proporcione será la más fiable.

Sin embargo, a pesar de que el clasificador lineal es uno de los clasificadores más simples, suele sufrir un error de generalización (R) y un error empírico (R_{emp}) bastante elevados en situaciones en las que la separabilidad de las muestras no sea tan precisa, tal y como puede apreciarse en la Figura 3.22. Este proceso de clasificación se trata de una transformación no lineal Φ del conjunto de datos de entrada χ que busca las condiciones ideales en la que los datos son separables. En otras palabras, en la Figura 3.23 puede apreciarse la tarea de una máquina de soporte vectorial; en el lado izquierdo los datos originales de entrada son reubicados a través del uso de una serie de funciones no lineales para luego encontrar la frontera óptima con los datos ya reubicados, todo ello a través de funciones matemáticas denominadas *kernels*. Estos kernels permiten el cálculo del producto escalar en el espacio de características de los datos de entrada de forma directa, lo cual supone una gran ventaja; el espacio de los datos χ no necesita ser una estructura vectorial, tan sólo un conjunto no vacío, lo que facilita la labor en muchas situaciones en las que no se tiene una repre-

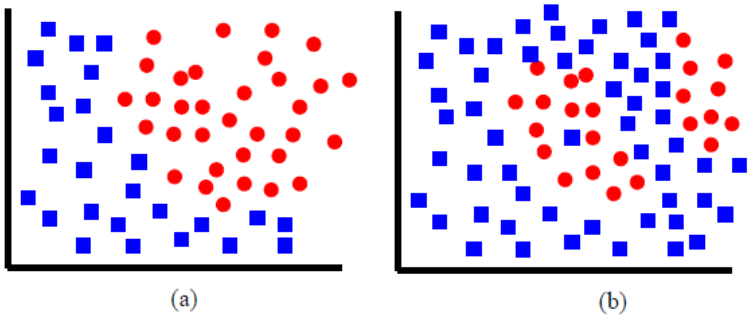


Figura 3.22: En el conjunto de muestras presentados en estas imágenes, la separabilidad de los datos no permite una clasificación óptima a través del uso de clasificadores lineales [66].

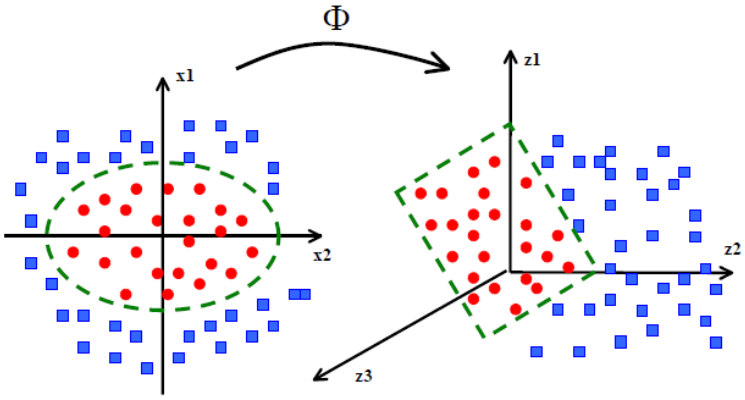


Figura 3.23: Ejemplo de transformación no lineal [66].

sentación vectorial de los datos y se trabaja con parejas de datos u objetos no vectoriales [109]. El proceso de reubicación de estas entradas en el espacio es conocido como mapear y se trata de una transformación no lineal de la forma descrita en la ecuación 3.6.

$$\begin{aligned} \Phi : &\chi \rightarrow \chi' \\ &x \mapsto x := \Phi(x) \end{aligned} \tag{3.6}$$

Resulta interesante destacar que, en la nueva configuración resultante de aplicar la transformación Φ, los datos son linealmente separables y, en lugar de construir una curva compleja que permita la separación de datos, tan sólo hace falta encontrar la frontera lineal que maximice la distancia entre los dos conjuntos de datos.

Al margen de las máquinas de soporte vectorial, también se ha considerado

un segundo método de clasificación en esta sección, el algoritmo de clasificación k-NN (*"k-Nearest Neighbour"* (k-NN)) [82]. Se trata de un algoritmo de aprendizaje no paramétrico, es decir, que no hace ninguna suposición sobre la distribución de los datos de entrada, lo cual dista con el enfoque anteriormente presentado para las máquinas SVM. Además, otra diferencia con respecto a las máquinas SVM es que se trata de un algoritmo clasificado como *lazy*; apenas existe fase de entrenamiento (o no existe o es mínima). El algoritmo mantiene todos los datos de entrenamiento disponibles durante la fase de prueba, pero no existe una fase previa de entrenamiento, tal y como ocurre con las máquinas de soporte vectorial. Sin embargo, aunque no existe una fase de entrenamiento, el posible coste computacional del entrenamiento se traslada a la fase de pruebas.

Antes de mostrar el enfoque matemático resulta esencial especificar que el algoritmo k-NN asume que los datos de entrada χ están ubicados en un espacio métrico, lo que le permite compararlos entre sí a través de una función de distancia. Por otro lado, al margen del conjunto de datos de entrada χ, también existe un espacio de etiquetas asociado γ. Por último, el parámetro de entrada que define el comportamiento del algoritmo ante χ es el parámetro k que decide cuántos vecinos influyen en la clasificación de un dato de entrada x_i. Estos vecinos son definidos sobre la base métrica de la distancia. Debe tenerse en cuenta que si $k = 1$ la aproximación del dato de entrada x_i se hará en función de la etiqueta y_j de su vecino más cercano.

Desde un punto de vista matemático, se trata de un método de clasificación supervisado que sirve para estimar una función de densidad. A diferencia de las máquinas SVM, en esta ocasión no se hace uso de funciones de kernel, sino que una estimación de densidad define la etiqueta de un elemento de entrada. De esta forma, para un elemento de entrada x_i se ubica un hipercubo con centro en x_i y aumenta su tamaño hasta que k vecinos son capturados en su interior. La densidad viene definida a través de la ecuación 3.7.

$$p(x_i) = \frac{\frac{k}{n}}{V} \tag{3.7}$$

Donde n es el número total de puntos de entrada y V hace referencia al volumen del hipercubo generado alrededor de x_i. Puede apreciarse que el numerador es una constante y que la densidad está influenciada por el volumen. Si la densidad es muy alta, la probabilidad de encontrar los k puntos cerca de x_i en un corto periodo de tiempo es muy alta. En la Figura 3.24 puede apreciarse la aplicación del algoritmo sobre un plano bidimensional, lo que reduce la ecuación 3.7 a través del uso de la expresión $V = A = \pi R^2$, donde R representa al radio de la circunferencia, en la ecuación 3.8.

$$p(x_i) = \frac{k}{n\pi R^2} \tag{3.8}$$

Por todo ello, si se considera que los datos de entrenamiento son vectores en un espacio multidimensional entonces cada muestra queda descrita en términos de w atributos, considerando γ posibles etiquetas para la clasificación. Los va-

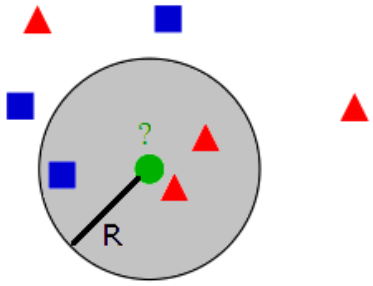


Figura 3.24: Aplicación del algoritmo k-NN en un plano bidimensional. En la imagen se muestra la aplicación del algoritmo para un valor de $k = 3$. El hipercubo se genera alrededor del elemento a etiquetar y se va ampliando hasta conseguir k muestras de entrenamiento dentro del mismo. Una vez se adquieren las k muestras, el elemento a etiquetar adquiere la etiqueta de la mayoría de los k elementos incluidos. En otras palabras, una muestra de entrenamiento dentro del radio del hipercubo supone un voto y se van sumando hasta llegar a k, al final la etiqueta que tenga más votos entre las k alternativas, etiqueta el elemento central. Sin embargo, no tiene por qué ser un criterio de simple votación, también es posible ponderar la votación en función de la distancia de cada una de las k muestras de entrenamiento en función de su distancia al elemento a etiquetar.

lores de los atributos del i-ésimo dato de entrada quedan definidos por el vector w-dimensional: $x_i = (x_{1i}, x_{2i}, ..., x_{si})/x_i \in X$

De esta forma, el espacio queda dividido en localizaciones y etiquetas de los datos de entrenamiento. Un punto en dicho espacio es asignado con la etiqueta y_i si esta es la clase más frecuente entre los k datos de entrenamiento más cercanos (lo que genera el hipercubo alrededor de la muestra ubicada en dicho punto). Por regla general, se usa la distancia euclídea, $d(x_i, x_j) = \sqrt{\sum_{r=1}^{s}(x_{ir} - x_{jr})^2}$, aunque también es posible hacer uso de otro tipo de distancias (por ejemplo, la distancia de Manhattan). En cambio, cuando se hace uso de histogramas, la función distancia considerada es la norma L1 que se especifica en la ecuación 3.9.

$$D(H1, H2) = ||H1 - H2|| = \sum_{k=0}^{n-1} |H1_k - H2_k| \tag{3.9}$$

Donde n es el número de elementos de cada uno de los histogramas $H1$ y $H2$ respectivamente.

3.2.1.4. Estudio de enfoques para la detección de sonrisas

Antes de considerar la posibilidad de explorar enfoques más complejos, es necesario realizar un estudio experimental para la detección de una expresión facial con los descriptores anteriormente descritos.

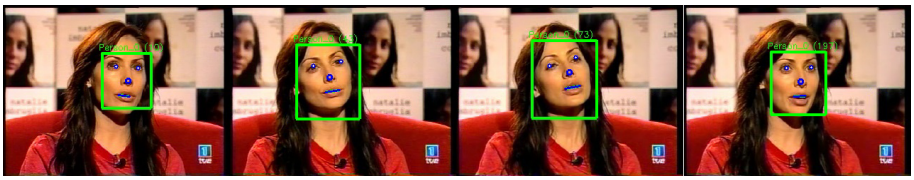


Figura 3.25: Captación de imágenes faciales en archivos audiovisuales.

Por otro lado, la búsqueda de un clasificador eficaz requiere de un elemento que cobra una mayor importancia en este experimento; el detector facial. Este detector resulta esencial, pues para determinar qué expresión está mostrando un sujeto antes es necesario poder detectar su rostro. En este sentido, recientemente se ha trabajado con distintos enfoques para la detección del rostro en tiempo real [132, 185, 216]. Todos estos trabajos tratan de minimizar el problema de las condiciones del entorno y están enfocados, fundamentalmente, a procesar un flujo de datos audiovisual en tiempo real. Considerando este aspecto, los detectores faciales basados en la combinación de varias señales del entorno de los fotogramas [37] obtienen mejores resultados que aquellos que se fijan en un solo aspecto del entorno como el trabajo desarrollado por Viola-Jones [216]. Sin embargo, el sistema de detección facial utilizado en este experimento para extraer imágenes faciales de los vídeos (véase la Figura 3.25 para más detalles), hace uso de la integración, entre otras señales, de diferentes clasificadores basados en el marco general desarrollado por Viola y Jones [216], color de piel, seguimiento multinivel, etc. Por último, con el objetivo de minimizar las falsas alarmas y maximizar la detección de rostros, las propiedades del detector facial fueron adaptadas para poder detectar, no sólo la cara, sino también los ojos, la boca y la nariz de un individuo. Otra propiedad es que el detector facial aprovecha la localidad de un rostro en un fotograma para buscarlo en fotogramas sucesivos.

Características del enfoque

En cuanto a las regiones faciales que van a ser consideradas en este estudio conviene especificar que se exploran varias alternativas. En la Figura 3.26 pueden apreciarse dos imágenes normalizadas. En la imagen de la derecha tres zonas han sido marcadas, ello es así porque en los experimentos que se describen en esta sección se ha considerado tanto el estudio del rostro en su totalidad así como el estudio de las distintas partes del rostro que son más relevantes para la sonrisa según Ekman [163] o Wu [229]. De esta forma, el experimento sugiere dos posibilidades, por un lado procesar el rostro normalizado en su totalidad o, por el contrario, procesar los distintos fragmentos extraídos de rostros normalizados y que, más concretamente, se corresponden con las regiones de ojos y boca respectivamente. Dentro de esta segunda posibilidad, también existen diversas configuraciones dependiendo de la forma en la que la zona ocular es extraída; los ojos pueden ser extraídos como una máscara entera (de color rojo en la Figura 3.26) o, por el contrario, pueden ser extraídos de forma independiente

Figura 3.26: Ejemplos de imágenes normalizadas a una escala de 59×65 píxeles. Se extrae el rostro de una imagen normal, se escala y se alinea para que la posición de los ojos quede ubicada siempre en los mismos puntos para todas las imágenes de la colección de imágenes considerada. Además, puede apreciarse las partes del rostro que son extraídas para algunos experimentos, siendo la zona de los ojos la que presenta la mayor variedad con dos posibilidades; la extracción de los ojos por separado (en color verde) o bien la extracción de una máscara conjunta con ambos ojos (en color rojo).

entre sí (de color verde en la Figura 3.26). Esta forma de extracción condiciona la cantidad de información que es utilizada a la hora de realizar el proceso de clasificación.

Antes de especificar las características extraídas para la realización del experimento, conviene definir que se trata de un proceso que consta de cuatro fases:

- **Fase 1.** Una vez que la imagen original en niveles de gris entra en el sistema se debe decidir si se trabaja con esta imagen original o si, por el contrario, se le aplica un preprocesado LBP (ULBP o SLBP). Y, en caso de aplicar el preprocesado, el resultado del mismo sería la entrada en la segunda fase, en caso contrario, se sigue trabajando con la imagen de grises original.

- **Fase 2.** Haciendo uso de la imagen proveniente de la primera fase, el segundo paso es determinar si se procesa el rostro completo o, si por el contrario, se deciden procesar regiones concretas del rostro, tal y como se ha comentado anteriormente (véase Figura 3.26). Si se trabaja con partes concretas del rostro, dichas partes son extraídas y conforman la entrada a la tercera fase del experimento, en caso contrario, la entrada a la tercera fase la conformaría la imagen proveniente de la primera fase.

- **Fase 3.** Se extraen las características de la entrada proveniente de la fase anterior. El vector de características puede estar conformado por los propios valores de píxeles de los datos provenientes de la segunda fase, por un histograma normalizado con estos datos, o bien, la proyección PCA de estos datos en el espacio de representación.

- **Fase 4.** Los descriptores resultantes de la tercera fase son las entradas de los clasificadores que determinarán qué etiqueta se le asigna a la imagen de entrada de la primera fase.

En cuanto al formato del conjunto de datos de entrada, se trata de un conjunto de imágenes en escala de grises, y se consideran las siguientes características en las pruebas:

- El espacio PCA (véase sección 3.2.1.2) considerando 40, 90 y 130 componentes respectivamente. Este espacio PCA es obtenido a partir de las imágenes originales de 59×65 píxeles.
- El espacio PCA calculado a raíz de preprocesar la imagen original con los dos enfoques de LBP descritos en la sección 3.2.1.1; el ULBP y el SLBP.
- La concatenación de histogramas de una rejilla de 3×3 obtenidos a partir de la imagen original en escala de grises.
- La concatenación de histogramas de una rejilla de 3×3 obtenidos a raíz de la imagen original preprocesada con ULBP y la concatenación de histogramas obtenidos a través del preprocesado de la imagen original con SLBP.
- La concatenación de valores de intensidad de la imagen original en escala de grises.
- La concatenación de valores de píxeles de la imagen original tras preprocesarla bien con ULBP o SLBP.

Conjuntos de datos

El conjunto de datos de trabajo lo conforman imágenes descargadas de Internet que, previamente, han sido separadas en sonrientes o no sonrientes. Esta primera dicotomía ha sido hecha a mano y, además, las imágenes han sido normalizadas a las dimensiones de 59×65 píxeles en base a la posición de los ojos. Dentro de esta colección de imágenes, un primer conjunto de imágenes contiene 2421 imágenes sonrientes, mientras que el segundo conjunto contiene 3360 imágenes no sonrientes. En el desarrollo de las pruebas se propone un *k-fold* con k=2 y diez repeticiones del experimento. En cada repetición, del número total de imágenes que componen la colección, 2000 imágenes son utilizadas para entrenar (1000 imágenes de cada clase) y otras 2000 imágenes son consideradas para las pruebas (1000 imágenes de cada clase).

Experimentos

Tal y como se ha comentado anteriormente, los descriptores para procesar las distintas regiones del rostro se construyeron a patir de la concatenación de la información extraída para cada una de las regiones analizadas. Así, para el experimento con el uso del clasificador k-NN, estos descriptores surgen desde la

Prueba k-NN	Valores de píxeles	Histograma
GRAY_M_1-NN	20.3 %	37.5 %
ULBP_M_1-NN	41.0 %	32.0 %
SLBP_M_1-NN	35.0 %	36.5 %
GRAY_M_5-NN	**18.7 %**	34.5 %
ULBP_M_5-NN	43.0 %	28.5 %
SLBP_M_5-NN	38.5 %	32.0 %
GRAY_ET+M_1-NN	26.0 %	31.0 %
ULBP_ET+M_1-NN	44.0 %	32.5 %
SLBP_ET+M_1-NN	42.0 %	34.5 %
GRAY_ET+M_5-NN	**23.5 %**	30.0 %
ULBP_ET+M_5-NN	43.0 %	27.5 %
SLBP_ET+M_5-NN	38.0 %	33.0 %
GRAY_SE+M_1-NN	24.0 %	33.0 %
ULBP_SE+M_1-NN	46.0 %	33.5 %
SLBP_SE+M_1-NN	42.0 %	34.0 %
GRAY_SE+M_5-NN	**21.0 %**	30.0 %
ULBP_SE+M_5-NN	42.0 %	30.0 %
SLBP_SE+M_5-NN	39.0 %	33.5 %

Tabla 3.5: Tasa de error con el clasificador k-NN considerando partes del rostro. Para cada entrada en la tabla hay una denominación que hace referencia al nombre del enfoque aplicado, las partes del rostro involucradas en la prueba y el método de clasificación aplicado. Así, para el enfoque aplicado existen tres posibilidades; ULBP, SLBP y GRAY que hacen referencia a los enfoques ULBP, SLBP e imagen en escala de grises respectivamente. Para las partes del rostro involucradas en la prueba también hay tres posibilidades; M, SE, ET que significan boca, ojos separados y máscara conjunta de ojos respectivamente (véase la Figura 3.26). Por último, el método de clasificación sugiere sólo dos posibilidades; 1-NN y 5-NN que hacen referencia a la clasificación cuando $k = 1$ y cuando $k = 5$, según sea el caso.

concatenación de los *bins* de un histograma normalizado o bien desde la concatenación de la intensidad de los píxeles de las áreas analizadas. La tabla 3.5 muestra los resultados considerando cada zona del rostro por separado y su combinación. Resulta evidente que, únicamente los ojos, no permiten dilucidar si una persona está mostrando una sonrisa o no, pero no es menos cierto que, según Ekman, existen expresiones que muestran la boca abierta pero no son una sonrisa, de ahí el uso de los ojos como parte del estudio [163]. De hecho, aunque no ha sido incluido en las tablas, la tasa de fallos considerando únicamente los ojos, giraba en torno al 40 %.

En cuanto a los resultados obtenidos haciendo uso del clasificador k-NN, puede apreciarse que los datos relativos a considerar los valores de píxeles de la propia imagen como descriptor proporcionaron mejores cifras cuando se trabajó

Prueba k-NN	Valores de píxeles	Histograma
GRAY_1-NN	**27.0 %**	43.0 %
ULBP_1-NN	45.0 %	43.5 %
SLBP_1-NN	43.0 %	43.0 %
GRAY_5-NN	**29.5 %**	41.5 %
ULBP_5-NN	46.3 %	41.0 %
SLBP_5-NN	38.4 %	46.0 %

Tabla 3.6: Tasa de fallos con el clasificador k-NN considerando todo el rostro. Para cada entrada en la tabla hay una denominación que hace referencia al nombre del enfoque aplicado y el método de clasificación aplicado. Así, para el enfoque aplicado existen tres posibilidades; ULBP, SLBP y GRAY que hacen referencia a los enfoques ULBP, SLBP e imagen en escala de grises respectivamente. Por último, el método de clasificación sugiere sólo dos posibilidades; 1-NN y 5-NN que hacen referencia a la clasificación cuando $k = 1$ y cuando $k = 5$, según sea el caso.

con escala de grises que cuando se hizo uso de la imagen resultante de preprocesar la imagen original con ULBP o SLBP. El motivo por lo que ocurre esta mejoría se explica desde una doble perspectiva; por un lado, no existe preprocesado previo en las imágenes en escala de grises, sino que se trabaja directamente con los valores de la imagen y esto favorece, en el caso de que existan zonas con mayor intensidad (sonrisas), que la clasificación sea más sencilla. De hecho, es apreciable cómo el hecho de añadir los ojos eleva esta tasa de fallos, lo que indica que el enfoque funciona de forma más precisa si sólo se considera la región bucal. Por otro lado, el preprocesado LBP está añadiendo un factor de compresión en los datos que los hace incompatibles para una comparación directa a través de la diferencia entre los vectores de características. Por el contrario, cuando se hace uso de un enfoque basado en histogramas, la tasa de fallos de los valores en la escala de grises empeora de forma considerable e, incluso, es peor que el uso de los enfoques LBP, que nuevamente vuelven a tener tasas de fallos muy similares entre sí.

Además, otro aspecto importante que merece ser destacado es la variación de comportamiento cuando se aplica el clasificador 1-NN ($k = 1$) y cuando el clasificador aplicado es el 5-NN ($k = 5$). El enfoque 5-NN proporciona mejores resultados que el 1-NN debido, fundamentalmente, a que maneja mayor información, pues la clasificación de un dato depende de las cinco muestras de entrenamiento más cercanas en el espacio de características.

La tabla 3.6 muestra los resultados al analizar todo el rostro usando el clasificador k-NN. Es apreciable cómo las cifras de error obtenidas a partir de los valores de las imágenes son bastante mejores que las obtenidas a raíz del análisis a través de los histogramas normalizados. Otro punto interesante es la causa por la cual se han obtenido, en este caso, peores tasas de error que las que había proporcionado el análisis único de las distintas partes de interés del rostro.

Prueba SVM	Valores Píx.	Hist.	PCA 40	PCA 90	PCA 130
GRAY_M	13.3 %	37.8 %	10.3 %	11.4 %	11.6 %
GRAY_ET+M	12.5 %	30.8 %	**09.7 %**	10.2 %	10.3 %
GRAY_SE+M	13.2 %	30.7 %	10.6 %	10.9 %	11.2 %
ULBP_M	23.5 %	25.3 %	27.2 %	26.3 %	25.1 %
ULBP_ET+M	32.1 %	22.3 %	29.9 %	36.6 %	36.0 %
ULBP_SE+M	31.2 %	**22.1 %**	29.0 %	32.2 %	30.1 %
SLBP_M	**16.4 %**	27.7 %	20.8 %	18.0 %	18.3 %
SLBP_ET+M	22.2 %	25.3 %	21.5 %	19.9 %	19.8 %
SLBP_SE+M	21.5 %	25.0 %	21.4 %	21.5 %	24.3 %

Tabla 3.7: Tasa de fallos con el clasificador SVM considerando partes del rostro. Para cada entrada en la tabla hay una denominación que hace referencia al nombre del enfoque aplicado así como de las partes del rostro involucradas en la prueba. En cuanto al enfoque aplicado existen tres posibilidades; ULBP, SLBP y GRAY que hacen referencia a los enfoques LBP uniforme, LBP simplificado e imagen en escala de grises respectivamente. Para las partes del rostro involucradas en la prueba también hay tres posibilidades; M, SE, ET que significan boca, ojos separados y máscara conjunta de ojos respectivamente (véase la Figura 3.26).

Los resultados sugieren que analizar sólo boca y ojos proporciona información más valiosa que analizar todo el rostro en su conjunto. Los peores datos son los proporcionados por el enfoque que hace uso del ULBP, siendo superado por el SLBP y el uso de las características de los valores de la imagen en escala de grises.

En la segunda parte de este experimento se ha hecho uso del clasificador SVM. Una primera ventaja apreciable respecto al uso de este tipo de clasificador radica en la drástica reducción del tiempo de ejecución durante el proceso de clasificación. Además, la prueba con el clasificador SVM incluye el enfoque PCA que permite obtener representaciones faciales a través de los datos de entrada que son extraídos de las proyecciones de PCA sobre las imágenes en escala de grises o sobre las imágenes preprocesadas con LBP. Por ello, en esta parte del experimento coexisten hasta cinco posibles enfoques de representación de la información: valores de píxeles de la imagen, histogramas normalizados y diversos números de componentes PCA (40, 90 y 130). Cuando se hace referencia a 40, 90 o 130 proyecciones, se está especificando la dimensión del espacio de representación, es decir, indica el número de autovectores usados para proyectar la imagen facial.

En cuanto a los resultados obtenidos, en la tabla 3.7 puede apreciarse cómo las tasas de fallos de las imágenes en escala de grises obtienen la mejor valoración en casi todas las pruebas realizadas. Ninguna prueba que hace uso del enfoque LBP supera esas cifras. No obstante, el enfoque basado en ULBP experimenta una notable mejora cuando se hace uso de los histogramas normalizados que,

Prueba SVM	Valores Píx.	Hist.	PCA 40	PCA 90	PCA 130
Imagen Gris	10.6 %	42.9 %	11.3 %	10.4 %	**10.1 %**
ULBP	27.2 %	35.7 %	**25.9 %**	27.6 %	27.0 %
SLBP	**26.2 %**	38.2 %	27.5 %	27.7 %	28.8 %

Tabla 3.8: Tasa de fallos con el clasificador SVM considerando todo el rostro.

salvo en esta ocasión, siempre es superado por los resultados cosechados cuando se aplica el preprocesado SLBP. Ello se explica a través del trabajo de sus autores, Tao y Veldhuis, que ya afirmaron que el enfoque SLBP proporcionaba un enfoque contextual mucho más rico dentro del análisis facial [202]. Por otro lado, restringir los datos de entrada a las zonas más expresivas para la sonrisa (ojos y boca) permite que, de nuevo, el enfoque en escala de grises obtenga el mejor rendimiento, superando de largo los resultados obtenidos usando el clasificador k-NN. Este resultado sugiere que enfocar la investigación hacia esas zonas de la cara permite eliminar ruido y así poder construir clasificadores más eficaces.

Por otro lado, cuando no se realiza ningún preprocesado (intensidad en escala de grises), se obtiene el mejor resultado mientras que ULBP presenta las peores tasas de fallos. En el caso de hacer uso del PCA para obtener el descriptor de características, el comportamiento es bastante similar al experimentado anteriormente, pues al no hacer uso de preprocesado alguno favorece la detección de sonrisas. Además, el enfoque del PCA sugiere que incrementar el espacio de representación no siempre produce los mejores resultados. Por último, cuando la representación de los datos se obtiene a través del histograma normalizado, hacer uso de un preprocesado ULBP permite obtener los mejores resultados para dicho preprocesado. Este enfoque parece modelar de forma más precisa la textura de la sonrisa aunque el histograma pierda la localización relativa de la información. Este resultado es similar al obtenido cuando se aplica el SLBP, cosa que no ocurre cuando no se considera preprocesado alguno. Aplicar únicamente la intensidad de los píxeles de la imagen en un histograma hace que el enfoque sea muy sensible a la localización relativa de la información. El motivo de ello es que la sonrisa presenta una alta intensidad en sus píxeles, pero el rostro también puede presentar una intensidad similar dependiendo de la tonalidad de la piel y la iluminación (la imagen está en escala de grises), en cambio, la aplicación de LBP no sugiere el mismo comportamiento pues, según lo descrito en la sección 3.2.1.1, detecta de forma más eficaz posibles cambios en las texturas locales y se muestra más robusto frente a cambios de iluminación.

Una vez analizados los resultados con el clasificador SVM es posible extraer algunas conclusiones interesantes. En primer lugar, el uso del preprocesado SLBP tiene un comportamiento intermedio por regla general al rendimiento obtenido entre el ULBP y el obtenido por la imagen original en escala de grises. De hecho, con el SLBP, cuando se hace uso de los histogramas, obtiene mejores tasas que el enfoque con escala de grises pero peores que el uso del ULBP. Sin

embargo, cuando se hace uso del PCA o de la extracción de valores de la imagen de forma directa, el SLBP se comporta mejor que el ULBP pero peor que la imagen sin preprocesar.

Además, por lo comentado anteriormente, se ha demostrado que procesar ciertas zonas del rostro (ojos y boca) suele proporcionar mejor rendimiento que el hecho de trabajar con el rostro completo (considerándolo como un único bloque a procesar). Precisamente, el trabajo de Li ya trabaja con esta hipótesis, en la que las partes del rostro proporcionan una información más relevante que la totalidad del mismo [131]. Para SVM, el hecho de usar regiones concretas del rostro se traduce en una drástica reducción de la dimensión del espacio de representación cuando se hace uso de los valores de la imagen, efecto similar que ocurre con el PCA. Si se observa con atención las diferencias entre los resultados obtenidos con ULBP y SLBP, puede apreciarse que la versión simplificada puede ser utilizada como un buen enfoque de representación para imágenes con texturas más o menos uniformes, el inconveniente es que la cara tiene una textura muy variable; ya no sólo en términos de resolución sino que también la posible presencia de componentes estructurales (vello facial, gafas, etc.) puede alterar de forma notable esta textura entre diferentes individuos. Por tanto, cuando el área se restringe a la zona concreta de la boca, el rendimiento mejora debido a que otras regiones, susceptibles de introducir este ruido, han sido eliminadas.

3.2.1.5. Combinación de enfoques para la detección de sonrisas

En la literatura se puede encontrar que, en múltiples ocasiones, la aplicación de una combinación de técnicas proporciona mejores resultados que la mera aplicación de un único enfoque. De hecho, existen dos formas de plantearlo. En la primera, la combinación sucede en etapas sucesivas y de forma progresiva. Es decir, primero se aplica un enfoque y el resultado de dicho enfoque sirve de entrada para un segundo método y así sucesivamente. Siguiendo este razonamiento Cristinacce y Cootes investigaron una técnica multi-etapa para la detección de características faciales [42]. Para ello aplican diversos detectores en cascada: el primero detecta la localización del rostro en la imagen, el segundo detecta una estimación de la localización de los puntos más relevantes de la cara, posteriormente se aplica un detector que analiza la fiabilidad de estos puntos y, por último, un enfoque basado en *Active Appearance Model* [198] para extraer las regiones de interés.

La segunda forma para obtener una aproximación a partir de una combinación de distintos clasificadores resulta de una mezcla de los resultados de cada uno de los clasificadores. A diferencia de la anterior combinación progresiva, en esta ocasión, no se tiene un clasificador para cada tarea, sino que varios clasificadores realizan la misma tarea en paralelo para que, finalmente, a través de un sistema de votación se pueda determinar un resultado final. A pesar de que la entrada de datos al sistema es única, es posible que estos clasificadores no trabajen con los mismos datos de entrada, de hecho existen dos posibilidades; la primera es combinar varios clasificadores en los que la entrada de datos sea idéntica (o, al menos, se solape) [206], o bien, una segunda posibilidad es a través

de la combinación de enfoques en los que se trabaja con subconjuntos disjuntos de los datos de entrada [131]. Sea cual sea el conjunto de datos de entrada, todos los clasificadores pertenecientes al nuevo enfoque emiten un veredicto en forma de etiqueta y todos ellos comparten exactamente el mismo dominio de posibles etiquetas.

Características de la propuesta

En esta ocasión, se considera el segundo esquema de combinación con el objetivo de lograr mejorar las cifras obtenidas de forma individual por cada uno de los clasificadores utilizados en la sección anterior. Para ello, se combinan los distintos clasificadores generados que han proporcionado una menor tasa de error en el experimento anterior. Además, cada uno de los enfoques considerados trabaja con una codificación distinta. El objetivo está enfocado en, no sólo lograr una menor tasa de error, sino que también se pretende encontrar un enfoque que responda de forma positiva ante posibles cambios de iluminación, por lo que preservar las codificaciones LBP es esencial de acuerdo con Ojala [158]. Esta es la razón por la que el nuevo enfoque está compuesto por la combinación de clasificadores que han sido generados a partir de las siguientes técnicas:

- **App1** - El espacio de PCA obtenido a raíz de las imágenes originales en escala de grises considerando las regiones de los ojos y la boca.
- **App2** - La concatenación de histogramas obtenidos a raíz de la imagen original preprocesada con ULBP considerando las regiones de los ojos y la boca.
- **App3** - La concatenación de valores de píxeles de la imagen preprocesada con SLBP considerando, únicamente, la región de la boca.

Tras analizar distintas posibilidades de combinación, como por ejemplo generar una ecuación en la que se atribuye distinto peso a las decisiones de los clasificadores elegidos para conformar una decisión final, se llegó a la conclusión de que el mejor resultado es aquel proporcionado por un sistema de simple votación. Así, el enfoque adoptado para determinar la etiqueta de una imagen es el propuesto en la ecuación 3.10.

$$Resp(t) = \sum_{app=1}^{n} Dt_{app}$$

$$\Gamma(t) = \begin{cases} +1 & \text{Si } Resp(t) > 0 \\ -1 & \text{Si } Resp(t) \leq 0 \end{cases} \tag{3.10}$$

Donde Dt_{app} es el resultado de aplicar el clasificador generado a partir del enfoque *app* con la imagen t. Los enfoques a los que hace referencia *app* son los

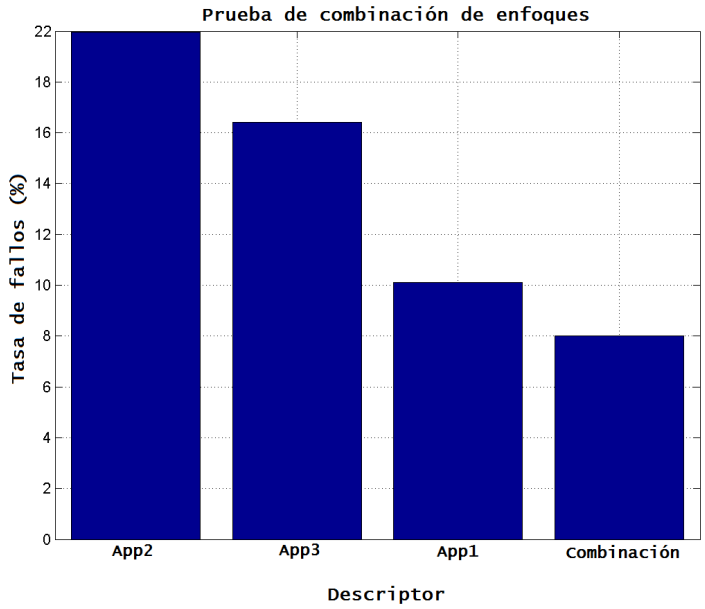


Figura 3.27: Comparativa de distintos enfoques para la detección de sonrisas.

previamente mencionados App1, App2 y App3 respectivamente. Debe tenerse en cuenta que, de acuerdo lo explicado en la sección 3.2.1.3, cada clasificador sólo puede generar salidas dentro del rango $[-1; +1]$. Quedando las salidas positivas para cuando el rostro es clasificado como sonriente y las negativas para aquellos rostros que no son detectados como no sonrientes.

Conjuntos de datos

Se utilizó la misma base de datos que en el experimento anterior. Además el experimento se ejecutó bajo los mismos parámetros que los especificados para el primer experimento de esta sección, esto es, se ejecutó diez veces la prueba en la que, para cada ejecución, de forma aleatoria, la mitad fue utilizada para entrenar los clasificadores mientras que la otra mitad fue utilizada para hacer la prueba del experimento.

Experimentos

Tal y como puede apreciarse en la Figura 3.27, la tasa de fallos registrada con este nuevo enfoque es del 8 %, en la que la tasa de falsos negativos se sitúa en torno al 4.4 % y la de falsos positivos en torno al 10.6 %. Si se observa el gráfico de dicha imagen con detenimiento, puede apreciarse como la mejora conseguida es considerable respecto a los enfoques LBP y bastante significativa respecto al

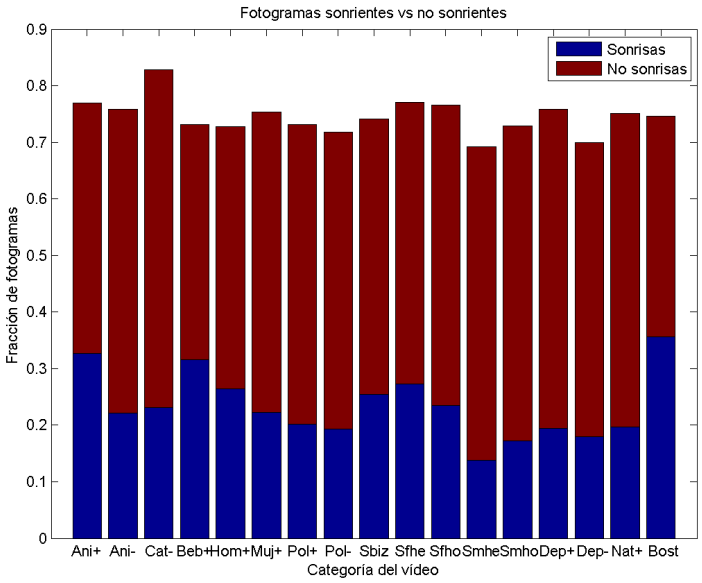


Figura 3.28: Aplicación del detector de sonrisas combinado sobre Vedana.

enfoque aportado por el App1. Esto se explica en términos de iluminación, es decir, la base de datos analizada recoge imágenes de la red en situaciones de iluminación bastante variable. Lo que sucede es que el enfoque App2 adquiere una especial relevancia cuando se presentan este tipo de cambios, lo que unido al equilibrio que proporciona el clasificador App3 y las buenas cifras proporcionadas por el de App1, permiten generar un clasificador conjunto que tiene una tasa de acierto en torno al 92 %.

De esta forma, si aplicamos el clasificador resultante a los distintos fotogramas de la base de datos Vedana, se pueden obtener una relación entre la categoría del vídeo observado y la presencia de sonrisas en base a la detección automática realizada. En esta ocasión sólo se analizaron los fotogramas de la cámara C3 (Webcam). Haciendo uso de esta combinación de detectores, la Figura 3.28 muestra una gráfica en la que cada columna representa una temática. La aplicación se hace sobre el número de fotogramas detectados por ENCARA2 para la normalización de las imágenes. En ocasiones, la fracción de fotogramas no llega a 1 porque existen fotogramas para los que no hubo detección facial y, por tanto, no fueron analizados. En el eje X pueden apreciarse las distintas categorías siguiendo el mismo orden que la tabla 3.2. De esta forma, Ani+ se corresponde con secuencias de vídeos positivas de animales, Cat- con catátrofes naturales, etc. Resulta apreciable cómo los vídeos de bebés (Beb+) causaron un impacto bastante positivo en las personas, pues casi la mitad de los fotogramas analizados se corresponden con personas sonrientes. Además, otros vídeos que

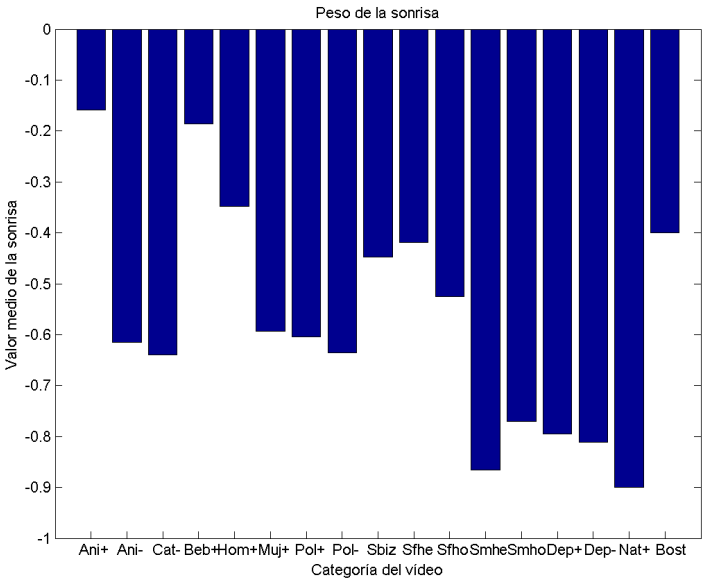


Figura 3.29: Valor promedio de la sonrisa sobre Vedana. Cuanto mayor es el valor, más sonrisas produjo categoría de la secuencia de vídeo.

causaron un efecto positivo en los participantes fueron el de animales positivo (Ani+) y el de los bostezos (Bos). En este último, salían personas incitando a bostezar a los participantes, lo que es posible que al contener las ganas de bostezar, los participantes sonrieran. Por otro lado, los vídeos con contenido sexual no generan muchas sonriesas (Sbiz, Sfhe, Sfho, Smhe y Smho), siendo aquellos con contenido homosexual masculino (Smho) los que más rechazo produjeron. La misma interpretación, desde un punto de vista distinto, puede observarse en la Figura 3.29. En esta ocasión, se muestra el valor promedio de la sonrisa de todos los participantes para cada una de las categorías. De esta forma, puede apreciarse de forma más clara, cómo los participantes respondieron de manera más, o menos, sonriente a las distintas categorías. De hecho, en dicha Figura 3.29, se ratifica lo comentado anteriormente; las categorías que más sonrisas produjeron fueron las de animales y bebés, mientras que algunas secuencias con contenido sexual, de deportes o de naturaleza, apenas generaron una respuesta sonriente.

3.2.1.6. La firma digital

Como se ha podido apreciar, en la sección anterior se describe un clasificador basado en la aplicación de distintos enfoques para el estudio de la sonrisa. Estos enfoques están basados en el uso de técnicas existentes (como las descritas en las secciones 3.2.1.1 y 3.2.1.2) con el fin de poder refinar técnicas de clasificación

Figura 3.30: Estas imágenes pertenecen a un mismo objeto (un corazón), sin embargo, no comparten atributos inherentes a una imagen tales como color, textura o bordes. Lo que sí comparten es una plantilla geométrica de similitud [187].

que proporcionen un resultado óptimo. Sin embargo, existe la posibilidad de desarrollar nuevas técnicas que sirvan para proporcionar descriptores válidos para la detección de elementos contextuales con una estructura propia y, la sonrisa, es una expresión que marca un cambio notable en la estructura facial del ser humano.

En esta ocasión se presenta un descriptor de imágenes fundamentado en el concepto de auto-similitud desarrollado por Irani y que es capaz de capturar la estructura general de una imagen [187]. En su trabajo, Irani presenta un método para medir la similitud entre entidades visuales (imágenes o vídeos) basándose en la correlación de auto-similitud. De hecho, según los autores de este trabajo, la distribución de patrones internos de una imagen permite capturar descriptores eficientes que mantienen una proporción similar independientemente de los cambios geométricos o distorsiones lumínicas que pueda sufrir la escena. En la Figura 3.30 puede apreciarse como, a pesar de la diversa contextualización, una figura puede compartir características inherentes a la distribución de patrones a lo largo de varias imágenes. De esta forma, el descriptor obtenido para una imagen es similar a otros descriptores de otras imágenes que compartan el mismo patrón estructural de la imagen del primer descriptor sin importar si la textura o el color difieren.

Características del enfoque

En el enfoque que se propone en esta sección, la imagen se divide en pequeños bloques denominados celdas y, a través de la comparación de estas celdas, se obtiene un vector de características que describe una correlación del aspecto local de una imagen. La firma digital (*Digital Signature* (DS)) es un descriptor calculado a partir de escalar y dividir una imagen en $n \times n$ celdas, en la que cada celda se corresponde con un área cuadrada (también denominado *patch*) de $m \times m$ píxeles. Dada una imagen, tanto el número de celdas como el tamaño de las mismas influyen en el grado de generalización del descriptor obtenido con respecto a dicha imagen; un bajo número de celdas puede no capturar múltiples detalles estructurales mientras que demasiadas celdas pueden generar un descriptor demasiado exhaustivo. Puede apreciarse que la propuesta del enfo-

que es cuadrada para el tamaño, tanto del patch como de la imagen, ello es así porque se ha trabajado con esta configuración del enfoque. Sin embargo, el enfoque también puede ser adaptado para calcular el descriptor en base a celdas no cuadradas. Una vez la imagen es dividida y seleccionado un patch principal de $m \times m$ píxeles (el cual puede, o no, corresponderse con una celda de la imagen dividida), éste es comparado con todas las celdas previamente calculadas en la imagen. Con el fin de aprovechar al máximo las características generales del descriptor, esta comparación se produce a través del cálculo de la suma de diferencias cuadradas (*Sum of Squared Differences* (SSD)) entre los valores de los píxeles de las regiones implicadas. Además, cada comparación, entre el "centro seleccionado" y las celdas de la imagen, se almacena en el denominado descriptor DS que tiene un total de $m \times m$ elementos consecutivos.

Desde una perspectiva matemática, una firma digital z de una determinada celda p, es una función que, a través de la comparación entre celdas, genera un valor que cae en una categoría disjunta (similar a lo ocurrido con los *bins* de un histograma). Así, si se considera *ncells* como el número total de celdas de la imagen, *Wcsize* la anchura para cada celda y *Hcsize* la altura para cada celda, entonces el descriptor para cada canal de la imagen queda definido según la ecuación 3.11.

$$DS_z = \sum_{p}^{ncells} \mid \sum_{i}^{Wcsize} \sum_{j}^{Hcsize} cell_z(i,j) - \sum_{u}^{Wcsize} \sum_{v}^{Hcsize} cell_p(u,v) \mid \qquad (3.11)$$

Este enfoque es robusto frente a colores, contrastes y texturas, pues no considera estas características durante su desarrollo. De hecho, las imágenes quedan descritas en función de su estructura general, algo parecido a lo que ocurre en el trabajo de Irani [187]. Por ejemplo, una imagen cuya mitad superior sea blanca y cuya mitad inferior sea negra va a generar un mismo descriptor que si la imagen tuviera la mitad superior negra y la mitad inferior blanca. La correspondencia de aspectos locales es exactamente la misma; la mitad superior difiere de la mitad inferior. Por otro lado, el descriptor de la firma digital es especialmente sensible para describir puntos definidos por un detector escalar de puntos salientes (como DoG [44] o SURF [15]). Además resulta interesante observar la Figura 3.31, en la que se expone el descriptor DS en forma de código de barras únicamente con el objetivo de mostrar una representación válida.

De esta forma, dada una imagen de entrada con un conjunto de puntos escalares salientes o una región estimada donde podría estar localizada una estructura de interés (como, por ejemplo, la boca), un número de n celdas y el tamaño de m píxeles que la define, la firma digital es calculada de la siguiente forma:

1. La imagen es escalada y dividida en diferentes celdas ubicadas dentro de una plantilla de $(n \times m) \times (n \times m)$ píxeles.

2. La plantilla queda dividida en $n \times n$ celdas independientes, cada una con un tamaño de $m \times m$ píxeles.

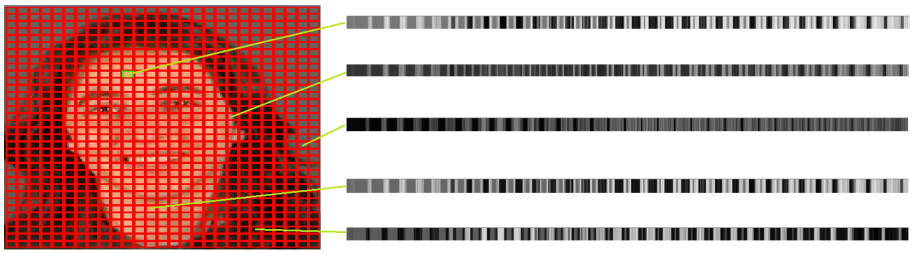


Figura 3.31: Creación de la firma digital. En el ejemplo puede observarse cómo la imagen ha sido dividida en una plantilla compuesta por celdas. Las imágenes con forma de código de barras representan todas las comparaciones entre cada celda (poseedora de dicho descriptor) con el resto de celdas de la plantilla. Cada celda tiene un descriptor asociado fruto de su comparación con el resto de celdas. Fuente de la imagen utilizada para este ejemplo: DaFEx Database [14].

3. Cada celda es comparada con el resto de celdas de la plantilla y, cada resultado de dicha comparación, se almacena en un vector de $n \times n$ elementos. De esta forma, cada celda genera su propio descriptor DS.

Adaptación del enfoque al estudio del caso

En las pruebas ejecutadas con la firma digital se ha podido apreciar el excelente rendimiento que ofrece tanto cuando se utiliza para la detección de sonrisas, como cuando sirve de base como método para segmentar una región de una imagen [84]. En cuanto al experimento de detección de sonrisas, es necesario describir de forma breve el proceso seguido para la incorporación de la firma digital a la base de pruebas ya desarrollada. Por regla general, una sonrisa tiene una forma bien diferenciada; las bocas presentan una región más clara (los dientes) que el resto de la zona y éste es, precisamente, el cambio que el descriptor debe, presumiblemente, recoger. Así, en lugar de hacer uso de la suma de diferencias cuadradas (SSD) para comparar celdas, las celdas se comparan a través de la suma de diferencias. En el caso de haber hecho uso de la SSD se hubiera generado un descriptor difuso, es decir, una boca cerrada generaría un descriptor muy similar al de una boca abierta; los labios de color rojizo están rodeados de piel con un color de piel diferente a los mismos y, de la misma forma, los dientes blanquecinos están rodeados de los labios que también tienen una tonalidad distinta. La Figura 3.32 muestra un ejemplo para el cálculo y obtención del descriptor DS con una partición de 11×11 celdas, cada una con un tamaño de 10×10 píxeles. Así, una vez que el detector facial devuelve la imagen normalizada de un rostro, el descriptor de la firma digital se aplica de la siguiente forma:

1. La región de la boca es escalada y dividida en una plantilla de $(n \times m) \times (n \times m)$ píxeles.

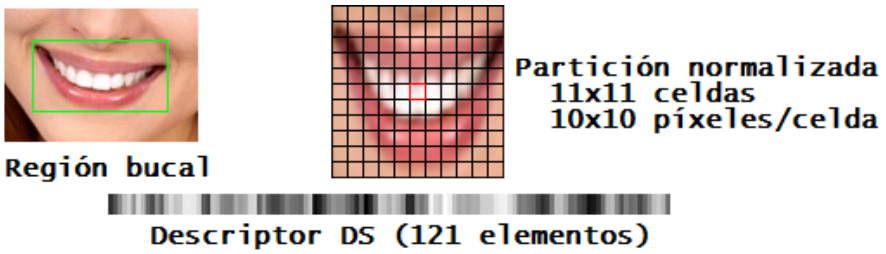


Figura 3.32: Se selecciona una celda principal para realizar el conjunto de comparaciones con el fin de obtener un único descriptor. Esta celda principal debe estar ubicada en un punto estratégico que ayude a diferenciar la estructura de la boca en caso de cambio por la sonrisa, de ahí a que el punto seleccionado sea el centro de la boca.

2. Dicha plantilla está dividida en $n \times n$ celdas, cada una con un tamaño de $m \times m$ píxeles.

3. La celda principal de $m \times m$ píxeles es seleccionada de entre todas las celdas de la imagen.

4. Esta celda principal es comparada con todas las celdas que conforman la plantilla y, cada resultado fruto de la comparación, se almacena en un vector de $n \times n$ elementos, siendo este vector el descriptor DS resultante.

Para el desarrollo del experimento de detección de sonrisas se ha comparado con los enfoques propuestos en la sección 3.2.1.4 conocidos como App1, App2 y App3, además de la nueva propuesta:

- **DS** - La concatenación de valores de píxeles de la imagen preprocesada con SLBP considerando, únicamente, la región de la boca.

Conjuntos de datos

En la fase de entrenamiento se ha optado por entrenar con imágenes estáticas, haciendo uso del conjunto de imágenes ya utilizado anteriormente. Para la fase de pruebas se hizo uso de la base de datos DaFEx [14]. En dicha colección de vídeos puede apreciarse ocho actores profesionales mostrando siete expresiones faciales (las 6 expresiones faciales básicas y una neutral) con tres niveles de intensidad (bajo, medio, alto). Por tanto, cada actor simuló, hasta en dos ocasiones, una expresión facial y un nivel de intensidad. Haciendo uso del detector facial, se extrajo el conjunto de fotogramas de los 48 vídeos que muestran una persona sonriente en dicha base de datos para, posteriormente normalizarlas de acuerdo al formato de 59×65 que acepta el conjunto de clasificadores.

DaFEx	Enfoque	Tasa Fallos	FP	FN
Base de datos Completa	App1	13.5 %	16.3 %	11.1 %
	App2	21.4 %	23.6 %	19.5 %
	App3	20.7 %	21.3 %	20.2 %
	DS	**11.5 %**	13.0 %	10.2 %
Intensidad BAJA	App1	16.2 %	19.1 %	13.6 %
	App2	24.4 %	24.0 %	24.9 %
	App3	22.9 %	23.8 %	22.1 %
	DS	**14.3 %**	16.6 %	12.3 %
Intensidad MEDIA	App1	13.9 %	14.4 %	13.6 %
	App2	21.1 %	22.5 %	19.9 %
	App3	20.6 %	20.4 %	20.9 %
	DS	**12.6 %**	13.7 %	11.6 %
Intensidad ALTA	App1	10.6 %	15.5 %	6.3 %
	App2	18.8 %	24.5 %	13.7 %
	App3	18.6 %	19.7 %	17.6 %
	DS	**7.7 %**	8.7 %	6.8 %

Tabla 3.9: Comparativa de la aplicación de la firma digital sobre la base de datos DaFEx en función de la intensidad de la expresión. La tasa de error está calculada sobre el número total de fotogramas y, para cada vídeo, la sonrisa se detectó siempre en, al menos, un fotograma.

Experimentos

Tal y como puede apreciarse en la tabla 3.9, la nueva propuesta, DS, proporciona el mejor resultado en cada uno de las pruebas ejecutadas. Ninguno de los enfoques basados en LBP supera las cifras proporcionadas por la firma digital. Puede apreciarse que las cifras, sin considerar una ventana temporal para los tres enfoques estudiados en la tabla 3.11, ofrecen un comportamiento bastante similar al observado en dicha tabla. El enfoque App1 obtiene mejores tasas que los enfoques basados en LBP, como ya habían sugerido los experimentos previos. Sin embargo, no mantiene un buen equilibrio entre las tasas de falsos positivos y de falsos negativos. Para la firma digital, la tasa de fallos es similar a la obtenida por el App1 y, de nuevo, las tasas de fallo más bajas se dan cuando no existe preprocesado LBP previo.

Se incluye un estudio pormenorizado de los distintos parámetros que afectan a la firma digital (el tamaño y el número de celdas) con el fin de encontrar la configuración que menor tasa de error genere. En la Figura 3.33 queda demostrado que, para la detección de sonrisa, un conjunto de 10×10 celdas con tamaño de 3×3 píxeles ofrecen la mejor tasa debido, principalmente, a la cercanía existente entre el tamaño de la plantilla generada (30×30 píxeles) y el tamaño original de la captura extraída de la zona de la boca (20×12 píxeles). En dicha gráfica también puede apreciarse que el peor resultado es el obtenido para las configuraciones con un número de celdas menor a 10×10 debido a la considerable pérdida

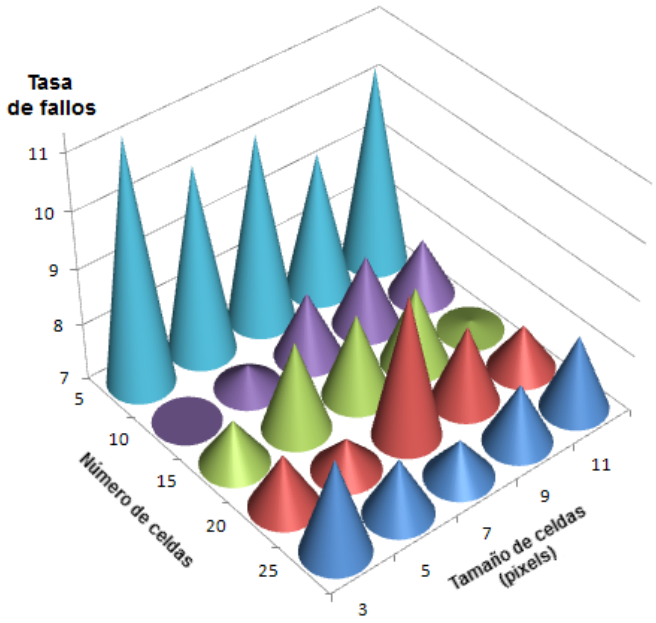


Figura 3.33: Exploración de posibles configuraciones de la firma digital para la detección de sonrisas.

de información durante la fase de escalado en la etapa de normalización. Otras configuraciones como, por ejemplo, mayor cantidad de celdas o mayor tamaño de celdas, ofrecen un comportamiento bastante irregular debido al hecho de que la información extraída se ve afectada por la morfología de la plantilla generada; cuando las imágenes son ampliadas para centrar la plantilla, se crea nueva y redundante información. Por otro lado, no se ha introducido solape para la comparación entre celdas vecinas porque, cuando se introdujo solape, las cifras fueron peores en torno a un 15 %. Esto es así porque las celdas son demasiado pequeñas como para que exista solape entre ellas y, de haber solape, la cantidad de información redundante se multiplica. Cabe resaltar que, a diferencia de lo que sucede con el PCA, la firma digital ofrece un mejor balance entre la tasa de falsos positivos y la de falsos negativos, lo que lleva a la conclusión de que este nuevo enfoque obtiene un comportamiento más positivo que los anteriormente presentados. Sin embargo, a diferencia de los enfoques anteriores, la firma digital no puede ser aplicada a detectores en sistemas de tiempo real. El procesado de la diferencia entre las distintas celdas incrementa el tiempo de cómputo en torno a un 50 % a diferencia de los enfoques presentados anteriormente.

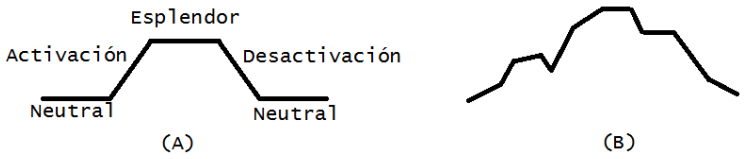


Figura 3.34: Ciclo de vida de una expresión facial. En la imagen de la izquierda puede apreciarse el ciclo de vida de una expresión facial genuina, mientras que en la imagen de la derecha puede apreciarse el irregular ciclo de vida de una expresión facial fingida [123].

3.2.2. Un enfoque dinámico

En la sección 2.3.1 se explicó que la expresión facial no se genera de forma discreta en un determinado instante sino que lo hace de forma continua a lo largo de un determinado periodo de tiempo. Koelstra denominó este fenómeno de formación de expresiones en el rostro humano como una secuencia temporal que, en función del momento en el que se encuentra la manifestación de la expresión, puede ubicarse en alguno de los siguientes segmentos: neutral, activación, esplendor y desactivación [123]. Lo que afirma esta teoría es que las expresiones, no se accionan a través de un interruptor, sino que la evolución del mapa facial de una persona determina la intensidad de la expresión y las fases que dicha expresión puede sufrir a lo largo de su ciclo de vida. En este contexto, en la Figura 3.34 pueden observarse dos líneas temporales de comportamiento que pueden sufrir las expresiones faciales a lo largo de su ciclo de vida natural; en la imagen situada a la izquierda puede apreciarse el comportamiento natural ya descrito. La etapa de activación comienza cuando los músculos comienzan a contraerse y aumentan la intensidad de su movimiento mientras que la etapa de esplendor es cuando la intensidad alcanza un estado estable. Por último, la etapa de desactivación nace de la relajación de los músculos hasta llegar a un estado de relax inicial que da lugar a una expresión facial neutral. Por otro lado, en la imagen situada en el lado derecho se muestra un conjunto de etapas en las que el movimiento muscular del rostro no es continuo ni espontáneo, dando lugar a una expresión facial fingida. Esto incluye una etapa de activación dubitativa, duración muy corta (o muy larga) de la etapa de esplendor y una desactivación muy brusca.

Otros autores afirmaron que la existencia del movimiento de un vídeo, con una determinada duración, permite analizar aspectos relevantes acerca de las expresiones faciales tales como acciones específicas de músculos, niveles de intensidad, asimetría de la expresión u otras características [105]. En la tabla 3.10 pueden apreciarse este conjunto de características así como especificaciones asociadas a las mismas.

Amparándose en esa teoría temporal de las expresiones faciales, esta parte de la investigación pretende analizar de forma meticulosa si, una vez encontrado un enfoque óptimo para la detección de sonrisas en estático, es posible aplicarlo

Características	**Especificaciones**
Acciones musculares específicas	Variables
Niveles de Intensidad	Pendiente de activación/desactivación Duración de cada etapa
Otras características temporales	Concurrencia con otras acciones (solape) Separación de otras acciones
Asimetría facial	Temporización Intensidad

Tabla 3.10: Características faciales temporales que son medibles.

en un entorno dinámico de expresiones aprovechando la coherencia temporal que ofrece el vídeo frente a la imagen.

Características del enfoque

En esta sección del capítulo de expresiones faciales se propone un método basado en la aplicación de derivadas temporales de las acciones faciales para cada uno de los enfoques destacados en las pruebas anteriores. De esta forma, el conjunto de enfoques que se han considerado para la aplicación del nuevo método son los propuestos en la sección 3.2.1.4 y que han sido etiquetados como App1, App2 y App3 respectivamente:

- **App1** - El espacio de PCA obtenido a raíz de las imágenes originales en escala de grises considerando las regiones de los ojos y la boca.
- **App2** - La concatenación de histogramas obtenidos a raíz de la imagen original preprocesada con ULBP considerando las regiones de los ojos y la boca.
- **App3** - La concatenación de valores de píxeles de la imagen preprocesada con SLBP considerando, únicamente, la región de la boca.

Como puede apreciarse, la firma digital no ha sido incluida dentro de esta combinación. Ello se debe a que el tiempo para procesar la firma digital es casi el doble que el resto de enfoques propuestos para la combinación. Este hecho ralentiza el sistema en exceso y, por ello, no ha sido considerado para el procesado dinámico de sonrisas.

En el nuevo método, una vez que se procesa un fotograma, se aplica una ponderación temporal. El esquema propuesto tiene dos ventajas principales sobre el reconocimiento de imágenes estáticas. En primer lugar, la precisión a la hora de reconocer una expresión facial será mejor debido a que se considera más de un fotograma para etiquetar un determinado fotograma. En segundo lugar, debido al análisis del fotograma principal y sus vecinos desde tres distintas perspectivas, la robustez del proceso de reconocimiento se presenta como un punto fuerte.

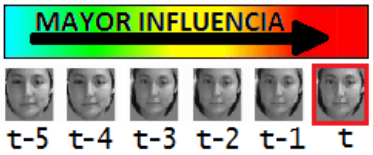


Figura 3.35: Influencia de los distintos fotogramas a la hora de clasificar un fotograma (enmarcado en rojo) en un instante t considerando $w_s = 5$.

Tal y como se expresa en la ecuación 3.12, antes de decidir si un fotograma se clasifica como sonriente o no sonriente, se analiza, tanto la respuesta de los fotogramas precedentes, como la del fotograma actual. Esta ventana temporal es el resultado de una combinación ponderada por parte de un clasificador en un determinado instante. Cuanto más cerca está el fotograma analizado del instante actual, mayor será el peso que le sea asignado (véase la Figura 3.35). De esta forma, el fotograma actual se procesa combinando las derivadas temporales de las accciones faciales para cada enfoque:

$$\begin{aligned} Dt_{app} &= \frac{\partial w(t) \partial \tau_{app}(t)}{\partial t} \\ &= \sum_{k=t-w_s}^{t} \left(\frac{1}{1 + sqrt(t-k)}\right) * (\tau_{app}(k)) \end{aligned} \tag{3.12}$$

Donde $\tau_{app}(k)$ es la decisión tomada por un clasificador sobre el fotograma en el instante k. De acuerdo lo explicado en la sección 3.2.1.3, cada clasificador sólo puede generar salidas dentro del rango $[-1, +1]$. Quedando las salidas positivas para cuando el rostro se clasifica como sonriente y las negativas para aquellos rostros que no son clasificados como no sonrientes. Por otro lado, w_s es el tamaño de la ventana que ha sido considerado. Este parámetro se define como el número de fotogramas previos que han sido tenidos en cuenta de cara a la clasificación del fotograma del instante t. Un tamaño de ventana elevado puede proporcionar una mayor seguridad a la hora de clasifica una sonrisa estable. Sin embargo, un valor elevado de este parámetro también puede afectar de forma negativa al proceso de clasificación durante las fases de activación y desactivación de la emoción (véase la Figura 3.34).

La segunda etapa consiste en la combinación de las derivadas temporales de los valores obtenidos por los clasificadores de la primera etapa con el objetivo de obtener un enfoque combinado que ofrezca una mayor precisión. Por lo tanto, el nuevo método está basado en una combinación ponderada de los clasificadores SVM basados en los tres enfoques mencionados anteriormente (véase Figura 3.36). La combinación de los enfoques se define como:

$$\Gamma(t) = \sum_{app=1}^{n} \psi_{app} * Dt_{app} \tag{3.13}$$

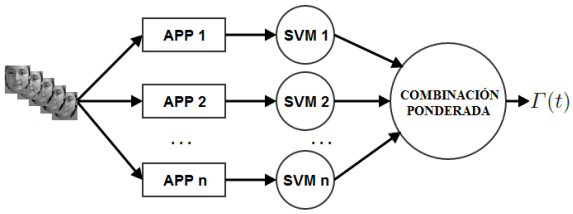


Figura 3.36: Esquema del clasificador dinámico

Donde ψ_{app} define el peso de cada clasificador. Un mayor peso se asigna a aquellos clasificadores que ofrezcan mejores tasas de acierto. Con este objetivo, un estudio experimental se lleva a cabo en la sección de los experimentos para definir los pesos de los enfoques considerados. En cualquier caso, esta asignación de pesos se realiza de forma que ningún clasificador puede decidir, de forma unilateral, en contra de los otros dos. Así, los pesos se asignan considerando que no puede haber un clasificador con la mayoría absoluta en el proceso de toma de decisión ($\Gamma(t)$). De esta forma, la influencia de un clasificador sobre un determinado fotograma depende de dos factores; el peso asociado con el clasificador (ψ_{app}) y la seguridad con la que el classificador *app* etiqueta el fotograma en cuestión. La idea consiste en asegurar que el mayor peso no garantice el éxito del clasificador que lo posee; los otros clasificadores pueden revocar una decisión débil. Así, el peso de cada clasificador se calcula de la forma descrita en la ecuación 3.14. Sea un conjunto de n clasificadores, el peso $\psi_i \in \mathbb{R}$ de un clasificador i viene definido como:

$$\forall i, j \in \mathbb{N} \implies \psi_i \leq \sum_{\substack{j=1 \\ j \neq i}}^{n} \psi_j \tag{3.14}$$

Conjuntos de datos

En cuanto al conjunto de datos utilizado en este experimento, es necesario redefinir un nuevo conjunto sobre el que trabajar, especialmente durante la etapa de pruebas. Resulta obvio que, si se pretende trabajar con secuencias de vídeo, se haga uso de una base de datos de expresiones faciales que incluya este formato y, además, sea muy extendida en la literatura. No obstante, en la fase de entrenamiento se ha optado por seguir un enfoque de imágenes estáticas, haciendo uso del conjunto de imágenes ya utilizado anteriormente con imágenes descargadas de Internet, además de imágenes de otras bases de datos faciales como la BIO-ID [90]. Todas estas imágenes fueron anotadas a mano para obtener la posición de los ojos además de ser etiquetadas de acuerdo con si sonríe o no sonríe el sujeto. Finalmente, estas imágenes han servido como entrada para entrenar los distintos clasificadores que son combinados en el enfoque propuesto.

Para la fase de validación, en primer lugar, se hizo uso de una base de datos

Figura 3.37: Distintas intensidades de emociones en la base de datos DaFEx. Del total de 48 vídeos con sujetos que simulan una expresión sonriente se extrajeron 12988 fotogramas, de los cuales 6783 son imágenes con un actor sonriente, mientras que las restantes 6205 son actores que no están sonriendo.

no pública del SIANI. El objetivo del uso de esta base de datos fue el de encontrar la configuración de la ventana temporal que resultase más apropiada para cada uno de los enfoques considerados. Esta base de datos contiene grabaciones espontáneas de 106 participantes. Se trata de una colección creada de forma explícita la detección de rostros [37]. Aunque no es posible distribuir esta colección de datos debido a cuestiones relacionadas con los derechos de imágenes de los participantes, los resultados obtenidos son incluidos con el fin de proporcionar al lector el rendimiento del enfoque sobre una base de datos menos restrictiva en cuanto a las condiciones del entorno de adquisición. De hecho, esta base de datos contiene 59386 fotogramas, de los cuales 17406 muestran rostros sonrientes y el resto, 41980 fotogramas, muestran rostros no sonrientes.

Una vez que se ha analizado la mejor configuración para cada enfoque, la base de datos DaFEx [14] proporciona un interesante marco porque su uso está muy extendido entre la comunidad a la hora de tratar el problema de la expresiones faciales. En dicha colección de vídeos puede apreciarse ocho actores profesionales mostrando siete expresiones faciales (las seis expresiones faciales básicas y una neutral) con tres niveles de intensidad (bajo, medio, alto). Por tanto, cada actor simuló, hasta en dos ocasiones, una expresión facial y un nivel de intensidad (véase la Figura 3.37).

Debido a que el sistema propuesto procesa secuencias de vídeo, se ha hecho uso de rostros [37] que, combinando distintos rasgos del rostro, ofrece un rendimiento superior a otros detectores que sólo analizan una característica en particular del mismo. Los rostros detectados son recortados y normalizados de acuerdo al formato de 59×65 que acepta el conjunto de clasificadores utilizados.

Experimentos

En primer lugar, se realizó el análisis de la ventana temporal sobre la base de datos SIANI con el objetivo de definir la dimensión más apropiada de la ventana para cada uno de los enfoques analizados. De esta forma, se obtienen los siguientes resultados para cada uno de los enfoques considerados; App1 (w_s=5)

con una tasa de fallos de 17.4 % (la tasa de falsos negativos es de 15.3 % y la tasa de falsos positivos de 22.3 %), App2 (w_s=15) con una tasa de fallos de 23.8 % (la tasa de falsos negativos es de 22.3 % y la tasa de falsos positivos de 27.2 %) y App3 (w_s=20) con una tasa de fallos de 26.7 % (la tasa de falsos negativos es de 25 % y la tasa de falsos positivos de 30.6 %). Como puede apreciarse, el estudio temporal sobre la base de datos del SIANI muestra que el tamaño de ventana más apropiado para cada enfoque varía. El enfoque que obtiene las mejores tasas (App1) requiere de un tamaño de ventana menor que los otros enfoques. Por el contrario, aquellos enfoques que consideran la textura de la imagen (App2 y App3) hacen uso de un tamaño de ventana mayor con el fin de obtener sus mejores resultados. Por lo tanto, la correlación entre el tamaño de la ventana y la precisión de cada enfoque sugieren que cuanto mayor es la precisión, mayor es la confianza del enfoque en sus resultados.

En la tabla 3.11 pueden apreciarse los resultados obtenidos para cada clasificador después de procesar el conjunto de vídeos sonrientes de DaFEx de acuerdo con la ecuación 3.12. En dicha tabla, las tasas de fallos son ligeramente diferentes a las obtenidad anteriormente con la base de datos del SIANI. Ello se debe a la mayor variedad de poses y movimientos rígidos de la cabeza que ocurren en la base de datos del SIANI, lo que introduce una mayor dificultad a la hora de computar la detección de la emoción. Sin embargo, de acuerdo a lo ocurrido anteriormente con la base de datos del SIANI, el mejor resultado con DaFEx se obtiene sin hacer ningún preprocesado LBP previo, haciendo uso directamente de las imágenes en escala de grises y del App1. Por su parte, el App2 presenta tasas realmente interesantes, modelando de forma apropiada las texturas de la sonrisa a través del uso de histogramas normalizados. Por último, y al contrario que el App2, el App3 consigue un mejor equilibrio entre la tasa de falsos positivos y negativos, aunque la tasa de fallos es bastante parecida al enfoque App2.

Una vez que se han obtenido diferentes conclusiones acerca del tamaño de ventana temporal óptimo para cada enfoque, se propone el nuevo enfoque de acuerdo con la ecuación 3.13. Los experimentos haciendo uso de esta ecuación han demostrado que el tamaño de ventana temporal de 5 fotogramas ha resultado el mejor para cada prueba (considerando un vídeo con 30 fps). En la Figura 3.38 pueden apreciarse dos gráficas del clasificador temporal aplicado sobre dos vídeos grabados para la base de datos SIANI. Las gráficas incluyen; 1) la anotación manual de las distintas secuencias de vídeo (+1 fotogramas imágenes y -1 imágenes no sonrientes), 2) los resultados obtenidos con cinco tamaños distintos de ventana temporal (rojo, verde, azul, celeste y negro para los tamaños de ventana temporal de 5, 8, 10, 15 y 18 respectivamente). Además, la línea horizontal marca la frontera entre el estado sonriente (por encima de dicha línea) y el no sonriente (por debajo de dicha línea). En primer lugar, el etiquetado a mano sólo tiene dos posibles estados, on (+1) / off (-1), y no permite discernir de forma clara las etapas de activación o desactivación, la pendiente es máxima sin que exista una evolución natural de la expresión de acuerdo con lo expuesto en la Figura 3.34. En la gráfica de la izquierda, el clasificador muestra un comportamiento más natural del ciclo evolutivo de la expresión facial asegurando una

DaFEx	Enfoque	Tam. Ventana	Tasa Fallos	FP	FN
	App1	5	11.6 %	16.2 %	7.3 %
Base Datos	App2	15	18.8 %	20.1 %	17.7 %
Completa	App3	20	20.1 %	21.1 %	19.6 %
	App1	5	13.4 %	14.4 %	12.0 %
BAJA	App2	15	20.9 %	20.0 %	21.6 %
Intensidad	App3	20	22.2 %	23.3 %	21.3 %
	App1	5	13.3 %	18.1 %	8.7 %
MEDIA	App2	15	19.9 %	21.5 %	18.0 %
Intensidad	App3	20	20.4 %	20.4 %	20.6 %
	App1	5	6.4 %	11.5 %	1.7 %
ALTA	App2	15	16.2 %	19.5 %	13.3 %
Intensidad	App3	20	17.8 %	19.7 %	16.6 %

Tabla 3.11: Tasa de fallos obtenida por cada enfoque sobre la base de datos DaFEx en función de la intensidad de la expresión. Existen distintos tamaño de ventana para cada uno de los enfoques de acuerdo con la ecuación 3.12. La tasa de error está calculada sobre el número total de fotogramas y, para cada vídeo, la sonrisa se detectó siempre en, al menos, un fotograma.

relativa estabilidad en la etapa de esplendor y que las etapas de activación y desactivación sean progresivas. Un comportamiento opuesto puede apreciarse en la gráfica de la derecha, en la que no existe estabilidad durante la detección de sonrisas. De acuerdo con Koelstra [123], este segundo comportamiento sugiere una sonrisa no genuina. Esta sugerencia no está basada en las características de la propia sonrisa en un fotograma ni en la configuración de la musculatura facial, sino que se basa en el ciclo de vida de la expresión facial.

Por otro lado, existe una mayor certeza sobre la decisión adoptada en función de la altura que toma la curva de respuesta del clasificador, en otras palabras, los clasificadores que componen el clasificador principal están más de acuerdo en su respuesta. Por último, y debido al tamaño de la ventana, se puede apreciar un pequeño desfase (β) al principio y al final de la expresión facial, justo después de las etapas de activación y desactivación, esto ocurre debido a la consideración del uso de la ventana temporal que, si bien dota de mayor estabilidad la respuesta y ayuda a que las fases de activación y desactivación sean progresivas, genera un pequeño desfase al comienzo de los cambios en el rostro de un individuo.

Además, un estudio empírico se realizó en base a la ecuación 3.14 con el fin de asignar el peso a cada uno de los enfoques propuestos. Como puede apreciarse, a dichos enfoques que ofrecieron una menor tasa de error en la tabla 3.11 les corresponde un mayor peso en el presente estudio. Los datos sugieren la existencia de una correlación entre la distribución de pesos para cada enfoque (ψ) y el tamaño de la ventana temporal (w_s). De esta forma, puede observarse como la distribución de pesos (ψ) muestra una mejoría en el enfoque App2 a medida que aumenta la intensidad de la sonrisa. Por otro lado, el enfoque App3

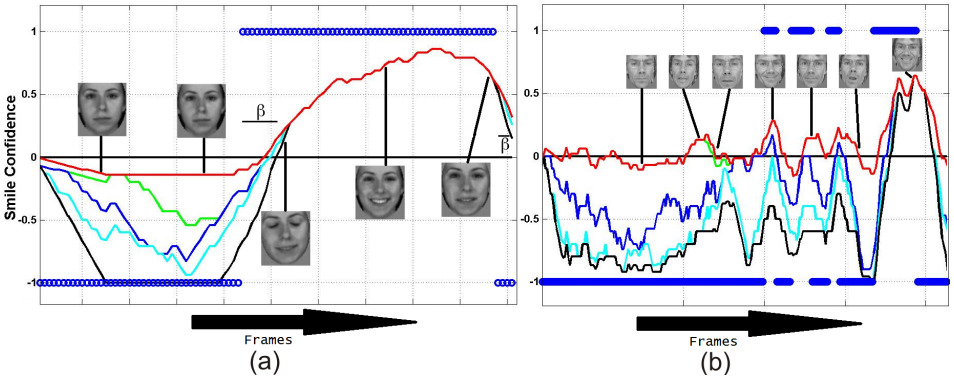


Figura 3.38: Gráficas del clasificador temporal sobre secuencias de vídeo. En dichas gráficas puede apreciarse que el eje X marca la frontera entre lo que se considera una escena con una persona sonriente (por encima del eje) y lo que se considera una escena carente de sonrisa (por debajo del eje). Además, las marcas de color azul determinan la clasificación de los fotogramas hecha a mano, mientras que las distintas líneas de colores determinan las salidas del clasificador para cada uno de los tamaños de ventana temporal de acuerdo a la ecuación 3.13. Por último, la señal β muestra un intervalo de retraso debido a las características inherentes del uso de la ventana de tiempo.

parece comportarse de forma opuesta. Sin embargo, el App1 obtiene el mayor peso, en la mayoría de las ocasiones, debido al rendimiento en la prueba mostrada en la tabla 3.11. La mejoría en la aplicación de este enfoque sobre la base de datos SIANI es notable; haciendo uso de la misma ventana temporal y la misma distribución de pesos mostradas en la tabla 3.12, la tasa de fallos es de 14.7 % (la tasa de falsos negativos es de 17.7 % y la tasa de falsos positivos de 13.44 %). Los resultados ofrecidos por el nuevo enfoque propuesto después de procesar el conjunto de datos DaFEx en busca de sonrisas son mostrados en la tabla 3.12. En ella puede apreciarse cómo el nuevo enfoque combina una baja tasa de fallos y una relación equilibrada entre la tasa de falsos positivos y la de falsos negativos.

La tabla 3.12 presenta los resultados para el nuevo enfoque después de ejecutarlo sobre la base de datos DaFEx. Como puede apreciarse, los enfoques que mejor tasa de fallos ofrecieron en la tabla 3.11 tienen asignado un mayor peso en este estudio. El nuevo enfoque aprovecha los beneficios entre la baja tasa de fallos y la relación equilibrada entre la tasa de falsos positivos y negativos. Además, puede apreciarse que el tamaño de ventana 5 ha sido el que mejor tasa de fallos ha proporcionado para cada una de las pruebas (considerando un muestreo de los datos de 30 fps), de ahí que sea la única dimensión de ventana de tiempo expuesta en la tabla.

DaFEx	Tam. Vent.	Tasa Fallos	FP	FN	ψ_i
Base Datos Completa	5	10.5 %	11.7 %	09.0 %	App1 - 2.0 App2 - 1.3 App3 - 0.7
BAJA Intensidad	5	13.8 %	17.1 %	11.2 %	App1 - 2.0 App2 - 1.3 App3 - 0.7
MEDIA Intensidad	5	11.3 %	14.4 %	07.8 %	App1 - 2.0 App2 - 1.5 App3 - 1.0
ALTA Intensidad	5	05.7 %	06.2 %	05.4 %	App1 - 1.8 App2 - 1.7 App3 - 0.5

Tabla 3.12: Tasa de fallos obtenida por el nuevo enfoque sobre la base de datos DaFEx en función de la intensidad de la expresión. Los pesos para cada enfoque han sido ajustados buscando la menor tasa de fallos posible. Se han considerado tres enfoques; *App*1, *App*2 y *App*3 que hacen referencia al enfoque PCA con imagen en escala de grises, histogramas con imagen codificada con ULBP y valores de píxeles considerando el enfoque SLBP respectivamente. La tasa de error está calculada sobre el número total de fotogramas y, para cada vídeo, la sonrisa se detectó siempre en, al menos, un fotograma.

3.2.3. Discusión

En el primer experimento, el estudio de enfoques para la detección de sonrisas, se ha podido analizar el comportamiento a través del uso de distintos descriptores de la literatura actual como el ULBP o el SLBP así como el comportamiento exhibido a través del uso de distintos métodos de representación (valores de píxeles, histogramas normalizados o PCA) combinados con dos tipos de clasificadores distintos, SVM y k-NN. En este estudio se demuestra la mejora de rendimiento que supone analizar sólo las áreas de interés respecto a una determinada emoción, evitando regiones faciales que pueden introducir ruido debido a su poca influencia en la emoción. En cuanto a los distintos enfoques de preprocesado, el SLBP ha demostrado ser bastante potente en la verificación de sonrisas cuando se hace uso del clasificador SVM. Por su parte, el enfoque ULBP no responde de forma positiva a la localidad espacial estadística de los patrones, en otras palabras, no existe un cambio gradual entre bloques adyacentes de información procesados con LBP. Esto es, atendiendo a la explicación de la sección 3.2.1.1, en función del valor del píxel central en el bloque de información a transformar en LBP , el resultado del preprocesado entre dos píxeles vecinos puede distar de forma notable (por ejemplo, un patrón ser 2 y el otro ser 9) a pesar de ser vecinos en el espacio de origen. Este efecto no ocurre en el caso del SLBP, en el que las variaciones de vecindad afectan en menor medida al valor codificado del píxel central y dos píxeles vecinos están relativamente

Técnica empleada	**Clasificador**	**Datos para Entrenar**	**Datos para Pruebas**	**Tasa de Fallos %**	**Ref.**
Flujo Óptico	Adaboost	BIOID [90]	FGnet [91]	11.5 %	[110]
Sony T300	Adaboost	+++	FGnet [91]	27.3 %	[110]
3D Face mapping	+++	+++	FGnet [91]	05.9 %	[217]
Filtros de Gabor	SVM	GENKI [224]	GENKI [224]	03.7 %	[224]
Intensidad de Píxeles	SVM	Imágenes locales	Imágenes locales	17.8 %	[233]
LIH + CS-LBP	SVM	FEED [104]	FEED [104]	02.1 %	[189]
PCA	SVM	BIOID [90]	DaFEx [14]	13.5 %	[89]
Histogramas ULBP	SVM	BIOID [90]	DaFEx [14]	21.4 %	[89]
SLBP	SVM	BIOID [90]	DaFEx [14]	20.7 %	[89]
Digital Signature	SVM	BIOID [90]	DaFEx [14]	11.5 %	[84]
Ventana Temporal	SVM SVM	BIOID [90] BIOID [90]	Base datos SIANI DaFEx [14]	14.7 % 10.5 %	

Tabla 3.13: Comparación de métodos para la detección de sonrisas en el rostro humano. El símbolo "+++" está relacionado con que no es posible conocer las condiciones exactas del parámetro debido a que se trata de un sistema comercial y sus especificaciones técnicas no están disponibles. La tasa de error está calculada sobre el número total de fotogramas y, para cada vídeo, la sonrisa se detectó siempre en, al menos, un fotograma.

cerca en el espacio transformado de SLBP.

En el segundo experimento se propone la combinación de los clasificadores, con cada uno de los tipos de descriptores, que mejor resultado individual presentaron en el experimento anterior, es decir, los que menor tasa de fallos obtuvieron. La propuesta no sólo disminuye la tasa de fallos sino que también pretende aprovechar las características individuales de cada uno de los clasificadores que lo compone y, por tanto, ser más robusto frente a cambios de iluminación en las escenas. El enfoque presentado en este segundo experimento obtuvo una mejora bastante significativa con respecto a los resultados obtenidos en el primer experimento, por lo que se siguió trabajando con este clasificador en el tercer experimento.

El tercer experimento muestra el desarrollo de un enfoque alternativo, la firma digital. Los resultados ofrecidos por este enfoque mejoran de forma consi-

derable los datos que ofrecen los enfoques anteriores. Además, esta alternativa se ha planteado como una técnica independiente de texturas, bordes o colores, siendo sensible a cambios geométricos de la propia imagen.

Por último, se propone un enfoque dinámico que aprovecha la coherencia temporal que proporciona un vídeo y el ciclo de vida de las expresiones faciales para determinar, de forma precisa, la activación en el caso de que un usuario sonría. Tanto el detector facial como el detector de sonrisas pueden trabajar como una aplicación en tiempo real; el detector proporciona la posición de ojos, boca y nariz, mientras que el detector de sonrisa es capaz de detectarla e, incluso, determinar la intensidad de la misma con menor error que los enfoques estáticos.

En la tabla 3.13 se ofrece una comparativa entre distintos métodos utilizados para la detección de sonrisas. En ella, se muestra un conjunto de técnicas utilizadas en los últimos años para la detección de sonrisas. Resulta especialmente relevante el trabajo de Huang, en el que analiza las tasas de acierto de dos sistemas de reconocimiento de sonrisas que están disponibles en el mercado; la cámara Sony T300 y el detector de sonrisas de la tecnología Omron [217]. Para ello, hace uso de una base de datos como la FGnet (*"Facial Expression and Emotion Database"*), la cual está compuesta por una serie de vídeos con distintas emociones grabados a través de una cámara frontal, muy parecida a la imagen que graba una webcam incorporada a la pantalla, con una calidad de imagen totalmente nítida [91]. Huang demostró que su detector de sonrisas ofrece unas cifras en torno al 88,5 % de acierto, superando en más de un 15 % las cifras de la cámara Sony T300 y aproximándose bastante al detector de sonrisas de Omron. Precisamente, este detector desarrollado por Omron hace uso de una técnica de mapeo tridimensional para determinar si el sujeto sonríe o no [217]. Por ello, Huang propone un método para la detección de sonrisas que está basado en la aplicación, en un entorno controlado, de técnicas de flujo óptico sobre la región bucal en exclusiva [110]. Su teoría está basada en la intuición de que, alrededor de la región bucal, en el caso de existir una sonrisa, los vectores de flujo óptico apuntan hacia fuera y hacia arriba. Sin embargo, este sistema no considera severas rotaciones o un escalado del rostro cuando, por ejemplo, ocurre que un sujeto se acerque o se aleje de la imagen. Por su parte, Whitehill presenta en su investigación una base de datos propia, GENKI, para mostrar un enfoque basado en la aplicación de Filtros de Gabor sobre el rostro de un sujeto con el fin de determinar si está, o no, sonriendo [224]. Las tasas que ofrece su trabajo son francamente buenas y se aproximan a las ofrecidas por Shimada en la investigación en la que combina un nuevo enfoque basado en LBP y los histogramas de intensidad local [189]. Shimada hace uso de máquinas SVM, pero las combina creando un detector en cascada con el objetivo de mantener el coste de los cálculos en concordancia con una baja tasa de fallos. Para su experimento, este autor hizo uso de la base de datos FEED (*"Facial Expression and Emotion Database"*) [104] además de distintas imágenes obtenidas de la televisión. Al igual que en el trabajo de Whitehill, Shimada considera todo el rostro para la detección de sonrisas. Por último, Yadappanavar, hizo un extenso estudio sobre imágenes locales en los que se combinó, de forma variada, descriptores SVM obtenidos únicamente a través del valor de la intensidad de los píxeles de la

región bucal [233].

Una vez descritos algunos de estos enfoques de la literatura, se puede entender mejor el contexto en el que se ubica la investigación desarrollada en esta sección del capítulo. Las cifras proporcionadas muestran una progresión de los resultados obtenidos a medida que se han ido combinando, desde un punto de vista dinámico, los distintos enfoques estáticos. El uso de una base de datos normalizada y públicamente accesible, como DaFEx, otorga una mayor validez a los resultados obtenidos. Por otro lado, se ha desarrollado un método alternativo denominado firma digital como un enfoque que, de forma autónoma, ofrece una tasa de aciertos similar a la combinación de enfoques y en el que se ha tratado de recoger aspectos inherentes a la fortaleza ofrecida por los LBP, como por ejemplo, la robustez frente a cambios lumínicos. Además, resulta interesante comentar que una opción que se barajó, con el fin de mejorar los resultados del sistema de detección de sonrisas, fue la integración la firma digital dentro del enfoque de ventana temporal como uno de los clasificadores que participaran en la toma de decisión. Sin embargo, debido a las caracteríticas de la firma digital, tal y como se explica en la sección 3.2.1.6, se trata de un enfoque basado en la superposición de parches sobre una imagen dividida en bloques. En términos temporales, no resulta actualmente factible incluir este enfoque en la ventana temporal si lo que se pretende es obtener un clasificador de sonrisas en tiempo real. De ahí y, con el propósito de evitar un cuello de botella, que se hayan desarrollado estos dos enfoques de forma independiente.

Capítulo 4

Extracción de descriptores demográficos fusionando información facial y de contexto

En el capítulo 2 se describieron distintos tipos de técnicas orientadas al reconocimiento una característica demográfica, como es el sexo de una persona, en base a su apariencia. Ocurre que, en la mayoría de los casos expuestos en dicha revisión, al reconocimiento del sexo se lleva a cabo en un entorno restringido o controlado (véase la tabla 2.4), o bien, en base al uso de dispositivos muy específicos que obtienen de imágenes por medio de sensores de un coste elevado [140].

De hecho, en la sección 2.4.1.2 se define un conjunto de experimentos realizados por distintos autores sobre bases de datos bastante extendidas pero que, salvo contadas excepciones [16], carecen de una variabilidad en cuanto a las poses de los sujetos, las características demográficas de estos sujetos (raza y edad), las condiciones lumínicas o la presencia de elementos oclusivos. El problema presentado en dichos experimentos radica en que se centraban exclusivamente en el rostro a la hora de determinar el sexo de las personas y, cuando se hacía uso de regiones no faciales, se trabajaba en entornos muy restrictivos; los sujetos son un grupo de personas con características similares, tras las cuales hay un fondo homogéneo para poder realizar la segmentación del sujeto con relativo éxito. En el presente capítulo se pretende abordar este problema a través del uso de nuevos mecanismos que faciliten la segmentación o extracción de las características de la región del sujeto (pelo, piel y ropa) con independencia de las características del fondo de la escena. Una ventaja inherente de la propuesta que se realiza es la flexibilidad al poder asumir la variabilidad de los datos procesables. Para ello, se propone una solución del problema en dos fases bien diferenciadas; en la primera

fase se trata el diseño de un algoritmo eficiente para segmentar las zonas relevantes de la escena y eliminar aquellas que introducen ruido. Con el objetivo de resolver la primera fase anteriormente mencionada, se describen los fundamentos básicos de los bautizados como tríxeles (del anglosajón *"trixels" (Triangular Superpixels)*) que, si bien guardan cierta relación con algunas técnicas de segmentación existentes en la literatura, es un enfoque novedoso que, a diferencia de estos, ofrece una solución optimizada y adaptable a sistemas de tiempo real. En la literatura existe otra aplicación para el término "Triangular Superpixels" que guarda relación con la segmentación de elementos compuestos pos cristal, pero no tiene relación directa con el trabajo presentado en este capítulo [220]. Una vez cubierta la primera fase, en la segunda fase resulta más sencillo extraer los descriptores que se deseen de las zonas de interés. De esta forma, se trata de un enfoque en el que hay un procesamiento bottom-up (obtención de los datos a través de los tríxeles) con top-down (a través del conocimiento de qué se busca y la información del rostro detectado).

El proceso de creación de la malla de tríxeles se describe en la sección 4.2, mientras que en la sección 4.3 se explica la adaptación del algoritmo GrabCut para la segmentación de dicha malla. Además, en la sección 4.4 se ofrece una comparativa entre la nueva propuesta del algoritmo GrabCut para la segmentación de tríxeles y la segmentación de píxeles (versión original del GrabCut). Por último, un conjunto de experimentos con el fin de aplicar la técnica desarrollada a la detección de una característica demográfica como el sexo se presenta en la sección 4.5.

4.1. Antecedentes

La técnica propuesta en este capítulo guarda una estrecha relación con dos de los enfoques más populares de la literatura en el ámbito de la segmentación de imágenes; los superpíxeles [151] y el *GrabCut* [176]. Se trata de una técnica que combina la eficacia de ambos enfoques pero que los mejora en términos de eficiencia. Esta técnica se desarrolla en dos fases; la primera es genérica para cualquier problema, se construye la malla inicial de tríxeles denominada TriToM (*"Trixels Topological Mesh"*) para, en una segunda fase, proceder con la segmentación de dicha malla con el objetivo de delimitar los tríxeles que pertenezcan al área de interés en nuestro problema. La primera fase guarda una relación con el desarrollo de los superpíxeles debido a que el planteamiento inicial es bastante similar en ambos casos. Sin embargo, para la fase de segmentación, el método propuesto opera de una forma bastante similar al método GrabCut, sólo que, a diferencia de éste, se procesan tríxeles en lugar de píxeles.

4.1.1. Superpíxeles

Los algoritmos basados en superpíxeles (del término anglosajón *"superpixels"*) agrupan los píxeles en regiones atómicas que tienen un significado perceptual propio. Además, estas regiones son dinámicas en el sentido que pueden

Figura 4.1: Imágenes segmentadas haciendo uso de la versión SLIC de superpíxeles considerando distinto tamaño de superpíxeles (64, 256 y 1024 píxeles respectivamente). Imagen tomada de [2].

absorber nuevos píxeles y seguir creciendo, o bien, estabilizarse. En la Figura 4.1 pueden apreciarse distintas iteraciones de este algoritmo en función del tamaño del superpíxel. Se trata de una alternativa a la rigidez ofrecida por la clásica estructura de píxeles que ofrece una imagen. Los superpíxeles encuentran la redundancia de información en la imagen y proponen una nueva estructura que reduce su complejidad de cara a las sucesivas tareas a realizar sobre dicha imagen. De hecho, los superpíxeles se han convertido en una técnica clave para el cómputo de muchos algoritmos de visión por computador que tienen objetivos tales como estimar la profundidad [242], segmentar [133], modelar el cuerpo humano [150] o la localización de objetos [93].

En la actualidad existen múltiples enfoques para generar superpíxeles, cada uno ofrece un conjunto de ventajas, además de algunas desventajas, que hacen que el desarrollo de estos algoritmos esté orientado hacia la aplicación específica a la que se pretende dar respuesta. Por ejemplo, existen ciertas aplicaciones en las que se prima más la precisión del proceso de segmentación en la frontera de las regiones, en cuyo caso el cómputo de los superpíxeles tiene lugar a través de un método basado en grafos [81]. En cambio, otras aplicaciones están más interesadas en trabajar con la malla generada por los superpíxeles en lugar de trabajar únicamente con las regiones que delimita, en este caso se hace uso de un método que permita la formación de esta malla o *lattice* [188].

Por tanto, si bien existe un conjunto de características distintivas por cada uno de los métodos desarrollados para generar los superpíxeles, Achanta [2] propone una serie de propiedades que son generalmente deseables para cada enfoque que se basa en la idea de agrupamiento de píxeles:

- Los superpíxeles deben ajustarse de forma precisa a los bordes existentes

en la imagen.

- Cuando se generan con el fin de reducir la complejidad computacional de la imagen, deben ser de rápido cálculo, eficientes en términos de memoria y fáciles de manipular.
- Cuando se generan con el fin de segmentar una imagen, los superpíxeles deben ofrecer una alta velocidad de cómputo, además de proporcionar calidad en los resultados obtenidos.

Por último, Achanta destaca que los métodos para generar los superpíxeles pueden ser categorizados en dos clases; los métodos basados en grafos y, por otro lado, los basados en gradiente. Aquellos basados en grafos consideran los superpíxeles como un nodo dentro de un grafo. En este caso, el peso de la arista entre dos nodos es proporcional a la similitud entre los píxeles vecinos; la creación de los superpíxeles se basa en la minimización de una función de coste. En cambio, los métodos basados en el gradiente empiezan con una distribución aleatoria de agrupamientos de píxeles y, a través de la aplicación iterativa de métodos ascendentes de gradiente, van refinando estos agrupamientos iniciales hasta que un criterio convergente da como resultado el conjunto de superpíxeles. La técnica que proponemos en este capítulo estaría enmarcada en los métodos basados en grafos pues, como se verá más adelante, se propone una topología en el que los nodos que componen el grafo es el centroide de cada uno de los tríxeles de la imagen.

4.1.2. GrabCut

Muchos problemas de la visión por computador pueden ser reformulados como un problema de minimización de energía. Con anterioridad a este enfoque, la forma de resolver este tipo de problemas era a través del uso de programación dinámica (la cual sólo sirve si se presentan casos muy simples) o mediante simulación (que es muy lento de procesar). Kolmogorov y Zabih [124] exploraron qué tipos de funciones pueden ser minimizadas a través del uso de cortes en los grafos (*Graph Cut* en inglés). De hecho, todas las técnicas de segmentación que hacen uso de grafos se aprovechan de algún algoritmo de corte para segmentar el grafo en dos regiones y, por tanto, para minimizar la energía.

La estructura básica del Graph Cut la conforma un grafo no dirigido $G = (V, E)$ que consiste en un conjunto de nodos V y las correspondientes aristas ponderadas E que los conectan entre sí. Cada nodo del grafo se corresponde con un píxel. Además, el grafo contiene un conjunto de nodos especiales denominados nodos terminales. En el contexto de la visión por computador, los nodos terminales están relacionados con el conjunto de etiquetas que pueden ser asignadas a los distintos píxeles. Normalmente, estos terminales son conocidos como fuente (*"source"*) y sumidero (*"sink"*) (véase la Figura 4.2). A todas las aristas del grafo se les asigna un valor (un peso o un coste). El coste de una arista dirigida (p, q) puede diferir del coste de la arista opuesta (q, p) en cualquier momento.

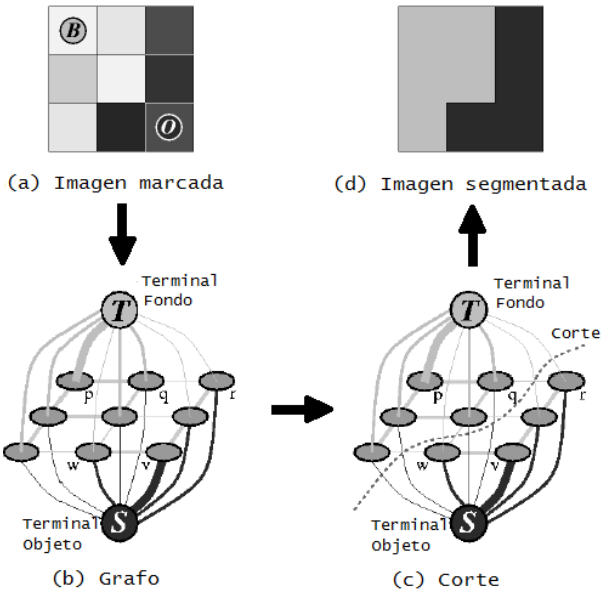


Figura 4.2: Fundamentos básicos sobre el proceso de segmentación del Graph Cut [101].

En la Figura 4.2 se muestra un simple ejemplo de dos grafos terminales desarrollados por Greig [101]. En su trabajo, este autor minimiza la energía de Potts sobre una rejilla de 3×3 elementos con dos posibles etiquetas. Dentro de dicha imagen pueden distinguirse dos tipos de aristas dentro del grafo; los N-enlaces y los T-enlaces.

Los N-enlaces (*"N-Links"*) conectan cada píxel con sus vecinos directos. Si bien en la imagen la conexión se realiza con sus 4 vecinos directos, en los últimos algoritmos desarrollados en este ámbito, se ha ampliado la relación a los 8 vecinos inmediatos a cada píxel [124]. El peso de estos enlaces indica la penalización existente por situar un corte entre los dos píxeles que separa. Se desea que la penalización sea muy alta en regiones con bajo gradiente y que, en cambio, sea muy baja en regiones con un alto gradiente.

Por su parte, los T-enlaces (*T-Links*) conectan cada píxel a los nodos fuente o sumidero que vienen a distinguir entre las clases relativas a la parte trasera de la imagen (*"background"*) o al objeto de interés (*"foreground"*) respectivamente. Precisamente, este tipo de enlaces determina la probabilidad de que un píxel pertenezca, o no, a la región que se desea buscar dentro de la escena. Estos nodos fuente son definidos como símbolos terminales en el grafo y, precisamente, son estos dos símbolos terminales los que permiten dividir el grafo en dos subconjuntos disjuntos $\mathcal{S}$ y $\mathcal{T}$ de forma que la fuente $\mathbf{s}$ está en $\mathcal{S}$ y el sumidero $\mathbf{t}$ está en $\mathcal{T}$. En la Figura 4.2c se aprecia un corte que da origen a la segmentación. Dicho corte $C = \{\mathcal{S}, \mathcal{T}\}$ viene definido como la suma de todos los costes de las

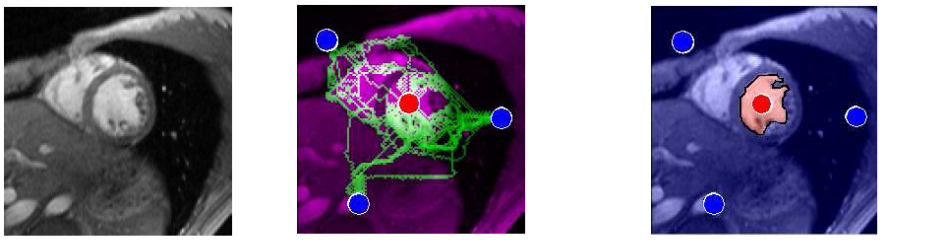


Figura 4.3: Relación entre el mínimo coste de corte y el flujo máximo. Los círculos rojos y azules están relacionados con enlaces muy pesados a la fuente s y al sumidero t respectivamente. Los costes de las aristas entre los píxeles (nodos del grafo) es menor en lugares en los que existe un alto contraste de intensidad, por lo que los cortes a lo largo de la frontera de este objeto en la imagen deben ser menores en cuanto a coste. Desde la perspectiva del flujo, tener menor coste implica que, si consideramos las aristas como una tubería de agua, la cantidad de agua que puede soportar es menor y, por tanto, actúan como un cuello de botella. En esta demostración, Boykov, muestra el flujo máximo desde s hasta t, saturando las aristas del grafo que se corresponden con el corte de coste mínimo [23].

aristas frontera (r, q), donde $r \in \mathcal{S}$ y $q \in \mathcal{T}$ respectivamente. Por lo tanto, el coste total se realiza en base a la suma de los pesos de todas las aristas frontera entre $\mathcal{S}$ y $\mathcal{T}$. Con ello, el problema del coste mínimo en el grafo es aquel que encuentra un corte que tiene el coste mínimo entre todos los posibles cortes.

Bajo esa perspectiva, uno de los aspectos más importantes en la optimización combinacional es que un corte mínimo $\mathcal{S}/\mathcal{T}$ puede ser resuelto encontrando el flujo máximo existente entre la fuente s y el sumidero t. Análogamente, si se considera que lo que fluye entre las aristas es agua y que, cada arista, tiene una capacidad máxima para dejar pasar el agua. Entonces, el problema se reduce a encontrar el máximo flujo de agua que puede pasar a través de las distintas tuberías (aristas), desde la fuente hasta el sumidero, interpretando las aristas como tuberías con una capacidad igual al peso de las mismas. El teorema de Ford establece que encontrar el flujo máximo en la ruta de s a t implica saturar un conjunto de aristas en el grafo que divide los nodos entre dos partes disjuntas $\{\mathcal{S}, \mathcal{T}\}$, y que se corresponde con el corte que tiene el coste mínimo [83]. De esta forma se relacionan los problemas de corte mínimo y el de búsqueda de flujo máximo dentro del grafo resultando, ambos, problemas equivalentes. En otras palabras, encontrar el máximo flujo en el grafo es equivalente a encontrar el coste mínimo en el corte. En la Figura 4.3 puede apreciarse esta equivalencia.

En cuanto a la relación existente entre el corte de coste mínimo y el problema de minimización de energía en relación al etiquetado de las distintas regiones, Boykov estableció una relación intuitiva [23]. Supóngase la Figura 4.2; el grafo que se genera se corresponde con una imagen de 3×3 píxeles. Cualquier corte

posible de los nodos en dos grupos disjuntos contendrá exactamente un símbolo terminal. Por tanto, cualquier corte está basado en la idea de asignación de píxeles (nodos del grafo) con las etiquetas (símbolos terminales). Además, el cálculo de los T-enlaces se realiza en base al desarrollo de modelos de mezcla de gaussianas (*"Gaussian Mixture Models"* (GMM)) entre los píxeles pertenecientes a cada símbolo terminal. Por tanto, si se tiene en cuenta que el peso de una arista está calculado en términos de energía, entonces un corte de coste mínimo se corresponde con un etiquetado con el valor mínimo de esta energía.

En resumen, desde una perspectiva formal, Graph Cut es una técnica apropiada para la segmentación de imágenes. Los nodos del grafo representan píxeles y las aristas establecen cualquier tipo de relación entre los píxeles. Una vez establecidas las relaciones a través de los N-enlaces y los T-enlaces, se produce el corte en el grafo. Como puede apreciarse en la Figura 4.2, la partición en el grafo se corresponde con la consiguiente segmentación de los píxeles de la imagen que representan los distintos nodos del grafo. Al tratarse de un problema de minimización de energía, el mínimo coste en el corte genera una segmentación óptima en función de los valores existentes en las aristas que separan los nodos del grafo. A menudo, suelen confundirse Graph Cut [22] y GrabCut, la principal diferencia es que GrabCut es una extensión de Graph Cut, la cual permite su extensión al procesado de imágenes con color además de simplificar la acción del usuario sobre la imagen permitiendo seleccionar, de forma directa, el objeto de interés que se desea segmentar [200]. El algoritmo GrabCut es, por tanto, una técnica iterativa de segmentación en dos dimensiones de cualquier objeto en una imagen si se conoce una estimación de la posición de dicho objeto en la misma escena [176].

4.2. Creación de la malla de tríxeles

En este apartado se describe el proceso para la generación de tríxeles. En la sección 4.1 se comentó la similitud existente entre los tríxeles y los superpíxeles. Un tríxel viene definido como una unidad atómica procesable que tiene forma triangular. Al igual que sucede con los superpíxeles, los tríxeles dividen la imagen en un conjunto finito de segmentos independientes. Comparten propiedades con los superpíxeles, como por ejemplo:

- Son computacionalmente eficientes; reducen la complejidad de una imagen de forma drástica, pasando de cientos de miles de píxeles a unos cientos de tríxeles.
- La información contenida dentro de los tríxeles es consistente y representativa de los píxeles que alberga en su interior.
- La representación es casi completa; la mayor parte de las estructuras de la imagen se conservan, sin embargo, existe una ínfima pérdida de información cuando se trabaja con la malla de tríxeles en lugar de trabajar con la rejilla de píxeles.

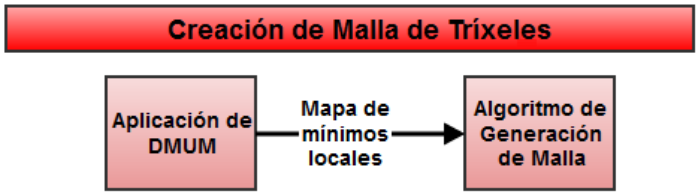


Figura 4.4: Fases para la creación de TriToM.

Por otro lado, aunque los tríxeles guardan muchas similitudes con los superpíxeles, se distinguen de ellos en dos aspectos fundamentales:

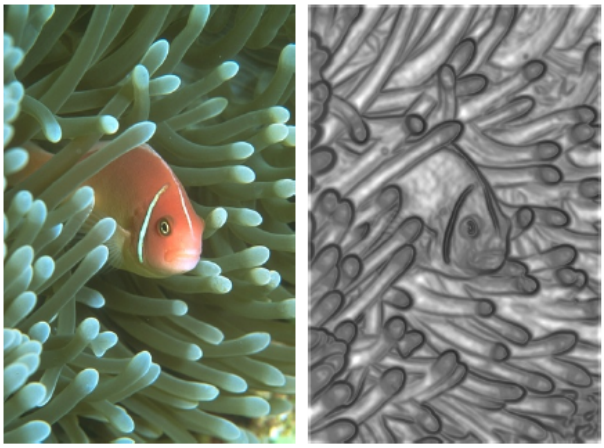
- Se genera una malla de tríxeles no homogénea desde una perspectiva estructural manteniendo, cada tríxel, un número de vecinos entre dos (como mínimo) y tres (como máximo). La malla también es conocida como TriToM (*"Trixels Topological Mesh"*).
- A diferencia de lo que ocurre con la técnica de los superpíxeles, los tríxeles no siguen un patrón de posicionamiento aleatorio en su primera fase (como ocurre en el algoritmo *k-means* [143]) ni la imagen se divide de forma homogénea (considerando bloques de $n \times n$ píxeles) sin considerar la información propia del contenido de la imagen durante el proceso de creación (como ocurre con los superpíxeles [2]). En el caso de los tríxeles, la colocación de los vértices de los triángulos que forman la malla se hace en función del algoritmo DMUM (*"Distance Maps from Unthresholded Magnitudes"*) descrito en la sección 4.2.1.

Se entiende por la creación del TriToM de una determinada escena, como un método de preprocesado con el fin de poder realizar una segunda fase de segmentación de la imagen a través del uso de esta malla. La primera fase puede ser apreciada en la Figura 4.4, en la que se distinguen dos pasos que se describen en las siguientes secciones.

4.2.1. Agrupamiento DMUM

El algoritmo DMUM es un método reciente que calcula los mapas de distancia de forma directa a partir de magnitudes no umbralizadas. Este método ejecuta los cálculos de forma que no pierde información crítica a lo largo del proceso sin la necesidad de binarizar la imagen [8]. Los valores de la salida representan el tamaño de la región más pequeña, alrededor de cada píxel, que encierra una cantidad relevante de acumulación de valores de magnitud que dependen únicamente de la imagen que se está procesando. Por último, para su cálculo, se necesita una única pasada sobre la imagen de magnitud, y la imagen integral [215] se utiliza para acelerar el proceso de sumar áreas.

En el enfoque más clásico de la distancia transformada (*Distance Transform (DT)*) de Rosenfeld y Platz [175], cada punto recibe un valor que representa la



(a) (b)

Figura 4.5: (a) Imagen original, (b) DMUM resultante a partir del cómputo del gradiente de Sobel sobre la imagen original. Los valores de baja intensidad reflejan los mínimos locales captados en la escena.

distancia a un punto del objeto más cercano. Sin embargo, Antón ha redefinido el término objeto. Normalmente, el propio objeto, se obtiene a partir de la aplicación de un umbral, o bien, a través de la detección de bordes cuyos resultados son clasificados como "objeto" o "no objeto". Sin embargo, en el DMUM, la expresión relativa a la *"distancia al punto del objeto más cercano"* ha sido relevada por la *"distancia a la zona de magnitud relevante más cercana"* en una función de mapeo que se basa directamente en magnitudes no umbralizadas.

Desde una perspectiva más formal se define un conjunto de puntos en el interior de una esfera de radio r centrada en el punto p:

$$B_r(p) = \{x \in \mathbb{S}^2 | d(x,p) \leq r\} \tag{4.1}$$

Donde $d(x,p)$ se corresponde con la función distancia. Además, si $\mathbb{S}^2 \equiv \mathbb{R}^2$ entonces $B_r(p)$ está definida en un espacio métrico continuo, mientras que si $\mathbb{S}^2 \equiv \mathbb{Z}^2$ entonces $B_r(p)$ está definida en un espacio métrico discreto [8]. Así, sea una imagen $i \in \mathbb{Z}^2$ y su imagen gradiente normalizada g' (siendo 1.0 su valor máximo), se considera la suma de gradientes para una determinada $B_r(p)$ en i como:

$$\Phi_r(p) = \gamma \cdot \sum_{\forall x \in B_r(p)} g'(x) \tag{4.2}$$

Siendo el parámetro γ un factor de escalado que suaviza los valores de salida del DMUM, haciendo innecesario un proceso de filtrado posterior. Por otro lado,

se define el radio q como:

$$q(p) = \underset{r}{\operatorname{argmin}}\{\Phi_r(p) > 1{,}0\} \tag{4.3}$$

Es decir, se define como el radio mínimo r de una esfera $B_r(p)$ que encierra una suma del gradiente $\Phi_r(p)$ mayor que 1.0. De esta forma, el valor de los mapas de distancia a partir de estas magnitudes sin umbral en un determinado punto p se define como:

$$\Gamma(p) = q(p) - \sum_{r=1}^{q} \Phi_r(p) \tag{4.4}$$

Donde el valor DMUM, $\Gamma(p)$, tiene un valor menor para los puntos situados en las regiones con un gradiente disperso y mayor para los puntos con mucha pendiente o puntos colocados en regiones de gradiente denso [8]. De acuerdo con Antón, su aplicación a una imagen binaria produce una salida similar a la producida por la distancia transformada, tal y como se puede apreciar en la Figura 4.5.

Se trata de un método que ofrece unas tasas de error por región bastante bajas en comparación con otro tipo de gradientes como el gradiente de Sobel o el gradiente de Deriche [8]. Además, este método permanece estable cuando procesa secuencias de vídeo, devolviendo los mismos valores si las condiciones en las que suceden las operaciones no han variado de forma significativa. El conjunto de puntos de mínimos locales que el mapa de distancias ofrece, se utiliza como parámetro de entrada para los vértices en la formación de la malla de tríxeles que se explica en la siguiente sección.

4.2.2. Topología de la malla

Haciendo uso de un mapa de puntos de mínimos locales se pretende generar una malla de tríxeles. Este proceso de triangulación divide el conjunto de píxeles de la imagen original en un conjunto de subregiones que cumplen ciertas pautas de homogeneidad en cuanto a la información contenida en cada subregión. Además, se trata de un proceso que tiene el mismo fundamento que la génesis de los superpíxeles (véase sección 4.1.1); simplificar la información contenida en la imagen para un procesado posterior óptimo. La topología que sigue la malla de tríxeles está basada en la técnica de la triangulación de Delaunay.

Para la triangulación de Delaunay se puede considerar un espacio multidimensional, sin embargo, en el caso de una imagen sólo se va a tener en cuenta un espacio bidimensional debido a los vértices proporcionados por el mapa de mínimos locales del DMUM en este espacio. El proceso de triangulación es muy utilizado en la literatura porque tiene aplicaciones en una variedad de trabajos en los que confluyen diversos campos de interés como la simulación de elementos finitos, la aproximación del cálculo numérico de superficies, o bien, los procesos de clasificación que hacen uso del vecino más cercano [194].

Dado un conjunto S de puntos $p_1, p_2, ..., p_n$ se puede obtener la triangulación de Delaunay del espacio S a través del cálculo de su diagrama de Voronoi [59].

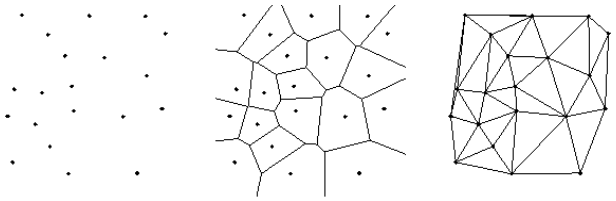


Figura 4.6: (a) Un conjunto de puntos, (b) su diagrama de Voronoi, y (c) su triangulación de Delaunay.

El diagrama de Voronoi divide el espacio bidimensional en regiones alrededor de cada punto, de forma que todos los puntos en la región alrededor del punto p_i están más cerca del punto p_i que de cualquier otro punto de S (véase la Figura 4.6b). De esta forma, una vez proporcionado el diagrama de Voronoi sobre S, la triangulación de Delaunay puede generarse a través de la conexión de los centros de cada par de regiones pertenecientes al diagrama de Voronoi (véase la Figura 4.6c). El motivo de haber utilizado esta topología a la hora de formar la malla de tríxeles se refleja en las particularidades que ofrece la triangulación de Delaunay; (1) el cálculo de la triangulación de Delaunay para un mismo conjunto de puntos es único, (2) un círculo que abarque los tres vértices de un triángulo formado a través de esta triangulación no contiene ningún otro punto externo al propio triángulo, (3) el ángulo mínimo entre todos los ángulos de todos los triángulos pertenecientes a una triangulación de Delaunay es mayor que el ángulo mínimo ofrecido por cualquier otra triangulación que haga uso de los mismos puntos [59]. En resumen, la primera propiedad asegura que esta configuración es estable y se puede utilizar para indexar los triángulos formados bajo esta técnica. La segunda propiedad implica que insertar un nuevo punto dentro de una triangulación de Delaunay sólo afecta a aquellos triángulos cuyos círculos circunscritos afecten a dicho punto. En otras palabras, el ruido sólo repercute de forma local a los triángulos calculados en la región afectada, lo que dota de estabilidad al resto de triángulos previamente procesados y, por tanto, de eficiencia al algoritmo. Por último, esta estructura garantiza la estabilidad de la vecindad de los triángulos; entendiendo que dos triángulos son vecinos si comparten una arista, en la Figura 4.6c puede apreciarse que el número máximo de vecinos por parte de un triángulo es de tres vecinos, mientras que el número mínimo de vecinos es de dos en aquellos triángulos que forman parte de los bordes del conjunto de puntos proporcionados.

Otro motivo importante a la hora de seleccionar esta topología se encuentra en el estudio comparativo con diversas estructuras topológicas conocidas que llevó a cabo Tuceryan [209]. En su estudio, este autor afirma que la triangulación de Delaunay ofrece la mejor estabilidad estructural frente a posibles perturbaciones posicionales de puntos. Por último, tanto el diagrama de Voronoi como el proceso de triangulación de Delaunay son algoritmos bastante eficientes debido a que el número de vértices para ambas técnicas es proporcional a una pequeña constante por el número de puntos procesados ($O(N)$). Además, cada

arista pertenece como máximo a dos triángulos, lo que implica que el número de triángulos generados a partir de la triangulación de Delaunay es lineal con respecto al número de puntos presentados para su triangulación.

4.2.3. Algoritmo de generación de la malla

El algoritmo que genera la estructura de tríxeles a través de la topología previamente descrita resulta de una importancia crítica; no sólo debe recoger la información interna de cada píxel perteneciente a cada tríxel sino que, además, debe resultar eficaz en términos computacionales. Por ello, el algoritmo que genera la malla tiene dos módulos diferenciados. El primer módulo es el encargado de generar los tríxeles de la imagen desde una perspectiva estructural mientras que el segundo módulo se ocupa de formar la estructura que da lugar a la malla de tríxeles. La malla de tríxeles ofrece una visión holística de la organización de los tríxeles así como de sus vecinos, permitiendo una mayor sencillez de acceso a la hora de procesarlos.

Previo al comienzo del proceso de generación de tríxeles, es necesario haber detectado los puntos a través del algoritmo DMUM explicado en la sección 4.2.1 y, con dichos puntos, haber generado una topología triangular de Delaunay que sirve de entrada para la creación de la malla. De esta forma, la primera parte del algoritmo se divide en dos bloques. El primer bloque se muestra a continuación y, en dicho bloque, en primer lugar busca la contrucción del tríxel inicial en base a la configuración de Delaunay previamente calculada. La idea es ir recorriendo los puntos según esta topología en busca de los triángulos detectados para poder formar los tríxeles. Así, en base a la localización de un primer tríxel, empiezan a recorrerse todas sus aristas para detectar triángulos vecinos y almacenarlos en la cola de tríxeles candidatos pendientes de procesar. Además, una arista que ha sido recorrida en dos ocasiones no vuelve a ser recorrida una tercera ocasión debido a que, como máximo, cualquier arista separa únicamente dos tríxeles (véase algoritmo 1).

Algoritmo 1: Generación de la malla de tríxeles

```
Data: Integer image img sized n x m
Data: DelaunayEdgeMem topology sized 1 x q
Result: Floating point image TriToM sized n x m
Result: Trixel List trixels sized 1 x w
Result: Edge List trixels_Edges sized 1 x p
TriToM[i, j] ← 0, i ∈ [0..n), j ∈ [0..m)
candidates_queue ← Find_First_Trixel(topology)
while candidates_queue ≠ empty do
    current_candidate ← candidates_queue.pop_front
    edge_list ← (current_candidates → get_edge_list())
    for edge ∈ edge_list do
        if ((edge → Num_Trixels) < 2 then
            if edge →
            New_Candidate(topology, current_candidate, candidates_queue)
            then
                trixels_Edges.push_back(edge)
            end
        end
    end
    trixels.push_back(current_candidate)
    Generate_Topological_Mesh(img, current_candidate, TriToM)
end
return TriToM, trixels, trixels_Edges
```

Como puede apreciarse en dicho algoritmo, los tríxeles se van formando a medida que se recorren las aristas de los triángulos almacenados en la topología de Delaunay. Debido a ello, para cada candidato, se recorren todas sus aristas en busca de posibles candidatos a tríxeles. En el algoritmo 2 se recorre el contenido de la topología de Delaunay, precisamente, a través de la estructura de aristas con el objetivo de ir generando estos nuevos candidatos. Dada una arista perteneciente a un candidato existente, se debe recorrer dicha arista en sentido opuesto para encontrar un posible nuevo candidato. Una vez se ha encontrado una nueva ruta de tres aristas que contengan a este nuevo candidato, se crea el nuevo tríxel candidato y se almacena en la cola de candidatos a la espera de ser procesado. Debido a que el recorrido se realiza a partir de las aristas de los tríxeles previamente identificados, una vez se encuentra un nuevo posible candidato, éste se inserta en la cola de nuevos candidatos y se actualiza el conjunto de vecinos de cada uno de los candidatos implicados en la operación. De esta forma, se garantiza la conectividad total entre tríxeles pertenecientes a la malla. Por último, se actualiza la lista de tríxeles pertenecientes a la arista en cuestión con el objetivo de no volver a recorrerla en el futuro si ya han sido localizados sus dos posibles candidatos a tríxeles.

Algoritmo 2: Crea nuevo candidato ($New_Candidate$)

```
Data: DelaunayEdgeMem mem sized 1 x m
Data: Edge this
Data: Trixel candidate
Result: Trixel List candidates_queue sized 1 x z
edge ← this
orientation ← RIGHT
new_candidate_edges.push_back(edge)
edge_counter ← 1
while edge_counter < 3 do
    if (orientation = RIGHT) then
        edge ← Next_Around_Right(edge, mem)
    else
        edge ← Next_Around_Left(edge, mem)
    end
    if (Belongto_Trixel(edge, candidate))OR(Outof_Bounds(edge)) then
        if orientation = RIGHT then
            orientaton ← LEFT; edge_counter ← 1
        else
            return FAIL
        end
    else
        new_candidate_edges.push_back(edge)
        edge_counter ← edge_counter + 1
    end
end
this → Update_Trixels_List(new_candidate)
new_candidate → new Trixel(new_candidate_edge)
candidate → Update_Neighbours(new_candidate)
new_candidate → Update_Neighbours(candidate)
candidates_queue.push_back(new_candidate)
return SUCCEED
```

Una vez se han procesado todas las aristas pertenecientes al candidato a tríxel en busca de posibles vecinos, se procede a registrarlo dentro de la malla. En el momento en el que se registra dentro de la malla, dejando de ser candidato. Este registro ocurre una vez se han recorrido todos los píxeles pertenecientes a dicho tríxel con dos objetivos; el primero es recoger la información interna del tríxel para realizar los cálculos que se estimen oportunos (media, varianza, cálculo de descriptores, etc.) mientras que el segundo es asociar cada píxel del mapa topológico al tríxel en cuestión. De esta forma, cuando se referencie a un píxel de la imagen original, es posible identificar de forma veloz y precisa el tríxel al que pertenece, así como a todos sus píxeles compañeros dentro de dicho tríxel.

Algoritmo 3: Genera mapa topológico ($Generate_Topological_Mesh$)

```
Data: Integer image img sized n x m
Data: Trixel trixel
Result: Floating point image TriToM sized n x m
edges ← trixel.get_edge_list()
Get_Max_Point(max_pt, edges)
Get_Min_Point(min_pt, edges)
for i ∈ [min_pt → x..max_pt → x) do
    for j ∈ [min_pt → y..max_pt → y) do
        if Pt_Belong_To_Trixel(i, j, edges) then
            trixel ← Get_Image_Info(img[i, j])
            TriToM[i, j] ← trixel
        end
    end
end
return TriToM
```

En el algoritmo 3 puede apreciarse cómo el recorrido de píxeles se hace de forma rectangular; se detectan los puntos extremos de los vértices del tríxel y se recorre el área que encierra dichos extremos. En el caso de que el píxel que se esté recorriendo pertenezca al área de dicho tríxel, entonces se incluye la información de este píxel para su consideración entre las distintos descriptores del tríxel y se actualiza el TriToM, en el punto de localización del píxel, con el tríxel en cuestión.

El algoritmo para crear TriToM resulta eficiente para el procesado de imágenes estáticas y, aunque no hay un procesado especializado para vídeo, su aplicación en secuencias de video ofrece un notable rendimiento. En la tabla 4.1 puede apreciarse un conjunto de imágenes, a partir de las cuales, se genera el TriToM de una escena previo al inicio del proceso de segmentación. Para dichas imágenes la segmentación mostrada ha sido realizada manualmente ya que se pretende mostrar cómo las distintas texturas de los tríxeles, así como su contenido, pueden condicionar las distintas regiones o clases en las que se pueden dividir la imagen.

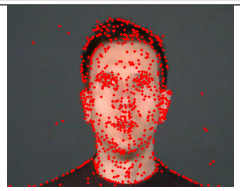
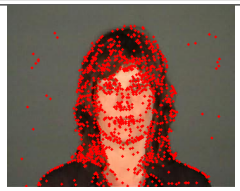
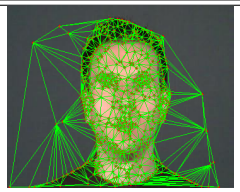
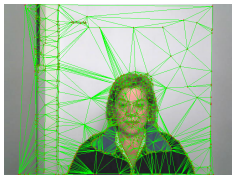
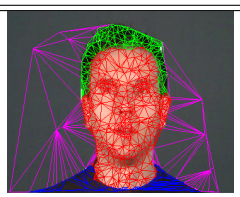
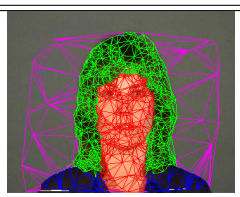
Fase	DaFEx	DaFEx	Vedana
Imagen			
Fase 1			
Fase 2			
Fase 3			

Tabla 4.1: En la secuencia de imágenes se muestra el proceso de generación de tríxeles. Los valores de mínimos locales proporcionados por el DMUM se calculan en la primera fase y se computan como los vértices en la triangulación de los tríxeles que siguen una topología de Delaunay (fase 2). La tercera fase muestra cómo se diferencia visualmente el contenido de los tríxeles que pertenecen a distintas clases. Puede apreciarse la forma en la que el color o la textura pueden caracterizar al tríxel, ya sea en cuanto a su contenido como en cuanto a su tamaño, forma, etc.

4.3. Segmentación de la malla

Una vez generada la malla de tríxeles, se pasa a la siguiente fase, la cual hace uso de dicha malla como parámetro de entrada. Al contrario de lo que sucede con los superpíxeles, el proceso de segmentación no se realiza de forma directa sobre la malla de tríxeles generada. Es decir, no se procesa a nivel atómico cada tríxel por los motivos descritos en la sección 4.4.2, sino que se considera la malla como un grafo ponderado en el que se adaptan los cortes del algoritmo GrabCut a la malla de tríxeles, considerando las características topológicas de la malla así como el procesado de los tríxeles como unidad básica de información. De esta forma, dada una determinada malla de tríxeles, se define como segmentación de dicha malla a la división de los tríxeles que componen la malla en varios agrupamientos en cuyo interior se hallen los tríxeles que comparten propiedades similares entre sí en cuanto a su contenido. Por tanto, en esta sección se describe la modificación del algoritmo GrabCut explicado en la sección 4.1.2.

A diferencia de lo que ocurre en el GrabCut, los distintos píxeles de la imagen no conforman los nodos del grafo, sino que la base sobre la que se sustenta el método de segmentación propuesto es la malla de tríxeles anteriormente descrita. De por sí, dicha malla ya divide distintas regiones de la imagen que guardan cierta homogeneidad en cuanto a la información contenida en el interior de los tríxeles. Se trata de aprovechar la simplificación ofrecida por la malla para reducir el coste computacional de aplicar el algoritmo de segmentación sobre toda la imagen. Para ello, se pretende capturar las fronteras que permitan discernir entre la figura (a partir de ahora, denominada como *objeto de interés*) y el fondo de la imagen. Este proceso de división entre clases puede apreciarse en en conjunto de imágenes que forman parte de la tabla 4.1, donde se muestran imágenes originales y sus correspondientes mallas de tríxeles.

Otro aspecto divergente con la versión original del GrabCut es la forma en la que se relacionan los tríxeles a lo largo del proceso de segmentación; anteriormente se había indicado que el número máximo de vecinos que tiene un tríxel es de tres elementos, por lo que la asignación de pesos en los N-enlaces se realiza considerando la información proporcionada por estos elementos no atómicos (los tríxeles), en lugar de los elementos atómicos de una imagen (los píxeles) como ocurre en el GrabCut original. Además, en dicha versión del algoritmo original, el número máximo de vecinos que tiene un píxel es de ocho, por lo que, la nueva propuesta, simplifica el grado de relación entre las unidades atómicas de cara al procesado de la segmentación. Por último, en el algoritmo propuesto siempre se utiliza la misma malla generada al inicio del proceso. En el GrabCut original se genera un grafo nuevo previo a la aplicación del algoritmo de segmentación, en el caso de los tríxeles, si bien se recalculan los enlaces asignados a los tríxeles a medida que van cambiando las relaciones topográficas entre los grupos, la estructura de tríxeles se reutiliza al completo.

La Figura 4.7 ilustra de forma sucinta el algoritmo para la segmentación de tríxeles. No obstante, a continuación se describen con un poco más de detalle los siete pasos fundamentales que se suceden a lo largo del proceso de segmentación:

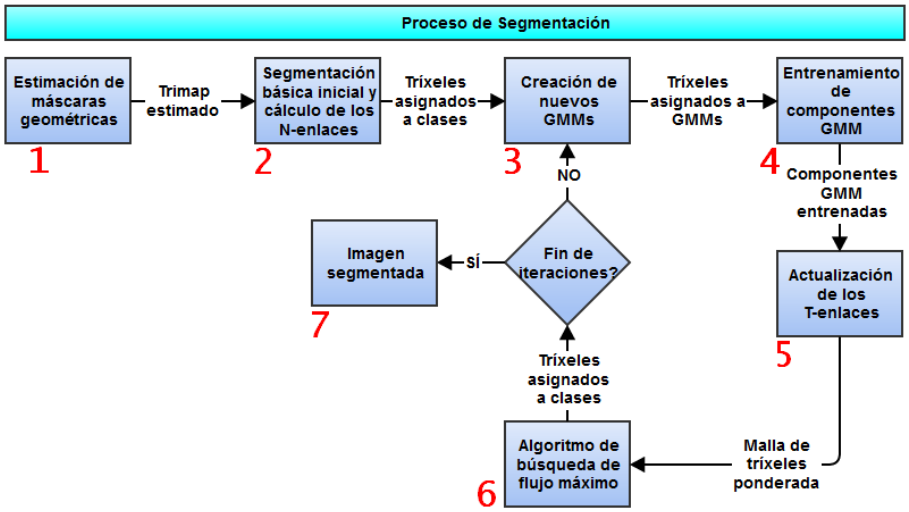


Figura 4.7: Diagrama de bloques del proceso para la segmentación de tríxeles.

1. Se genera una estimación de las regiones de interés. El GrabCut es una técnica que incluye el *"trimap"* y modelos probabilísticos de color para conseguir una segmentación robusta. El trimap actúa como un parámetro de entrada con el fin de inicializar los modelos probabilísticos. Viene definido por la tupla $T = \{T_U, T_F, T_B\}$, consistiendo en una máscara inicial en la que los tríxeles pendientes de analizar son marcados como desconocidos (T_U) y los que proporcionan la información discriminante son marcados como fondo (T_B) u objeto de interés (T_F). Aquellos tríxeles cuyos centroides formen parte de cada una de las regiones pertenecientes a cada una de las clases anteriormente comentadas, son marcados en consecuencia como parte de dicha clase al principio del proceso. Precisamente, en este punto interviene el conocimiento top-down; a partir de saber qué se pretende buscar se enfoca la solución del problema de manera oportuna en base a la información facial pertinente.

2. En el caso de que el tríxel fuera marcado como parte del *fondo*, sería clasificado dentro de dicho grupo. Por el contrario, si el tríxel fuera marcado como parte del *objeto de interés*, o simplemente no fuera marcado (grupo desconocido), entonces se asociaría en principio al grupo de tríxeles que pertenecen a la clase *objeto de interés*. Estos últimos, los asociados a la clase *objeto de interés*, son los que serán analizados durante el proceso de segmentación para determinar cuáles de ellos serán clasificados como pertenecientes a la clase *fondo*, quedando el resto clasificados como para del *objeto de interés*.

3. En el tercer paso del algoritmo de segmentación, se genera un conjunto de modelos de mezcla de gaussianas (*"Gaussian Mixture Models"* (GMM))

con el fin de desarrollar modelos que sirvan, en virtud de la información relacionada entre dichos modelos para la clase perteneciente al tríxel y la propia información del tríxel, para dotarlo de un valor que se relacionará con la fuente o el sumidero con el objetivo de calcular los T-enlaces. Se trabaja con dos GMM; uno para la descripción del objeto de interés y otro para la descripción del fondo de la imagen respectivamente.

4. A continuación, cada tríxel de la clase del *objeto de interés* se asigna a la componente gaussiana más similar del GMM perteneciente al objeto de interés. Por el contrario, aquellos tríxeles etiquetados como *fondo* de la imagen son asignados a la componente gaussiana más similar del GMM perteneciente al fondo de la imagen. Posteriormente, se recalculan los GMM con los nuevos tríxeles asignados, desechando la versión previa de los GMM.

5. Se completa la información de la malla asignando pesos a las aristas que separan los tríxeles de acuerdo a lo explicado en la sección 4.3.1.

6. Se aplica el algoritmo de maximización de flujo descrito en la sección 4.3.2 donde algunos tríxeles del objeto de interés podrían ser reubicados.

7. Los pasos cuatro a siete son repetidos hasta que el proceso de clasificación finalice según las iteraciones especificadas antes del comienzo de todo el proceso.

4.3.1. Inicialización de la malla

La inicialización de la malla la componen los primeros tres pasos mencionados anteriormente. En el primer momento, se selecciona el área de interés que se desea buscar dentro de la escena proporcionada. Este proceso de selección se realiza fuera del entorno del algoritmo debido a que se trata de un paso previo el cual resulta un requisito de entrada indispensable, tanto para el algoritmo propuesto, como para el GrabCut. Si bien este requisito está sujeto al criterio del usuario, es posible adaptar la búsqueda de acuerdo al tratamiento de un problema en concreto.

En el problema que nos ocupa, se pretende segmentar personas dentro de una escena con el objetivo de poder analizar elementos no faciales que la caracterizan. Para ello, se ha hecho uso de una selección geométrica adaptativa. Es un esquema automático que hace procesamiento top-down, para ello es requisito de partida disponer de la información de la detección facial. Del conjunto de detectores disponibles se ha elegido el detector ENCARA2 [37] para realizar esta tarea. Una vez conocida la posición de los ojos, se aplica un esquema geométrico como el que se describe en esta sección. Se trata de un modelo adaptativo que depende de dos factores; el primero es la distancia existente entre la posición de ambos ojos y el segundo es la alineación existente entre ambos ojos, lo que permite que el modelo sea adaptable frente a cambios de pose.

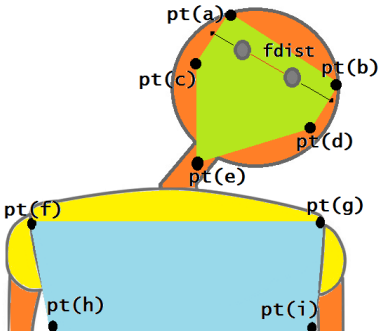


Figura 4.8: Selección geométrica del *trimap*.

Sea la distancia entre los ojos f_{dist}, se identifican los distintos puntos geométricos $pt(a), pt(b), ..., pt(k)$ como se refleja en la ecuación 4.5 e ilustra en la Figura 4.8.

$$pt(k) = (f_{dist} + w * \vec{v}) + \vec{p_e} * w_k \tag{4.5}$$

Donde $\vec{v}$ representa la dirección del vector en función de la desviación entre el alineamiento de los ojos, w es la distribución simétrica del peso para conseguir la misma distancia a ambos lados de la imagen facial y w_k es la distribución de pesos individual para cada uno de los puntos. El vector $\vec{p_e}$ es el vector de rotación de 90º que orienta los puntos a lo largo de la imagen. En base a esta distribución geométrica se consigue señalizar una serie de puntos que encierran un área. Tal y como puede apreciarse en la Figura 4.8, las área estimadas de forma automática son la facial con el color verde y la corporal con el color azul respectivamente.

En segundo lugar se definen los modelos de mezcla de gaussianas (GMM), desde una perspectiva formal, como un conjunto de modelos probabilísticos que asumen que todos los puntos de una imagen pueden ser generados a través de un número finito de distribuciones gaussianas con una serie de parámetros desconocidos. Podría considerarse que la mezcla de modelos se trata de una generalización del agrupamiento *kmeans* en la medida que incorpora información sobre la estructura de la covarianza de los datos así como de los centros de las gaussianas latentes. Supóngase que existen K grupos de datos y que, cada grupo, es una distribución gaussiana parametrizada por μ_k y Σ_k. Denotar los datos por X / X $\in \mathbb{R}^d$. La densidad de la componente k es:

$$p_k(\chi) = \frac{1}{(2\pi)^d \mid \Sigma_k \mid} \exp\left(\frac{-(\chi - \mu_k)^t \Sigma_k^{-1} (\chi - \mu_k)}{2}\right) \tag{4.6}$$

Además, la probabilidad de la componente k es π_k y la mezcla de densidad es:

$$p(\chi) = \sum_{k=1}^{K} \pi_k p_k(\chi) \tag{4.7}$$

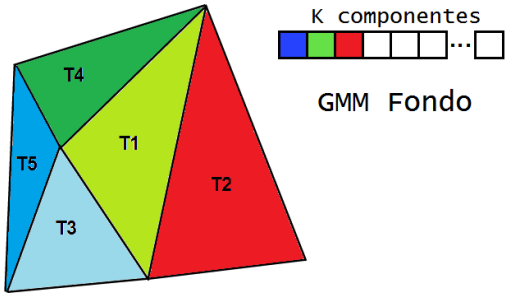
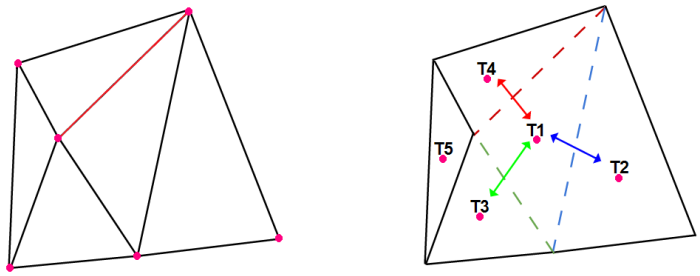


Figura 4.9: Asociación de los tríxeles con las componentes de la gaussiana a la que pertenecen. En el ejemplo puede apreciarse el modelo de mezcla de gaussianas con K componentes. La similitud mostrada por las características de los tríxeles, para dicho ejemplo, haría que los tríxeles etiquetados como T3 y T5 sean relacionados con la componente primera del modelo GMM de la clase *fondo*. De la misma forma, los tríxeles T1 y T4 están asociados con la segunda componente o el T2 con la tercera componente respectivamente.

De esta forma, en el tercer paso del algoritmo propuesto, se generan K componentes para cada uno de los dos GMM; uno para el fondo de la imagen y otro para el objeto de interés que se quiere destacar. En otras palabras, se deben crear un total de $2 \cdot K$ componentes. Una vez son creadas, se dividen los tríxeles de cada una de las regiones en K componentes (mismo número de componentes para fondo y objeto), tal y como se muestra en la Figura 4.9. En ese momento se inicializan las componentes gaussianas de cada región con la información que va a ser determinante de cara a realizar el proceso de segmentación. En este caso, se ha escogido el color, considerando el espacio RGB, como característica discriminante. El color interno a un tríxel viene dado por la media del color de los píxeles que conforman dicho tríxel. Para que exista un buen contraste entre el fondo de la imagen y el objeto de interés, es necesario generar componentes gaussianas con poca variación. El siguiente paso es procesar los GMM a través del uso de autovectores de la matriz de covarianza para encontrar brechas de elementos entre los grupos generados. Es, precisamente, a través de estos grupos cómo se procesan los GMM. La media y la matriz de covarianza se calculan a raíz de los datos proporcionados por los tríxeles. Posteriormente, se completa el modelo de mezcla de gaussianas mediante el determinante obtenido a partir de la matriz de covarianza. Estos modelos se utilizan para realizar el agrupamiento de los tríxeles asociados a cada clase y así determinar la fortaleza existente entre un determinado tríxel y su T-enlace a las clases del fondo (sumidero) o del objeto de interés (fuente). Sin embargo, en cada fase de aprendizaje del algoritmo, es decir, en cada iteración del mismo, existe una etapa que redefine los nuevos GMM para el fondo y para el objeto de interés. Ello es así porque, en cada iteración, existen tríxeles que pueden pasar de ser parte del objeto interés a ser parte fondo, y viceversa. Cuando esto ocurre, es necesario actualizar los GMM



(a) GrabCut (b) TriToM

Figura 4.10: En estas dos imágenes se puede apreciar en color fucsia los núcleos de información necesarios para generar los T-enlaces y los N-enlaces. Para la versión clásica del GrabCut, los píxeles son los que ocupan estos centros y, además de formar los vértices del grafo, dotan de información a ambos tipos de enlace. En el caso de TriToM, la información necesaria para abastecer los enlaces la proporciona el propio tríxel y, si bien se hace uso de las aristas que separan los tríxeles, la información contenida en las mismas es fruto del cálculo del contenido de los dos tríxeles que divide a través de la ecuación 4.8.

para reflejar la nueva distribución de color tanto del fondo como del objeto de interés.

Una vez que los modelos se recalculan, cada tríxel se asocia con la componente gaussiana con la que guarde más similitud de las K componentes del GMM de la clase a la que pertenece. Esto se realiza de forma sencilla; evaluando la ecuación de los GMM con la información proporcionada por el tríxel (véase la Figura 4.9). Es importante matizar que cada tríxel sólo se asocia con la componente de los GMM al que pertenece (fondo u objeto de interés). De esta manera, el par del tríxel *(clase a la que pertenece, componente GMM)* permite indexar uno de los $2 \times K$ componentes de forma rápida para acceder a los datos necesarios con el fin de generar los T-enlaces.

En la última fase, previa a la aplicación del algoritmo de maximización de flujo, se añade la información complementaria a los enlaces. En el GrabCut, los nodos (píxeles) son considerados los vértices y están unidos a través de las aristas. Además, la información contenida, tanto en los T-enlaces (aristas entre vértices) como en los N-enlaces (aristas con la clase), surge a raíz de la información contenida en los propios vértices (véase la Figura 4.10a). En el caso de TriToM los vértices siguen existiendo sólo desde un punto de vista físico, no son funcionales. En otras palabras, sirven para mantener la estructura que une aristas pero, la información contenida en los T-enlaces y en los N-enlaces, no surge a raíz de la propia información de los vértices, sino que lo hace del propio tríxel (véase la Figura 4.10b).

Como se ha comentado anteriormente, se distinguen los N-enlaces de los T-enlaces. Los primeros conectan a los tríxeles con sus vecinos en la malla. Un tríxel tiene, como máximo, tres N-enlaces. Este tipo de enlaces describen la

Tipo de tríxel	T-enlace Fondo	T-enlace Objeto Interés
$T_i \in TrimapFondo$	λ	0
$T_i \in TrimapObjeto$	0	λ
$T_i \in TrimapDesconocido$	$D_{Fondo}(T_i)$	$D_{Obj}(T_i)$

Tabla 4.2: Donde λ es un valor constante que representa el mayor peso posible de la malla y que, formalmente, se define como $\lambda = \text{máx}_{T_i} \sum_{n:(T_i,T_j)\in E} N(T_i, T_j)$ donde E es el conjunto de todas las aristas que unen tríxeles. Además, D_{Fondo} y D_{Obj} es la similitud existente para que el tríxel T_i pertenezca al *fondo* o al *objeto de interés*. Desde una perspectiva formal, esta similitud viene definida de la forma descrita por la ecuación 4.9.

penalización que existiría en caso de poner una frontera, a la hora de segmentar, entre los dos tríxeles implicados. Recordemos que el objetivo es que esta frontera sea alta en regiones de bajo gradiente y baja en regiones de alto gradiente. De esta forma, el N-enlace asignado a una arista que separa el par de tríxeles (T_i, T_j) viene definido por la ecuación 4.8.

$$N(T_i, T_j) = \frac{50}{dist(T_i, T_j)} \mathrm{e}^{-\beta \| z_{T_i} - z_{T_j} \|^2} \tag{4.8}$$

Donde $\beta = \frac{1}{2 \times avg(\sqrt{\| z_{T_i} - z_{T_j} \|})}$ para todo par de tríxeles vecinos en el grafo, z_{Ti} hace referencia a la información suministrada por el tríxel i, y $dist(T_i, T_j)$ es la distancia euclídea existente entre los centroides de ambos tríxeles.

Por último, los T-enlaces conectan cada tríxel con los nodos fuente y sumidero o, como han sido previamente redefinidos, los nodos pertenecientes a la clase del objeto de interés o a la clase del fondo respectivamente. Este tipo de enlaces describe la probabilidad de que un tríxel pertenezca a cualquiera de estas dos clases. En el algoritmo, dicha probabilidad se calcula a partir de los parámetros almacenados dentro de los dos GMM previamente entrenados. A medida que se van redistribuyendo los tríxeles entre una clase y otra, los GMM deben ser reentrenados y estas probabilidades van cambiando. De esta forma, el T-enlace de un determinado tríxel T_i es definido en función de la etiqueta que tiene el tríxel en cuestión (véase la tabla 4.2) y responde a la ecuación 4.9.

$$\begin{aligned} D(T_i) = -\log \sum_{j=1}^{H} \Bigg[& \pi(\alpha_{T_i}, j) \frac{1}{\sqrt{det \sum(\alpha_{T_i}, j)}} \\ & \times \exp\left(\frac{1}{2} [z_{T_i} - \mu(\alpha_{T_i}, j)]^\top \sum(\alpha_{T_i}, j)^{-1} [z_{T_i} - \mu(\alpha_{T_i}, j)] \right) \Bigg] \end{aligned} \tag{4.9}$$

Se trata de una ecuación implementada por Talbot [200] para el modelo clásico del GrabCut y que, según dicho autor, corrige errores de la fórmula original en la que no se consideraba la suma de componentes necesaria para el

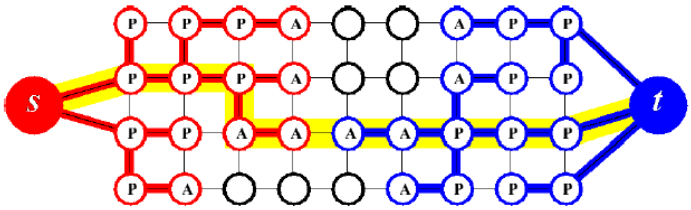


Figura 4.11: Ejemplo de una escena entre los árboles de búsqueda S (nodos rojos) y T (nodos azules) al final de la primera fase del algoritmo de maximización de flujo. En ella puede apreciarse una ruta (en color amarillo) entre la fuente (s) y el sumidero (t). Los nodos activos están etiquetados con la letra A, mientras que los pasivos con la letra P. Por último, los nodos libres carecen de etiqueta [23]. En el contexto en el que se desarrolla la malla de tríxeles, es importante identificar estos nodos de los árboles de búsqueda como los tríxeles de la malla (de acuerdo con el enfoque de enlaces entre tríxeles proporcionado en la Figura 4.10).

procesado de los GMM. La constante H se identifica como el peso de mayor valor en el grafo y se calcula de la forma:

$$H = \max_{m} \sum_{n:(m,n)\in E} N(m,n) \tag{4.10}$$

Donde E representa el conjunto de todas las aristas que unen todos los tríxeles que son vecinos.

4.3.2. Algoritmo de maximización de flujo

El algoritmo para realizar la segmentación sobre la malla está basado en el desarrollado para la versión clásica de GrabCut (véase la sección 4.1.2). Por tanto, el objetivo es ir recorriendo las aristas de los tríxeles, con la información previamente calculada, e ir delimitando las fronteras entre lo que se considera *fondo* de la imagen y lo que se considera el *objeto de interés*. Según afirma Boykov en su trabajo acerca del GrabCut [23], explorar todas las posibles opciones para determinar una frontera entre ambas clases suele ser una opción extendida en muchos métodos de la visión por computador. Se trata de un proceso de ramificación de búsqueda que, al final, acaba con la exploración de la mayoría de los píxeles de la escena. Sin embargo, Boykov propuso un método bastante más eficiente basado en ir aumentando una ruta entre la fuente (objeto de interés) y el sumidero (fondo) de forma paulatina, sin necesidad de explorar todas las posibles alternativas y dotando el proceso de segmentación de la malla con un comportamiento autónomo que facilita dicha tarea. Precisamente, este es el enfoque adaptado en este trabajo.

Según puede apreciarse en la Figura 4.11 se mantienen dos árboles de búsqueda, definidos como S y T, cuyas raíces son las dos clases que se distinguen: la fuente s para el árbol S y el sumidero t para el árbol T respectivamente. En el árbol S ninguna de las aristas de cada tríxel padre a su tríxel hijo están saturadas (es decir, peso distinto de 0), mientras que en el árbol T es en sentido opuesto; todas las aristas desde un tríxel hijo a su tríxel padre no están saturadas. Aquellos tríxeles que no forman parte de S o T se denominan tríxeles libres.

Por otro lado, los tríxeles en los árboles de búsqueda pueden estar en dos posibles estados; activo o pasivo. Los tríxeles activos representan la capa exterior del árbol de búsqueda, mientras que los tríxeles pasivos son los tríxeles internos de dicho árbol. Esto significa que sólo los tríxeles activos permiten al árbol de búsqueda expandirse adquiriendo nuevos hijos (del conjunto de tríxeles libres) a través de enlaces no saturados. En cambio, los tríxeles pasivos no pueden crecer debido a que se encuentran completamente rodeados por otros tríxeles de su mismo árbol.

El momento clave del algoritmo ocurre cuando los tríxeles activos de ambos árboles de búsqueda entran en contacto, de hecho, una *ruta de aumento* es detectada cuando un tríxel activo de un árbol de búsqueda detecta, entre sus vecinos, un tríxeles activo del otro árbol de búsqueda. A partir de ese instante, el algoritmo cambia de fase. Así, existen tres fases bien diferenciadas [23]:

- Fase de crecimiento: En esta fase los tríxeles activos de los árboles de búsqueda exploran entre sus vecinos libres nuevos hijos para ser adoptados. Cuando adquieren un tríxel libre, el nuevo tríxel pasa a ser un tríxel activo del árbol de búsqueda al que pertenece. Cuando un tríxel activo ha adquirido a todos sus vecinos, entonces este tríxel pasa a ser un tríxel pasivo. Esta fase concluye cuando un tríxel activo de uno de los árboles encuentra, entre sus vecinos, un tríxel que pertenezca al otro árbol. En ese momento se ha encontrado una ruta de aumento y se pasa a la siguiente fase.

- Fase de aumento: Se ha encontrado una ruta de aumento, los árboles de búsqueda pueden romperse en *bosques*; debido a que se sustrae el máximo flujo posible de la ruta a las aristas pertenecientes a dicha ruta, existe la posibilidad de que se sature algún enlace que mantenga unidos un grupo de tríxel a uno de los árboles y estos tríxeles sean "descolgados" de su árbol de búsqueda original formando un bosque de tríxeles huérfanos. En ese instante, s y t seguirían siendo las raíces de sus respectivos árboles, pero también existiría un bosque de huérfanos que, si bien no están enlazados de forma directa a ninguno de esos dos conjuntos, mantendrían su etiqueta del conjunto al que pertenecían antes de la ruptura.

- Fase de adopción: Se restituyen los árboles de búsqueda S y T. El objetivo de esta fase es restaurar la estructura única de árbol de los conjuntos S y T con sus raíces correspondientes. Para ello se inicia la búsqueda de un padre para cada tríxel huérfano. El nuevo padre debe pertenecer al mismo conjunto (S o T) que el huérfano y, además, la unión de este padre

con el huérfano debe producirse a través de un enlace no saturado. En el caso de que no exista la posibilidad de unir un tríxel huérfano bajo las condiciones anteriormente especificadas, dicho tríxel se elimina de su árbol de pertenencia y pasaría a ser un tríxel libre. De darse esta circunstancia y dicho tríxel tuviera hijos, todos sus hijos también pasarían a ser huérfanos y serían analizados en consecuencia. Esta fase concluye cuando no quedan tríxeles huérfanos que analizar, lo cual significa que los árboles S y T han sido restituidos. Debido a que existe la posibilidad de que los árboles pierdan tríxeles, esta fase trae como resultado una posible contracción de los árboles de búsqueda para seguir explorando mejores alternativas en busca de una segmentación más precisa.

Las fases se ejecutan siguiendo el mismo orden en el que han sido descritas anteriormente y lo hacen de forma cíclica; es decir, después de la fase de adopción comienza de nuevo la fase de crecimiento para adquirir los nuevos posibles tríxeles libres por parte de los árboles de búsqueda. El algoritmo concluye cuando los árboles de búsqueda no pueden crecer debido a que los tríxeles activos se encuentran entre sus propios tríxeles pasivos y los enlaces saturados que les separa de los tríxeles del árbol de búsqueda opuesto (marcando la frontera de segmentación).

Desde un punto de vista algorítmico, de acuerdo a lo descrito con anterioridad, es posible detallar un poco más cada una de estas tres fases. De esta forma, si se asume que se mantiene una lista A con los tríxeles activos y una lista O con los tríxeles huérfanos, el proceso se define en el algoritmo 4.

Algoritmo 4: Algoritmo de flujo máximo

Data: $S = \{s\}$, $T = \{t\}$, $A = \{s, t\}$, $O = 0$
while *true* **do**
  grow S or T to find an augmenting path P from s to t
  if $P = 0$ **then**
    Exit
  end
  augment on P
  adopt orphans
end

Antes de explicar cada fase, es conveniente aclarar la nomenclatura. De esta forma, para especificar la clase a la que pertenece un tríxel es necesaria la función $TREE(p)$, siendo p el tríxel evaluado. Esta función sólo puede retornar tres posibles valores; S si $p \in S$, T si $p \in T$ o 0 si p es libre. En el caso de que el tríxel p pasara a formar parte de alguno de los árboles de búsqueda, entonces su tríxel padre sería almacenado como $PARENT(p)$. Las raíces de los árboles de búsqueda así como los tríxeles huérfanos y, también, los libres, no tienen padre ($PARENT(p) = 0$). Por último, se hace uso de la notación $NLink(p \rightarrow q)$ para indicar el peso de la arista entre los tríxeles (p, q). Estos enlaces deben estar no saturados (tener un valor distinto de 0) para garantizar la posibilidad de que el tríxel p sea un padre del tríxel hijo q en un árbol de búsqueda.

En la fase de crecimiento descrita en el algoritmo 5, se van recorriendo todos

los tríxeles activos del árbol de búsqueda. A medida que se recorren todos los vecinos de un tríxel p, pueden darse tres posibles circunstancias; si todos sus vecinos son de su misma clase, entonces el tríxel p deja de ser activo. La segunda opción es que el tríxel p encuentre un tríxel libre, en cuyo caso pasaría a ser el padre del nuevo tríxel y el nuevo tríxel pasaría a pertenecer a la clase de p. El último caso es que encuentre un vecino del árbol opuesto, en cuyo caso se saltaría a la fase de aumento.

Algoritmo 5: Fase de crecimiento

```
while A ≠ 0 do
    pick an active trixel p ∈ A
    for every neighbor q such that NLink(p → q) >0 do
        if TREE(q) = 0 then
            TREE(q) = TREE(p)
            PARENT(q) = p
            A = A∪ {q}
        end
        if TREE(q) ≠ 0 and TREE(q) ≠ TREE(p) then
            return P = PATH_{s→t}
        end
    end
    remove p from A
end
return P = 0
```

En la segunda fase, ilustrada en el algoritmo 6, se aprovecha la ruta de aumento proporcionada desde s hasta t. Se busca la capacidad que supone un cuello de botella entre todos los N-enlaces de las aristas de la ruta para actualizar, haciendo uso de dicho valor, al conjunto de aristas pertenecientes a la ruta. Aquellas aristas que terminen con un valor saturado rompen el árbol y, los tríxeles descolgados, pasan a ser huérfanos.

Algoritmo 6: Fase de aumento

```
Data: find the bottleneck capacity △ on P
Data: update the graph by pushing flow △ through P
for each edge (p, q) ∈ P that becomes saturated do
    if TREE(p) = TREE(q) = S then
        PARENT(q) = 0
        O = O ∪ {q}
    end
    if TREE(p) = TREE(q) = T then
        PARENT(p) = 0
        O = O ∪ {p}
    end
end
```

La fase de adopción se ejecuta mientras existan tríxeles en la lista de huérfanos (véase algoritmo 7). Cada tríxel perteneciente a dicha lista intenta permanecer en el árbol de búsqueda al que pertenecía sólamente si es capaz de encontrar un nuevo padre en dicho árbol. En el caso de no encontrarlo, ese tríxel huérfano pasaría a ser un tríxel libre.

Algoritmo 7: Fase de adopción

```
while O ≠ 0 do
    pick an active orphan trixel p ∈ O
    remove p from O
    PARENT(p) ← 0
    for every neighbor q such that NLink(p → q) >0 do
        if TREE(q) = TREE(p) then
            PARENT(p) ← q
        end
    end
    if PARENT(p) = 0 then
        for every neighbor q such that TREE(q) = TREE(p) do
            if NLink(q → p) >0 then
                A ← A ∪ {q}
            end
            if PARENT(q) = p then
                O ← O ∪ {q}
                PARENT(q) ← 0
            end
            TREE(p) ← 0
            A ← A − {p}
        end
    end
end
```

Esta adaptación del algoritmo para los tríxeles ha seguido los principios básicos propuestos por Boykov [23]. De hecho, se adoptaron particularidades en los detalles de implementación que hacen que, si bien según Boykov no permite obtener la mejor segmentación posible, sí obtiene la más eficiente en términos de tiempo de ejecución y resultado. Entre estas particularidades, destacan tres que son especialmente singulares. En primer lugar, se implementa una cola FIFO para las listas de tríxeles activos y de tríxeles huérfanos que permiten obtener la ruta de aumento más corta posible. En segundo lugar los tríxeles se marcan para calcular su distancia con respecto a la fuente, o al sumidero, con el objetivo de aumentar la velocidad del algoritmo a la hora de asignar padres de los tríxeles huérfanos. Por último, los tríxeles tienen una marca de tiempo que permite evaluar cada tríxel con respecto a la iteración actual con el fin de no repetir acciones.

4.4. Análisis del rendimiento

En esta sección se ofrece un conjunto de experimentos con vistas a comparar el rendimiento del algoritmo propuesto con respecto a la versión clásica del GrabCut que parte de los píxeles. Para ello, se han realizado las pruebas en base a la propuesta realizada en la sección 4.3 con el objetivo de segmentar a las personas presentes en una imagen. Un experimento similar lo propone Rother en su trabajo acerca del GrabCut para comparar el resultado de dicho

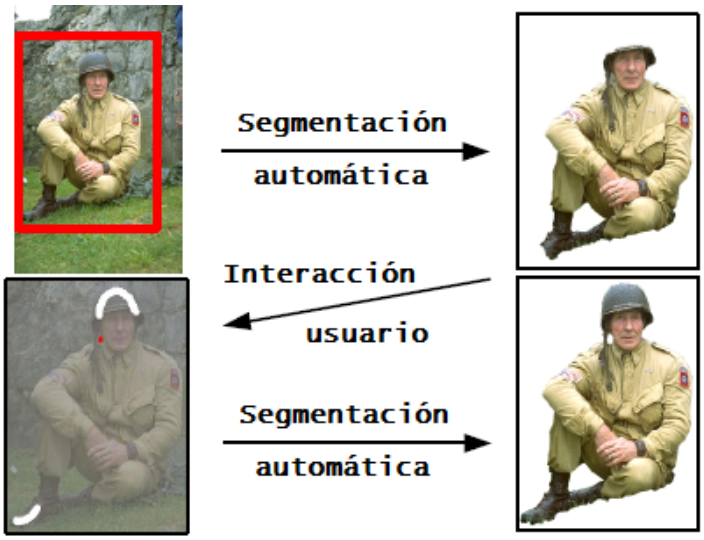


Figura 4.12: Propuesta desarrollada por Rothen para el uso del GrabCut. En esta imagen puede apreciarse como, después de una primera interacción con el usuario en la que se hace una selección inicial rectangular, es necesario un proceso de refinado para poder complementar los elementos contextuales que rodean al sujeto segmentado [176].

algoritmo con otros existentes (entre ellos el de Boykov) [176]. Sin embargo, dicho trabajo no podría ser comparable de forma directa con el enfoque de tríxeles propuesto anteriormente por el siguiente motivo; en la selección del *trimap* existe la interacción de un usuario externo. Es decir, no se trata de un proceso automático de selección que pretenda dilucidar, en primera instancia, la región de interés inicial frente a la selección del fondo inicial (véase Figura 4.12). Por tanto, si bien dicho trabajo establece que el GrabCut es un enfoque más eficaz que el Graph Cut propuesto por Boykov (disminuyendo en un 0.8 % la tasa de error), no es posible comparar su estudio con este trabajo de forma directa.

La alternativa que se plantea para analizar el enfoque del GrabCut con el trabajo presentado en este capítulo, es a través de la réplica del mismo experimento sobre una base de datos ya existente y disponible en Internet. Como se comentó en el capítulo 2, la base de datos PASCAL [77] contiene imágenes de personas con una buena variedad de poses, fondos variables y a diferentes escalas. Debido a la complejidad de la base de datos en cuanto a la enorme variedad de fondos y de poses, se estimó que ofrece el conjunto de datos más adecuado a la hora de realizar las pruebas. De este conjunto de imágenes, se extrajeron aquellas en las que aparecen posando alguna de las 1500 personas que se han identificado en la base de datos. Las pruebas son evaluadas desde una doble perspectiva; por un lado aprovechar las imágenes segmentadas (*goundtruth images*) que ofrece la

	Tasas Obtenidas	
	Píxeles	Tríxeles
Tasa de Verdaderos Positivos	79 %	71 %
Tasa de Verdaderos Negativos	84 %	80 %
Tasa de Falsos Positivos	16 %	20 %
Tasa de Falsos Negativos	21 %	29 %
Exactitud	81.5 %	75.5 %

Tabla 4.3: Comparativa entre el uso de píxeles o de tríxeles para una prueba ejecutada sobre la base de datos PASCAL [77]. La prueba está evaluada sobre los píxeles pertenecientes a cada imagen a través del uso del subconjunto de imágenes de validez que la base de datos PASCAL ofrece para el análisis de datos.

propia base de datos para evaluar el rendimiento de los algoritmos y, por otro lado, evaluar la velocidad del algoritmo propuesto con respecto a su homónimo original variando el número de iteraciones.

La primera prueba permite evaluar el rendimiento del método propuesto. De esta forma, en primer lugar se genera una máscara trimap de acuerdo a lo explicado en la sección 4.3.1 acerca de la selección geométrica, con el objetivo de separar el objeto de interés del fondo de la imagen. Una vez que este *trimap* queda definido, entonces se aplica el algoritmo de segmentación que se desee analizar y se extraen los resultados en forma de máscara. Estos resultados son analizados (píxel a píxel) con la información de anotación mencionada que proporciona la propia base de datos para generar las estadísticas de la técnica analizada. En cuanto a la segunda prueba, el algoritmo de GrabCut permite añadir un parámetro que permite forzar un número de iteraciones mínimas sobre el grafo. La idea consiste en realimentar el proceso con la segmentación generada por el algoritmo de flujo máximo (véase sección 4.3.2) para volver a aplicar el proceso de segmentación. De esta forma se refina el proceso y la imagen generada con más iteraciones tiende a tener una segmentación mucho más ajustada que aquella en la que se aplica una única iteración.

En la tabla 4.3 puede apreciarse la diferencia que existe entre la versión original de GrabCut (que procesa píxeles) y la nueva propuesta realizada (que procesa tríxeles). Tal y como puede distinguirse, el enfoque propuesto pierde un 6 % en cuanto a la exactitud (*accuracy*) sobre el número total de píxeles con respecto al enfoque original. No obstante, esta exactitud ofrece una cifra bastante aceptable si se tiene en cuenta las principales limitaciones que ofrece la adaptación de los tríxeles en función de la forma en la que se ha simplificado la imagen. En la Figura 4.13 puede observarse la imagen que procesaría el algoritmo considerando píxeles frente a una generalización de la imagen que procesaría si, por el contrario, se consideran los tríxeles. Como puede apreciarse, dicha simplificación se basa en que cada uno de los tríxeles se trata como un elemento atómico. No se introducen todos los píxeles en el modelo, sólo el valor de la

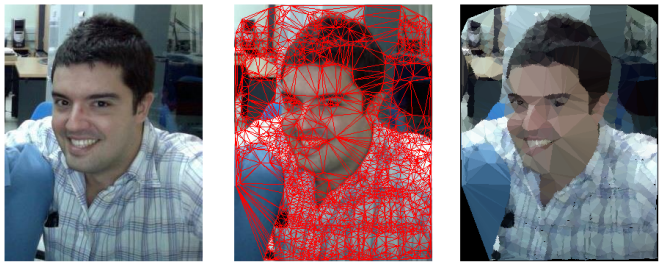


Figura 4.13: Simplificación de elementos de procesado desde el modelo GrabCut original que procesa píxeles (a) al nuevo modelo que procesa tríxeles. La imagen simplificada (c) ha sido generada para la ocasión con el objetivo de mostrar el resultado de extrapolar el valor cromático medio de cada tríxel a los píxeles que lo componen. Resulta importante matizar que la nueva propuesta no procesa la imagen (c), sino que tan sólo procesa el valor cromático medio de cada tríxel que se muestra en la imagen (b).

media de color contenida para cada uno de los tríxeles, sin embargo, a efectos de visualización, si ese valor cromático medio se extrapola a todos los elementos pertenecientes a cada uno de los tríxeles, el resultado es el de la Figura 4.13c. En otras palabras, existe una notable simplificación de la imagen a nivel de elementos atómicos de proceso; en el caso de GrabCut se procesan 94230 píxeles para la imagen de dicha Figura, por el contrario, en el caso del enfoque propuesto son 4403 tríxeles a procesar para dicha imagen. Precisamente, esta reducción afecta a la exactitud debido a que, al procesar la imagen simplificada, se tiene menos información para poder delimitar de forma precisa el corte entre el objeto de interés o el fondo.

Sin embargo, el nuevo enfoque de tríxeles ofrece una faceta positiva en cuanto a la velocidad de ejecución del algoritmo. Esta mejora se sitúa en torno al 80 % del tiempo necesario empleado por el GrabCut original. También es cierto que existen factores determinantes a la hora de poder cuantificar esta mejoría ya que la simplificación es dependiente del número de tríxeles que se formen. A priori, y aunque la formación de tríxeles es determinista, el número de tríxeles que se forma para una escena depende del número de mínimos que detecta el algoritmo DMUM (véase sección 4.2.1). El algoritmo DMUM calcula los mínimos en función del contenido de la imagen, por lo que no existe una regla que ayude a dilucidar una fórmula precisa que relacione la imagen con la cantidad de tríxeles que se pueden generar en su interior. No obstante, se ha realizado un estudio analizando los tiempos de ejecución del algoritmo y comparándolo con la versión clásica del GrabCut (véase la Figura 4.14). En dicho estudio se puede apreciar cómo el proceso de refinar los resultados de la segmentación a lo largo de múltiples iteraciones ofrece una notable mejoría, partiendo de un 62.5 % de mejora en el tiempo inicial para una sola iteración del algoritmo y llegando al 80 % para 10 iteraciones. Además, un factor a considerar es la penalización que

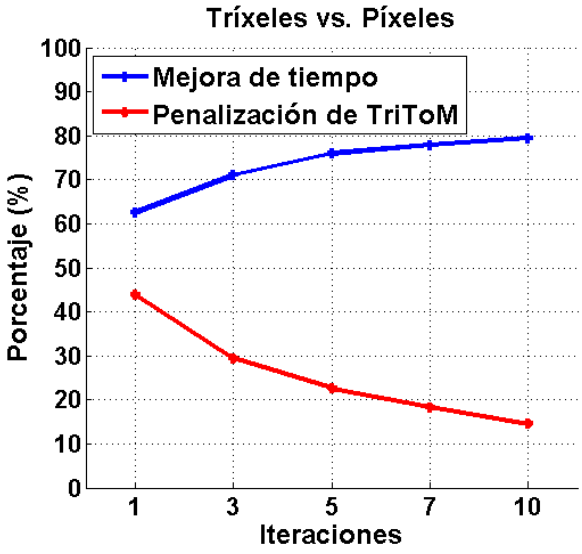


Figura 4.14: Comparación temporal entre el uso de píxeles y tríxeles en función del número de iteraciones del algoritmo aplicado. La línea azul marca el porcentaje de mejora que ofrece el uso de tríxeles frente al uso de píxeles. Puede apreciarse una curva ascendente a medida que aumenta el número de iteraciones. De la misma forma, en color rojo puede apreciarse como, a medida que se aumenta este número de iteraciones, la penalización que supone generar la malla de tríxeles (TriToM) se va atenuando cada vez más.

supone generar la malla de tríxeles, pero a medida que se va incrementando el número de iteraciones dicha penalización tiene menos influencia sobre el tiempo total de ejecución.

En resumen, la simplificación de la información de la imagen a través de la aplicación el método propuesto genera resultados que son ligeramente peores en términos de rendimiento de segmentación ($-6\,\%$), pero este empeoramiento de las tasas se ve contrarrestado con la notable mejoría ($> 60\,\%$) que ofrece el algoritmo en cuanto a tiempos de ejecución.

La Figura 4.15 ofrece una comparación entre la ejecución de las distintas fases descritas a lo largo del presente capítulo en función del elemento a procesar. La fase de creación de tríxeles es la que consume más tiempo dentro de la propuesta realizada. Sin embargo, este consumo no es mucho mayor que el ofrecido por esta misma fase para la versión original del GrabCut. Esto es así porque, durante la creación del grafo, el GrabCut hace uso del algoritmo *kmeans* a la hora de dividir los distintos píxeles entre las K componentes de las GMMs según lo especificado por el trimap. La carga de trabajo a realizar en dicha distribución no es la misma en el caso de trabajar con píxeles que si se procesaran tríxeles. Por ejemplo, en una imagen de 270×349 píxeles de la Figura 4.13, mientras que el GrabCut original procesaría 94230 píxeles como entrada en el algoritmo del

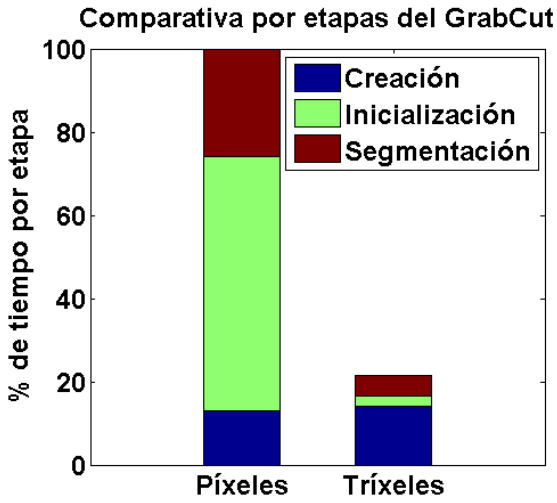


Figura 4.15: Comparación temporal de la ejecución de las fases del algoritmo.

kmeans (y en el resto de fases del proceso), la propuesta realizada a través de TriToM sólo haría uso de 4403 tríxeles, justamente un 4.6 % de lo que necesita el GrabCut original. Sin embargo, en la primera fase, la creación de la malla de tríxeles supone un aumento en el tiempo de ejecución de la nueva propuesta, de ahí a que esta propuesta sea ligeramente superior a la versión que procesa los píxeles durante esta primera fase. Por otro lado, en la fase de inicialización, así como en la de segmentación, la mejoría que ofrece la nueva propuesta queda reflejada de forma notable.

En la tabla 4.5 se muestra un conjunto de imágenes con el objetivo de reflejar las principales diferencias existentes entre los dos enfoques a nivel de resultados. Como puede apreciarse en dichas imágenes, los resultados mostrados por ambos métodos no ofrecen una diferencia muy evidente a simple vista. De hecho, en algunas ocasiones es posible que el enfoque de los tríxeles mejore ligeramente el ofrecido por el GrabCut clásico (por ejemplo, en la imagen de la cuarta fila de dicha tabla) aunque no es la norma (véase la imagen segunda fila). En dichas imágenes pueden observarse algunos errores cuyo origen no se encuentra en la técnica de segmentación, sino que se encuentra en la máscara original seleccionada para ambas técnicas (el *trimap*) en el momento de definir la entrada al modelo de segmentación. Se han desarrollado este tipo de pruebas con una máscara genérica, lo cual puede ser beneficioso en algunas ocasiones, pero perjudicial en otras. Por ejemplo, la máscara propuesta está preparada para segmentar personas de pie o que ocupen la mayor parte de la escena. Sin embargo, ¿qué ocurriría si la persona está sentada tras una mesa? El resultado de la imagen de la sexta fila muestra ese hecho.

Además, si la máscara tiene un carácter más dinámico de búsqueda y, una vez encuentra la región que desea buscar, se le da la alternativa a que la busque

	Tasas Obtenidas	
	Piel	Ropa
Tasa de Verdaderos Positivos	93 %	82 %
Tasa de Verdaderos Negativos	97 %	93 %
Tasa de Falsos Positivos	3 %	7 %
Tasa de Falsos Negativos	7 %	18 %
Precisión	96 %	88 %

Tabla 4.4: Análisis del proceso de segmentación de la base de datos Vedana usando tríxeles (véase sección 3.1).

en el resto de la escena podrían darse errores semánticos cómo el ocurrido en la imagen de la primera fila de la tabla 4.5. Otro tipo de error semántico es aquel en el que ocurren circunstancias en las que, difícilmente, se puede distinguir un elemento oclusivo frente al sujeto de interés (véase la imagen de la tercera fila) ¿podría diferenciarse un gato real de un gato dibujado en una camiseta sin información adicional?. Por otro lado, en la sección 4.3.1 se comentó acerca del uso de ENCARA2 [37] para localizar la posición de los ojos con el objetivo de generar el *trimap* en base a la localización ocular. Como puede apreciarse en la imagen de la quinta fila, la mujer del fondo que no ha sido localizada por el detector ENCARA2, pero sí por el enfoque propuesto cuando se ha rastreado el resto de la imagen en busca de la piel del objeto de interés.

En este contexto, si analizamos la aplicación de los tríxeles en un entorno mucho más controlado como el desarrollado para la base de datos Vedana (véase sección 3.1), los resultados son acordes a la dificultad del escenario presentado. En la tabla 4.4 se ofrecen las cifras obtenidas de segmentar los 1524 fotogramas etiquetados pertenecientes a dicha base de datos. Puede apreciarse, tanto la segmentación de la piel, como los resultados obtenidos de la segmentación de la indumentaria. En dicha tabla puede apreciarse un mayor acierto de la piel. La indumentaria, ofrece resultados notables, sin embargo, accesorios etiquetados de forma independiente como las gafas de sol (que podían portarlas en el cuello de la camisa) o la confusión entre el cabello largo que cae sobre la camisa/camiseta hacen que la tasa de fallos sea ligeramente superior a la ofrecida por la piel.

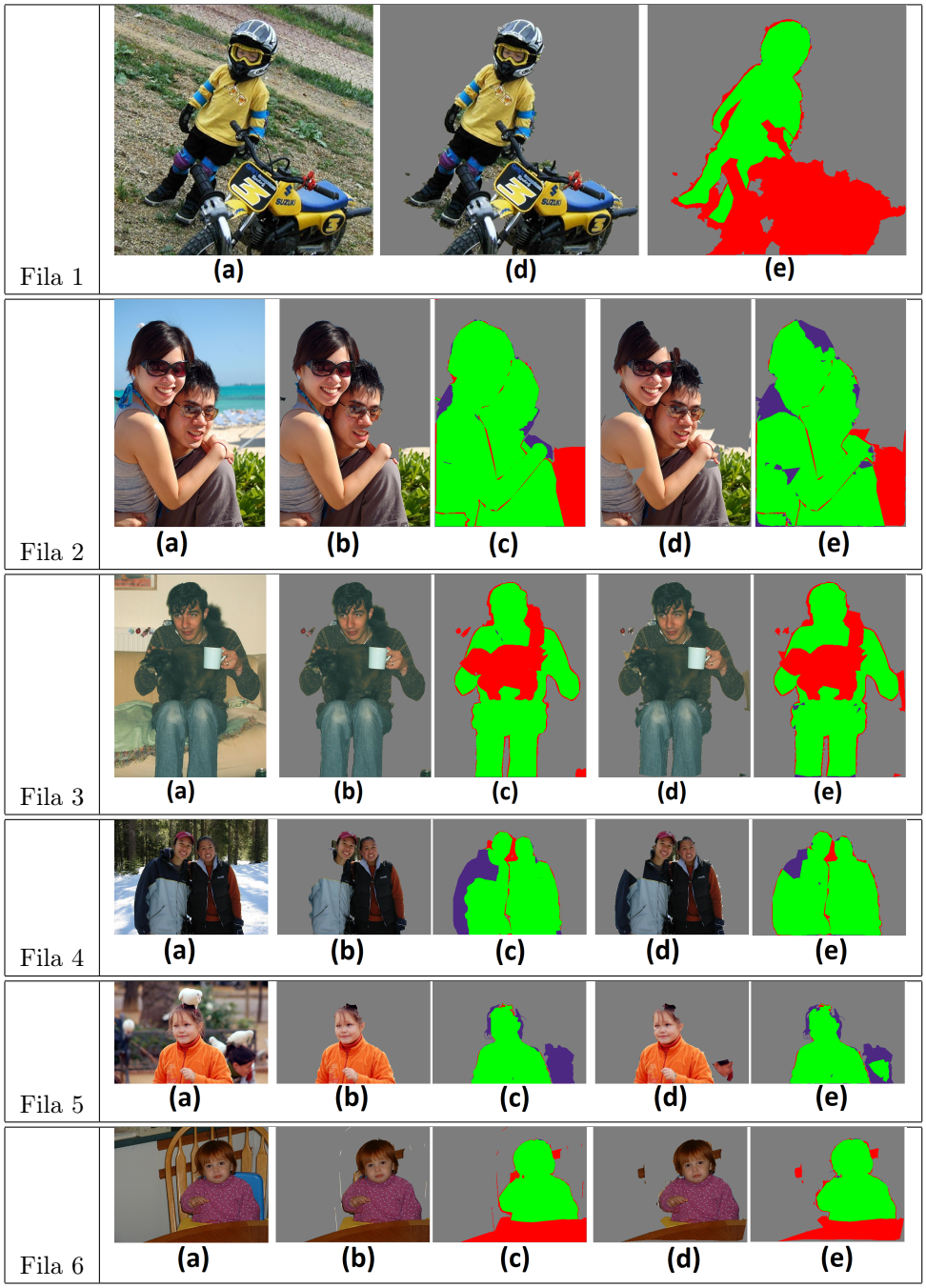


Tabla 4.5: Comparación entre la aplicación del GrabCut clásico que procesa píxeles (b) y la nueva propuesta que procesa tríxeles (d) sobre un conjunto de imágenes (a) de la base de datos PASCAL. Además, es apreciable la comparación píxel a píxel entre las imágenes segmentadas para GrabCut (c) y TriToM (e) con las propias imágenes segmentadas que ofrece la propia base de datos. En color verde se marcan los píxeles que han acertado para la selección del objeto de interés, en color morado aquellos píxeles que no fueron marcados pero debieron haber sido marcados, en color rojo los píxeles que fueron marcados pero no debieron haber sido marcados y, por último, en color gris las regiones que forman parte del fondo y se acertó marcándolas como fondo.

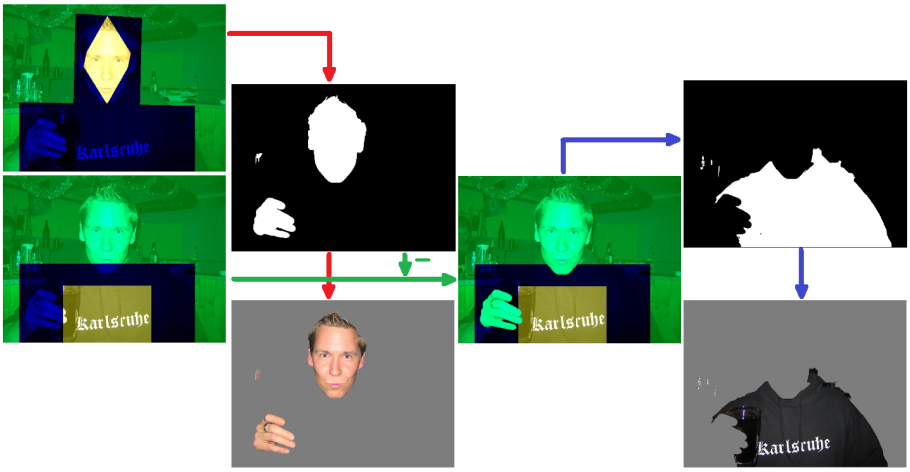


Figura 4.16: Retroalimentación de la máscara de la piel para la generación del trimap de la ropa. La Figura se muestra cómo el trimap es definido para la extracción de la ropa. Las regiones marcadas como fondo (en verde) y las regiones marcadas como objeto de interés (en amarillo) contribuyen a la distribución de los tríxeles pertenecientes a la región pendiente de analizar (en azul) en estos dos grandes clusters durante la etapa de segmentación.

4.4.1. Indice de calidad (Qindex)

El análisis de las técnicas de segmentación descritas en las secciones 4.1.1 y 4.1.2, requiere la evaluación de la robustez de los algoritmos de segmentación. Si bien en la tabla 4.3 se ofrece una comparación entre las tasas de fallos obtenidas por parte de la versión original del GrabCut y de TriToM, no se ha determinado la segmentación propuesta por TriToM. Por este motivo se introduce un índice de calidad que, a través de una heurística aplicada al modelo geométrico adaptativo definido en la sección 4.3.1, ofrece una estimación de lo precisa que puede haber sido la segmentación resultante. De esta forma, en la Figura 4.16 puede observarse cómo se hacen uso de regiones aproximadas para la detección de componentes. Estas regiones se calculan a través del modelo geométrico previamente comentado. En dicha Figura 4.16 puede apreciarse, en la imagen para la segmentación de la piel en el rostro, una máscara de color amarillo. Esta máscara está compuesta por píxeles de los que se tiene la certeza que pertenecen a la piel (con valores 1 o 0 dependiendo de si pertenece, o no, a la piel). En ocasiones sucede que, en la segunda fase del proceso de segmentación (la fase de segmentación de ropa), se localiza ropa en píxeles donde está ubicada la máscara de piel anteriormente comentada (falsos positivos). La estrategia se basa en analizar, para ambas salidas, esta región de interés (de $m \times n$ píxeles) de la piel:

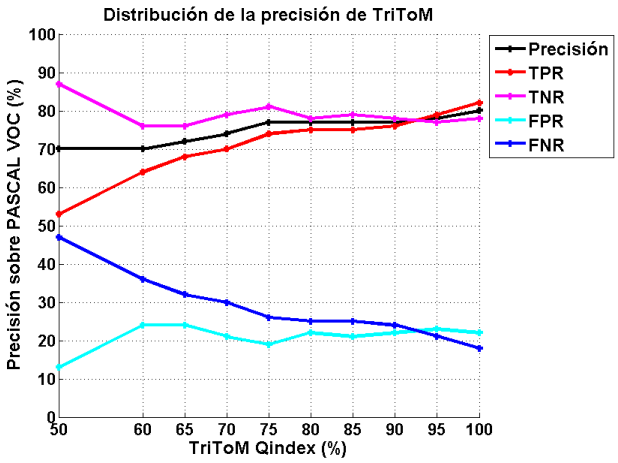


Figura 4.17: Distribución de la precisión según el Qindex.

$$Qindex = \sum_{\substack{j=1 \\ j\in\rho}}^{m} \sum_{\substack{i=1 \\ i\in\rho}}^{n} \frac{skinmask(i,j) - clothesmask(i,j)}{l} \tag{4.11}$$

Donde se recorren los l píxeles pertenecientes a la región de la piel marcada como de interés (ρ) y se calcula la diferencia entre ambas respuestas obteniendo el índice de calidad ($Qindex$). Cuanto menor sea el resultado, significaría que se han detectado mayor número de píxeles de la región de piel por parte de la ropa que por parte de la propia piel, en cuyo caso el valor resultante del $Qindex$ sería 0. Redistribuyendo las imágenes según el factor Qindex se puede volver a analizar el rendimiento del algoritmo de segmentación TriToM con las imágenes segmentadas de la base de datos VOC PASCAL [77] pero, en esta ocasión, distribuidas según el valor del $Qindex$.

Los resultados de la Figura 4.17 sugieren que un alto Qindex tiende a garantizar un mejor resultado del proceso de segmentación y, considerando únicamente las imágenes con un Qindex del 100 % se obtiene una tasa de fallos muy cercana a la del GrabCut original (véase tabla 4.3). Por tanto, se trata de un método eficaz y simple para evaluar la propia calidad de la segmentación.

4.4.2. Otro caso de estudio: Supertríxeles

Antes de combinar el algoritmo de maximización de flujo con los tríxeles, se investigaron diversas técnicas para agrupar los tríxeles. Se estudió desde la aplicación de modelos basados en Mahalanobis hasta diversos enfoques que aprovechaban las singularidades geométricas de los triángulos. Sin embargo, merece mención especial la idea basada en la posibilidad de usar los tríxeles de forma

Figura 4.18: Evolución de la segmentación de tríxeles basada en la aplicación de la adaptación del algoritmo propuesto por Achanta para el procesado de tríxeles [2].

similar a los superpíxeles [2]. En su implementación, Achanta divide la imagen original en un conjunto de superpíxeles, calculando para cada uno de ellos un kernel asociado a la información contenida en los superpíxeles. Posteriormente se evalúan todos los píxeles pertenecientes de un superpíxel con el kernel del propio superpíxel y con los kernels de los superpíxeles vecinos. Si dicho píxel tiene una menor distancia con un superpíxel vecino entonces el superpíxel original pierde a dicho píxel, que pasa a formar parte del superpíxel vecino más próximo. Por último se recalculan los kernels y se sigue iterando hasta alcanzar una convergencia como la ofrecida por la Figura 4.1. Una idea similar se implementó para TriToM; cada tríxel genera su propio kernel basándose en la información cromática contenida en su interior y se evalúan los píxeles en su interior entre el tríxel central y sus dos o tres vecinos. Al final se obtiene una imagen parecida a la de la Figura 4.19 en la que se puede apreciar cómo se dividen los píxeles de un tríxel en función de la distancia a su propio kernel y el kernel de sus vecinos.

Como puede apreciarse en la Figura 4.18, esta idea no funcionó por dos motivos; en la versión de Achanta los superpíxeles parten de una estructura inicial en la que todos tienen el mismo tamaño, mientras que en la versión de tríxeles es posible que, desde el primer momento, existan tríxeles muy pequeños (de 2 o 3 píxeles) que, debido a su tamaño, tienen un kernel tan fuerte que son incapaz de asociarse con otros tríxeles. Además, Achanta permite que los superpíxeles pierdan su forma original, deformándose a medida que pierden píxeles. Para la metodología propuesta, se permitió asociar tríxeles pero no separarlos de sus píxeles individuales. Si un tríxel $t0$ tiene un 40 % de sus píxeles que tienen menor distancia con el tríxel $t1$, un 30 % con el tríxel $t2$, un 5 % con el tríxel $t3$ y un 15 % con menor distancia a su propio kernel, entonces $t0$ pasa a formar un cluster con $t1$, con independencia del resto de distribuciones (véase Figura 4.19); no se permite romper el tríxel. El motivo de esta regla es que si se maneja una imagen con un número determinado de tríxeles, se debe manejar la estructura con el fin de manipularla de forma sencilla sin necesidad de volver a crecer en el número de tríxeles. La segmentación no debe determinar el número

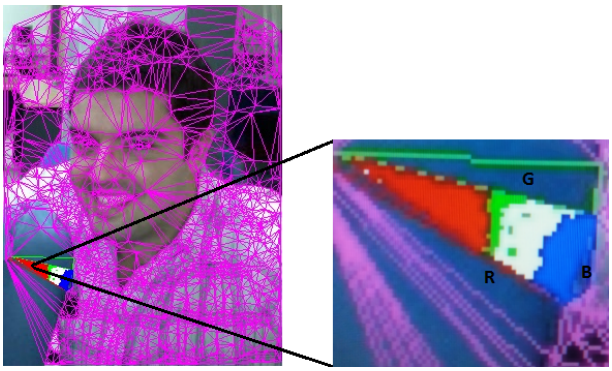


Figura 4.19: En esta imagen puede apreciarse la distribución de píxeles de un tríxel en función de los vecinos que le rodean. En color blanco se marcan aquellos píxeles que tienen menor distancia cromática con el propio kernel del tríxel analizado. Por su parte el resto de distribuciones responden a los siguientes criterios; color rojo aquellos que tienen mayor similitud con el tríxel inferior, en color azul aquellos que la tienen con el tríxel de la derecha y en verde con el tríxel superior.

de elementos atómicos a manejar. Por estos motivos la técnica propuesta por Achanta no ha sido aplicable a la solución de tríxeles propuesta.

4.5. Evaluación del rendimiento de extracción de ropa y aplicaciones

La biometría analiza diversos rasgos biométricos que permiten identificar a las personas. La voz o el rostro son rasgos clásicos a la hora de llevar a cabo esta tarea de identificación. Sin embargo, la disponibilidad de dichos rasgos no resulta siempre evidente. De hecho, aplicar técnicas para reconocer estos rasgos puede suponer una estrategia invasiva para el individuo; por ejemplo, solicitar al sujeto que muestre su rostro para analizarlo, o bien, que pronuncie unas palabras para analizar su patrón de voz. En ambos casos, se hace necesaria la cooperación de las personas a las que se pretende identificar. Por el contrario, en el caso de optar por una aproximación no invasiva, es posible encontrarse con otro tipo de inconvenientes. Por ejemplo, que los sistemas de reconocimiento facial pudieran no acertar en ocasiones debido a múltiples factores como un mal enfoque del rostro, que la pose facial no es la apropiada o, incluso, cuando la resolución del dispositivo de adquisición no es la adecuada.

La información de la ropa puede ayudar a superar este tipo de inconvenientes. La ropa puede contribuir, no sólo a identificar a las personas, sino también a apoyar la identificación de otros aspectos relevantes del individuo; el sexo, el tipo de evento al que se pretende acudir, etc. Sin embargo, para poder obtener un

Figura 4.20: Generación de la malla de tríxeles, de acuerdo con lo explicado en la sección 4.2, antes de comenzar el proceso de segmentación de ropa. En la imagen de la izquierda puede apreciarse la imagen original. Por el contrario la imagen central muestra el conjunto de puntos detectados por el algoritmo DMUM (véase sección 4.2.1). Por último, el proceso de triangulación se completa en la imagen de la derecha a partir de los puntos previamente calculados.

conjunto de descriptores representativos de la indumentaria antes es necesario segmentar las regiones que son relevantes para dicho análisis [137]. Muchos enfoques que abordan este problema lo hacen orientando sus esfuerzos a la mejora de las propias técnicas empleadas con el fin de obtener descriptores de mejor calidad y, para ello, hacen uso de métodos de segmentación poco elaborados. Por ejemplo, si se procesan fotogramas de vídeos procedentes de cámaras de vigilancia, se hace una sustracción de fondo, o bien, se hace una diferencia de imágenes con el objetivo de limpiar la imagen y segmentar personas [97, 237]. Este conjunto de técnicas no es aplicable en el caso de las imágenes estáticas. En este caso, el método más popular es aprovechar la detección facial en primer lugar, para luego ajustar una región cercana a la posición del rostro que permita delimitar la zona en la que se pudiera ubicar cuerpo del sujeto de forma aproximada [7]. Este tipo de técnicas han sido refinadas con el paso del tiempo y han surgido enfoques más novedosos. Por ejemplo, técnicas que ajustan la posición de la región de interés para el cuerpo de forma que se minimiza el tamaño de esta región con el objetivo de reducir al máximo el riesgo de solape con otros sujetos cercanos al sujeto analizado [195]. Sivic propuso una alternativa que trata de ajustar diferentes regiones de interés a partes concretas del individuo como el pelo, la cara y algunas zonas del cuerpo, obteniendo de esta manera, un modelo estructural del sujeto [193]. Por último, Gallagher y Chen proponen una presegmentación de la imagen a través del uso de superpíxeles para, posteriormente, hacer uso de los superpíxeles en la zona facial con el fin de lograr una segmentación más precisa de la región que pertenece a la ropa del individuo [96].

En esta sección se pretende hacer uso de la técnica de tríxeles desarrollada con el objetivo de poder segmentar la ropa en entornos no controlados. Además, en una segunda etapa, se ofrece una aplicación práctica para el reconocimiento del sexo de las personas haciendo uso de la información extraída en el paso previo.

4.5.1. Configuración del experimento

La propuesta para el reconocimiento del sexo se compone, esencialmente, de un método que se aplica en tres fases. En estas fases se particulariza el proceso descrito en las secciones previas, a fin de adaptarlo al problema en cuestión. En primer lugar, la creación de un modelo geométrico en base la información del rostro proporcionada por ENCARA2 siguiendo el procedimiento explicado en la sección 4.3.1. En segundo lugar, se aplican un conjunto de extractores de características sobre las máscaras previamente obtenidas para calcular los descriptores en el espacio de representación apropiado. Por último, estos vectores de características son analizados por un clasificador SVM con el fin de resolver un problema de clasificación binaria; hombre o mujer (véase la sección 3.2.1.3 para obtener una descripción más precisa de la clasificación binaria por parte de las SVM [30]).

4.5.2. Modelo geométrico adoptado

En los últimos años, el análisis de la pose humana ha tenido una mayor presencia en los trabajos de investigación relacionados con la temática [235, 191]. Realmente, si se conoce la pose de una persona, resulta mucho más sencillo poder delimitar las regiones donde pudiera estar ubicada su ropa; por ejemplo, los brazos no siempre tienen la misma pose y suelen ser una región determinante para conocer los límites del cuerpo humano. Además, en muchas ocasiones, la parte inferior del cuerpo de un individuo está oculta o no es del todo visible en algunas imágenes. Por ello, estimar la pose del cuerpo de una persona, en dos dimensiones, supone un auténtico reto a la hora de definir la región de interés (el trimap) dentro de la escena para que los tríxeles la computen y, de esta forma, extraer la segmentación.

En múltiples situaciones, el tono de la piel y el color de la ropa no están cerca en el espacio de representación. Esto puede ocasionar que los modelos probabilísticos ocupen un espacio bastante amplio dentro de este espacio de representación y, en el peor de los casos, generar la aparición de múltiples falsos positivos a la hora de detectar la región de interés en torno al sujeto. Con el fin de evitarlos, se propone un proceso de segmentación a través del uso de tríxeles en dos etapas. En la primera etapa se segmenta la piel del individuo. Los resultados cosechados en la primera fase de segmentación sirven para realimentar el proceso en la localización de la ropa en la segunda etapa. Así, sea una imagen de entrada, un detector facial ENCARA2 [37] localiza la región facial y proporciona el conjunto de datos necesarios para estimar la localización del trimap de acuerdo a lo explicado en la sección 4.3.1. Después de generar el TriToM de la imagen de entrada (véase la Figura 4.20) y, de acuerdo a lo especificado en esta primera fase del trimap en la que se delimita la región de la piel, se generan los modelos probabilísticos para rellenar la información del grafo. Por último, se ejecuta el algoritmo de maximización de flujo para extraer la máscara de los tríxeles que pertenecen a la región de piel. En la segunda fase, esta máscara de piel se utiliza para sustraer del segundo trimap, las regiones pertenecientes a

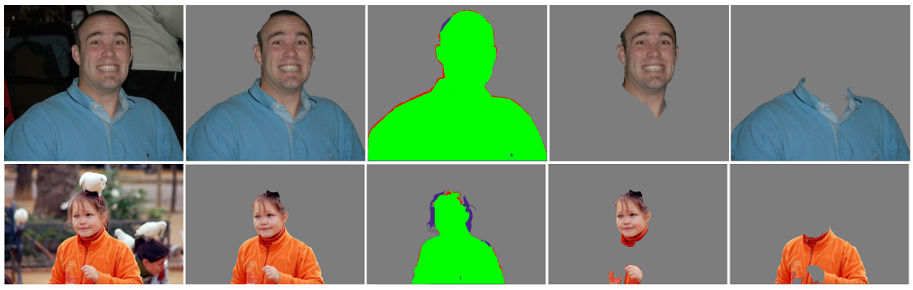


Figura 4.21: Imágenes piel y ropa segmentadas con TriToM. En esta secuencia se puede apreciar, de derecha a izquierda, las siguientes imágenes; la imagen original, la imagen segmentada, la comparación con las segmentadas de PASCAL (color verde y color gris señalan acierto con respecto a figura y fondo respectivamente), la imagen con segmentación de la piel y, por último, la imagen con la ropa segmentada.

Figura 4.22: Problemática a la hora de seleccionar el trimap.

la piel y, de esta forma, segmentar la región de ropa haciendo uso de un modelo más preciso (véase el proceso en la Figura 4.16). En la Figura 4.21 puede apreciarse el resultado de la segmentación de la piel y la ropa, comparando el resultado de dicha segmentación con la plantilla ofrecida por la base de datos de PASCAL [77].

Las posibles limitaciones del modelo geométrico son las mismas que expuso Chen en su trabajo acerca de el reconocimiento del sexo en base a la segmentación de la ropa. De hecho, este autor omite en su trabajo el tren inferior de los humanos porque afirma no poseer la certeza de la presencia de dicha región en todas las imágenes de la base de datos [39]. En la Figura 4.22 se puede observar el problema propuesto por Chen en función de las dimensiones del trimap calculado. Por ello, es importante considerar que, en el enfoque propuesto en esta sección, se aprecia una limitación en cuanto a la longitud del modelo geométrico

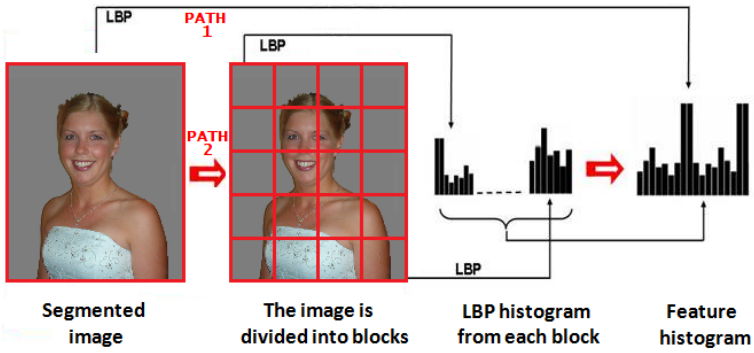


Figura 4.23: Aplicación de LBP para el reconocimiento del sexo. Se cosideran dos alternativas; considerar la imagen como un único conjunto o dividirla en bloques de 5×5 píxeles.

que define el trimap. En este caso se ha preferido por optar a segmentar el tren inferior con el fin de poder obtener una información valiosa de la indumentaria para futuros experimentos.

4.5.3. Representación de los datos

En cuanto a los descriptores de las imágenes resultantes, se ha optado por el uso de descriptores locales; tanto los patrones binarios locales (LBP) [158] como los histogramas de gradientes orientados (HOG) [44].

Los patrones binarios locales, descritos en la sección 3.2.1.1, han resultado francamente útiles en el análisis facial a través de la concatenación de histogramas de una imagen dividida en bloques (véase la Figura 4.23). Este enfoque, propuesto por Ahonen, divide el rostro en regiones a las que se les aplica el operador LBP para, posteriormente, concatenar los histogramas correspondientes a cada una de las regiones. Obteniendo, de esta forma, un único histograma representativo de toda la imagen analizada que no pierde información de localización.

Por el otro lado, los HOG, explicados en la sección 2.4.1.2 ya encierran un histograma dentro del descriptor de forma implícita. Tal y como puede apreciarse en la Figura 4.24, las orientaciones del gradiente así como sus magnitudes forman parte de los histogramas construidos a partir de la información proporcionada por la imagen. Posteriormente, la imagen es representada a través de una concatenación de los histogramas obtenidos a partir de cada celda. Además, la influencia lumínica es reducida normalizando la información de cada celda con respecto a las celdas vecinas y, formando de esta manera, bloques de celdas. El tamaño de cada celda es de 8×8 píxeles, mientras que los bloques están formados por 2×2 celdas. Por último, el número de bins considerado para los histogramas es de 9 unidades.

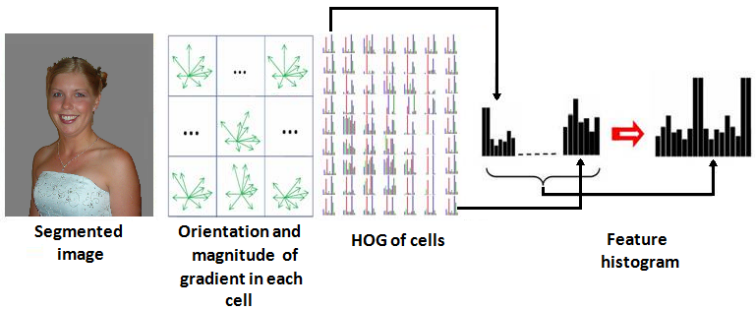


Figura 4.24: Aplicación de HOG para el reconocimiento del sexo.

4.5.4. Resultados

Para la clasificación del sexo se ha hecho uso de la base de datos denominada *ClothesDB* propuesta por Chen [39] y que ya ha sido previamente evaluada para resolver el problema del reconocimiento del sexo combinando la información del rostro y de la ropa. Aunque en el artículo de Chen no se ofrecen datos específicos acerca del tipo de experimento realizado, el enfoque que se ofrece en esta sección repite 10 veces un experimento de validación cruzada con 5 conjuntos.

De acuerdo con la respresentación de los datos anteriormente comentada, se hace uso de diversos clasificadores. Por un lado, los clasificadores del rostro (LBP y HOG) que hacen uso de un patrón normalizado de 59×65 píxeles (en la Figura 3.26 se expone un ejemplo de dicha normalización). En el HOG se ha usado una rejilla de 8×8 píxeles. Por otro lado, la cabeza y los hombros (CH) también son utilizados para generar un clasificador a partir del descriptor HOG y, en esta ocasión, las imágenes tienen unas dimensiones de 64×64 píxeles (véase Figura 4.24). Por último, se incorpora la segmentación de la ropa para generar un último clasificador. En este caso, se ha estudiado el uso de la máscara en blanco y negro generada por el proceso de segmentación de los tríxels, así como las imágenes segmentadas en escala de grises generadas por dicha técnica. Para el primer caso, se ha aplicado el descriptor LBP sobre toda la imagen (sin dividirla en celdas) de acuerdo a lo expuesto en la Figura 4.23. Para el segundo caso, se ha dividido la imagen segmentada en escala de grises en una rejilla de 5×5.

Con el fin de poder comparar los resultados, se incluyen los datos proporcionados por Chen [39] y también los obtenidos, más recientemente, por un enfoque que combina la información facial con la información contextual que rodea a la cabeza [36].

Como puede apreciarse en la tabla 4.6, los descriptores obtenidos únicamente a través de la ropa ofrecen una tasa de acierto del 83.9 %. Estos datos mejoran los proporcionados por Chen haciendo uso exclusivamente de la ropa [39]. Resulta interesante matizar que no basta sólo con delimitar, a través del proceso de segmentación, la región de la ropa. Obtener un análisis más preciso sobre el reconocimiento del sexo requiere de un análisis más exhaustivo sobre la infor-

Tabla 4.6: Reconocimiento de sexo en la base de datos ClothesDB.

Referencia	Datos	Precisión
[39]	Rostro	71.5 %
	Ropa	81.0 %
	Rostro+Ropa	84.9 %
[36]	Rostro HOG	78.3 %
	Rostro LBP	82.5 %
	CH HOG	75.9 %
Propuesta [85]	Ropa (LBP)	67.3 %
	Ropa (HOG)	83.9 %
	Ropa + CH	84.5 %
	Ropa + CH + Rostro (LBP + HOG)	**86.8 %**

mación semántica contenida en la propia indumentaria. Por ejemplo, en la tabla 4.7 pueden apreciarse imágenes que hacen referencia a esta cuestión; mientras en la primera imagen (fila 1) resulta evidente que la ropa ofrece una máscara en la que se puede afirmar el sexo de la persona que la lleva puesta, en la segunda imagen (fila 2) se puede apreciar cierta ambigüedad si sólo se considera el marco de la indumentaria. Por ello, Chen incluye información relativa a atributos semánticos de la imagen que incorporan con mayor exactitud el análisis de la pose de forma manual antes de extraer los atributos.

Sin embargo, la mejoría experimentada cuando se incluye información relativa a la cabeza y los hombros (CH) es mínima, el acierto sube en torno a un 1 %. Este nuevo enfoque considera un conjunto de datos más amplio y relevante; por ejemplo, el contorno del pelo del sujeto analizado o bien si tiene los hombros del mismo color que la piel (hombros descubiertos). De esta forma, la contribución de estos parámetros resulta útil para el reconocimiento del sexo. Aunque la mejora no sea notable, resulta esencial relativizar los datos obtenidos a través de la aplicación de la técnica a la cabeza y los hombros de los sujetos. En otras palabras, si el ámbito de aplicación fuera un escenario en el que las mujeres llevaran vestidos, la mejoría podría verse incrementada debido a que podrían ser fácilmente identificables a través de los hombros descubiertos.

Por último, la mejor tasa obtenida es la que se corresponde con la combinación de la ropa, la cabeza, los hombros y el rostro (86.8 %). En esta ocasión, las tasas obtenidas mejoran ampliamente las ofrecidas anteriormente y mejoran los datos obtenidos por Castrillón [36] y Chen [39]. Características faciales como la presencia de vello o la presencia de maquillaje ayudan a determinar con mayor exactitud el sexo de la persona.

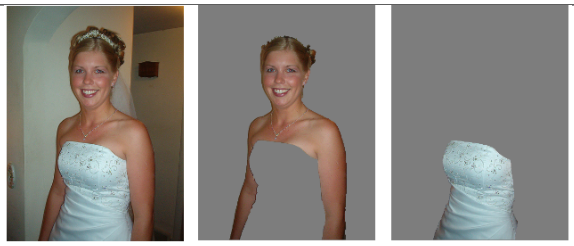
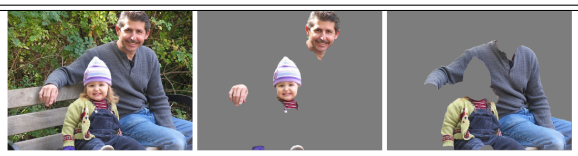
Fila 1	
Fila 2	
Fila 3	
Fila 4	

Tabla 4.7: Segmentación de piel y ropa para el reconocimiento del sexo.

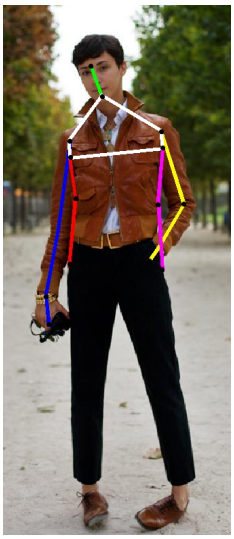


Figura 4.25: Etiquetado manual de la colección de datos ClothesDB.

4.5.5. Detección de componentes específicas en la indumentaria

La detección de componentes pretende detectar elementos específicos que pueden presentarse en la indumentaria de una persona en un momento determinado. Con esta prueba se aumenta la complejidad pues se ofrece una comparativa del algoritmo de tríxeles haciendo uso de características particulares de la vestimenta de los sujetos.

Para ello, a partir del conjunto de datos proporcionado por ClothesDB [39], se ha etiquetado de forma manual una serie de 14 puntos (véase Figura 4.25) con el objetivo de comparar con los datos extraídos a través de los tríxeles. A partir de los puntos del etiquetado manual es posible conocer los límites del tronco superior de una persona. Además, para el estudio presentado en esta ocasión, se ha generado un pentágono en la zona del cuello a fin de poder obtener los resultados para dos tipos de experimentos; por un lado determinar si la forma del cuello de la vestimenta permite dilucidar si se trata de una camisa y, por otro lado, si la persona está llevando una corbata.

En el caso de las pruebas haciendo uso de la información extraída por los tríxeles, se consideró el rectángulo que circunscribe la zona marcada como ropa con los tríxeles y luego se normalizó la imagen a 32×24 píxeles para el caso de la corbata y a 32×32 para el experimento del cuello.

Al igual que en los experimentos anteriores, se ha hecho uso de LBP y HOG como enfoques básicos para extraer las características de la información enmarcada en la zona de interés. En el caso de LBP no se normaliza sino que se utilizó toda la máscara ya que sólo se calcula el LBP para los puntos que pertenecen a la máscara.

	10-CV		Train/Test	
	SVM Lineal	Random Forest	SVM Lineal	Random Forest
HOG	66.2 % (78.4 %)	64.1 % (74.3 %)	66.6 % (78.7 %)	63.5 % (71.9 %)
LBP	63.3 % (66.9 %)	59.0 % (60.0 %)	70.6 % (69.7 %)	56.9 % (63.6 %)

Tabla 4.8: Comparativa para la detección de cuello de camisa. Entre paréntesis la tasa de aciertos del etiquetado manual [137].

	10-CV		Train/Test	
	SVM Lineal	Random Forest	SVM Lineal	Random Forest
HOG	76.7 % (82.4 %)	70.5 % (80.4 %)	73.3 % (84.0 %)	79.0 % (86.3 %)
LBP	62.7 % (72.6 %)	56.2 % (65.4 %)	59.7 % (76.0 %)	65.8 % (73.3 %)

Tabla 4.9: Comparativa para la detección de corbatas. Entre paréntesis la tasa de aciertos del etiquetado manual [137].

Debido a la problemática que supone no tener el mismo número de muestras positivas y de muestras negativas, los experimentos se han realizado de la siguiente forma; en primer lugar, se ha extraído el 80 % de muestras positivas para combinarlas con el mismo número de muestras negativas. Por ejemplo, si existen 288 imágenes en las que aparecen sujetos con corbatas y 242 sujetos sin corbata, entonces se escogen 230 sujetos con corbata y 230 sujetos sin corbata. En el primer experimento, se ha realizado una validación cruzada con 10 conjuntos considerando el mismo número de muestras positivas y de muestras negativas. En el segundo experimento, se ha hecho uso de la extracción realizada anteriormente para entrenar el clasificador y, posteriormente, se ha usado el resto de imágenes (en el ejemplo considerado anteriormente, serían las 58 de muestras positivas y el resto de muestras negativas) para hacer las pruebas. En las tablas 4.8 y 4.9 se puede identificar a esta segunda prueba como *Train/Test*, mientras que la primera prueba es denominada como *10-CV*.

Como puede apreciarse en ambas tablas, los resultados obtenidos a partir del etiquetado manual son mejores en casi todas las pruebas realizadas. Analizar el cuello, o la corbata, exige un nivel de detalle mucho mayor; ser capaz de segmentar de forma muy precisa los límites de la región de interés. En la detección del cuello, el LBP ofrece cifras muy similares a las ofrecidas en el etiquetado manual, lo que indica que los bordes del proceso de segmentación no son imprecisos. Sin embargo, en el caso de la detección de corbatas tanto el LBP como el HOG ofrecen unas pérdidas de precisión en torno al 6 % y el 10 %. Esto sugiere que es necesario un análisis más preciso de la zona de interés cuando se pretende localizar los elementos de la indumentaria considerados. Por tanto, resulta de vital importancia seguir estudiando una fórmula adaptativa a la pose de los sujetos con el fin de poder mejorar la calidad del proceso de segmentación del cuello y, de esta forma, acercar los resultados ofrecidos por

los tríxeles al etiquetado manual. Además, la aparición de la corbata sugiere que, en la imagen de tríxeles, aparecen un patrón bien definido justo en la zona del cuello (generalmente, todas las corbatas tienen la misma forma), por lo que entender los patrones de tríxeles formados y extrapolar este entendimiento a los descriptores podrá ayudar a mejorar estas cifras.

Capítulo 5

Conclusiones

En esta tesis doctoral se han analizado e implementado técnicas orientadas a capacitar a las máquinas para describir mejor a las personas con las que interactúan. Se han estudiado desde características personales e intransferibles del ser humano, hasta técnicas que facilitan la extracción de elementos contextuales que definen distintos rasgos de estas personas. En cualquiera de los casos, el objetivo principal de este trabajo es la aplicación en sistemas de tiempo real que permitan su uso en entornos sin restricciones para la obtención de información relevante acerca de personas.

En la primera parte del trabajo se ha tratado el tema de las emociones humanas y se ha analizado, a partir del rostro, la evolución temporal de la formación de las expresiones, aprovechando dicha información para determinar el estado facial del sujeto con respecto a la emoción estudiada. La sonrisa fue seleccionada como emoción de estudio debido a su indudable identificación con independencia del usuario o el entorno cultural. Los resultados obtenidos, tanto haciendo uso de descriptores habituales así como del enfoque desarrollado para la ocasión, han demostrado la aplicabilidad de dichas técnicas en un escenario poco restrictivo.

En la segunda parte de esta tesis se ha desarrollado una técnica de segmentación novedosa que combina dos técnicas existentes en la literatura con la ayuda del algoritmo DMUM [8]. Los resultados alcanzados no mejoran la tasa de aciertos por el proceso de segmentación ofrecida por el GrabCut, siendo ligeramente inferior. Sin embargo, su rendimiento es muy superior en términos de tiempo a su predecesor, lo cual se ajusta a la necesidad del desarrollo de un sistema en tiempo real para el análisis de características en su ámbito de aplicación. Para el desarrollo de esta parte del trabajo, se ha combinado la información de alto nivel (la detección facial) con el procesamiento de bajo nivel (a través del uso de los tríxeles). Resulta importante destacar que el objetivo que subyace a las distintas técnicas desarrolladas en esta tesis doctoral no es otro que reducir el canal comunicativo entre máquinas y humanos.

5.1. Principales aportaciones

A raíz de todo lo comentado en este documento, se pueden enumerar las siguientes aportaciones por parte de esta tesis doctoral:

1. En el segundo capítulo se ha hecho una extensa revisión del estado del arte de aspectos comunicativos existentes entre máquinas y humanos a través de los siguientes puntos; las bases de datos existentes en la literatura para entrenar máquinas que detecten personas, interpreten las emociones humanas o analicen la información contextual que rodea a las personas, el reconocimiento de las expresiones faciales por parte de estas máquinas y, por último, la extracción de componentes contextuales para la clasificación de un atributo demográfico, como lo es el sexo de los humanos. Dentro de dicho estudio, llama poderosamente la atención los siguientes puntos que, a su vez, proporcionan los retos más importantes que se abordan en este trabajo:

    - La mayoría de las bases de datos con expresiones faciales tienen su génesis en la simulación de las emociones frente a una cámara por parte de actores. Algunas de estos conjuntos de imágenes hacen uso de profesionales, mientras otros no. Capturar las emociones de esta forma permite controlar y acotar la prueba de forma más eficiente a través de distintas tomas. El principal aspecto negativo es que se va a entrenar una máquina para reconocer emociones con emociones que son simuladas y que, como se apuntaba en dicho capítulo, un símil podría ser el de entrenar una máquina para reconocer billetes de euro con billetes falsos. Además, también se han presentado bases de datos que ofrecen una perspectiva contextual de las personas que han sido capturadas.
    - En el reconocimiento de expresiones faciales se puede considerar el uso del rostro al completo, o bien, de regiones de interés a la hora de capturar una determinada expresión. En el caso de delimitar la zona de estudio, el análisis proporcionado por el FACS (*"Facial Action Coding System"*) permite entender qué partes son las más relevantes para el estudio de una determinada emoción.
    - En muchos trabajos, se propone el uso de una segmentación manual y/o de bases de datos normalizadas para la extracción de características contextuales, más allá del rostro, que permitan la descripción de una persona. Este reconocimiento de características demográficas se puede realizar en función del rostro de una persona, pero también se ha ofrecido una revisión al reconocimiento de características demográficas considerando elementos de contexto (pelo y ropa).

2. Una novedosa base de datos con expresiones faciales capturadas de forma inducida a través de la proyección de contenido multimedia. La base de datos Vedana incluye una serie de particularidades que hace atractivo su uso

en varias líneas de investigación acerca del reconocimiento de expresiones faciales:

- La captura de expresiones faciales se hace desde tres cámaras con distintas configuraciones. Una webcam, una cámara HD y una cámara con una alta frecuencia de captura de imágenes para la captura de microexpresiones. No se capturan instantáneas, sino que se ofrece la secuencia de vídeo completa.
- Se propone un esquema físico y un conjunto de contenidos adecuados para poder capturar expresiones faciales a través de un proceso de inducción.
- Los usuarios aparecen en dos sesiones con distinta indumentaria y accesorios para su análisis.
- La base de datos incorpora 1524 fotogramas de la cámara frontal que han sido anotados a mano, diferenciando las zonas de la piel, el pelo, la ropa, las gafas de sol y el fondo.

3. Un estudio exhaustivo acerca del reconocimiento de sonrisas haciendo uso de descriptores novedosos en la literatura, considerando dos tipos de clasificadores distintos así como diversas regiones de la estructura facial para ayudar a determinar cuáles son más relevantes a la hora de identificar dicha expresión. Este estudio ha dado lugar a la combinación de clasificadores para mejorar los datos proporcionados por cada uno de ellos individualmente. Además, las mejores técnicas se seleccionaron para el estudio dinámico de la expresiones faciales en el que, a través de la entrada de secuencias de vídeo, se estudia una clasificación en dos dimensiones; a través una combinación pesada de los enfoques individuales que componen el nuevo clasificador y considerando la respuesta de dicho clasificador en fotogramas anteriores para la clasificación del fotograma actual.

4. Se ha desarrollado descriptor novedoso, denominado *firma digital*, basado en el concepto de autosimilitud y que es capaz de capturar la estructura general de una imagen. A través de la división de una región de celdas independientes y de la comparación de la subregión de interés con el resto de las celdas se obtiene un interesante descriptor de cara al análisis de la estructura captada. Los resultados en el análisis de la sonrisa han permitido obtener tasas mejores que las obtenidas por cada uno de los enfoques individuales del experimento anteriormente comentado.

5. Una nueva técnica para resolver el problema de extracción de una figura en un fondo complejo. En esta propuesta se combinan dos enfoques muy utilizados en la literatura de los últimos años; los superpíxeles [151] y el GrabCut [176]. De esta forma se trabaja con una malla de tríxeles manipulable que ofrece el siguiente conjunto de características distintivas:

   - La ubicación de los tríxeles a lo largo de la imagen no se realiza de forma geométrica ni aleatoria, sino que se parte de una configuración

dependiendo de las características de la imagen a través del algoritmo DMUM [8]. De esta forma, para una determinada imagen, la topología de tríxeles es única y determinista gracias al uso del DMUM y de una topología basada en la triangulación de Delaunay para la distribución de estos tríxeles.

- La unidad atómica de procesado es el tríxel. Los tamaños de los tríxeles no son homogéneos pero, cada uno de ellos, agrupa regiones de píxeles que comparten características similares.
- El proceso de segmentación respeta la estructura de malla propuesta; los tríxeles no son divididos ni redistribuidos, sólo se agrupan en clusters. Esto permite obtener una malla estable y manipulable a lo largo de todo el proceso de segmentación.
- Cada tríxel puede ser manipulado a nivel local, sin embargo, el algoritmo de segmentación aprovecha la topología de la malla y lo hace en forma de grafo para realizar los cortes de la misma forma que se hace en el GrabCut.
- La técnica propuesta ha demostrado tener un rendimiento notable en términos de velocidad, eficiencia de memoria e impacto en el rendimiento durante el proceso de segmentación. La posible limitación en los resultados ha venido más determinada por desajustes en la selección del *trimap* de acuerdo con la pose del sujeto así como discrepancias semánticas sobre el objeto segmentado debidas a una no estandarización de los criterios de búsqueda sobre qué se está buscando.
- El índice de calidad permite evaluar la robustez del sistema de segmentación propuesto. Para ello, se aprovecha el modelo geométrico estimado en base a la información facial para analizar las regiones segmentadas de piel y de ropa.

6. Un estudio en el que se han extraído sujetos sobre un conjunto de escenas heterogéneas con el fin de poder analizar, con mayor detalle, los componentes no faciales del sujeto añadidos a la información del rostro. Para ello, se ha aplicado la técnica de los tríxeles, obteniendo una máscara de cada uno de los sujetos de la base de datos. Posteriormente, se han analizado distintos métodos para el reconocimiento aprovechando elementos contextuales así como faciales. De esta forma, este estudio ofrece:

   - Un trabajo sobre el reconocimiento del sexo de una persona aprovechando información facial y contextual, combinada o por separado.
   - Un análisis sobre la detección de componentes específicas en la indumentaria tales como corbatas o cuellos de camisas.

5.2. Ponencias y publicaciones

En esta sección se enumeran y describen, de forma breve, el conjunto de publicaciones que se han derivado del desarrollo de esta tesis doctoral. En primer lugar se muestran las publicaciones derivadas del estudio de las expresiones faciales expuestas en el capítulo 3:

- **VISAPP (2009) [89]:** Se describen los resultados obtenidos a raíz del estudio de distintos enfoques para la detección de sonrisas expuesto en la sección 3.2.1.4. Tal y como se describe en dicha sección, los distintos enfoques estudiados son comparados entre sí haciendo uso de distintas características faciales y considerando dos tipos de clasificadores; SVM y k-NN.
- **EUROCAST(2009) [88]:** Considerando los mejores resultados obtenidos del estudio previo, se hace una combinación de los mejores enfoques para la detección de sonrisas tal y como se describe en la sección 3.2.1.5.
- **LNCS (2009) [86]:** En el congreso Multicomm BioID se expuso la firma digital, descrita en la sección 3.2.1.6 para la detección de sonrisas.
- **INTECH (2011) [84]:** Una versión extendida de dicho enfoque se publicó en INTECH.
- **JRBP (2012) [87]:** La base de datos Vedana (véase sección 3.1.2) se presentó en las Jornadas de Reconocimiento Biométrico de Personas a fin de presentar a la comunidad científica una nueva alternativa para el reconocimiento biométrico de características faciales.
- **IJPRAI (2014):** En la actualidad, se encuentra en segunda revisión el trabajo descrito en la sección 3.2.2 acerca del reconocimiento dinámico de la sonrisa, en el que se aprovecha la coherencia temporal que ofrece el vídeo frente a la imagen.

En segundo lugar, se muestran las publicaciones relacionadas con la detección de sexo en las personas y con el contenido relacionado en el capítulo 4:

- **ICIP (2010) [35]:** Se presentó un trabajo acerca de la detección de sexo en base a la experiencia. Se trata de una aproximación inicial a la detección de esta característica biométrica a través del aprendizaje automático de la máquina en base a la retroalimentación con los resultados obtenidos.
- **ICIP (2014) [85]:** Se presenta la técnica de tríxeles descrita en las secciones 4.2 y 4.3, así como los resultados para la detección de sexo de la sección 4.5.4 que hace uso de dicha malla de tríxeles.
- **VAAM (2014) [137]:** En este trabajo se expone una comparativa de detección de componentes específicas en la indumentaria como la que se expone en la sección 4.5.5.

5.3. Trabajo futuro

Esta tesis abre nuevas perspectivas que se describen a continuación:

1. La base de datos Vedana puede ser mejorada desde múltiples puntos de vista:
   - El principal reto es ampliar las posibilidades de la base de datos introduciendo una mayor cantidad de metadatos acerca de las características de los usuarios así como añadir la codificación FACS a las expresiones de los sujetos capturados.
   - Conseguir articular un mecanismo para la distribución eficaz de los vídeos capturados con la cámara con una alta tasa de captura (210 fotogramas por segundo) a través de la red.
   - Aumentar la calidad y la aplicabilidad. Aunque la base de datos fue planteada como una colección destinada al reconocimiento de expresiones faciales, puede ser aplicable para tareas destinadas al reconocimiento facial. Sin embargo, el tamaño de la colección (63 sujetos) sugiere la necesidad de ampliarla para poder hacer, más apetecible, su aplicabilidad en una mayor variedad áreas.

2. En el análisis dinámico de la sonrisa se ha realizado de acuerdo al uso de ventanas de tiempo para estimar la clasificación de un fotograma en base a la evolución natural de la expresión facial. Se ha mostrado que existe un factor β que ofrece un margen de error durante la clasificación, y que este factor depende del tamaño de ventana. Se pueden estudiar alternativas para el análisis evolutivo de las expresiones faciales, como por ejemplo, los modelos ocultos de Markov (*"Hidden Markov Models (HMM)"*). Con estos modelos es posible lograr la evaluación de la probabilidad (o verosimilitud) de una secuencia de observaciones dado un modelo HMM específico.

3. En este documento se ha aplicado el algoritmo DMUM para el posicionamiento inicial de los vértices de los triángulos que forman los tríxeles. El estudio de alternativas para la ubicación de estos puntos (como por ejemplo, SIFT [138] o SURF [15]) puede dar lugar a un estudio más extenso que evite la formación de micro-tríxeles.

4. Profundizar en las características geométricas que ofrece la malla de tríxeles. A simple vista se pueden inferir propiedades que pueden ayudar, aún más, a simplificar el proceso de segmentación. Esto es posible a través de la aplicación de un paso previo a la ejecución del algoritmo de flujo máximo. Para ello, existen evidencias de la existencia de ramificaciones en los vértices de muchos de los tríxeles, es decir, a partir de un vértice salen un gran número de líneas que se amontonan alrededor de un determinado número de grados angulares. Por ejemplo, de un vértice salen seis aristas en un espacio que comprende un ángulo de 30^{o}. Resulta que, empíricamente, se ha demostrado que todas las aristas que comparten esta situación forman

parte de tríxeles que comparten la misma clase. El análisis y el diseño de un método de preprocesado que aproveche estas particularidades pueden simplificar el proceso de segmentación de forma notable.

5. La propuesta de tríxeles sólo considera el espacio de color RGB. Se considera necesario explorar alternativas con otros espacios de color. Por ello, después de generar la malla de tríxeles, se adaptaría la información contenida dentro de cada tríxel variando el espacio de color de su contenido.

6. En la sección 4.2.2 se ha mencionado que, una de las propiedades de la triangulación de Delaunay, radica en que insertar un nuevo punto dentro de una triangulación sólo afecta a aquellos triángulos cuyos círculos circunscritos afecten a dicho punto. De la misma forma, si se considera la malla de tríxeles, incorporar nuevos puntos sólo afectaría a aquellos tríxeles cuyos círculos circunscritos afecten a dicho punto. En otras palabras, se puede acelerar el procesado en tiempo real aprovechando esta propiedad. En lugar de generar la malla en cada fotograma, sólo se regenerarían los tríxeles a los cuales la nueva configuración de mínimos proporcionados por el DMUM afectase. Además, una de las principales ventajas del uso del DMUM es su robustez ya que los valores que proporcionan permanecen iguales si no hay variaciones significativas en la imagen.

7. El uso del índice de calidad (Qindex) permite detectar fallos en la segmentación de las partes de la persona, por ello sería posible definir un esquema iterativo que ajustara mejor la máscara a la persona en sucesivas iteraciones con el fin de conseguir una máscara más precisa.

8. Por último, los tríxeles han abierto una importante alternativa en los modelos geométricos para el reconocimiento de expresiones faciales. El posicionamiento de estos tríxeles en localizaciones invariantes del rostro (como los ojos, la nariz o la boca) pueden ayudar a encontrar nuevas vías para clasificar las expresiones faciales.

Bibliografía

[1] J. Abascal. La interacción persona-computador. *Horizonte*, pages 24–26, 2010.

[2] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012.

[3] Fuente Internet: Affectiva. http://www.affdex.com/, 2014. [Online; accessed 31-August-2014].

[4] J. Aghajanian, J. Warrell, S. Prince, J. Rohn, and B. Baum. Patch-based within-object classification. In *IEEE 12th International Conference on Computer Vision*, page 1125–1132, 2009.

[5] T. Ahonen, A. Hadid, and M. Pietikäinen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12), December 2006.

[6] L.A. Alexandre. Gender recognition: A multiscale decision fusion approach. In *Pattern Recognition Letters*, volume 11, page 1422–1427, 2010.

[7] D. Anguelov, K.C. Lee, S. B. Gokturk, and B. Sumengen. Contextual identity recognition in personal photo albums. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2007.

[8] L. Antón-Canalís, M. Hernández-Tejera, and E. Sánchez-Nielsen. Distance maps from unthresholded magnitudes. *Pattern Recognition*, 45(9):3125 – 3130, 2012.

[9] Fuente Internet: Diario AS. http://www.as.com/, 2010. [Online; accessed 14-May-2010].

[10] Fuente Internet: Psychological Image Collection at Stirling (PICS). University of Stirling Psychology Department. http://pics.psych.stir.ac.uk/, 2010. [Online; accessed 07-April-2010].

[11] Fuente Internet: Yale Face Database B. ftp://plucky.cs.yale.edu/cvc/pub/images/yalefacesb/tarsets/, 2000. [Online; accessed 06-April-2010].

[12] S. Baluja and H. A. Rowley. Boosting sex identification performance. In *International Journal of Computer Vision*, volume 71, page 111–119, 2007.

[13] M.S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Fully automatic facial action recognition in spontaneous behavior. In *In Proceedings of the Seventh IEEE International Conference on Automatic Face and Gesture Recognition*, 2006.

[14] A. Battocchi, F. Pianesi, and D. Goren-Bar. A first evaluation study of a database of kinetic facial expressions (dafex). *Proceedings of Int. Conf. on Multimodal interfaces.*, pages 214–221, 2004.

[15] H. Bay and T. Tuytelaars. Surf: Speeded up robust features. In *Proceedings of the Ninth European Conference on Computer Vision*, May 2006.

[16] J. Bekios-Calfa, J. M. Buenaposada, and L. Baumela. Revisiting linear discriminant techniques in gender recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(4):858–864, 2011.

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

[18] P. Booth. *An introduction to human-computer interaction.* Hilldale, NJ: Erlbaum, 1989.

[19] D. Bos, B. Reuderink, B. van de Laar, and A. Nijholt. Human-computer interaction for bci games: Usability and user experience. *Proceedings of the International Conference on Cyberworlds (CW)*, pages 277–281, 2010.

[20] L. Bourdev, S. Maji, and J. Malik. Describing people: A poselet-based approach to attribute classification. In *IEEE International Conference on Computer Vision*, pages 1543–1550, 2011.

[21] F. Bourel, C.C. Chibelushi, and A.A. Low. Robust facial expression recognition using a state-based model of spatially-localised facial dynamics. In *In Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, 2002.

[22] Y. Boykov and M. Joly. Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. *IEEE International Conference on Computer Vision (ICCV)*, pages 105–112, 2001.

[23] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1124–1137, 2004.

[24] M. M. Bradley and P. J. Lang. *Motivation and emotion.* Cambridge University Press, 2006.

[25] V. Bruce and S. Langton. The use of pigmentation and shading information in recognising the sex and identities of faces. *Perception*, (23):803–822, 1994.

[26] V. Bruce and A. Young. *The eye of the beholder. The Science of face perception.* Oxford University Press, 1998.

[27] R. Brunelli and T. Poggio. Face recognition: features versus templates. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 10, page 1042–1052, 1993.

[28] J. M. Buenaposada, E. Muñoz, and L. Baumela. Efficiently estimating facial expression and illumination in appearance-based tracking. In *Proceedings of BMVC*, pages 57–66, 2006.

[29] S. Burak, C. Tomasi, B. Girod, and J. Bouguet. Model-based face tracking for view-independent facial expression recognition. In *In Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, 2002.

[30] C.J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[31] A.M. Burton, V. Bruce, and N. Dench. What's the difference between men and women? Evidence from facial measurement. *Perception*, (22):153–176, 1999.

[32] Y. Cai and J. Abascal. Ambient intelligence in everyday life. *Lectures Notes in Computer Science (LNAI 3864)*, 2006.

[33] S. Candela, M. Castrillón, D. Freire, A. González, A. Salgado, L. Doreste, S. Kemchandani, and D. Hernández. *Fundamentos de Informática y Programación para Ingeniería. Ejercicios Resueltos para C y Matlab.* Paraninfo, 2011.

[34] L. Cao, M. Dikmen, Y. Fu, and T. S. Huang. Gender recognition from body. In *Proceeding of the 16th ACM International Conference on Multimedia*, page 725–728, 2008.

[35] M. Castrillón, J. Lorenzo, D. Freire, and O. Déniz. Learning to recognize gender using experience. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 4533–4536, 2010.

[36] M. Castrillón, J. Lorenzo, and E. Ramón. Improving gender classification accuracy in the wild. In *18th Iberoamerican Congress on Pattern Recognition (CIARP)*, 2013.

[37] M. Castrillón Santana, O. Déniz Suárez, M. Hernández Tejera, and C. Guerra Artal. ENCARA2: Real-time detection of multiple faces at different resolutions in video streams. *Journal of Visual Communication and Image Representation*, pages 130–140, April 2007.

[38] M. Castrillón, O. Déniz, D. Hernández, and A. Domínguez. Identity and gender recognition using the encara real-time face detector. In *Conferencia de la Asociacion Espanyola para la Inteligencia Artificial*, volume 3, 2003.

[39] H. Chen, A.C. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012.

[40] I. Cohen, N. Sebe, L. Chen, A. Garg, and T.S. Huang. Facial expression recognition from video sequences: Temporal and static modelling. *Computer Vision and Image Understanding: Special Issue on Face Recognition*, pages 160–187, 2003.

[41] M. Collins, J. Zhang, and P. Miller. Full body image feature representations for gender profiling. In *IEEE 12th International Conference on Computer Vision Workshops*, pages 1235–1242, 2009.

[42] D. Cristinacce, T. Cootes, and I. Scott. A multi-stage approach to facial feature detection. In *Proceedings of BMVC*, pages 231–240, 2004.

[43] E. Curran and M. Stokes. Learning to control brain activity: A review of the production and control of eeg components for driving brain-computer interface (bci) systems. *Brain Cogn.*, 51:326–336, 2008.

[44] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, page 886–893, 2005.

[45] A. R. Damasio. *Descarte's Error: Emotion, Reason and the Human Brain.* Picador, 1994.

[46] C. Darwin. *The expression of emotion in man and animals.* 1998.

[47] Fuente Internet: Cohn-Kanade AU-Coded Facial Expression Database. http://vasc.ri.cmu.edu/idb/html/face/facial_expression/index.html, 2000. [Online; accessed 11-April-2010].

[48] Fuente Internet: Equinox Infrared Face Database. http://www.equinoxsensors.com/products/hid.html, 2004. [Online; accessed 07-April-2010].

[49] Fuente Internet: Harvard Robotics Lab (HRL) Database. ftp://cvc.yale.edu/cvc/pub/images/hrlfaces, 2000. [Online; accessed 07-April-2010].

[50] Fuente Internet: NIST Mugshot Identification Database. http://www.nist.gov/srd/nistsd18.htm, 2000. [Online; accessed 08-April-2010].

[51] Fuente Internet: Olivetti Research Lab (ORL) Database. http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html, 1994. [Online; accessed 08-April-2010].

[52] Fuente Internet: The AR Face Database. http://cobweb.ecn.purdue.edu/~aleix/aleix_face_db.html, 1998. [Online; accessed 09-April-2010].

[53] Fuente Internet: The BANCA Database. http://www.ee.surrey.ac.uk/cvssp/banca/, 2004. [Online; accessed 07-April-2010].

[54] Fuente Internet: The FERET Database. http://www.itl.nist.gov/iad/humanid/feret/, 2003. [Online; accessed 08-April-2010].

[55] Fuente Internet: The Japanese Female Facial Expression (JAFFE) Database. http://kasrl.org/jaffe.html, 1999. [Online; accessed 11-April-2010].

[56] Fuente Internet: The Sheffield Face Database. http://www.shef.ac.uk/eee/research/vie/research/face.html, 1998. [Online; accessed 07-April-2010].

[57] Fuente Internet: Yale Face Database. ftp://plucky.cs.yale.edu/cvc/pub/images/yalefaces/, 1997. [Online; accessed 06-April-2010].

[58] D. Datcu and L.J.M. Rothkrantz. Automatic recognition of facial expressions using bayesian belief networks. In *In Proceedings of IEEE Systems, Man and Cybernetics*, 2004.

[59] B. Delaunay. Sur la sphere vide. a la mémoire de georges voronoi. *Bulletin of Academy of Sciences of the USSR*, 7(1):793–800, 1934.

[60] M. Demirkus, M. Toews, J.J. Clark, and T. Arbel. Gender classification from unconstrained video sequences. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, page 55–62, 2010.

[61] L. J. Denes, P. Metes, and Y. Liu. Hyperspectral face database. Technical Report CMU-RI-TR-02-25, Robotics Institute, Pittsburgh, PA, October 2002.

[62] O. Déniz, M. Castrillón, J. Lorenzo, L. Antón, and G. Bueno. Smile detection for user interfaces. *Advances in Visual Computing. Lecture Notes in Computer Science*, 5359:602–611, 2008.

[63] G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski. Classifying facial actions. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 10, page 974–989, 1999.

[64] F. Dornaika, E. Lazkano, and B. Sierra. Improving dynamic facial expression recognition with feature subset selection. *Pattern Recogn. Lett.*, 32(5):740–748, 2011.

[65] E. Douglas-Cowie, R. Cowie, and M. Schröder. A new emotion database: considerations, sources and scope. *Proceedings of ISCA Workshop on Speech and Emotion*, 2000.

[66] E. García Díaz. Boosting support vector machines, 2005. Tesis magister en ingeniería electrónica. Universidad de los Andes.

[67] J. Schwartz E. Kandel and J. Jessell. *Neurociencia y Conducta*. Prentice Hall, 1997.

[68] P. Ekman. Facial expression and emotion. *American Psychologist*, 48(1):384–392, 1993.

[69] P. Ekman. *Handbook of Cognition and Emotion*. T. Dalgleish and M. Power, 1999.

[70] P. Ekman. *Emotions revealed (2nd ed.)*. New York: Times Books, 2003.

[71] P. Ekman and W. Friesen. *Manual for the Facial Action Coding System*. Consulting Psychologists Press, 1977.

[72] P. Ekman and W.V. Friesen. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1(1):49–98, 1969.

[73] P. Ekman and W.V Friesen. Felt, false, and miserable smiles. *Journal of Nonverbal Behavior*, 6(4):238–252, 1982.

[74] D. Enlow. *Handbook of facial growth*. Saunders company, 1982.

[75] M. Ernst and H. Buelthoff. Merging the senses into a robust percept. *Cognitive Science*, (8):127–159, 2004.

[76] I. Essa and A. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 7, page 757–763, 1997.

[77] M. Everingham, L. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. In *International Journal of Computer Vision*, volume 88, pages 303–338, 2009.

[78] Fuente Internet: MMI Facial expression database. http://www.mmifacedb.com/, 2004. [Online; accessed 10-April-2010].

[79] Fuente Internet: The CAS-PEAL face database. http://www.jdl.ac.cn/peal/home.htm, 2004. [Online; accessed 07-April-2010].

[80] J.M. Fellous. Gender discrimination and prediction on the basis of facial metric information. In *Vision Research*, volume 24, page 1961–1973, 1997.

[81] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision (IJCV)*, 59(2):167–181, 2004.

[82] E. Fix and J. Hodges. An important contribution to nonparametric discriminant analysis and density estimation. *International Statistical Review*, 57(3):233–238, 1989.

[83] L. Ford and D. Fulkerson. *Flows in Networks*. Princeton University Press, 1962.

[84] D. Freire-Obregón, M. Castrillón-Santana, and O. Déniz-Suárez. New techniques applied for facial expression detection. *Journal on Biometrics Systems, Design and Applications*, 1:93–108, 2011.

[85] D. Freire-Obregón, M. Castrillón-Santana, L. Ramón-Balmaseda, and J. Lorenzo-Navarro. Automatic clothes segmentation for soft biometrics. In *Proceedings of IEEE International Conference on Image Processing (ICIP)*, 2014.

[86] D. Freire-Obregón, L. Antón-Canalis, and M. Castrillón-Santana. A study for the self similarity smile detection. In *Lecture Notes in Computer Science 5707*, pages 97–104, 2009.

[87] D. Freire-Obregón and M. Castrillón-Santana. A video database for facial behaviour understanding. In *Proceedings of JRBP*, number 1, 2012.

[88] D. Freire-Obregón, M. Castrillón-Santana, and O. Déniz-Suárez. A novel approach for smile detection combining lbp and pca. In *Proceedings of EUROCAST2009*, 2009.

[89] D. Freire-Obregón, M. Castrillón-Santana, and O. Déniz-Suárez. Smile detection using local binary patterns and support vector machines. In *Proceedings of Computer Vision Theory and Applications (VISAPP'09)*, 2009.

[90] R. W. Frischholz and U. Dieckmann. Bioid: A multimodal biometric identification system. *IEEE Computer*, 33(2), February 2000.

[91] Face Fuente Internet: FGNET and Gesture Recognition Working Group. http://wwwprima.inrialpes.fr/fgnet/html/home.html, 2009. [Online; accessed 25-August-2013].

[92] M. Fujita. Aibo: Towards the era of digital creatures. *International Journal of Robotics Research*, 20(10):781–794, 2001.

[93] B. Fulkerson, A. Vedaldi, and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[94] J.F. Fulton. *Functional Localization in the Frontal Lobes and Cerebellum*. Oxford,ClarendonPress, 1949.

[95] A. Gallagher and T. Chen. Understanding images of groups of people. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2009.

[96] A. C. Gallagher and T. Chen. Clothing cosegmentation for recognizing people. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2008.

[97] N. Gheissari, T. B. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2006.

[98] D. Goleman. *Inteligencia Emocional*. Kairos, 1996.

[99] M. Graw, A. Czarnetzki, and H.T. Haffner. The form of the supraorbital margin as a criterion in identification of sex from the skull: investigations based on modern human skulls. *J. Phys. Anthropology*, (108):91–96, 1999.

[100] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Performance Evaluation of Tracking and Surveillance*, 207.

[101] D. Greig, B. Porteous, and A. Seheult. Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society*, 51(2):271–279, 1989.

[102] G. Guo, G. Mu, and Y. Foo. Gender from body: A biologically-inspired approach with manifold learning. In *Proceedings of the 9th Asian Conference on Computer Vision-Volume Part III*, number 1, page 236–245, 2009.

[103] S. Gutta, J. Huang, P. Jonathon, and H. Wechsler. Mixture of experts for classification of gender, ethnic origin, and pose of human faces. In *IEEE Transactions on Neural Networks*, volume 11, pages 948–960, 2000.

[104] Wallhoff H. Fuente internet: Facial expressions and emotion database, 2006.

[105] J.C. Hager. *The inner and outer meanings of facial expression*. Guilford, 1983.

[106] J.C. Hager and P. Ekman. Long distance transmission of facial affect signals. *Journal of Nonverbal Behavior*, 4(1):61–62, 1979.

[107] P. Hallinan. *A Deformable Model for Face Recognition Under Arbitrary Lighting Conditions*. PhD thesis, Harvard University, 1995.

[108] H. Hill, V. Bruce, and S. Akamatsu. Perceiving the sex and race of faces: the role of shape and colour. In *Procs. of the Real Society London Biology Science*, number 261, page 367–373, 1995.

[109] T. Hoffmann and J. Buhmann. Pairwise data clustering and deterministic annealing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(1):1–25, 1997.

[110] Y. H. Huang and C. S. Fuh. Face detection and smile detection. In *Proceedings of IPPR*, number 1, 2009.

[111] Fuente Internet: Pose Illumination and Expression (PIE) Database. http://www.ri.cmu.edu/research_project_detail.html?project_id=418, 2000. [Online; accessed 07-April-2010].

[112] R. Isukapalli, A. Elgammal, and R. Greiner. Learning to identify facial expression during detection using markov decision process. In *In Proceedings of the Seventh IEEE International Conference on Automatic Face and Gesture Recognition*, 2006.

[113] A. Ito, X. Wang, M. Suzuki, and S. Makino. Smile and laughter recognition using speech processing and face recognition from conversation video. In *Procs. of the 2005 IEEE Int. Conf. on Cyberworlds (CW'05)*, 2005.

[114] A. Crespo-Abelleira J. A. Suárez-Quintanilla, E. Crespo-Vázquez and Angeles Rodríguez-Cobos. *Morfología de la cabeza humana*. Universidad de Santiago de Compostela, 2000.

[115] C. Jacobsen. The function of the frontal association areas in monkeys. In *Studies on cerebral function in primates*, number 13, pages 1–60, 1936.

[116] Z. Ji, X. Lian, and B. Lu. Gender classification by information fusion of hair and face. *I-Tech*, pages 215–229, 2009.

[117] R.E. Kaliouby and P. Robinson. Real-time inference of complex mental states from facial expressions and head gestures. In *IEEE CVPR Workshop on Real-time Vision for Human–Computer Interaction*, 2004.

[118] T. Kanade, J. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Proceedings of the 4th IEEE Int Conf on Automatic Face and Gesture Recognition*, 2000.

[119] M. Karg, A. Samadani, R. Gorbet, K. Kuhnlenz, J. Hoey, and D. Kulic. Body movements for affective expression: A survey of automatic recognition and generation. *IEEE Transactions on Affective Computing*, 2011.

[120] Y. Kirby and L. Sirovich. Application of the Karhunen-Loéve procedure for the characterization of human faces. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 12(1), July 1990.

[121] A. Kleinsmith and N. Bianchi. Affective body expression perception and recognition: A survey. *International Journal of Affective Computing*, 4:15–38, 2013.

[122] B. Knappmeyer, I. M. Thornton, and H. H. Bulthoff. The use of facial motion and facial form during the processing of identity. *Vision Res.*, 43(1):1921–1936, 2003.

[123] S. Koelstra, M. Pantic, and I. Patras. A dynamic texture-based approach to recognition of facial actions and their temporal models. In *Proceedings on Pattern Analysis and Machine Intelligence*, volume 32, pages 1940–1954, 2010.

[124] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):254–262, 2004.

[125] U. Kowalik, T. Aoki, and H. Yasuda. Broaference - a next generation multimedia terminal providing direct feedback on audience's satisfaction level. In *INTERACT*, pages 974–977, 2005.

[126] Y. Kamitani L. Shams and S. Shimojo. What you see is what you hear. *Nature*, 408(1):788, 2000.

[127] J.I. Lacey. *Psychophysiological approaches to the evaluation of psychotherapeutic process and outcome*. Washington, DC: National Publishing., 1958.

[128] A. Lapedriza, D. Masip, and J. Vitria. Are external face features useful for automatic face classification? In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005.

[129] R. Lazarus. *Emotion and adaptation*. Oxford University Press, 1991.

[130] C.S. Lee and A. Elgammal. Facial expression analysis using nonlinear decomposable generative models. In *IEEE International Workshop on Analysis and Modeling of Faces and Gestures (AMFG)*, 2005.

[131] B. Li, X. Lian, and B. Lu. Gender classification by combining clothing, hair and facial component classifiers. *Neurocomputing*, 1(76):18–27, 2012.

[132] S. Z. Li, L. Zhu, Z. Zhang, A. Blake, H. Zhang, and H. Shum. Statistical learning of multi-view face detection. In *European Conference Computer Vision*, pages 67–81, 2002.

[133] Y. Li, J. Sun, C. Tang, and H. Shum. Lazy snapping. *ACM Transactions on Graphics (SIGGRAPH)*, 23(3):303–308, 2004.

[134] H.C. Lian and B.L. Lu. Multi-view gender classification using local binary patterns and support vector machines. In *Advances in Neural Networks-ISNN*, page 202–209, 2006.

[135] S.W. Littlejohn. *Theories of human communication*. Wadsworth, 2002.

[136] H. Liu and P. Wu. Comparison of methods for smile detection by training au6 and au12 simultaneously. In *Proceedings of ICIP*, number 3, pages 1805–1808, 2012.

[137] J. Lorenzo, M. Castrillón, L. Ramón, and D. Freire. Evaluation of LBP and HOG descriptors for clothing attribute description. In *Proceedings of Video Analytics for Audience Measurement in Retail and Digital Signage (VAAM)*, 2014.

[138] D.G. Lowe. Object recognition from local scale-invariant features. *IEEE International Conference on Computer Vision (ICCV)*, pages 1150–1157, 1999.

[139] L. Lu and P. Shi. A novel fusion-based method for expression-invariant gender classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, page 1065–1068, 2009.

[140] X. Lu, H. Chen, and A. Jain. Multimodal facial gender and ethnicity identification. In *Advances in Biometrics*, volume 3832, pages 554–561, 2005.

[141] M. Lyons, S. Akamastu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. *Proceedings of Third IEEE Conf. on Face and Gesture Recognition*, pages 200–205, 1998.

[142] M. Lyons, J. Budynek, and S. Akamatsu. Automatic classification of single facial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12):1357–1362, 1999.

[143] J. MacQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, page 281–297, 1967.

[144] A.M. Martínez and R. Benavente. The ar face database. *CVC Technical Report*, 1998.

[145] D. Matsumoto and B. Willimgham. Spontaneous facial expressions of emotion of congenitally and noncongenitally blind individuals. *Journal of Personality and Social Psychology*, 96(1):1–10, 2009.

[146] D. McNeill. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, 1992.

[147] K. Messer, J. Kittler J. Matas and, J. Luettin, and G. Maitre. Xm2vtsdb: the extended m2vts database. In *Proceedings of the 2nd International Conference on Audio and Video-based Biometric Person Authentication*, 1999.

[148] Fuente Internet: MIT Mood Meter. http://arts.mit.edu/fast/fastlight/mit-mood-meter/, 2012. [Online; accessed 08-April-2012].

[149] K. Miller. *Communication Theories: Perspectives, processes, and contexts*. McGraw-Hill, 2005.

[150] G. Mori. Guiding model search using segmentation. *IEEE International Conference on Computer Vision (ICCV)*, 2005.

[151] G. Mori, X. Ren, A. Efros, , and J. Malik. Recovering human body configurations: Combining segmentation and recognition. *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2004.

[152] Fuente Internet: XM2VTSDB multi-modal face Database. http://www.ee.surrey.ac.uk/cvssp/xm2vtsdb/, 2003. [Online; accessed 09-April-2010].

[153] M. Nakano, F. Yasukata, and M. Fukumi. Age and gender classification from face images using neural networks. In *Signal and Image Processing*, 2004.

[154] F. Netter. *Anatomía Humana*. Elsevier, 2011.

[155] Fuente Internet: Database of Human Facial Expressions. http://tcc.itc.it/research/i3p/dafex/index.html, 2004. [Online; accessed 12-April-2010].

[156] Fuente Internet: University of Texas Video Database. http://bbs.utdallas.edu/facelab/perceptionlab.htm, 2005. [Online; accessed 06-April-2010].

[157] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29:51–59, 1996.

[158] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[159] A. V. Oppenheim and J. S. Lim. The importance of phase in signals. In *Proceedings of the IEEE*, 1981.

[160] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio. Pedestrian detection using wavelet templates. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 193–199, 1997.

[161] A. O'Toole, J. Harms, S. Snow, D. R. Hurst, M. Pappas, and H. Abdi. A video database of moving faces and people. *Vision Res.*, 39(1):3145–3155, 2003.

[162] A.J. O'Toole, T. Vetter, and V. Blanz. Three-dimensional shape and two-dimensional surface reflectance contributions to face recognition: An application of three-dimensional morphing. *Vision Res.*, 39(1):3145–3155, 1999.

[163] R. Davidson P. Ekman and W. Friesen. The duchenne smile: emotional expression and brain physiology. In *Journal of Personality and Social Psychology*, number 2, 1990.

[164] C. Padgett and G. Cottrell. Representing face images for emotion classification. In *Advances in Neural Information Processing Systems (NIPS)*, 1997.

[165] M. Pantic and I. Patras. Dynamics of facial expression: Recognition of facial actions and their temporal segments form face profile image sequences. *IEEE Transaction on System, Man and Cybernetics*, 36(1):433–449, 2006.

[166] M. Pantic and L. Rothkrantz. Expert system for automatic analysis of facial expression. In *Image and Vision Computing*, volume 11, page 881–905, 2000.

[167] M. Pantic and L. Rothkrantz. Facial action recognition for facial expression analysis from static face images. In *IEEE Transactions on Systems, Man, and Cybernetics*, volume 3, page 1449–1461, 2004.

[168] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *Proceedings of IEEE International Conference Multmedia and Expo (ICME05)*, 2005.

[169] P.J. Philips, H. Moon, P. Rauss, and S.A. Rizvi. The feret evaluation methodology for face-recognition algorithms. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition*, 1997.

[170] R. W. Picard. *Affective Computing*. MIT Press, 1997.

[171] R. W. Picard. Affective computing for hci, 1999. [Online; accessed 24-Jun-2013].

[172] S. Pigeon and L. Vandendrope. The m2vts multimodal face database. In *Proceedings of the 1st International Conference Audio and Video Based Biometric Person Authentication*, 1997.

[173] L. Reddy, P. Wilken, and C. Koch. Face-gender discrimination is possible in the near-absence of attention. *Journal of Vision*, (4):106–117, 2004.

[174] Research Group. ROC SIANI, 2012. [Online; accessed 20-Jul-2013].

[175] A. Rosenfeld and J. Pfaltz. Distance functions on digital pictures. *Pattern Recognition*, 1(1):33–61, 1968.

[176] C. Rother, V. Kolmogorov, and A. Blake. Grabcut - interactive foreground extraction using iterated graph cuts. *Proceedings of ACM Siggraph*, 2004.

[177] M. Ruiz. *Escritura y comunicación social: historia de la comunicación escrita*. Alfar, 1996.

[178] Y. Rodríguez S. Marcel and G. Heusch. On the recent use of local binary patterns for face authentication. *International Journal of Image and Video Preprocessing, Special Issue on Facial Image Processing*, 2007.

[179] Y. Saatci and C. Town. Cascaded classification of gender and facial expression using active appearance models. In *Automatic Face and Gesture Recognition*, pages 393–398, 2006.

[180] J. Sadr and P. Sinha. Exploring object perception with random image structure evolution. massachusetts institute of technology, artificial intelligence laboratory memo, 2001.

[181] J. Sadr and P. Sinha. Random image structure evolution (RISE). *Vision Sciences Society Abstracts*, 83(1):128–142, 2001.

[182] F.S. Samaria. *Face Recognition using Hidden Markov Models*. PhD thesis. Cambridge University Press, 1994.

[183] A. Savran. Bosphorus database for 3d face analysis. *Proceedings of First European Workshop on Biometrics and Identity Management Workshop*, pages 200–205, 2008.

[184] W. Schiff, L. Banka, and G. de Bordes. Recognizing people seen in events via dynamic mug shots. *Amer. J. Pychol*, 99(1):219–231, 1986.

[185] H. Schneiderman and T. Kanade. A statistical method for 3D object detection applied to faces and cars. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1746–1759, 2000.

[186] M. Schuller. Emotion recognition in the noise applying large acoustic feature sets. *Speech Prosody*, 2006.

[187] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *IEEE Conference on Computer Vision and Pattern Recognition 2007 (CVPR'07)*, June 2007.

[188] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[189] K. Shimada, T. Matsukawa, Y. Noguchi, and T. Kurita. Appearance-based smile intensity estimation by cascaded support vector machines. In *Proceedings of the 2010 international conference on Computer vision*, pages 277–286, 2010.

[190] Y. Shinohara and N. Otsu. Facial expression recognition using Fisher weight maps. In *Procs. of the IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2004.

[191] J. Shotton, A. Fitzgibbon, M. Cook, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2011.

[192] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE*, 94(1):1948–1962, 2006.

[193] J. Sivic, M. Everingham, and A. Zisserman. Person spotting: video shot retrieval for face sets. In *Proceedings of the International Conference of Image and Video Retrieval (CIVR)*, 2005.

[194] S. Skiena. *The algorithm design manual*. Springer-Verlag, 1998.

[195] Y. Song and T. Leung. Context-aided human recognition-clustering. In *Proceedings of the European Conference Computer Vision (ECCV)*, 2006.

[196] Z. Sun, G. Bebis, X. Yuan, and S. Louis. Genetic feature subset selection for gender classification: a comparison study. In *IEEE Proceedings on Applications of Computer Vision*, pages 165–170, 2002.

[197] K.K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):39–51, 1999.

[198] G. J. Edwards T. F. Cootes and C. J. Taylor. Active appearance models. In *Proc. of the 5th European Conference on Computer Vision*, volume 2, page 484–498, 1998.

[199] Fuente Internet: SONY DSC T200. http://www.sony.es/product/dsc-t-series/dsc-t200, 2007. [Online; accessed 05-August-2013].

[200] J. Talbot and X. Xiaoqian. *Implementing GrabCut*. Tech Report, 2006.

[201] X. Tan. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *Image Processing, IEEE Transactions on Biometrics Compendium*, 19:1635 – 1650, 2010.

[202] Q. Tao and R. Veldhuis. Illumination normalization based on simplified local binary patterns for a face verification system. In *Proc. of the Biometrics Symposium*, pages 1–6, 2007.

[203] Fuente Internet: Top Tenz. http://www.toptenz.net/top-10-insane-plots-to-kill-celebrities.php, 2013. [Online; accessed 28-June-2013].

[204] Y. Tian. Evaluation of face resolution for expression analysis. In *In Proceedings of CVPR Workshop on Face Processing in Video*, 2004.

[205] Y. Tian, L. Brown, A. Hampapur, S. Pankanti, A. Senior, and R. Bolle. Real world real-time automatic recognition of facial expression. In *IEEE Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, 2003.

[206] Y. Tian, T. Kanade, and J. Cohn. Recognizing facial actions by combining geometric features and regional appearance patterns. Technical Report CMU-RI-TR-01-01, Robotics Institute, Pittsburgh, PA, January 2001.

[207] Y. Tian, T. Kanade, and J. Cohn. Facial expression analysis. *Handbook of Face Recognition*, (2):247–275, 2005.

[208] P. Tome, J. Fierrez, R. Vera-Rodriguez, and M.S. Nixon. Soft biometrics and their application in person recognition at a distance. *Information Forensics and Security, IEEE Transactions on*, 9(3):464–475, 2014.

[209] M. Tuceryan and T. Chorzempa. Relative sensitivity of a family of closest-point graphs in computer vision applications. *Pattern Recognition*, 24(5):361–373, 1991.

[210] M. Turk and A. Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3(1):71–86, 1991.

[211] Anatomía UNAM. http://anatomiaunam.blogspot.com.es/, 2013. [Online; accessed 06-May-2013].

[212] M. Valstar and M. Pantic. Fully automatic facial action unit detection and temporal analysis. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, volume 3, pages 149–156, 2006.

[213] M. Valstar, I. Patras, and M. Pantic. Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, volume 3, page 76–84, 2005.

[214] S. P. Vecera. Affective, cognitive, and social aspects of face recognition, 1997. [Online; accessed 24-Jun-2013].

[215] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 511–518, 2001.

[216] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):151–173, May 2004.

[217] Fuente Internet: OKAO Vision. http://www.omron.com, 2009. [Online; accessed 25-August-2013].

[218] B. Vlasenko. Combining frame and turn-level information for robust recognition of emotions within speech. *Proceedings of INTERSPEECH*, 2007.

[219] J. Wang, J. Li, C.Y. Lee, and W.Y. Yau. Dense sift and gabor descriptors-based face representation with applications to gender recognition. In *11th International Conference on Control Automation Robotics and Vision*, page 1860–1864, 2010.

[220] Tao Wang, Xuming He, and N. Barnes. Glass object localization by joint inference of boundary and depth. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 3783–3786, Nov 2012.

[221] Y. Wang, H. Ai, B. Wu, and C. Huang. Real time facial expression recognition with adaboost. In *In Proceedings of International Conference on Pattern Recognition*, 2004.

[222] M. Weiser. The world is not a desktop. *Interactions*, pages 7–8, 1994.

[223] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan. Developing a practical smile detector. Submitted to Transactions on Pattern Analysis and Machine Intelligence. Last accessed: March 2008., 2008.

[224] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan. Towards practical smile detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 31(1):2106–2111, 2009.

[225] H.A. Wild, S.E. Barrett, M.J. Spence, A.J. O'Toole, Y.D. Cheng, and J. Brooke. Recognition and sex categorization of adults' and children's faces: examining performance in the absence of sex-stereotyped cues. *J. Exp. Child Psychology*, (77):269–291, 2000.

[226] Fuente Internet: Problems with smile shutter. http://www.ehow.com, 2012. [Online; accessed 05-August-2013].

[227] D. Wright, S. Gutwirth, and M. Friedewald. Safeguards in a world of ambient intelligence. *Springer, Dordrecht*, 2008.

[228] T. Wu, C. Lin, and R. Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5(1):975–1005, 2004.

[229] X. Wu and X.C. Lian. Multi-view gender classification using symmetry of facial images. In *Neural Computing and Applications*, page 1–9, 2011.

[230] B. Xia, H. Sun, and B. Lu. Multi-view gender classification based on local gabor binary mapping pattern and support vector machines. In *IEEE International Joint Conference on Neural Networks*, page 3388–3395, 2008.

[231] Y. Yacoob and L. Davis. Detection, analysis and matching of hair. In *The tenth IEEE International Conference on Computer Vision*, page 741–748, 2005.

[232] Y. Yacoob and L.S. Davis. Recognizing human facial expression from long image sequences using optical flow. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 6, page 636–642, 1996.

[233] H. Yadappanavar and S. Shylaja. Machine learning approach for smile detection in real time images. In *International Journal of Image Processing and Vision Sciences (IJIPVS)*, pages 32–37, 2012.

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

[235] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2011.

[236] M. Yeasin, B. Bullot, and R. Sharma. From facial expression to level of interests: a spatio-temporal approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[237] K. Yoon, D. Harwood, and L. Davis. Appearance-based person recognition using color/path-length profile. *Journal of Visual Communication and Image Representation*, 17:605–622, 2006.

[238] Y. Zhang and Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequences. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 5, pages 1–16, 2005.

[239] Y. Zhang, Z. Liu, B. Guo, and H. Shum. Geometry-driven photorealistic facial expression synthesis. In *SIGGRAPH Symposium on Computer Animation*, number 22, pages 1–11, 2003.

[240] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu. Comparison between geometry-based and gabor wavelets based facial expression recognition using multi-layer perceptron. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 1998.

[241] J. Zheng and B.L. Lu. A support vector machine classifier with automatic confidence and its application to gender classification. In *Neurocomputing*, volume 11, page 1926–1935, 2011.

[242] L. Zitnick and S. Kang. Stereo for image-based rendering using image over-segmentation. *International Journal of Computer Vision (IJCV)*, 75(1):49–65, 2007.