DEPARTAMENTO DE LENGUAJES Y SISTEMAS INFORMÁTICOS
Máster en Sistemas Inteligentes

CARACTERIZACIÓN DE PERSONAS EN IMÁGENES DE FIGURAS HUMANAS: ESTUDIO DE DESCRIPTORES VISUALES Y ESQUEMAS DE CLASIFICACIÓN AUTOMÁTICA

Presentado por
Carlos Serra Toro

Supervisado por
Vicente Javier Traver Roig y **Raúl Montoliu Colás**

Castellón de la Plana, 20 de septiembre de 2010

Resumen

El presente trabajo de final de máster trata el estudio del problema del reconocimiento de género de las personas a partir de imágenes estáticas de las mismas. Este problema se enmarca dentro de lo que se podría llamar biometría blanda (*soft biometrics*), es decir, clasificar a una persona como perteneciente a un grupo, en lugar de verificar o identificar completamente su identidad. En el caso particular que nos ocupa se estudia la clasificación de las personas como pertenecientes a los grupos de hombres o mujeres.

Aunque existe algo de trabajo hecho en esta línea relacionado con secuencias de vídeo (e.g. determinar el género de una persona según su forma de caminar), por lo que se sabe, hasta la fecha sólo tres artículos han tratado este problema usando como fuente de datos las imágenes estáticas.

La presente memoria comienza realizando una revisión bibliográfica de la biometría blanda, centrándose en las tres aportaciones existentes en la bibliografía que tratan el problema concreto de reconocimiento de género a partir de imágenes estáticas. Partiendo de la primera propuesta existente en la literatura para tratar este problema, se estudia este primer enfoque y se realizan varias experimentaciones, poniendo especial énfasis en la comparación de los métodos de clasificación automática usados para resolver este problema, así como las distintas distribuciones de las ventanas sobre las imágenes y su impacto sobre el resultado final. Experimentalmente se comprueba que se consiguen mayores tasas de acierto cuanto más densa (hasta cierto punto) es la rejilla que configura las ventanas que se disponen sobre la imagen, al contrario de lo que sucede con el algoritmo clásico de reconocimiento de personas, donde la rejilla no necesariamente debe ser densa. Se ha comprobado que rejillas más densas (10×19, 10×35 o 12×35) requieren de algoritmos menos complejos para superar las tasas de acierto proporcionadas por los dos primeros enfoques de este problema existentes en la literatura, para la primera base de datos de imágenes etiquetada para el reconocimiento de género.

Abstract

The current master degree project studies the problem of gender recognition of individuals from static images. This problem is part of what might be called soft biometrics (e.g. classifying a person as belonging to a group, instead of completely verifying or identifying his/her identity). In this particular case, we study the classification of individuals as belonging to the groups of men or women.

Although there is some work done in this line related to video sequences (e.g. to determine the gender of a person according to his/her gait), to the best of our knowledge only three articles have addressed this problem using as data source static images.

This document begins with a literature review of soft biometrics, focusing on the three existing contributions in the literature concerning gender classification from static images. Starting from the first proposal in the literature addressing this problem, we study the first approach and perform several experiments, with particular emphasis on the comparison of the automatic classification methods used to solve this problem, and the different distributions of the windows over the images and their impact on the classification results. Experimentally, it is found that denser grids (to a certain extent) over the image give raise to more accuracy than sparser grids. This is contrary to the classical algorithm used for the detection of people, where the grid need not to be so dense to increase the results. Denser grids (10×19, 10×35 or 12×35) need lees complex algorithms to exceed the success rate reported by the first two previous approaches in the literature using the first labeled image database for gender recognition.

Índice de contenido

Capítulo 1 Introducción

La caracterización de personas (e.g. por rango de edad, género, grupo social, etc.) en imágenes y videos digitales resulta de gran relevancia en numerosas aplicaciones de interés científico y social. Por ejemplo, en estudios de mercado, la monitorización de clientes (potenciales) de una empresa puede aportar información relevante sobre su perfil, de modo que la empresa pueda tomar ciertas decisiones. Otros ámbitos de interés son las interfaces persona-ordenador, la recuperación de imágenes/video de (extensas) bases de datos, etc.

El presente trabajo final de máster se centra en el reconocimiento de género, sobre el que no existe mucho trabajo, y lo poco que se ha investigado ha sido, casi siempre, en vídeo (secuencias de imágenes). Sin embargo, resulta interesante estudiar el problema en imágenes (estáticas), básicamente por dos motivos. Por un lado, el análisis en imágenes permitiría mejorar las prestaciones de los sistemas basados en, por ejemplo, la forma de andar (Yu et al., 2009). Aunque, en general, la información temporal disponible en los videos aporta importantes pistas para la clasificación, hay otra información (forma, apariencia, etc.) que puede resultar complementaria. Por otro lado, existen situaciones donde esta información temporal no está disponible. Por ejemplo, incluso en secuencias de imágenes, puede ocurrir que la persona esté parada (no camine). En muchas otras situaciones, se puede disponer de imágenes, pero no de vídeos (e.g. colecciones, privadas o públicas, de fotos), por lo que el uso de la información dinámica es, simplemente, imposible.

Además, y como indican Guo et al. (2010), el reconocimiento a partir de imágenes de cuerpo entero tiene ciertas ventajas sobre el reconocimiento a partir de imágenes de la cara, como son: 1) la posibilidad de usar imágenes de baja resolución donde la cara es difícilmente distinguible, 2) la dificultad de determinar el género de una persona si ésta oculta su cara o la disimula de alguna forma, 3) es imposible reconocer el género de una persona a partir de su cara si ésta se encuentra de espaldas mientras que una vista trasera de una persona aporta información sobre su género, y 4) el uso de cámaras para la obtención de imágenes de alta resolución de las caras de las personas puede ser una práctica intrusiva.

El problema planteado en este proyecto consiste, pues, en reconocer el género de una persona a partir de *una* única imagen de la misma, en posición vertical (de pie). Además de importante y relevante, este problema resulta complejo: deducir el género de una persona es difícil, puede que incluso para los propios humanos. Aunque existen una serie de heurísticas que pueden guiar (parcialmente) el diseño del sistema (las mujeres *suelen* tener el pelo largo; los hombres *pueden* tener los hombros más anchos; hay más hombres que llevan pantalones que mujeres, etc.), existen excepciones que dificultan que tales heurísticas resulten fiables en general. Por ejemplo, en zonas o épocas de frío, tanto hombres como mujeres pueden llevar prendas de abrigo largas, lo que, no sólo les hace más parecidos, sino que oculta información (anchura de hombros o caderas) que podría ayudar a discriminar.

El enfoque que se le ha dado al problema se basa en la generación de descriptores visuales genéricos y la exploración de técnicas de aprendizaje automático que, a partir de

ejemplos de ambas clases (imágenes de hombres y mujeres), obtenga un modelo que permita predecir el género de imágenes no vistas anteriormente. Para ello, además de considerar nuevos planteamientos y estrategias, se estudian y comparan los pocos trabajos publicados sobre el reconocimiento de género en imágenes (Cao et al. 2008, Collins et al., 2009). A continuación se hace un resumen del estado del arte de este enfoque.

1.1 Estado del arte

Actualmente se le está dando bastante importancia al problema de la caracterización de las personas como pertenecientes a un grupo, desde un enfoque biométrico, esto es, reconocer no sólo que hay una persona, sino identificarla unívocamente tomando ciertas métricas sobre la misma, tales como su configuración facial o la forma de su mano, entre muchas otras (Jain y Ross, 2004).

1.1.1 Biometría blanda (*soft biometrics*)

Un caso particular de la biometría es la llamada biometría blanda (*soft biometrics*) que trata de clasificar una persona como perteneciente a determinado grupo (por ejemplo, en función de su edad, raza o género). En cuanto a este tipo de clasificación, se ha realizado bastante investigación en lo que respecta a imágenes de caras tomadas frontalmente, sobre todo en cuanto a la clasificación de género en el que por ejemplo Moghaddam y Yang (2002) consiguen tasas de acierto del 96%.

La determinación de la edad partiendo de imágenes de caras ha sido también tratada en profundidad: sirva como ejemplo reciente el artículo de Guo et al. (2008) con errores medios de determinación de la edad de unos 5 años en el mejor de los casos.

La subclasificación en razas ha sido menos estudiada: Gutta y Wechsler (1999) han tratado este tema trabajando con imágenes de caras de cuatro razas distintas, consiguiendo tasas de acierto de hasta el 94%, y Lu y Jain (2004) consiguieron elevar la tasa de acierto hasta el 96,3%, aunque con sólo dos "razas" (asiática y no-asiática). Actualmente se está dando cierta importancia a la clasificación del género de las personas por su forma de andar, consiguiendo Yu et al. (2009) hasta un 95,97% de acierto en la clasificación.

1.1.2 Reconocimiento de género a partir de imágenes estáticas

Para el tema que nos ocupa, en el momento de la finalización de la redacción de esta memoria de proyecto sólo existen tres propuestas en la literatura que abordan el problema de la clasificación de género por medio de imágenes estáticas. Cao et al. (2008) fue el primer intento, con una tasa media de acierto del 75% sobre una base de datos de imágenes de personas de frente y de espaldas en las que las personas se encontraban situadas de pie y en las que las imágenes sólo contenían una persona (Oren, 1997). Esta base de datos se detalla en la Sección 4.1. Cao et al. (2008) usan un enfoque basado en *boosting* con características obtenidas mediante HOG (Sección 2.1). Puesto que este enfoque se ha estudiado en la experimentación realizada en el presente proyecto, para una explicación más detallada se remite al lector a la Sección 3.2 de esta memoria de proyecto.

Collins et al. (2009) fue el segundo artículo que trató el tema del reconocimiento de género en imágenes estáticas. Su estudio es más exhaustivo que el realizado por Cao et al. (2008) ya que usan más extractores de características, así como distintas bases de datos

de imágenes diferentes de la usada por Cao et al. (2008). El clasificador que usan en este artículo es una máquina de soporte vectorial (SVM, por sus siglas en inglés) en vez de una estrategia de *boosting* como Cao et al. (2008), y la distribución de las ventanas también varía en este artículo respecto al de Cao et al. (2008).

No queda claro hasta qué punto los resultados obtenidos por Collins et al. (2009) son comparables a los del primer estudio puesto que, para los resultados que obtienen para la base de datos de este primer estudio, previamente a la aplicación de su método recortan la imagen. De esta forma eliminan una porción rectangular del fondo de la imagen de manera que la misma imagen es una caja de recubrimiento mínimo de la persona que muestra. La Figura 1 muestra un ejemplo del tipo de recorte que realiza Collins et al. (2009).

Figura 1: Ejemplo de recorte realizado por Collins et al. (2009) sobre las imágenes de la base de datos de imágenes del MIT (Oren, 1997). Imagen extraida de Collins et al. (2009).

Además, únicamente tienen en cuenta las imágenes que muestran una vista frontal. El resultado que obtienen es de casi un 72% para esta vista, es decir un 4% inferior al obtenido por Cao et al. (2008). Los autores lo explican diciendo que el desbalance entre las clases hace que el método de Cao et al. (2008) acierte prácticamente todas las instancias de la clase mayoritaria y unas pocas de la clase minoritaria, aunque esto no justifica el hecho de que obtengan una tasa de aciertos menor que aquéllos. Experimentalmente, en este proyecto se ha comprobado que Collins et al (2009) se halla en lo cierto (ver Sección 4.6.1). Collins et al. (2009) balancean la base de datos descartando imágenes al azar de la clase mayoritaria, de forma que ambas clases tiene el mismo número de muestras, y así obtienen una tasa máxima de aciertos de 72,28 ± 8,07% y de 76,00 ± 8,13% para el caso de las imágenes originales y las recortadas, respectivamente. Es decir, que incluso recortando las imágenes no superan la tasa obtenida por Cao et al. (2008) para su misma base de datos (se recuerda que su resultado para la vista frontal es de 76,0 ± 1,2%).

Para obtener estos resultados usa una combinación de características obtenidas mediante los descriptores de Local HSV Colour Histogram (LHSV) y PixelHOG (PiHOG). El LHSV es un descriptor de color que usan puesto que los autores afirman que, de acuerdo con otro estudio anterior (Historical Boys' Clothing Web, 2007), las mujeres prefieren vestir ropas de colores más vivos que los hombres. El LHSV es en realidad un Histograma de Tonos de Color (*Hues*) Orientados: convierten la imagen al espacio de color HSV y dividen el espacio H es una serie de entradas para el histograma, siendo el valor de S los votos. El PiHOG es un descriptor de formas que calcula el gradiente basado según una detección de bordes con un sólo umbral inferior (al contrario que el detector de bordes de Canny (1986), que usa dos umbrales -uno inferior y otro superior).

En el momento de la finalización de la experimentación del presente proyecto, Guo et al. (2010) han publicado un método que alcanza el 80,6 ± 1,2% de acierto. Guo et al. (2010) proponen un método basado en la extracción de características basadas en la forma en que los objetos se representan en el córtex visual. Según un estudio previo (Riesenhuber y Poggio, 1999), el proceso consiste en alternar capas de unidades llamadas Simple (S) y Compleja (C), en tantos niveles como sean necesarios. Guo et al. (2010) han descubierto que, para la detección de personas, basta con un sólo nivel de capas (es decir, una sóla capa Simple S1 seguida de otra Compleja C1). En realidad su proceso es similar al HOG en el sentido de que también usa las orientaciones obtenidas a partir de la aplicación de un filtro, en este caso de detección de bordes (un filtro de Gabor). El proceso que proponen se basa en aplicar en la etapa S1 el filtro de Gabor para 6 orientaciones con 12 escalas diferentes. Partiendo de este resultado, cada región de las obtenidas en la etapa S1 se une por separado a cada una de sus regiones adyacentes, formando así ocho bandas para cada una de las cuatro orientaciones. En la etapa C1 que sigue lo que se hace es coger los valores máximos de las 16 escalas consideradas para cada banda. Los autores del estudio comentan que esto permite al algoritmo ser robusto a pequeños movimientos en la imagen y a cambios en la escala de las mismas.

Para la clasificación usan una máquina de soporte vectorial. Únicamente usan la base de datos del MIT (Oren, 1997) usando el etiquetado proporcionado por Cao et al. (2008), al contrario que Collins et al. (2009) quienes usan su propio etiquetado. Para reducir la dependencia de la vista (frontal o trasera) del clasificador, lo que hacen es clasificar primeramente la vista de la imagen (si es frontal o si es trasera, o si no puede determinarse), y entonces clasifican el género de la imagen con un clasificador entrenado específicamente para esa vista en particular.

Las mejores tasas de acierto que obtienen para cada vista por separado son: 79,5 ± 2,6% para la vista frontal, 84,0 ± 5,2% para la vista trasera, y un 79,2 ± 1,4% para ambas vistas combinadas. Usando su método de clasificar primeramente la vista y posteriormente usar el clasificador específico para dicha vista, el reconocimiento global que alcanzan es de 80,6 ± 1,2%, similar al obtenido por Collins et al. (2009) aunque ligeramente superior. No obstante, dado que estos dos autores (Collins et al. 2009, y Guo et al. 2010), usan bases de datos de imágenes diferentes, no queda claro hasta qué punto ambos resultados son comparables entre sí.

1.1.3 Uso de las aportaciones existentes para la detección de personas

Estos enfoques (Collins et al., 2009, y Cao et al., 2008), hacen uso de las herramientas existentes en la literatura para la detección de personas, y tratan de usarlas de forma que les permitan determinar el género de una persona dada una imagen de ésta. Durante los últimos años la detección de personas mediante técnicas de visión por ordenador ha sido un campo muy estudiado. Aunque a lo largo de la literatura se han usado varios enfoques para detectar personas tanto en imágenes estáticas como en secuencias de vídeo, en la mayoría de los estudios recientes se aborda el problema mediante el entrenamiento supervisado por medio de grandes bases de datos de imágenes o vídeos (según el caso). Actualmente, y a juzgar por la literatura existente, las dos características más usadas son los histogramas de gradientes orientados (HOG, por sus siglas en inglés) y las *wavelets* de Haar.

Los HOG fueron popularizados por Dalal y Triggs (2005) al usarlos como característica principal para su detector de personas con el que conseguía tasas de acierto cercanas al 100% y con una tasa de falsos positivos por ventana cercana a 10^{-4}. Este trabajo atrae todavía hoy mucho interés y ha habido bastantes propuestas de mejora, como Pedersoli et al. (2009) que, haciendo uso de modelos de cascada (es decir: una combinación de

clasificadores encadenados cada uno de los cuales sólo se ejecuta si el clasificador anterior a uno dado no pudo clasificar la muestra), mejora la velocidad del mismo entre 10 y 20 veces manteniendo un nivel de reconocimiento similar al propuesto originalmente por Dalal y Triggs (2005).

Las características de *wavelets* de Haar han inspirado algoritmos de detección de objetos (Viola y Jones, 2001) que posteriormente han sido adaptados para la detección de personas, como es el caso de Viola et al. (2003), también con altas tasas de acierto aunque su popularidad en el reconocimiento de personas ha decaído desde la invención del HOG.

En cuanto a los clasificadores más usados en la detección de personas, los que mejor resultados ofrecen de acuerdo con la literatura son AdaBoost (Freund y Schapire, 1995), como en Viola et al. (2003), o las máquinas de soporte vectorial, usadas por ejemplo en Dalal y Triggs (2005).

La mayor parte de artículos publicados hasta ahora orientados al reconocimiento de personas adoptan un enfoque holístico, esto es, consideran como un todo a la imagen que posiblemente envuelve a una persona. Recientemente se está profundizando en otro tipo de enfoque (las *pictorial structures*) que considera a las personas como una suma de sus componentes físicos. Tal es el caso de Felzenszwalb y Huttenlocher (2005), cuyo sistema, partiendo de la silueta de una persona obtenida mediante sustracción de fondo, devuelve las diez partes en las que los autores dividen a la persona, así como sus relaciones entre ellas, haciendo uso de un modelo estadístico complejo. Otro enfoque más reciente (Lin y Davis, 2010) simplifica esta idea para "acoplar" la figura de una persona a una serie de siluetas predefinidas que, de acuerdo a una estructura jerárquica, permite modelar las diversas posturas que adopta un cuerpo cuando se encuentra de pie. Que se sepa, este tipo de enfoques no han sido usado todavía puestos a prueba en la determinación del género de las imágenes de personas, por lo que podría ser una cuestión a explorar.

1.2 Organización de la presente memoria de proyecto

El resto de la presente memoria de proyecto de final de máster se organiza como sigue. El Capítulo 2 explica la forma de extraer las características del HOG (Histograma de Gradientes Orientados, por sus siglas en inglés) de una imagen. Se considera conveniente explicar este método de extracción de características ya que, como se ha visto en el presente capítulo, es el método más usado actualmente en caracterización de personas, así como el método usado para obtener las características que usa Cao et al. (2008), cuyos algoritmos se estudian en esta memoria. Estos algoritmos se explican en el Capítulo 3 y se ponen a prueba experimentalmente en el Capítulo 4. Se termina con las conclusiones del Capítulo 5, donde también se apunta el posible trabajo futuro más inmediato para seguir investigando en la línea tratada por esta memoria.

Capítulo 2 Histograma de Gradientes Orientados

El método de extracción de características más usado actualmente para caracterizar la figura humana (aunque no necesariamente su uso se limita a ello) es el Histograma de Gradientes Orientados (HOG, por las siglas en inglés de Histogram of Oriented Gradients).

El HOG fue propuesto por Dalal y Triggs (2005) como un método para la detección de objetos. No obstante, el ámbito donde más destaca su uso, ya desde el artículo en el que se presentó el método, es en la detección de personas en imágenes estáticas. Dalal y Triggs (2005) lo usaban como característica principal para su detector de personas con el que conseguían tasas de acierto cercanas al 100% y con una tasa de falsos positivos por ventana cercana a 10^{-4}. Estos resultados se obtuvieron para una base de datos de imágenes en las que aparecían personas sin oclusiones, o con muy poca oclusión.

En esta sección se describe el método original propuesto por Dalal y Triggs (2005). Puesto que este método da lugar a muchas variantes según distintas configuraciones posibles (tipo de ventana, número de orientaciones, tipo de normalización del vector, etc.), se describe únicamente la configuración tenida en cuenta en la implementación de este método para la obtención de las características que se han usado en la fase de experimentación del presente trabajo. En el caso de que Dalal y Triggs (2005) propusieran diversas variantes para la realización de cada uno de los pasos de su algoritmo, éstas simplemente se citan.

2.1 Descripción del detector de personas basado en HOG

El detector de personas basado en HOG propuesto por Dalal y Triggs (2005) se basa en la división de la imagen en ventanas de acuerdo con una configuración en rejilla. Cada una de estas ventanas se sintetiza en un histograma cuyas entradas se corresponden a las distintas orientaciones de los gradientes. El proceso a grandes rasgos de obtención del HOG para una ventana dada es el siguiente:

1. Cálculo de los gradientes de la ventana.
2. Discretización mediante histograma de los gradientes calculados.

A continuación se describe en detalle cada uno de estos tres procesos.

2.1.1 Cálculo de los gradientes de la ventana

Dalal y Triggs (2005) prueban con distintas máscaras para la obtención de los gradientes. El uso de una máscara de suavizado previa a la aplicación de las máscaras para la obtención de los gradientes no mejora significativamente los resultados obtenidos y penaliza la velocidad de ejecución, por lo que no se usa.

De todas las máscaras probadas para la obtención de los gradientes de la ventana, Dalal y Triggs (2005) recomiendan la más simple unidimensional [–1 0 1], que es además la que mejores resultados obtiene. Para calcular el gradiente de cada ventana, pues, se aplican los dos filtros unidimensionales siguientes: [–1 0 1] y $[-1\ 0\ 1]^T$. Esto da lugar a gradientes con signo, que serán convertidos a sin signo en el paso siguiente.

2.1.2 Discretización mediante histograma de los gradientes calculados

Dalal y Triggs (2005) se dieron cuenta de que, para el reconocimiento de personas, elevar el número de orientaciones de los gradientes más allá de nueve no mejoraba los resultados. Según ellos, supuestamente esto se debe a la gran variabilidad que aporta la ropa a la figura de las personas. Debido a esto, la dirección del gradiente no aporta tampoco información significativa, por lo que únicamente se tiene en cuenta la orientación de éste. La rejilla que sitúan sobre la imagen tiene un tamaño de celda óptimo de 6×6 píxeles.

Los histogramas de orientaciones se crean acumulando un voto por cada posible orientación, previamente discretizada a cierto rango. Este voto tiene como valor la magnitud de la orientación ya que aunque Dalal y Triggs (2005) probaron con variantes (el cuadrado de la magnitud, su raíz cuadrada, etc.) la magnitud por sí misma es la que mejores resultados proporciona.

La Figura 2, extraída de Dalal y Triggs (2005), muestra de forma gráfica una imagen de prueba y su descriptor HOG asociado.

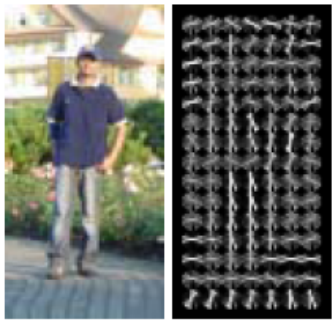

Figura 2: Imagen de prueba (izquierda) y su descriptor HOG asociado (derecha). Imagen extraída de Dalal y Triggs (2005).

Capítulo 3 Algoritmos de reconocimiento de género estudiados

En esta sección se describe el algoritmo de Cao et al. (2008) de reconocimiento de género en imágenes estáticas, cuyos autores llaman Part-Based Gender Recognition (PBGR). Este algoritmo puede considerarse una variante del algoritmo de AdaBoost (Freund y Schapire, 1996) por lo que primeramente se discute éste (Sección 3.1) y posteriormente se pasa a comentar el PBGR (Sección 3.2).

3.1 Algoritmo AdaBoost diseñado por Cao et al. (2008)

El algoritmo AdaBoost es un meta-algoritmo de aprendizaje automático propuesto por Freund y Schapire (1995, 1996) que hace uso de varios clasificadores para mejorar su aprendizaje. Básicamente, lo que hace es crear un clasificador que es la combinación lineal de los resultados de otros muchos clasificadores llamados *débiles*. Estos clasificadores débiles son clasificadores sencillos y rápidos cuyo grado de acierto es ligeramente superior al 50%, lo que hace que su capacidad de sobreajuste a los datos sea mínima.

El algoritmo AdaBoost fue formulado teóricamente por Freund y Shapire (1995) y posteriormente estudiado experimentalmente por estos mismos autores (Freund y Shapire, 1996). AdaBoost fue propuesto originalmente en dos variantes, llamadas M1 y M2. La principal diferencia entre ambas es que en M1 si un clasificador tiene un error superior al 50% en alguna etapa el algoritmo deja de calcular más clasificadores débiles aunque no se haya alcanzado el número de iteraciones máximas (número de clasificadores débiles) que se ha especificado. En cambio la variante M2 permite indicar más de una etiqueta para cada clasificador débil, cada una de ellas acompañada por una probabilidad de que esta etiqueta sea la elegida.

La versión de AdaBoost elegida por Cao et al. (2008) para su primera experimentación es una variante de la propuesta M1 por Freund y Shapire (1995). En realidad, la variante elegida por Cao et al. (2008) es la generalización del algoritmo propuesta por Shapire y Singer (1999) usando como clasificador débil los *decision stumps* más discriminativos usados a su vez por Viola y Jones (2004). A continuación se describe en detalle la implementación de AdaBoost usada por Cao et al. (2008). El clasificador débil que usan, el *decision stump*, se describe en la Sección 3.1.2.

3.1.1 Descripción del algoritmo

Se disponen de n_+ imágenes de hombres y de n_- imágenes de mujeres. Cada una de estas imágenes se representa mediante un vector de características $\mathbf{x}_i = (x_i(1), x_i(2), \ldots,$

$x_i(d)$), donde i es el número de imagen (siendo $1 \leq i \leq n_+ + n_-$), $x_i(g)$ es la característica g del vector de características de la imagen i, (con $1 \leq g \leq d$). Cada vector asociado a una imagen tiene asociada una etiqueta verdadera y_i que indica si es hombre ($y_i = +1$) o mujer ($y_i = -1$).

Primero se explica el clasificador resultante del entrenamiento mediante el AdaBoost y posteriormente se explica el proceso de entrenamiento seguido para obtenerlo.

El clasificador final entrenado es $H(\mathrm{x}) = sign\left(\sum_{t=1}^{T}(\alpha_t h_t(\mathrm{x}))\right)$. Siendo **x** el vector de características de la imagen a clasificar y h_t el clasificador débil para la etapa t, con $1 \leq t \leq T$. Cada clasificador tiene asociado un peso α_t que indica la importancia del mismo. A continuación se indica el proceso de entrenamiento seguido para obtener el clasificador H.

En el entrenamiento, primeramente, el vector de pesos de AdaBoost, D, se inicia para la iteración $t = 1$ de forma que D_1 tiene una entrada para cada posible imagen i de la base de datos, de la forma siguiente: $D_1(i) = 1 / |n_+|$ si $y_i = +1$, y $D_1(i) = 1 / |n_-|$ si $y_i = -1$, siendo $D(i)$ la entrada en el vector D para la imagen i. Esto da lugar a un vector de pesos normalizado para la primera iteración. Para asignar un peso a cada imagen de la base de datos, en cada iteración t del algoritmo se usa un vector de pesos D_t. Este vector de pesos se normaliza al final de cada iteración de forma que, a mayor valor del peso, más difícil resulta clasificar una imagen. Esto permite que el algoritmo se centre en las instancias más difíciles de clasificar, como se verá a continuación.

Para cada iteración t, con $1 \leq t \leq T$, se repite el siguiente proceso:

Basándose en D_t, seleccionar un clasificador débil $h_t(x_i) = h_t(x_i(k_t))$ de forma que k_t sea la característica óptima. Esto quiere decir que cualquier imagen x_i se clasificará de acuerdo a únicamente la característica de la posición k_t. Para elegir la característica óptima se usan los clasificadores débiles *decision stumps* de la misma forma que en Viola y Jones (2004). Este tipo de clasificador, así como el proceso seguido para obtener la característica óptima a partir de ellos, se discute a la finalización de la explicación del presente algoritmo, en la Sección 3.1.2.

El clasificador débil entrenado, h_t, tiene un error de entrenamiento ε_t. Este error se usa para calcular el peso asociado a dicho clasificador débil, es decir el peso α_t, que se calcula como $\alpha_t = 0.5 \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$.

Para finalizar la iteración t se prepara el vector de pesos D para la iteración siguiente, es decir el vector D_{t+1}. Esto se hace de la siguiente forma: $D_{t+1}(i) = \exp(-\alpha_t y_i h_t(\mathbf{x}_i))$. Por último, el vector de pesos D_{t+1} se normaliza de forma que la suma de sus componentes sea 1.

3.1.2 Descripción de los clasificadores débiles *decision stump*

Para elegir la característica óptima k_t, Cao et al. (2008) siguen la misma estrategia de Viola y Jones (2004) de usar *decision stumps*.

Un *decision stump* es un clasificador débil que puede verse como un árbol de decisión de un sólo nivel. Este clasificador, tal y como se define en Viola y Jones (2004), se define con un valor de umbral (que determina hacia qué rama del árbol bajar) y una polaridad (que indica en qué rama se encuentran las etiquetas positivas y negativas). Se hace notar que este clasificador nunca va a tener una tasa de error superior al 50% gracias a la posibilidad de cambiar la polaridad y que el hecho de que el árbol resultante conste de un sólo nivel hace que el clasificador no pueda sobreajustarse a los datos. Esto lo hace un

clasificador débil ideal.

Una forma rápida de obtener un clasificador con *decision stump* para una dimensión dada consiste en tomar un valor aleatorio para el umbral que se encuentre entre los valores mínimo y máximo de las características para dicha proyección y elegir la polaridad que dé la tasa de acierto igual o superior a 50%. No obstante, Cao et al. (2008) no usa esta forma de elegir un *decision stump* puesto que ellos buscan el *decision stump* más discriminativo, es decir, el óptimo. El proceso que siguen para obtenerlo es el descrito por Viola y Jones (2004), cuya idea es como sigue:

Se itera sobre cada una de las d características y, para cada una de ellas, se ordenan los valores de las características g de todas las instancias ($1 \leq g \leq d$). Una vez ordenadas, se itera sobre el vector resultante y, para la polaridad positiva (por ejemplo), se calcula el error resultante de etiquetar todas las muestras positivas de lo que sería la rama izquierda como negativas, y todas las de la rama derecha que sean negativas como positivas. Para la polaridad negativa sería al contrario. El error de una rama viene determinado por la suma de los pesos $D_t(i)$ para la muestra i siempre y cuando ésta requiera un cambio de etiqueta, y el error total es la suma del error de ambas ramas.

Viola y Jones (2004; p. 142) ofrecen un método que permite calcular todo el proceso anterior con una sóla pasada repetida d veces (una por cada una de las características de los vectores) resultando en un coste temporal $O(n \log_2 n)$, acotado por el coste de la ordenación de las características, siendo n el número de muestras.

3.2 Algoritmo Part-Based Gender Recognition de Cao et al. (2008)

La idea en la que se basa el algoritmo propuesto por Cao et al. (2008), llamado Part-Based Gender Recognition (PBGR) es que, aunque seguramente es difícil obtener una regla general por la que una imagen de una persona es de un hombre o de una mujer, hay partes de la figura humana que son más fáciles de clasificar como pertenecientes a cada uno de los géneros. Por ejemplo: las mujeres suelen tener el pelo más largo que los hombres, éstos es más probable que tengan los hombros más anchos, etc. Para capturar estas diferencias dividen la imagen en una rejilla de 6×19 partes, solapadas entre sí. Aunque el artículo no lo indica, por el intercambio de correos con los autores se sabe que el solapamiento entre ambas ventanas es del 50% en ambas dimensiones. Para cada una de estas partes se obtiene un vector de características mediante el descriptor HOG (Dalal y Triggs, 2005). A diferencia del detector de Dalal y Triggs (2005), cada parte se considera de forma individual en vez de concatenarse en un sólo vector de características.

En el algoritmo que proponen, cada una de estas partes aporta una clasificación al género de la imagen, y todas estas opiniones se combinan mediante un algoritmo de *boosting*. El algoritmo propuesto por Cao et al. (2008) se basa en el AdaBoost explicado en la sección anterior y es el siguiente:

El propósito del algoritmo consiste en construir un clasificador como el de AdaBoost: una sucesión de clasificadores débiles que, combinados linealmente, indican a qué clase se cree que pertenece una imagen. El clasificador adopta, pues, esta forma: $H(I) = sign\left(\sum_{t=1}^{T} (\alpha_t h_t)\right)$, siendo I la imagen de la persona que se quiere clasificar, h_t el clasificador de la etapa t (con $1 \leq t \leq T$), y α_t el peso asociado a dicho clasificador. Hasta aquí, todo es igual que en AdaBoost. Lo que cambia es el proceso de entrenamiento del clasificador H, que se comenta a continuación.

Antes de detallar el proceso seguido en el entrenamiento, se explica la representación de las imágenes elegida por Cao et al. (2008). Cada imagen i tiene asociado un conjunto

de vectores $\mathbf{x}_i$, con tantos elementos como partes P en que se ha dividido la imagen. Es decir, $\mathbf{x}_i = \{\boldsymbol{x}_i(p)\}$, con $1 \leq p \leq P$, y siendo $\boldsymbol{x}_i(p)$ el vector de características obtenido para la parte p mediante HOG (ver el Capítulo 2 para una descripción de este extractor de características)

El algoritmo de entrenamiento del algoritmo PBGR es como sigue:

Para cada iteración t, con $1 \leq t \leq T$, se repite el siguiente proceso:

Se selecciona la parte p_t más discriminativa.

Una vez elegida p_t, se entrena un clasificador débil de la misma forma que se describió cuando se hablaba de AdaBoost (ver la Sección anterior), pero considerando únicamente los vectores de la parte seleccionada, esto es $\{\boldsymbol{x}_i(p_t)\}$ para todas las imágenes i: $h_t(\mathbf{x}) = h_t(\boldsymbol{x}(p_t))$.

Se calcula el peso α_t asociado al clasificador de la etapa t de la misma forma que AdaBoost.

Capítulo 4 Experimentación

Se describe a continuación la experimentación realizada a lo largo del trabajo de final de máster. Se comienza describiendo la base de datos de imágenes utilizada para la realización de la experimentación (Sección 4.1), se pasa después a comentar brevemente los recursos *software* utilizados (Sección 4.2), y se sigue con la descripción en detalle de las pruebas realizadas (Sección 4.3).

4.1 Base de datos de imágenes utilizada

Tanto Cao et al. (2008) como Collins et al. (2009) (aunque éste último sólo parcialmente), y posteriormente Guo et al. (2010) usan la base de datos de peatones del MIT (Oren, 1997). Esta base de datos consta de 924 imágenes en formato PPM de dimensiones 64×128 píxeles, mostrando vistas de peatones tanto de frente como de espaldas. La Figura 3 muestra ejemplos de imágenes de esta base de datos.

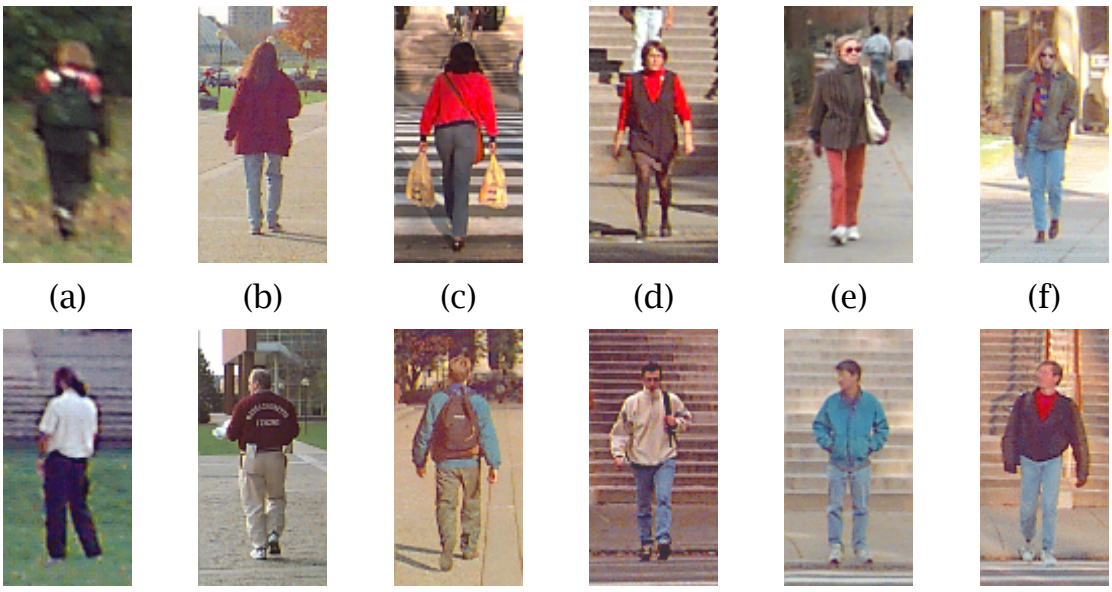

Figura 3: Diversos ejemplos de imágenes de la base de datos de imágenes del MIT (Oren, 1997) etiquetadas por Cao et al. (2008) como mujeres de espaldas (a, b, c), de frente (d, e, f), hombres de espaldas (g, h, i), y de frente (j, k, l)

Esta base de datos se creó inicialmente para evaluar algoritmos de detección de personas por lo que no existe originariamente un etiquetado de las imágenes de acuerdo con el género de la persona que aparece en las mismas. El primer etiquetado de género de esta base de datos, y el que se usa en este proyecto, es el proporcionado por Cao et al. (2008) en el que se distinguen 600 imágenes de hombres y 288 imágenes de mujeres, descartando el resto de imágenes por ser imposible incluso para ojos humanos el determinar el género de las personas que aparecen. De estas imágenes, un 47% corresponden a vistas frontales y un 53% a vistas traseras.

Collins et al. (2009) usan un etiquetado diferente de esta base de datos, aunque obtienen un número similar de hombres y mujeres al propuesto por Cao et al. (2008) en su etiquetado. El etiquetado distinto se debe a que, en el momento de la redacción de este artículo, Cao et al. (2008) todavía no habían hecho público su etiquetado. Collins et al. (2009) comentan que probablemente la diferencia entre su etiquetado y el propuesto por Cao et al. (2008) seguramente se debe a las imágenes que unos y otros han decidido dejar fuera del etiquetado debido a la dificultad para determinar si eran hombres o mujeres.

Antes de que el etiquetado de Cao et al. (2008) fuera disponible se realizaron las experimentaciones con un etiquetado propio que se creó para la experimentación en el presente proyecto. Para este etiquetado se marcaron 139 imágenes como de género inclasificable, o bien difícilmente clasificable. Cao et al. (2008) sólo marcan 31 imágenes como inclasificables. Esto da una idea de la dificultad de la base de datos usada y como muestra se enseñan las imágenes de la Figura 4 que han sido etiquetadas por Cao et al. (2008) y cuyo género no nos parece evidente.

(a) (b) (c)

Figura 4: Ejemplos de imágenes problemáticas *de la base de datos de Cao et al. (2008) que han sido etiquetadas como a) mujer de espaldas, y b) y c) hombre de frente.*

Durante toda la experimentación descrita en este capítulo se usa el etiquetado proporcionado por Cao et al. (2008) con la finalidad de poder comparar los resultados obtenidos con los informados por los autores del primer estudio.

4.2 Recursos informáticos *software* usados

El desarrollo del proyecto se ha realizado usando los siguientes recursos informáticos de *software*:

- **Lenguaje de programación MATLAB**: Este lenguaje se usó en las primeras fases del desarrollo del proyecto para el prototipado rápido de algunas ideas.

Su uso se descartó en fases tempranas del proyecto puesto que su lentitud de ejecución lo hacía inadecuado para la experimentación intensiva para el caso particular de este proyecto. La experimentación con clasificadores uni-clase (Sección 4.7) se realizó también usando este lenguaje.

- **Lenguaje de programación C++**: Es el lenguaje de la implementación del HOG (ver Capítulo 2). Inicialmente el HOG estaba implementado en MATLAB pero, por cuestiones de velocidad de ejecución, tuvo que reescribirse usando C++.
- **Librería OpenCV (Bradski y Kaehler, 2008)**: OpenCV (Open Source Computer Vision) es una librería para visión en tiempo real, cuya API se encuentra disponible en C, C++ y Python. Se usó en su versión 1.1pre1 para la implementación del HOG en C++.
- **Lenguaje de programación C#**: Es un lenguaje de programación orientado a objetos con una sintaxis muy similar a Java. Se eligió para poder implementar el método de Cao et al. (2008) (ver Sección 3.2) con mayor rapidez de la que hubiera sido posible usando C++ al tiempo que se mantenía una velocidad de ejecución comparablemente superior a la que proporciona MATLAB. Se eligió este lenguaje de programación, además, por la posibilidad de usar desde él la librería Emgu CV[1], que es un reconocido *wrapper* para .NET de la librería OpenCV, aunque no hizo falta usar Emgu CV durante el proyecto.
- **Librería NUnit**: NUnit es un *framework* de pruebas unitarias para la plataforma .NET, inspirado en JUnit en sus primeras versiones. El código del proyecto desarrollado usando C# ha sido testeado mediante las pruebas unitarias escritas usando esta librería. Esto da la seguridad de que los cambios introducidos durante el desarrollo del mismo, así como el cumplimiento de los requisitos funcionales del código, se cumplen.
- **Librería DDtools (Tax, 2010)**: Es una librería para MATLAB que proporciona "herramientas, clasificadores y funciones de evaluación para la investigación de clasificadores uni-clase" (Tax, 2010). Hace uso de la librería PRtools (Heijden et al., 2004) para el reconocimiento de patrones en MATLAB.

4.3 Diseño de los experimentos

Cao et al. (2008) usan una validación cruzada de 5-fold (Alpaydin, 2010; pp. 486-488) para obtener los resultados de los que informan. En general, la realización de una validación cruzada *k*-fold implica dividir el conjunto de muestras en *k* grupos disjuntos, de forma que *k-1* grupos se usan para entrenamiento y el grupo restante para validación, y los grupos se rotan de forma que cada muestra del conjunto se usa una única vez como muestra de validación durante todo el proceso.

En el desarrollo de las experimentaciones se ha comprobado que este procedimiento puede dar lugar a resultados demasiado pesimistas u optimistas, dependiendo de la "suerte" en la creación de los *folds*. Por ello, y salvo cuando se indica lo contrario, todos los experimentos realizados cuyos resultados se muestran en esta memoria han sido obtenidos mediante un procedimiento inspirado en el propuesto por Bouckaert (2003) que se basa en repetir varias veces una validación cruzada y proporcionar la media de los resultados obtenidos. Bouckaert (2003) indica que el método óptimo para obtener un resultado que sea lo más fácilmente reproducible por otros investigadores de forma que se reduzca la variabilidad entre resultados obtenidos es repetir 10 veces un 10-fold, sin embargo aquí, por falta de tiempo, se ha decidido repetir tres veces un 5-fold, lo que da un resultado menos preciso que el propuesto por él, pero sí más realista que si se

1 http://www.emgu.com/

realizara únicamente un 5-fold tal y como hace Cao et al. (2008).

4.4 Experimentación con el AdaBoost elegido por Cao et al. (2008)

Se experimentó con el AdaBoost elegido por Cao et al. (2008), cuya explicación del algoritmo se encuentra en la Sección 3.2 de la presente memoria de máster. Se eligió el *grid* que Cao et al. (2008) sugieren que usan, 3×3, lo que resulta en una dimensión para los vectores $\mathbf{x}_i$ de $d = 72$ (9 ventanas debidas al *grid* 3×3 y 8 posibles orientaciones para el HOG).

Cao et al. (2008) indican unas tasas de acierto del 70,9 ± 4,4% para la vista frontal y de 63,0 ± 4,1% para la vista trasera, lo cual tiene cierto paralelismo con la mayor facilidad que tenemos los humanos para determinar el género de una persona según si ésta se nos muestra de frente o nos da la espalda. No indican ninguna tasa de acierto para la vista mixta (combinando las imágenes frontales y traseras) y tampoco el número máximo de iteraciones realizado por el algoritmo AdaBoost para llegar a ese porcentaje.

Para reproducir el experimento se usó un número creciente de iteraciones. Es decir, se creó un reconocedor que usaba cierto número de clasificadores débiles del tipo *decision stump* (ver Sección 3.1.2). Se hicieron pruebas para 400, 600 y 800 clasificadores débiles combinados. La cota inferior para el número de clasificadores se determinó experimentalmente hasta que más o menos se alcanzó la tasa comunicada por Cao et al. (2008). Esto tuvo que hacerse así puesto que los autores no indican en su artículo el número de clasificadores que usan. No se probó con un número mayor de clasificadores por requerir éstos un tiempo computacional elevado y para no sobreajustarse a los datos.

Los resultados obtenidos para la ejecución de la implementación propuesta por Cao et al. (2008) de AdaBoost se recogen en la Tabla 1.

Tabla 1: Tasas de acierto obtenidas en la reproducción de la experimetnación y resultados comunicados por Cao et al. (2008) para el algoritmo AdaBoost con las características de HOG y una rejilla de 3×3, con T iteraciones para AdaBoost ($T \in \{400, 600, 800\}$)

		Vista		
		Frontal	**Trasera**	**Mixta**
Resultado obtenido con implementación propia	**T = 400**	67,6 ± 3,6	58,6 ± 4,7	61,9 ± 3,6
	T = 600	68,6 ± 4,0	58,1 ± 5,6	62,2 ± 3,0
	T = 800	69,2 ± 3,4	58,8 ± 5,4	61,8 ± 2,4
Resultado comunicado por Cao et al. (2008)		70,9 ± 4,4	63,0 ± 4,1	No comunicado

4.5 Experimentación con Part-Based Gender Recognition de Cao et al. (2008)

Cao et al. (2008) prueban con una configuración de 6×19 ventanas sobre la imagen, con un 50% de solapamiento en ambas dimensiones y obtienen 75,0 ± 2,9 % de acierto.

Los resultados obtenidos mediante la reproducción del algoritmo comentado en la Sección 3.2 se muestran en la Tabla 2.

Tabla 2: Tasas de acierto (%) obtenidas en la reproducción de la experimentación y resultados conunicados por Cao et al. (2008) para el algoritmo PBGR con las características de HOG y una rejilla de 6×19, para 400 iteraciones del algoritmo.

	Vista frontal	**Vista trasera**	**Vista mixta**
Resultado obtenido con implementación propia	75,2 ± 3,8	72,6 ± 4,6	73,5 ± 2,9
Resultado comunicado por Cao et al. (2008)	76,0 ± 1,2	74,6 ± 3,4	75,0 ± 2,9

Los resultados obtenidos se consideran comparables a los comunicados por Cao et al. (2008) ya que, aunque las tasas obtenidas son ligeramente superiores a las comunicadas por Cao et al. (2008), éstos son debidos a la media de los resultados de los tres 5-folds realizados, mientras que los resultados de Cao et al. (2008) se obtienen a partir de un único 5-fold.

4.6 Experimentación con AdaBoost usando distintas configuraciones de ventanas

Resulta interesante comparar los resultados obtenidos por Cao et al. (2008) con los obtenidos usando un algoritmo de aprendizaje más sencillo (AdaBoost). Cabe preguntarse, sobre todo comparando los resultados obtenidos en esta sección con los obtenidos en la siguiente, si el reconocimiento de género por personas no es más

dependiente del número y situación de las ventanas sobre las que se extraigan las características que del algoritmo usado para aprender a partir de éstas.

En esta sección se estudia el efecto de la situación de las ventanas y se comparan distintas configuraciones de las mismas contra los resultados obtenidos por el algoritmo PBGR de Cao et al. (2008) (Sección 4.5).

Para comprobar esta hipótesis, es decir, si importa más la distribución de las ventanas sobre la imagen que los algoritmos utilizados, se han realizado algunas pruebas variando la densidad de la malla de las ventanas.

Dado que lo que interesa es la obtención de un algoritmo de reconocimiento de género general (es decir, tanto de vista frontal como de vista trasera), únicamente se han comparado los resultados obtenidos para esta combinación de vistas (en esta sección denominada "vista mixta").

4.6.1 Pruebas con rejillas más densas sobre toda la imagen

Se probaron rejillas más densas que la propuesta por Cao et al. (2008). en concreto se probó con una rejilla el doble de densa (de 12×38 ventanas) y otra de densidad intermedia (de 9×29 ventanas). Los resultados para ambas configuraciones, para la base de datos de vistas mixtas (frontales y de espaldas combinadas), se muestran en la Tabla 3.

Tabla 3: Tasas de acierto (%) para la vista mixta (combinando imágenes de las vistas frontales y traseras) para distintas configuraciones de las ventanas.

Tamaño de la rejilla	Tasa de acierto (%)
6×19	73,5 ± 2,9
9×29	77,4 ± 2,4
12×38	76,1 ± 2,4

Como puede observarse, se obtienen tasas de acierto ligeramente superiores a la proporcionada por Cao et al. (2008). Parece que una densidad excesiva de la rejilla que configura la disposición de las ventanas hace que el resultado se degrade. Es por ello que se ha realizado otro estudio más exhaustivo para determinar el tamaño óptimo de rejilla, variándolo desde el 6×19 original a 12×35, en intervalos de 2 en número de ventanas dispuestas a lo ancho y en intervalos de 8 en número de ventanas dispuestas a lo alto, siempre manteniendo un 50% de solapamiento en ambas dimensiones entre las ventanas. La Tabla 4 muestra las tasas de acierto obtenidas para las distintas configuraciones. Por cuestiones de tiempo, en este caso, únicamente se ha realizado un 5-fold para la obtención de cada resultado mostrado, aunque la configuración de cada partición fue la misma para todas las pruebas, por lo que los resultados son comparables entre sí.

Como puede verse, prácticamente todas las configuraciones de rejilla más densas que la usada por Cao et al. (2008), es decir, 6×19, proporcionan mejores resultados (a excepción de 6×35). Parece que hay cierta tendencia a elevarse la tasa de acierto cuanto más densa es la ventana, si bien el tope parece situarse en torno al 77% de acierto. Esto es un 2% más elevado que el resultado proporcionado por Cao et al. (2008) para la vista mixta, aunque todavía es un 4% más bajo que el resultado comunicado por Guo et al. (2010), si bien éstos usan características diferentes.

Tabla 4: Tasas de acierto (%) para distintas configuraciones de rejilla $N \times M$ para la experimentación usando AdaBoost.

$N \times M$ ventanas		M		
		19	*27*	*35*
N	*6*	71,6 ± 2,2	76,4 ± 3,1	73,8 ± 3,5
	8	75,3 ± 1,9	76,9 ± 3,2	75,5 ± 0,7
	10	77,1 ± 3,5	76,6 ± 1,5	**77,1 ± 1,9**
	12	75,8 ± 1,6	75,8 ± 4,3	77,1 ± 3,3

A la vista de los resultados obtenidos en la Tabla 4 se ha determinado como configuración óptima de rejilla 10×35, ya que obtiene la tasa de acierto mayor con menor variabilidad: 77,1 ± 1,9 %.

Se ha usado esta configuración de rejilla para obtener las tasas de acierto para la vista frontal y de espaldas de las imágenes, con el fin de compararse con la tasa ofrecida por Cao et al. (2008) para estas dos vistas, y para compararse con el enfoque de Collins et al. (2009), quienes sólo informan de su tasa de acierto para el caso de las imágenes con vista frontal.

Los resultados para estas dos vistas se muestran en la Tabla 5, así como los resultados para la vista mixta (el resultado proporcionado difiere del de la Tabla 4 puesto que aquí se ha usado un 3×5-fold para la obtención de dicha tasa de acierto, en vez de un único 5-fold como en el caso anterior)

Tabla 5: Tasas de acierto (%) para una rejilla óptima de 10×35 con solapamiento del 50% en cada dimensión, para AdaBoost con características HOG.

Vista	**Tasa de acierto**		
	Global	**Hombres**	**Mujeres**
Frontal	77,8 ± 3,4	90,2 ± 2,0	43,9 ± 13,8
Trasera	77,2 ± 3,1	84,9 ± 4,4	64,4 ± 6,5
Mixta	76,2 ± 3,0	86,0 ± 2,7	55,8 ± 8,2

Observando la Tabla 5, que también muestra la tasa de acierto para cada una de las dos clases a considerar (hombres y mujeres), cabe plantearse si la base de datos que se está usando actualmente (Sección 4.1) para el estudio de los algoritmos de reconocimiento de género, es la adecuada. Puede observarse una tendencia clara a reconocer muy bien las muestras de la clase mayoritaria (hombres), pero muy mal a las muestras de la clase minoritaria (mujeres). La Sección 4.7, más adelante, tratará de sacar partido de esta observación para crear un clasificador de una única clase que trate de reconocer a las mujeres por contraposición a los hombres.

4.7 Experimentación con el enfoque *one-class classification*

A la vista del desequilibrio en cuanto al número de muestras existente entre ambas clases (hombres y mujeres) de la base de datos de imágenes usada (Sección 18), y habiendo observado las tasas de acierto para ambos sexos (Tabla 5), se planteó la cuestión de si no sería interesante entrenar al clasificador para que fuera capaz de

reconocer muy bien a la clase más numerosa (en este caso, los hombres), y que por lo tanto mejor sabe clasificar y, en caso de que una muestra no fuera lo bastante parecida a un hombre, descartarlo y decir que era de la clase minoritaria (en este caso, la clase mujer).

Este enfoque se conoce en la literatura como "clasificación de una clase", *one-class classification*, (Tax, 2001) y se basa en el hecho de que únicamente los objetos de una clase (o de varias) están disponibles para realizar el entrenamiento, y esta(s) clase(s) forman lo que se llaman los *objetos objetivo*, y a los otros objetos que no pertenencen a esta(s) clase(s) se les llama *objetos outliers*.

Se pensó que, para este proyecto en particular, sería interesante probar con este enfoque. Para ello se usó la librería *DDtools* (Tax, 2010), que implementa varios clasificadores uni-clase que clasifican de acuerdo al enfoque comentado anteriormente. Para la obtención de los resultados mostrados a continuación se usó una única repetición de una validación cruzada de 10-fold y se muestra el área bajo la curva (AUC, por su siglas en inglés) de una curva ROC (*Receiver Operating Characteristic*) (Fawcett, 2006) obtenido para cada clasificador de los que se han probado.

La Tabla 6 muestra los resultados obtenidos para una rejilla de 3×3, sin solapamiento, y la Tabla 7 hace lo propio para la rejilla de Cao et al. (2008) (es decir 9×19 con solapamiento del 50% en cada dimensión). La Tabla 8 muestra los resultados para la rejilla óptima encontrada en la Sección 8 (10×35 con 50% de solapamiento en cada dimensión).

Para la realización de los experimentos se probó con varios clasificadores (aunque no se probó con las máquinas de soporte vectorial), y sólo se muestran tres de los clasificadores que dieron resultados más altos:

- ***k-centers***. (Ypma y Duin, 1998) Es similar a *k-means*. Se crean *k* grupos de forma que la distancia máxima de los elementos de un grupo está minimizada.
- **Vecino más cercano simple**. Es como el vecino más cercano, pero la distancia de una instancia *d* de la instancia *A* a su vecino más cercano *B* se normaliza según la distancia de *B* a su vecino más cercano *C* (Tax, 2001).
- **Vecino más cercano** (considerando un sólo vecino).

Tabla 6: Áreas bajo la curva (AUC) de las curvas ROC para algunos clasificadores uni-clase usando las características del HOG para la vista mixta (frontal y trasera) con una rejilla de 3×3 sobre las imágenes, sin solapamiento.

Clasificador	AUC
k-centros	0,51 ± 0,11
Vecino más cercano simple	0,52 ± 0,06
Vecino más cercano	0,49 ± 0,09

Tabla 7: Áreas bajo la curva (AUC) de las curvas ROC para algunos clasificadores uniclase usando las características del HOG para la vista mixta (frontal y trasera) con una rejilla de 9×19 sobre las imágenes, con solapamiento del 50% en cada dimensión (rejilla usada por Cao et al. (2008))

Clasificador	**AUC**
k-centros	0,49 ± 0,11
Vecino más cercano simple	0,54 ± 0,07
Vecino más cercano	0,48 ± 0,06

Tabla 8: Áreas bajo la curva (AUC) de las curvas ROC para algunos clasificadores uniclase usando las características del HOG para la vista mixta (frontal y trasera) con una rejilla de 10×35 sobre las imágenes, con solapamiento del 50% en cada dimensión (rejilla considerada óptima según los resultados obtenidos en la Sección 4.6.1)

Clasificador	**AUC**
k-centros	0,53 ± 0,07
Vecino más cercano simple	0,52 ± 0,07
Vecino más cercano	0,54 ± 0,07

De acuerdo con Fawcett (2006), un área bajo la curva (AUC) cercana a 0,5 indica que el clasificador es prácticamente aleatorio, y esto es lo que se observa a la vista de los resultados. Se aprecia cierta mejora si se usa la configuración de ventanas determinada como óptima (Tabla 8) en la Sección 4.6.1, lo que de alguna forma valida esta elección de rejilla. No obstante, los resultados siguen siendo tan bajos que, o bien este enfoque de *one-class classification* no sirve para este problema, o bien no se le ha sabido sacarle todo el partido, algo que se deja como posible trabajo futuro.

Capítulo 5 Conclusiones y trabajo futuro

A la vista de los resultados experimentales obtenidos en el Capítulo 4, las principales conclusiones que se desprenden del trabajo son las siguientes:

5.1 Disposición de la rejilla de ventanas sobre las imágenes

Experimentalmente se ha comprobado que una disposición de ventanas de acuerdo a una rejilla densa aporta mejores resultados en clasificación de género que una rejilla menos densa. Cao et al. (2008) optan por una rejilla de 6×19 con solapamiento del 50% en cada dimensión para la clasificación usando su algoritmo PBGR (Sección 3.2). Experimentalmente se ha comprobado que rejillas más densas (10×19, 10×35 o 12×35) requieren de algoritmos menos complejos (AdaBoost, ver Sección 3.1) para superar las tasas de acierto proporcionadas por Cao et al. (2008) y Collins et al. (2009), para la base de datos de imágenes del MIT (Oren, 1996).

Resulta curioso que para la detección de personas no haga falta una rejilla tan densa como para determinar el género de estas mismas personas. El detector de personas de Dalal y Triggs (2005) que hace uso de las características HOG (Capítulo 2) alcanza su resultado óptimo con una rejilla de 10x21 (usando grupos de 3×3 bloques, cada uno de los cuales de 6×6 píxeles, según los resultados óptimos comunicados por Dalal y Trigs (2005), y considerando la base de datos de imágenes del MIT (Oren, 1997)). Esto contrasta con la rejilla de 10×35 que se usa en la Sección 4.6 para determinar el género de las personas con tasas de acierto aceptables (~ 77%).

Da la sensación de que una rejilla más densa le permite al algoritmo centrarse en las zonas concretas de la imagen que son más discriminantes, obviando el contexto que las rodearía si las ventanas fueran más grandes. Este hecho puede deberse a, al menos, una de las dos causas siguientes, o bien a una combinación de las mismas:

1) El género de una persona depende de detalles concretos de su anatomía, tal y como supone Cao et al. (2008) y parece reforzar Yu et al. (2009); por lo que
2) se está usando una rejilla equivocada en el sentido de que hay partes de la imagen que no aportan información sobre el género de una persona y se están teniendo en cuenta igualmente.

Se cree que la primera hipótesis es correcta ya que la anatomía de hombres y mujeres es notablemente diferente en ciertas partes del cuerpo y, por ejemplo, a las personas nos es más fácil reconocer el género de una persona si la miramos de cintura para arriba que si miramos de cintura para abajo (Yu et al., 2009). Como trabajo futuro habría que considerar rejillas de ventanas no uniformes, por ejemplo: más grandes en zonas poco importantes, y con más detalle en las zonas que más permitan discriminar el género de una persona (e.g. cabeza, hombros, tórax y caderas), así como estudiar la forma de obtener estas ventanas de forma automática (e.g. a partir de muchas ventanas en

posiciones y tamaños aleatorios). Falta por hacer un estudio de la importancia de cada una de las ventanas de la rejilla óptima considerada en la experimentación de la Sección 4.6.1.

5.2 Clasificadores de género estudiados

En el presente trabajo se han estudiado dos algoritmos de clasificación: AdaBoost (Sección 3.1) y el algoritmo propuesto por Cao et al. (2008), basado también en AdaBoost, llamado Part-Based Gender Recognition (PRBG) (Sección 3.2).

En la experimentación de la Sección 4.6 se ha visto que la configuración de la rejilla que dispone las ventanas sobre la imagen es más importante que el propio algoritmo de clasificación, al menos para el algoritmo propuesto por Cao et al. (2008), ya que una rejilla densa usando AdaBoost supera en tasa de acierto la obtenida para distribuciones de ventanas menos densas usando PBGR. Falta por estudiar el impacto de ventanas más densas para este último algoritmo, algo que se deja para trabajo futuro.

5.3 Base de datos de imágenes usada

La base de datos de imágenes del MIT (Oren, 1997) que se está usando mayoritariamente en la literatura se encuentra tan desequilibrada en cuanto a muestras de cada clase, y algunas de las imágenes tienen tan poca calidad (ver Figuras 3a y 4a), que se duda de que sea adecuada para seguir estudiando este problema. Tanto Cao et al. (2008) como Guo et al. (2010) no tienen en cuenta esta característica, ni tratan el desbalance al realizar sus experimentos. Experimentalmente se ha comprobado en la Sección 4.6 que las tasas de acierto con esta base de datos se deben, sobre todo, a la gran cantidad de muestras de la clase "hombre" de la base de datos, que hace que casi siempre se clasifiquen bien las muestras de esta clase. Collins et al. (2009) se da cuenta de esto y propone bases de datos de imágenes alternativas que deberían ser consideradas en trabajos futuros.

Otro enfoque a considerar puede ser asumir el desequilibrio entre las muestras de cada clase y seguir usando la base de datos del MIT (Oren, 1997), aunque tratando este balanceo de una forma similar a como hacen Martín-Félez et al. (2010) en su trabajo de clasificación de género partiendo de secuencias de vídeo, en el que crean hasta 25 clasificadores entrenados con grupos formados por todas las muestras de la clase minoritaria y el mismo número de muestras elegidas aleatoriamente de la clase mayoritaria.

Capítulo 6 Referencias

E. Alpaydin, *Introduction to Machine Learning*, 2nd edition, The MIT Press, United States, 2010.

R.R. Bouckaert, "Choosing between Two Learning Algorithms Based on Calibrated Tests", *International Conference on Machine Learning*, pp. 51-58, 2003.

G. Bradski y A. Kaehler, *Learning OpenCV: Computer Vision with the OpenCV Library*, O'Reilly, United States, 2008. (La página de descarga de la librería OpenCV es http://opencv.willowgarage.com/wiki)

J. Canny, "A Computational Approach To Edge Detection", *IEEE TPAMI*, Vol. 8, No. 6, pp. 679-698, 1986.

L. Cao, M. Dikmen, Y. Fu, y T.S. Huang, "Gender recognition from body", *MM'08: Proceeding of the 16th ACM international conference on Multimedia*, pp. 725-728, New York, NY, USA, 2008.

M. Collins, J. Zhang, P. Miller, y H. Wang, "Full Body Image Feature Representations For Gender Profiling", *ICCV 2009: IEEE Workshop on Visual Surveillance*, 2009.

N. Dalal y B. Triggs, "Histograms of oriented gradients for human detection", *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

T. Fawcett, "An introduction to ROC Analysis", *Pattern Recognition Letters*, Vol. 27, Issue 8, pp. 882-891, 2006.

P.F. Felzenszwalb y D.P. Huttenlocher, "Pictorial structures for object recognition", *International Journal of Computer Vision*, Vol. 61, No. 1, pp. 55-79, 2005.

Y. Freund y R.E. Schapire, "A decision-theoretic generalization of online learning and an application to boosting", *Computational Learning Theory: Second European Conference, EuroCOLT '95*, pp. 23-37, Springer-Verlag, 1995.

Y. Freund and R.E. Schapire, "Experiments with a new boosting algorithm," *International Conference on Machine Learning*, 1996.

G. Guo, Y. Fu, C.R. Dyer, y T.S. Huang, "Image-based human age estimation by manifold learning and locally adjusted robust regression", *IEEE Transactions on Image Processing*, Vol. 17, No. 7, pp. 1178-118, 2008.

G. Guo, G. Mu, y Y. Fu, "Gender from Body: A Biologically-Inspired Approach with Manifold Learning", *ACCV 2009*, Part III, LNCS 5996, pp. 236-245, 2010.

S. Gutta y H. Wechsler, "Gender and ethnic classification of human faces using hybrid classifiers", *International Joint Conference on Neural Networks*, Vol. 6, pp. 4084-4089, 1999.

F. van der Heijden, R.P.W. Duin, D. de Ridder y D.M.J. Tax, *Classification, parameter estimation and state estimation - an engineering approach using Matlab*, John Wiley & Sons, England, 2004. http://www.prtools.org/book.html

Historical Boys' Clothing Web, *Color and Gender, 2007. http://histclo.com/gender/color.html*

A.K. Jain y A. Ross, "An introduction to biometric recognition", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 14, No. 1, 2004.

Z. Lin y L. S. Davis, "Shape-based human detection and segmentation via hierarchical part-template matching", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 32, No. 4, 2010.

X. Lu y A.K. Jain, "Ethnicity identification from face images", *Proceedings of SPIE Conference of Biometric Technology for Human Identification*, Vol. 5404, pp. 114-123, 2004.

R. Martín-Félez, R.A. Mollineda y J.S. Sánchez, "A gender recognition experiment on the CASIA gait database dealing with its imbalance nature", *International Conference on Computer Vision Theory and Applications (VISAPP 2010)*, pp 439-444, 2010.

M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, y T. Poggio, "Pedestrian detection using wavelet templates", *Computer Vision and Pattern Recognition*, pp. 193-199, 1997. (La base de datos de imágenes del MIT está disponible en la página web: http://cbcl.mit.edu/cbcl/software-datasets/PedestrianData.html.)

M. Pedersoli, J. Gonzàlez, y J.J. Villanueva, "High-speed human detection using a multiresolution cascade of histograms of oriented gradients", *IbPRIA, Lecture Notes in Computer Science*, Vol. 5524/2009, pp. 48-55, 2009.

R. E. Shapire y Y. Singer, "Improved Boosting Algorithms Using Confidence-rated Predictions", *Machine Learning*, Vol. 37, pp.297-336, 1999.

D.M.J. Tax, *One-class classification*, PhD Thesis, Delft University of Technology, Delft, 2001.

D.M.J. Tax, *DDtools, the Data Description Toolbox for Matlab*, (version 1.8.0), 2010.

P. Viola y M.J. Jones, "Rapid object detection using a boosted cascade of simple features", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 511-518, 2001.

P. Viola y M.J. Jones, "Robust Real-Time Face Detection", *International Conference on Computer Vision*, Vol. 57, No. 2, pp. 137-154, 2004.

P. Viola, M.J. Jones, y D. Snow, "Detecting pedestrians using patterns of motion and appearance", *Proceedings of the 9th International Conference on Computer Vision*, Vol. 2, pp. 734-742, 2003.

A. Ypma y R.P.W. Duin, "Support obects for domain approximation", *ICANN'98*, 1998.

S. Yu, T. Tan, K, Huan, K. Jia, y X. Wu, "A study on gait-based gender classification", *IEEE TPAMI*, Vol. 18, No. 8, pp. 1905-1910, 2009.