UNIVERSIDAD AUTÓNOMA DE MADRID
ESCUELA POLITÉCNICA SUPERIOR

Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación

TRABAJO FIN DE GRADO

DETECCIÓN DE PERSONAS MEDIANTE REDES CONVOLUCIONALES

Esther Sánchez Atienza
Tutor: Álvaro García Martín
Ponente: José María Martínez Sánchez

Julio 2017

DETECCIÓN DE PERSONAS MEDIANTE REDES CONVOLUCIONALES

Esther Sánchez Atienza
Tutor: Álvaro García Martín
Ponente: José María Martínez Sánchez

Video Processing and Understanding Lab
Dpto. de Tecnología Electrónica y de las Comunicaciones
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Julio 2017

Trabajo parcialmente financiado por el Ministerio de Economía y Competitividad del Gobierno de España bajo el proyecto TEC2014-53176-R (HAVideo) (2015-2017)

Resumen

Actualmente vivimos un periodo en el que la detección de personas está a la alza, siendo muy importante en algunos aspectos de la sociedad como la vídeo seguridad. De hecho, muchos investigadores crean sus propios algoritmos para la detección de objetos obteniendo resultados excelentes sobre bases de datos elegidas por ellos mismos. Pero el problema surge, principalmente, a la hora de generar un modelo que abarque cuantas más características mejor, para hacerlo capaz de sobreponerse a cambios de postura, de movimiento, de iluminación, de interacción entre personas o cambios en el punto de vista. Para entrenar dichos modelos, es cada vez más popular el uso de un nuevo tipo de aprendizaje, el *Deep Learning*, que maneja grandes cantidades de datos con el que se consigue solventar dichos problemas intentando simular el comportamiento de las neuronas del cerebro humano. En este aprendizaje es en el que se basan los nuevos algoritmos para la detección de personas a los que podemos denominar detectores basados en redes convolucionales.

Por todo esto dicho, el objetivo de este trabajo de fin de grado es el de ofrecer una comparativa entre los algoritmos de detección de personas «tradicionales» (*Deformable Parts Model*, DPM, *y Aggregate Channel Features*, ACF) y los nuevos detectores «modernos» basados en redes convolucionales (*Faster Region-based Convolutional Network*, FASTER R-CNN), estudiados en el estado del arte. Para ello, se diseñarán e implementarán diferentes modelos de persona y se evaluarán los resultados conseguidos por los detectores elegidos sobre una base de datos y métrica en común en igualdad de condiciones.

Palabras clave

Detección de personas, base de datos, modelo, Deep Learning, redes convolucionales y métrica de evaluación.

Abstract

Nowadays we live in a period in which the people detection is on the rise, being very important in some aspects of society like video security. As a matter of fact, many researchers create their own algorithms for object detection obtaining excellent results over databases chosen by themselves. But the problem arises, mainly, at the time of generating a model that includes as much characteristics as possible, to make it capable of overcoming changes in posture, movement, illumination, the interaction among people or changes in the point of view. To train the models, is becoming more popular the usage of a new kind of learning, know as *Deep Learning*, this handles huge amounts of data that solves those problems trying to simulate the behaviour of neurons of the human brain. New algorithms are based in *Deep Learning* for people detection, which are named as detertors based on convolutional networks.

With all this being said, the aim of this final degree project is to offer a comparative between the tradicional algorithms of people detection (*Deformable Parts Model*, DPM, and *Aggregate Channel Features*, ACF) and the modern ones based on convolutional networks (*Faster Region-based Convolutional Network*, FASTER R-CNN), studied in the state of the art. For the achievements of this, different models of people will be designed and implemented and the obtained results will be evaluated by the chosen detectors over a database and metrics in common, in equal conditions.

Keywords

People detection, database, model, *Deep Learning*, convolutional networks and evaluation metric.

Agradecimientos

Agradecer a mi tutor, Álvaro, por toda la ayuda y apoyo recibido durante el desarrollo de este trabajo.

Dar las gracias a mi familia y amigos, en especial a mis padres y hermana. Papá, gracias por apoyarme y por «vivir» conmigo esta experiencia que es cursar Teleco, por hacer tuyos mis problemas siendo tú el estudiante y no yo. Mamá, gracias por confiar en mi, por ser mi pilar de apoyo, por creer muchas veces más que yo. Sandra, gracias por darme la «lata» dándome fuerzas para seguir, que te juro que no me volveré a meter contigo porque ya se lo que sientes.

No olvidarme tampoco de las personas que he conocido durante estos cuatro años, a los que puedo llamar familia. Que con vosotros este camino se hace mucho más fácil.

Gracias, a todos.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. Motivación

Estamos en un momento en el que la detección de personas en cualquier sistema de vídeo está a la orden del día, pues cada vez toman más importancia en determinados ámbitos de la sociedad. Esto conlleva que numerosos investigadores desarrollen algoritmos propios para la detección de personas, consiguiendo con ellos grandes resultados sobre bases de datos elegidas por los propios autores.

Por otro lado, cada vez es más común que muchos investigadores de cualquier ámbito busquen conseguir un aprendizaje continuo por parte de las máquinas, siendo éstas capaces de aprender por si mismas, lo llamado aprendizaje automático. Dicho aprendizaje automático, *Machine Learning*, es muy popular actualmente gracias a un nuevo tipo de aprendizaje denominado *Deep Learning*. Con él conseguimos, a partir de una serie de datos e instrucciones para solucionar problemas, hacer ser a la máquina un instrumento enormemente exitoso puesto que será capaz de solucionar el problema modificando los datos iniciales evitando así futuras complicaciones. Un método de implantación del *Deep Learning* es a partir de redes convolucionales que intentarán imitar las funciones de las neuronas del cerebro humano.

Estas nuevas ideas también hacen incursión en el ámbito de la detección y seguimiento de objetos, haciendo a éstos basarse en las redes convolucionales. Al basarnos en estas redes convolucionales se consiguen excelentes resultados, [5], en tareas como la detección de rostros o la clasificación de letras escritas a mano, [8]. El algoritmo seleccionado de detección de personas basado en redes convolucionales es el *Faster Region-based Convolutional Network*, FASTER R-CNN.

Por ello, en este trabajo de fin de grado, se propone evaluar detectores "tradicionales" y detectores "modernos" basados en redes convolucionales para comprobar los

resultados obtenidos y ofrecer una comparativa entre ellos.

1.2. Objetivos

La detección de personas es fundamental en cualquier sistema de vídeo seguridad. La complejidad de la detección de personas se encuentra, principalmente, en la dificultad para definir un modelo de las mismas, debido a la gran variabilidad en la apariencia física, poses, puntos de vista, movimiento e interacción entre las personas y los objetos e incluso otras personas.

Actualmente se ha extendido el uso de grandes cantidades de datos de entrenamiento o big data para modelar cualquier tipo de objeto y su detección. En concreto, el uso de redes convolucionales, con la finalidad de aprender de forma automática las características más importantes de los modelos, presenta en muchos casos mejoras considerables en la tarea de detección de cualquier tipo de objeto, y en este caso, personas.

Según la motivación anterior, el objetivo de este trabajo de fin de grado es el de diseñar e implementar diferentes modelos de objetos o personas mediante el uso de redes convolucionales y comparar los resultados obtenidos con algoritmos más tradicionales del estado del arte, definiendo un marco de evaluación común, *dataset*, métrica de evaluación y sobre todo el modelado de los objetos, en este trabajo personas, en igualdad de condiciones.

1.3. Organización de la memoria

La memoria consta de los siguientes capítulos:

- Capítulo 1. Motivación, objetivos del trabajo y organización de la memoria.
- Capítulo 2. Estado del arte.
- Capítulo 3. Diseño.
- Capítulo 4. Desarrollo.
- Capítulo 5. Integración, pruebas y resultados.
- Capítulo 6. Conclusiones y trabajo futuro.
- Bibliografía.

Capítulo 2

Estado del arte

2.1. Introducción

Para la realización de este trabajo se ha estudiado en profundidad la detección de personas teniendo en cuenta el estado del arte de diversos artículos [9, 10, 11, 5].

Tanto la detección de personas como el seguimiento de éstas en entornos complejos demuestran ser una dificultad a tener en cuenta. Estas complicaciones vienen dadas por diversos problemas [12]. Uno de ellos son las propias escenas donde se realiza la detección de personas puesto que éstas pueden incluir múltiples personas, así como posibles oclusiones de las mismas durante ciertos periodos de tiempo. Otra de las dificultades de la detección de personas son las determinadas poses que toma la persona a localizar, puesto que no siempre puede ser la misma, puede estar sentada, de pie, en silla de ruedas, tumbada, etc. Si estamos localizando personas en escenas en las que haya múltiples individuos podemos toparnos con otro problema que viene dado por las propias interacciones entre los diferentes individuos de la imagen.

Este capítulo lo dividiremos en varias secciones, en la sección 2.2 se hablará del estado del arte en cuanto a la detección de personas basándonos en los artículos citados anteriormente y se englobará el sistema básico o arquitectura común para todos los detectores; posteriormente, en la sección 2.3, se describirá la clasificación propuesta para los algoritmos de detección de personas empleados durante la realización del trabajo; y por último en la sección 2.4 se introducirán los detectores basados en redes convolucionales.

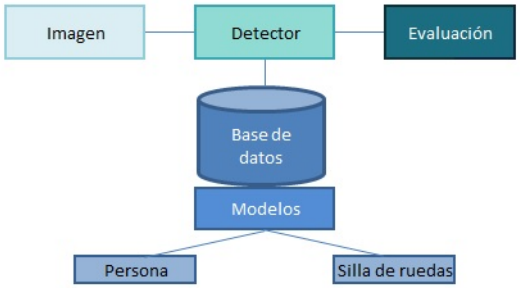


Figura 2.1: Arquitectura de los sistemas de detección de personas

2.2. Arquitectura basada en el estado del arte

Haciendo referencia a los artículos citados anteriormente, en la sección 2.1, éstos abordan el estado del arte centrado en la detección de personas.

En el artículo [9], el autor manifiesta un estado del arte de última generación en la detección de personas.

El autor del artículo [10] expone un estado del arte centrado principalmente en algoritmos de detección de personas con ventanas deslizantes.

En el artículo [11] describe un estado del arte basado en el estudio de los diversos detectores de personas y la posterior implementación de dichos detectores en sistemas totalmente completos. Los pasos que sigue a la hora de realizar la detección de personas son, en un principio, definir las áreas de interés o regiones, *Region of Interest*, ROI, posteriormente clasificación, y por último seguimiento.

Y por último, en el artículo [5] se combinan dos ideas clave, la aplicación de las redes convolucionales de alta capacidad a las propuestas de región de abajo hacia arriba para localizar y segmentar objetos, y el aumento del rendimiento cuando los datos de entrenamiento son escasos mediante el pre-entrenamiento supervisado para una tarea auxiliar seguido de un ajuste fino.

Todos los sistemas de detección de personas que empleamos en este trabajo siguen el mismo esquema, puesto que su arquitectura es la misma para todos. En la figura 2.1 se observan las tres etapas comunes a estos sistemas, de las que hablaremos en las subsecciones 2.2.1, 2.2.2 y 2.2.3.

2.2.1. Entrada

Esta etapa es la primera de nuestro sistema. Analizaremos todas las imágenes que componen el vídeo elegido a las que se procederá a realizar la detección de personas. Nuestra entrada utilizada en este trabajo serán listas de imágenes a color de igual

tamaño y resolución.

Aunque en nuestro trabajo utilicemos este tipo de entrada en concreto, los múltiples algoritmos existentes para la detección de personas pueden emplear diferentes formatos de entrada como secuencias de vídeo, imágenes en escala de grises, en color, imágenes en 2D o 3D, etc.

2.2.2. Detector

El objetivo de esta etapa es seleccionar las regiones posibles en la imagen a ser persona. No es común ya que cada algoritmo de detección de personas empleará sus propios métodos para la detección.

En nuestro caso, habrá tres tipos de detectores correspondientes al detector *Deformable Parts Model*, DPM, el detector *Aggregate Channel Features*, ACF, y el detector *Faster Region-based Convolutional Network*, FASTER R-CNN. De ellos se hablará en detalle en el capítulo 3.

Cada uno de estos detectores emplearán diferentes tipos de modelos que fijarán las características que deben tener las regiones candidatas para poder considerarlas persona.

Estos modelos, en nuestro trabajo, están generados para que, a la hora de comparar todos los detectores, estén en igualdad de condiciones. Se entrenan con una base de datos común, consiguiendo los modelos de persona de pie y en silla de ruedas.

2.2.3. Evaluación

Una vez realizada la comparativa entre las posibles regiones y los modelos se pasará a determinar cuáles son más probables a ser personas.

Estos resultados obtenidos por cada uno de los detectores y modelos mediante el uso de anotaciones manuales, *Ground Truth*, GT, se evaluarán en el capítulo 5.

2.3. Clasificación de los algoritmos de detección de personas

Para clasificar los algoritmos de detección de personas nos podemos guiar por múltiples criterios. En este trabajo vamos a centrarnos en las clasificaciones que se pueden hacer en las dos fases críticas, que son la detección de las posibles regiones de interés y los modelos de objetos (en nuestro caso personas).

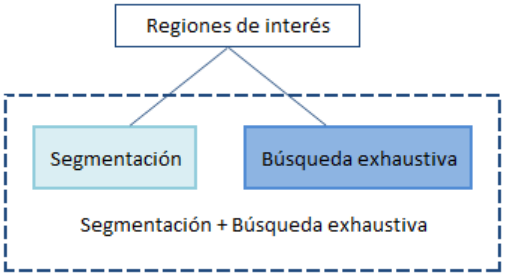


Figura 2.2: Clasificación dependiendo de las regiones de interés.

2.3.1. Regiones de interés

Las posibles regiones de interés se pueden obtener de tres maneras diferentes, [9, 10, 11]. La primera basada en la segmentación, la segunda en búsqueda exhaustiva y la tercera es una combinación de las dos primeras. La figura 2.2, muestra esta división.

Basándonos en [9, 10, 11], los detectores que usan búsqueda exhaustiva suelen dar mejores resultados.

En este trabajo, puesto que nuestros detectores solamente emplean búsqueda exhaustiva, no se tendrá en cuenta ni segmentación ni la mezcla de búsqueda exhaustiva más segmentación.

2.3.1.1. Búsqueda exhaustiva

La búsqueda exhaustiva consiste en, mediante una ventana con un tamaño determinado, ir recorriendo la imagen de entrada y comparar estas dos zonas. Obtenemos así un montón de regiones de interés con determinados grados de parecido.

Éste es el método más popular entre los detectores de personas aunque conlleve una carga computacional muy alta al comparar la ventana con todos los trozos de la imagen de igual tamaño.

Esta técnica de búsqueda exhaustiva se puede dividir en dos métodos. Un primero, [1, 13], que estará basado en la utilización de ventanas deslizantes que muestrearán y evaluarán diversas ventanas con un clasificador. Y el segundo, los detectores de personas basados en características que crean ascendentemente el volumen de densidad mediante votos probabilísticos por coincidencia en las características locales, [12, 14].

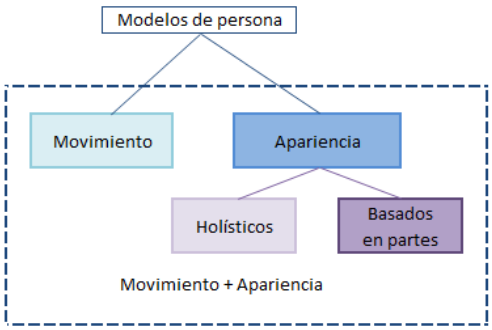


Figura 2.3: Clasificación dependiendo de los modelos de persona.

2.3.2. Modelos de persona

Otros de los puntos críticos surge a la hora de estimar si la región detectada coincide con el modelo entrenado para poder clasificarla como persona. Por lo tanto, los modelos de personas pueden tener en cuenta el movimiento, estar basados solamente en la apariencia o una mezcla de los dos. La figura 2.3, muestra esta división.

En el estado del arte, [9, 10, 11], los detectores cuyos modelos de persona que emplean la información de apariencia son los que logran mejores resultados.

En este trabajo, puesto que nuestros detectores solamente emplean modelos de personas basados en apariencia , no se van a tener en cuenta aquellos modelos dependientes del movimiento ni aquellos que juntan estos dos.

2.3.2.1. Basado en la apariencia

Los modelos basados en la apariencia se pueden dividir en dos clases, [9, 10, 11], los holísticos y los basados en partes.

Los holísticos están creados mediante una sola forma, es decir, generan el modelo asumiendo un sola región. Éstos son los más simples pero muestran problemas, puesto que si la persona está parcialmente oculta o si se producen cambios en la postura, el detector tiende a fallar.

Los basados en partes crean el modelo para la detección a partir de múltiples partes del cuerpo. Éstos son los más complejos pero a la vez muestran mejores resultados aunque se produzcan oclusiones o cambios de pose.

De los detectores que empleamos en el trabajo, son holíticos el ACF, [15], y el FASTER R-CNN, [7]. El único detector basado en partes que utilizamos es el DPM, [2].

2.4. Detectores basados en redes convolucionales

Dentro del estado del arte y teniendo en cuenta el objetivo principal de este trabajo de fin de grado, capítulo 1, debemos tener en cuenta los algoritmos basados en redes convolucionales, [5].

A diferencia de los enfoques más tradicionales, los sistemas o detectores de personas basados en redes convolucionales hacen uso del aprendizaje profundo o *Deep Learning* para la extracción y selección de las características discriminantes de las personas y su modelado, en vez del uso tradicional de características *Hand Craft* o de diseño "manual". Un ejemplo de estos detectores son los *Region-based Convolutional Network*, R-CNN, *Fast Region-based Convolutional Network*, FAST R-CNN y FASTER R-CNN, los que se explicarán en profundidad en el capítulo 3.

Éstos también siguen la estructura descrita en las secciones anteriores, en concreto los utilizados en este trabajo se basan en búsqueda exhaustiva y modelo holístico.

Capítulo 3

Diseño

3.1. Introducción

En este capítulo hablaremos sobre los diferentes detectores de personas utilizados durante la realización de este trabajo de fin de grado.

Todos nuestros sistemas de detección de personas poseen la misma arquitectura. El esquema de la figura 2.1, del capítulo 2, se compone de cuatro etapas o elementos. Todo sistema de detección de personas que se utilizará en este trabajo tomará como entrada una imagen en cada instante temporal de una secuencia de vídeo. A continuación se realizará la tarea de detección, para ello es necesario definir un modelo de detección (en este trabajo será o bien un modelo de persona de pie o persona en silla de ruedas). Por último se evaluará la salida del detector a partir de la generación de las curvas *Curve Presicion-Recall*, CPR, y su área bajo la curva, *Area Under Curve*, AUC.

De las tareas mencionadas, en este capítulo nos centraremos en la detección, en concreto de los detectores usados. El modelado o entrenamiento de los modelos de persona se realizará en el capítulo 4, y por último la evaluación y comparación de resultados se realizará en el capítulo 5.

3.2. Detectores

Una vez se ha analizado el estado del arte, capítulo 2, vamos a pasar a explicar con mayor detalle los diferentes algoritmos de detección de personas seleccionados del estado del arte para este trabajo. Debido a que el objetivo principal de este TFG es la comparativa entre detectores basados en redes convolucionales, hemos seleccionado dos de los mejores algoritmos "tradicionales" de detección del estado del arte DPM, 3.2.1, y ACF, 3.2.2, y un algoritmo más actual basado en redes convolucionales,

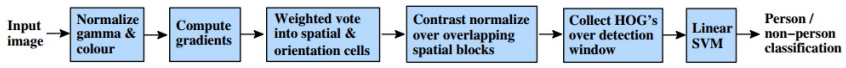


Figura 3.1: Funcionamiento del HOG. Procedente de [1].

FASTER R-CNN, 3.2.3.

3.2.1. *Deformable Parts Model* (DPM)

El algoritmo de detección de personas *Deformable Parts Model*, DPM, [2], surgió a partir de variaciones del histograma de gradientes orientados (*Histogram of Oriented Gradients*, HOG) [1]. Para poder entender este detector primeramente se introducirá el HOG, 3.2.1.1, y a continuación expondremos el DPM, 3.2.1.2.

3.2.1.1. Histograma de gradientes orientados (HOG)

Si nos guiamos por la clasificación propuesta en el capítulo 2, estamos ante un algoritmo que, al realizar las detecciones de las regiones de interés, se puede considerar de búsqueda exhaustiva, y emplea un modelo de persona holístico.

El histograma de gradientes orientados es un método de detección de personas basado en un modelo de características de forma y apariencia definidos mediante los gradientes de intensidad y sus distribuciones de direcciones de borde [1]. Sabiendo ésto, el algoritmo parte de un modelo de persona entrenado por una *Support Vector Machine*, SVM, en el que el autor clasifica cuando la ROI se considera persona o no, reparando en un conjunto total y no a partes de este. Para la implementación de este modelo, se divide la imagen en pequeñas celdas o regiones y para cada una de estas celdas se acumula un histograma con los gradientes de intensidad así como las direcciones de borde de los píxeles de la misma.

El autor de [1] divide el funcionamiento del HOG en 6 etapas mostrada en la figura 3.1. En la primera etapa se realizara la normalización Gamma/Color y a continuación se calculan los gradientes mediante un suavizado. El tercer paso es el cálculo del voto ponderado de las contribuciones espaciales así como la orientación de las celdas. Posteriormente se realiza una normalización de los gradientes en las regiones solapadas. El quinto paso consiste en recoger los HOG situados en la ventana de detección, y por último, se evaluará con una SVM lineal para determinar si el objeto detectado es una persona o no.

3.2.1.2. DPM

El algoritmo DPM [2] esta basado en la mezcla de diferentes modelos multi-escala y deformables determinados similares al modelo del HOG. Éste está formado por un filtro de raíz gruesa, que cubre todo el objeto o persona entera, es decir, considera a la persona como la componente principal, y otros filtros P más pequeños para cubrir partes del objeto más pequeñas. Por lo tanto, este detector DPM, según la clasificación realizada en el capítulo 2, es también un algoritmo de búsqueda exhaustiva pero con un modelado basado en partes.

En contraposición con el HOG, que caracterizaba a las personas mediante SVM lineal, el autor de [2] propone un cambio en las variables latentes de la SVM, las no observables directamente a las que denomina *Latent SVM*, LSVM. Y las utiliza para la creación del modelo de persona sin tener un etiquetado fiable de las partes.

En la figura 3.2 se observa el funcionamiento del detector DPM. En ésta, se observa que el autor obtiene las puntuaciones de los filtros por separado, por un lado las del filtro raíz y por otro, las de los filtros por partes, haciendo uso de la técnica de HOG. Los mapas de características de los filtros por partes se realizan al doble de resolución que las del filtro raíz, y además, estos filtros por partes permiten la "deformación" de sus posiciones. Tras la extracción de estos mapas de características se procede a emplear el clasificador LSVM para obtener la puntuación final de las personas.

Este tipo de detector DPM conseguirá mejores resultados que el HOG puesto que utiliza modelos de personas más complejos que se comportan mejor frente a los cambios de postura.

3.2.2. *Aggregate Channel Features* (ACF)

Este algoritmo de detección de personas llamado ACF, *Aggregate Channel Features*, introducido en [3, 15], es un detector que realiza búsqueda exhaustiva y en cuanto al modelo de persona que emplea es holístico.

El autor emplea lo llamado *Crosstalk Cascades*, [15], teniendo en cuenta que existe una correlación entre las escalas y las zonas adyacentes en las respuestas del detector, consiguiendo una reducción del coste computacional.

En un principio, con una imagen de entrada dada, I, se calcularán diferentes canales, $C = \Omega(I)$. Posteriormente, todos los bloques de píxeles en C se sumarán y se suavizarán realizando un sub-muestreado. Las características vienen dadas entonces por la búsqueda de los píxeles en los diferentes canales agregados. Finalmente, con el impulso entrenaremos y combinaremos los árboles de decisión sobre estas características para distinguir el objeto del fondo utilizando ventanas multi-escala deslizantes.

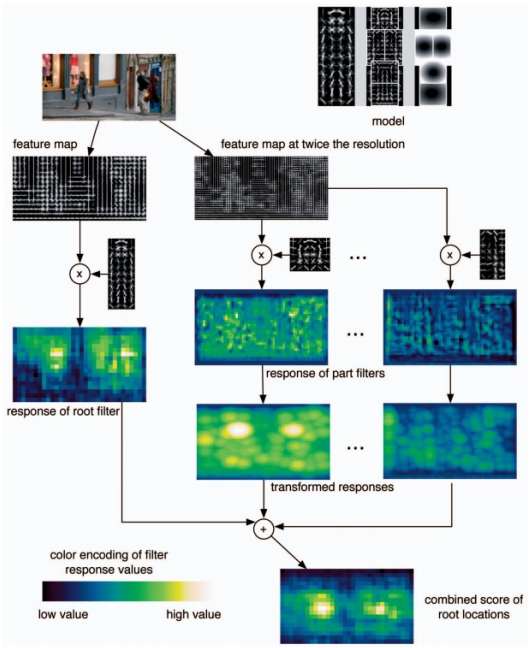


Figura 3.2: Funcionamiento del algoritmo DPM. Procedente de [2].

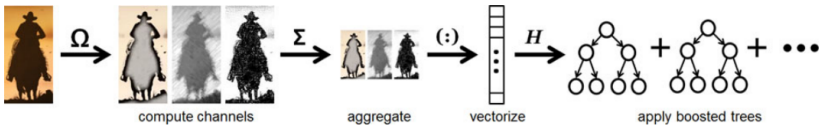


Figura 3.3: Esquema del detector ACF. Procedente de [3].

La siguiente figura, 3.3, muestra un esquema de los diferentes pasos que realiza el detector ACF.

En cuanto a los canales empleados, ACF utiliza 6 canales para histogramas de gradientes orientados, un canal para la magnitud de gradiente, y por último 3 canales de color LUV, los mismos que [4]. Estos 10 canales, aplicados sobre I, combinados es lo que se utilizará para determinar las regiones de interés. En la figura 3.4, se pueden ver estos canales por separado así como la suma final de ellos en la izquierda.

3.2.3. FASTER R-CNN

Por último, el detector FASTER R-CNN, [7], es el detector elegido basado en redes convolucionales. Este detector es a su vez una variación de una versión anterior denominada *Fast Region-based Convolutional Network*, FAST R-CNN y el original *Region-based Convolutional Network*, R-CNN, [6]. En esta etapa, por lo tanto, primero

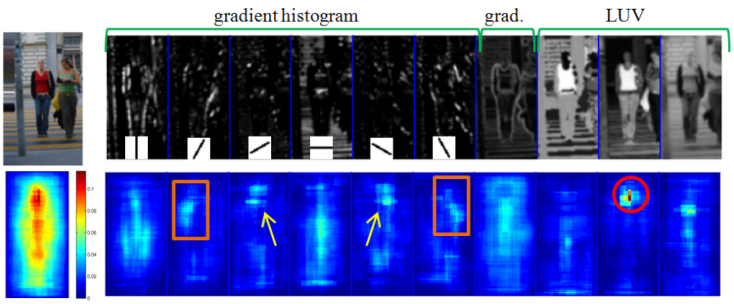


Figura 3.4: Descomposición de los canales del ACF. Procedente de [4].

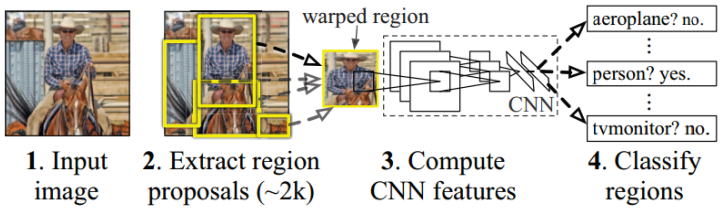


Figura 3.5: Etapas del detector R-CNN. Procedente de [5].

se explicará brevemente el detector R-CNN, 3.2.3.1, y posteriormente el FAST R-CNN, 3.2.3.2, terminando con las variaciones o mejoras que implementa el FASTER R-CNN, 3.2.3.3.

3.2.3.1. R-CNN

Actualmente, las redes convolucionales han demostrado ser lo suficientemente importantes, tanto, que aplican mejoras en la clasificación de imágenes así como en la detección de objetos y personas. Eso sí, si comparamos la clasificación con la detección, ésta última ofrece una mayor dificultad a la hora de su realización, lo que conlleva la utilización de detectores mucho más complejos que los utilizados anteriormente en otros tipos de detectores.

Este detector R-CNN [5], está dividido en 4 etapas. La primera etapa es la imagen de entrada al sistema. En la segunda etapa se generan las propuestas a regiones de interés. El tercer módulo consiste en una gran red neuronal convolucional que extrae un vector de características de longitud fija para cada región. Y por último, la cuarta etapa consiste en un conjunto de clases lineales SVM. La figura 3.5, muestra estas etapas.

Sin embargo, este algoritmo de detección de personas conlleva un alto coste computacional que lo hace extremadamente lento debido a que el cálculo para las propuestas

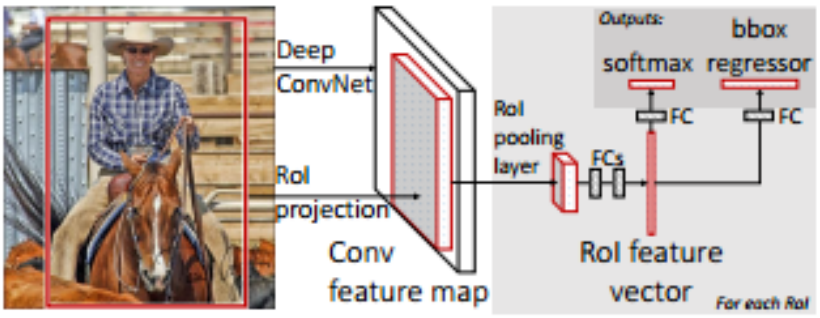


Figura 3.6: Arquitectura del detector FAST R-CNN. Procedente de [6].

de objeto no se comparten. El autor de [16] propone la utilización del método *Spatial Pyramid Pooling Network*, SPPnet, para reducir el tiempo compartiendo este cálculo.

3.2.3.2. FAST R-CNN

El nuevo modelo de detección de personas basado en redes convolucionales, en el método R-CNN, subsubsección 3.2.3.1, llamado FAST R-CNN, que consigue solventar los problemas que ocasionan la utilización de dichas redes convolucionales [6].

En la figura 3.6, se puede apreciar la arquitectura que sigue el detector FAST R-CNN. En un primer paso, este detector toma una imagen de entrada, I, y una serie de regiones de interés. Posteriormente, el detector introduce a I y a dichas regiones a una red convolucional y las agrupa en capas de agrupación máximas para crear el mapa de las características de convoluciones. Luego, para cada propuesta de región de interés, obtenidas en el primer paso, se le extraerán vectores de características de tamaño fijo procedentes del mapa de características. Cada vector es añadido a continuación de otros para crear una secuencia totalmente conectada, FC, que finalmente se divide en dos capas iguales. Una de ellas produce estimaciones de probabilidades *softmax* sobre K clases de objetos más una clase de tipo fondo. La otra de las capas genera cuatro números de valores reales para cada una de las K clases de objetos. Cada conjunto de 4 valores codifica las posiciones de la caja delimitadora para una de las clases K. Todo este proceso viene mucho más desarrollado en el artículo [6].

3.2.3.3. Mejoras y avances del FASTER R-CNN

Este último detector es el elegido basado en redes convolucionales, llamado FASTER R-CNN, [7].

El autor propone un sistema de detección de personas formado por dos módulos.

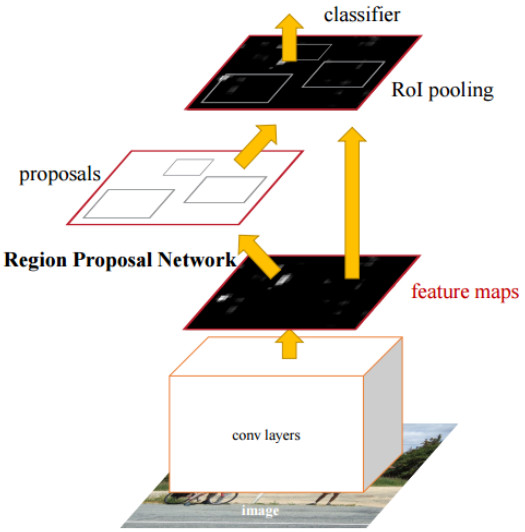


Figura 3.7: Arquitectura del detector FAST R-CNN. Procedente de [7].

Un primer módulo compuesto por una red completamente convolucional, cuya finalidad es la de proponer las regiones de interés; más un segundo módulo coincidente con el detector FAST R-CNN, [6], que usa esas regiones de interés propuestas. La figura 3.7, muestra estas dos etapas que se pueden considerar como una sola red unificada para la detección de personas.

El primer módulo, al que podemos llamar *Region Proposal Network*, RPN, toma cualquier imagen como entrada y devuelve un conjunto de rectángulos de las zonas de interés cada uno de ellos con una puntuación objetiva. Para generar las regiones, figura 3.8, al mapa de características convolucionales se le aplican una serie de redes pequeñas mediante ventanas deslizantes. A cada una de estas ventanas deslizantes se le asignará una característica de menor dimensión, en la figura representadas en la capa intermedia. Ésta alimentará a dos capas hermanas conectadas, una capa con la clasificación de la caja, cls, y otra con las regresiones de caja, reg. Simultáneamente, para cada una de las localizaciones que tome la ventana deslizante se calcularán posibles regiones de interés siendo el máximo número de regiones k. Por lo tanto, la capa reg contendrá las coordenadas de las posibles regiones, mientras que la capa cls contendrá las probabilidades de que dichas regiones sean objeto o no.

Una vez calculadas estas regiones de interés se pasará a la detección mediante el algoritmo FAST-CNN propuesto en [6].

Por lo tanto, el algoritmo FASTER R-CNN está compuesto por RPN más FAST-CNN, siendo este primero el que indica donde buscar al segundo.

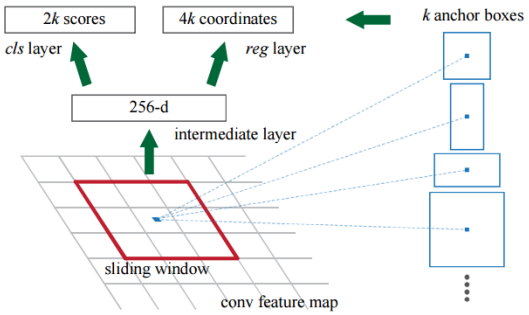


Figura 3.8: *Region Proposal Network*, RPN. Procedente de [7].

Capítulo 4

Desarrollo

4.1. Introducción

Una vez se han introducido los detectores, capítulo 3, empleados en el trabajo, en este capítulo veremos los diferentes métodos seguidos para entrenar los modelos de personas empleados en cada detector. Estos modelos tienen que estar entrenados en las mismas condiciones unos con otros para que podamos compararlos en el capítulo 5. En total hemos empleado 6 modelos de persona distintos, 3 modelos de personas de pie correspondientes a los 3 detectores, y otros 3 para los modelos de persona en silla de ruedas.

Este capítulo está dividido entonces en protocolo de aprendizaje del modelo de persona, sección 4.2, en la que definiremos los parámetros y elementos comunes y no comunes utilizados; y por último, en la sección 4.3, se indicará el código empleado para los diferentes detectores a la hora del entrenamiento de sus modelos.

4.2. Protocolo de aprendizaje del modelo de persona

4.2.1. Introducción

A la hora de entrenar los diferentes modelos de personas debemos de tener 3 elementos comunes para todos los detectores. Estos elementos son las imágenes de entrenamiento o *train* junto con las imágenes de validación o *test*, las anotaciones para entrenar dichas imágenes y un listado con las imágenes de *train* y *test*.

Las imágenes de *train* son las que utilizaremos para entrenar nuestro modelo, tanto de persona de pie como en silla de ruedas. Éstas imágenes de *train* se entrenaran con una serie de anotaciones recogidas manualmente que contienen las posiciones de los diferentes elementos que aparecen en las imágenes. Durante el entrenamiento se

controlará la eficacia de dicho modelo con las imágenes de *test*. Ambos grupos de imágenes, de *test* y de *train*, están recogidos conjuntamente y por separado en una serie de listados.

Aunque estos elementos tienen que ser comunes a la hora de generar los diferentes modelos, cada uno de los algoritmos utilizados de detección usan un formato diferente de entrada. Para crear los modelos empleados en la detección DPM y la detección FASTER R-CNN, estos elementos son iguales, mientras que para el detector ACF, las anotaciones tienen otro formato.

4.2.2. *Datasets* de entrenamiento

Como ya hemos dicho en el apartado anterior, en nuestro trabajo se generarán seis modelos diferentes. La mitad de éstos empleados en la detección de personas de pie, y la otra mitad en la detección de personas en silla de ruedas.

4.2.2.1. Personas de pie

La base de datos, es decir, las imágenes empleadas para la generación de este modelo de persona es la misma para todos nuestros detectores. Ésta es un *dataset* público creado por Inria (http://pascal.inrialpes.fr/data/human/) y descrito en [1], denominado *INRIA Person Dataset*. Esta compuesto por un grupo de imágenes para train y otro para test. Ambos grupos están divididos a su vez en una serie de imágenes positivas y negativas con sus correspondientes anotaciones. El dataset consta de 1832 imágenes para el entrenamiento y 741 para validación, con resoluciones diferentes. Para conseguir la misma resolución en todas las imágenes, éstas se han normalizado, tanto las originales como sus reflexiones, por Inria y se incluyen en el *dataset*. Algunas de las imágenes que componen este dataset se pueden observar en la figura 4.1.

Este *dataset*, por lo tanto, nos proporciona todos los elementos para realizar el entrenamiento del modelo de persona, las imágenes, las anotaciones y el listado de las imágenes. Ahora bien, estas anotaciones vienen dadas en formato ".txt" compatibles para el detector ACF. Para conseguir generar las anotaciones válidas para los detectores DPM y FASTER R-CNN generamos un *script*, denominado ParserTxtToXml, en el entorno Matlab que nos convierte estas anotaciones en otras en formato ".xml", compatibles para estos detectores. La siguiente figura 4.2 muestra el contenido de estos dos formatos de anotaciones.

(a)

(b)

Figura 4.1: Frames del *dataset* de Inria. (a) Entrenamiento (b) Validación.

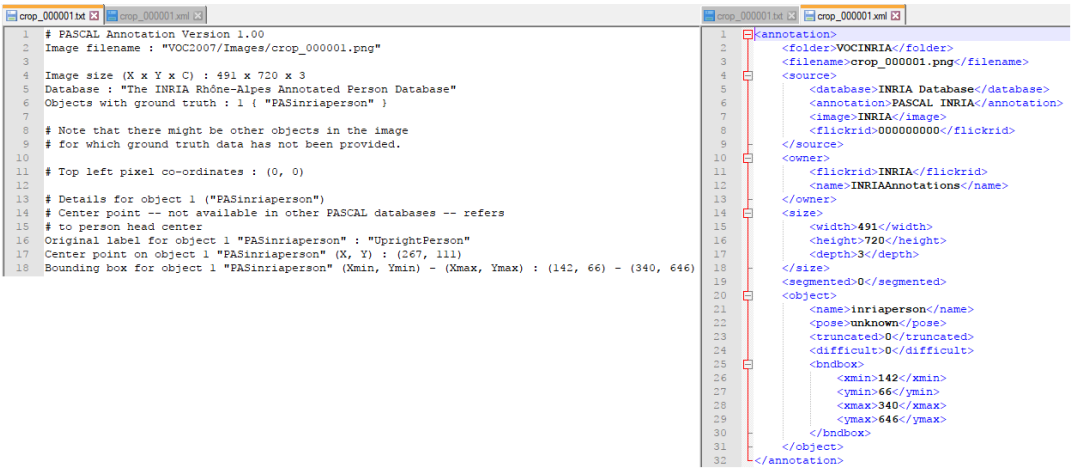
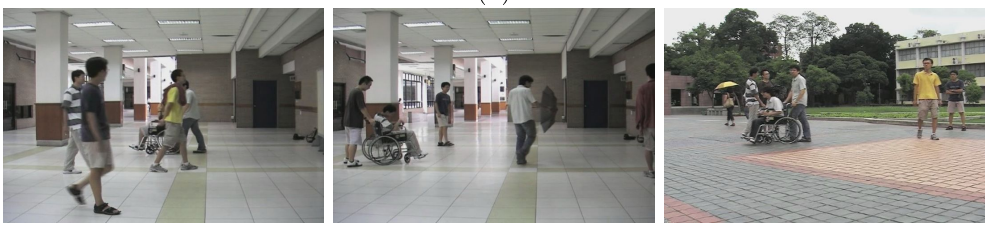


Figura 4.2: Contenido de las anotaciones (Derecha) ".txt" (Izquierda) ".xml".

(a)



(b)

Figura 4.3: Frames del *dataset* de Smile. (a) Entrenamiento (b) Validación.

4.2.2.2. Personas en silla de ruedas

La base de datos, es decir, las imágenes empleadas para la creación del modelo de persona en silla de ruedas, es común para todos nuestros detectores. Ésta es un *dataset* privado llamado *Smile Lab wheelchair dataset* desarrollado por Smile Lab (http://smile.ee.ncku.edu.tw/). Se compone de dos grupos de imágenes, las utilizadas para entrenar el modelo y las empleadas para la validación de dicho modelo. En cuanto a las imágenes para el entrenamiento, éstas están agrupadas en 8 secuencias con un total de 3650 imágenes, mientras que las imágenes para la validación están agrupadas en 4 secuencias con un total de 1317 imágenes. En la figura 4.3, se pueden ver algunas imágenes de este *dataset*. Todas éstas tienen una resolución de 720x480 píxeles. Las anotaciones no están disponibles por lo que se crearon manualmente por el grupo *Video Processing and Understanding Lab*, VPULab, estando disponible su descarga en la página web (http://www-vpu.eps.uam.es/DS/WUds/).

Ahora bien, nos pasa un caso similar al del modelo de persona de pie, 4.2.2.1, ya que las anotaciones proporcionadas solamente nos son compatibles con los detectores DPM y FASTER R-CNN, sin embargo, para el detector ACF no son válidas. Por ello, generamos otro script en Matlab, similar al anterior, ParserXmlToTxt, que convierte el contenido de estas anotaciones dadas en fomarto ".xml" al formato ".txt" para el detector ACF. Estas nuevas anotaciones están compuestas por una serie de líneas, en las que cada una indicará una persona en silla de ruedas, la posición en la que se sitúa dicha persona tanto en ancho como en alto que marcará la esquina superior izquierda del contorno de la persona y el tamaño de dicha región en ancho y alto. En la figura

LR45_0221.xml

```
<annotation>
    <folder>VOCVPU</folder>
    <filename>LR45_0221.jpg</filename>
    <source>
        <database>Wheelchair Database</database>
        <annotation>PASCAL VOCVPU</annotation>
        <image>VPU</image>
        <flickrid>000000000</flickrid>
    </source>
    <owner>
        <flickrid>VPU</flickrid>
        <name>VPUAnnotations</name>
    </owner>
    <size>
        <width>720</width>
        <height>480</height>
        <depth>3</depth>
    </size>
    <segmented>0</segmented>
    <object>
        <name>wheelchairuser</name>
        <pose>LR45</pose>
        <truncated>0</truncated>
        <difficult>0</difficult>
        <bndbox>
            <xmin>505</xmin>
            <ymin>80</ymin>
            <xmax>619</xmax>
            <ymax>227</ymax>
        </bndbox>
    </object>
</annotation>
```

LR45_0221.txt

```
% bbGt version=3
wheelchairuser 505 80 114 147 0 0 0 0 0 0 0
```

Figura 4.4: Contenido de las anotaciones (Derecha) ".xml" (Izquierda) ".txt".

4.4, se observa el contenido de estos dos formatos de anotaciones.

4.3. Entrenamiento de los modelos de persona

El código empleado, así como las instrucciones o directrices, para la generación de los modelos de persona, una vez establecidos los elementos comunes, 4.2, es el recomendado por el autor de cada detector. Para el caso del DPM en (https://github.com/rbgirshick/voc-dpm), el de ACF en su página web (https://github.com/pdollar/toolbox) y para el detector FASTER R-CNN son las indicadas en (https://github.com/ShaoqingRen/faster_rcnn).

Capítulo 5

Integración, pruebas y resultados

5.1. Introducción

En este capítulo se introducirá la base de datos empleada para la evaluación de la eficacia de los diferentes detectores seleccionados, propuestos en el capítulo 3, así como la métrica empleada para dicha comparativa en la sección 5.2. En la siguiente sección 5.3, pasaremos a analizar los diferentes resultados obtenidos. Finalizaremos el capítulo en la sección 5.4 con una breve conclusión.

5.2. Marco de evaluación

5.2.1. *Dataset*

El *dataset* o base de datos utilizada para la evaluación de los detectores propuestos en este trabajo es una serie de secuencias de vídeo creadas por el grupo *Video Processing and Understanding Lab*, VPULab, en colaboración con una residencia de ancianos para la vigilancia de los miembros de ésta, con el fin de conseguir un entorno lo más realístico posible. Estos vídeos están grabados en una sola sala mediante dos cámaras GoPro colocadas en dos posiciones diferentes para conseguir dos puntos de vista de la sala, figura 5.1.

El dataset está compuesto entonces por 11 secuencias, desde S1 hasta S11, cada una de ellas tomadas desde los 2 puntos de vista distintos, V1 y V2, dando lugar a 22 secuencias con cómputo de 29732 imágenes, todas ellas con la misma resolución de 768x432 píxeles. A parte, en ella, se puede apreciar la aparición de algunos de los miembros del VPULab tanto de pie como en silla de ruedas. En la siguiente tabla, 5.1, se puede observar el número de imágenes de cada secuencia así como el número de miembros que aparecen separados en personas de pie y en silla de ruedas. Para

Figura 5.1: Diferentes puntos de vista del *dataset*.

Secuencias de vídeo	Personas de pie	Personas en silla de ruedas	Número de imágenes
S1	0	1	1318
S2	0	1	916
S3	1	1	860
S4	1	1	1166
S5	0	2	1638
S6	0	2	723
S7	2	2	1083
S8	2	2	745
S9	2	2	2102
S10	2	2	2460
S11	0	4	1855

Tabla 5.1: Características del *dataset* empleado.

cada una de las 22 secuencias que componen este *dataset* existen unos ficheros de anotaciones, a los que denominamos *Ground Truth*, GT, con las posiciones exactas de todas las personas, tanto de pie como en silla de ruedas incluso si están parcialmente ocultas, que aparecen en las diversas imágenes que componen estas secuencias.

En la figura 5.2 se pueden apreciar algunos *frames* correspondientes a estos vídeos.

5.2.2. Métricas

Para la evaluación de los algoritmos se han empleado las *Curve Presicion-Recall*, CPR, y su área bajo la curva, *Area Under Curve*, AUC.

Por lo tanto, la métrica empleada serán las CPR que vienen definidas por:

$$Precision = \frac{\#Verdaderos\,positivos}{\#Verdaderos\,positivos + \#Falsos\,positivos}$$

$$Recall = \frac{\#Verdaderos\,positivos}{\#Verdaderos\,positivos + \#Falsos\,negativos}$$

Figura 5.2: *Frames* de algunas de las secuencias del *dataset*.

En la que cada término significa:

- Verdadero positivo cuando la detección es verdaderamente una persona, es decir, coincide con alguna detección del GT.
- Falso positivo cuando la detección no es una persona, es decir, no coincide con ninguna detección del GT.
- Falso negativo cuando la detección indica que no es persona cuando en realidad es persona, es decir, la detección del GT indica que hay persona pero nuestra detección no.

Por lo tanto, este tipo de métrica comparará los resultados obtenidos con los detectores empleados en el trabajo con el GT de cada uno de los vídeos analizados.

Una vez realizada la CPR calcularemos su área bajo la curva, AUC, para poder estimar mejor la eficacia de dichos algoritmos de detección.

5.3. Resultados

Una vez establecida la base de datos sobre la que se van a evaluar los diferentes detectores y la métrica empleada para esto, en esta sección se mostrarán los resultados obtenidos. Éstos se han dividido en dos comparaciones diferentes, la eficacia de los detectores con el modelo de persona de pie 5.3.1 y con el modelo de persona en silla de ruedas 5.3.2.

5.3.1. Para modelo de persona de pie

Todos los resultados en cuanto al área bajo la curva obtenidos en las detecciones conseguidas para nuestra base de datos comparándolos con los diferentes GT (para

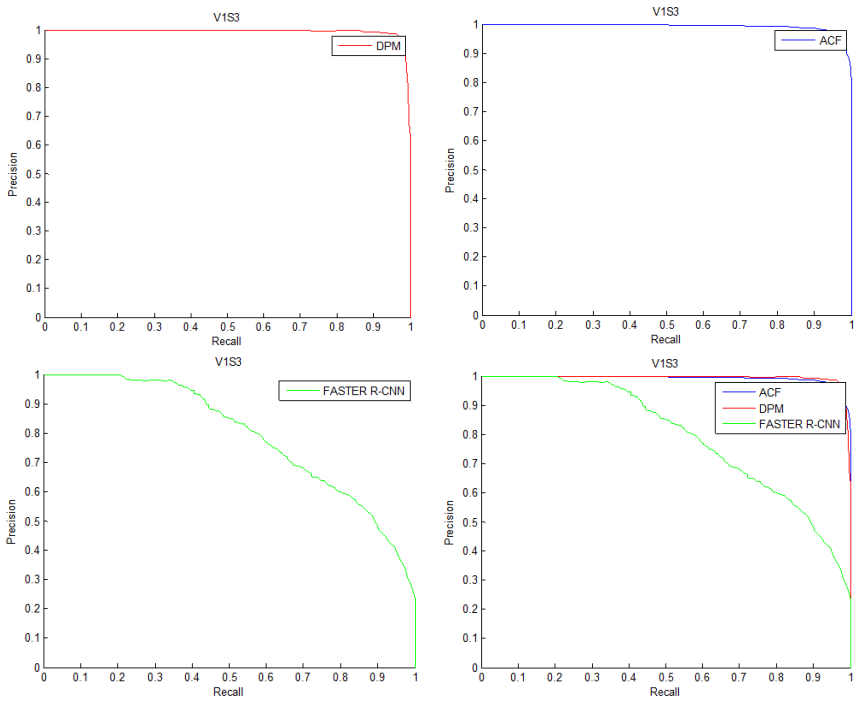


Figura 5.3: *Curve Presicion-Recall* para la secuencia V1S3 frente GT de personas de pie.

personas de pie, en silla de ruedas y ambos) son los mostradas en la tabla 5.2.

Algunos de los resultados obtenidos al realizar las CPR se pueden observar en la figura 5.3. En ellas se ve el comportamiento de cada uno de los detectores sobre la secuencia V1S3, por separado y conjuntamente, comparándolos con el GT de personas de pie.

En los resultados de la tabla 5.2 se puede apreciar que los detectores que ofrecen mejores resultados son los detectores DPM y ACF. En este caso, el FASTER R-CNN para la detección de persona de pie no es tan efectivo puesto que está detectando tanto personas de pie como personas en silla de ruedas. En las figura 5.4 y 5.5, se muestra una imagen de la secuencia V1S3 y V2S10 en la que se observan las dos detecciones que se han obtenido y su rendimiento en términos de CPR.

Otro de los factores que se tienen en cuenta a la hora de comparar los resultados es el número y las poses de las personas que aparecen en las secuencias. En las secuencias S1, S2, S5, S6 y S11 al no tener personas de pie, nuestro GT de personas de pie está vacío por lo que el detector no podrá comparar los resultados obtenidos con nada, no consiguiendo así las curvas CPR frente GT de personas de pie. Es decir, los resultados que obtiene nuestro detector serán falsos positivos en este caso.

Además, si enfrentamos los resultados obtenidos mediante el modelo de persona

Secuencias de vídeo	Detectores								
	DPM			ACF			FASTER R-CNN		
	GT de pie	GT silla	GT conjunto	GT de pie	GT silla	GT conjunto	GT de pie	GT silla	GT conjunto
V1S1	-	0.6170	0.6170	-	0.9318	0.9318	-	0.7518	0.7518
V1S2	-	0.4834	0.4834	-	0.6280	0.6280	-	0.7927	0.7927
V1S3	0.9950	0.3861	0.6498	0.9932	0.6597	0.8342	0.7957	0.4898	0.7918
V1S4	0.9249	0.4278	0.5924	0.9061	0.6088	0.7079	0.6731	0.4137	0.6147
V1S5	-	0.5981	0.5981	-	0.7370	0.7370	-	0.8182	0.8182
V1S6	-	0.5102	0.5102	-	0.6784	0.6784	-	0.6850	0.6850
V1S7	0.9354	0.3703	0.6051	0.9094	0.5421	0.6979	0.7125	0.4709	0.6570
V1S8	0.9137	0.2636	0.7460	0.8712	0.3652	0.8315	0.8206	0.3728	0.8178
V1S9	0.9350	0.2150	0.6506	0.9472	0.3588	0.8431	0.8014	0.3315	0.7564
V1S10	0.8446	0.2508	0.7267	0.5929	0.4914	0.8319	0.7416	0.3750	0.8251
V1S11	-	0.5175	0.5175	-	0.6459	0.6459	-	0.6380	0.6380
Media	0.9248	0.4218	0.6088	0.8700	0.6043	0.7607	0.7575	0.5581	0.7408

Secuencias de vídeo	Detectores								
	DPM			ACF			FASTER R-CNN		
	GT de pie	GT silla	GT conjunto	GT de pie	GT silla	GT conjunto	GT de pie	GT silla	GT conjunto
V2S1	-	0.2914	0.2914	-	0.7709	0.7709	-	0.7768	0.7768
V2S2	-	0.2311	0.2311	-	0.5578	0.5578	-	0.7138	0.7138
V2S3	0.9564	0.4473	0.5523	0.9848	0.6954	0.7903	0.8375	0.4787	0.7629
V2S4	0.8822	0.4525	0.5198	0.9496	0.6059	0.6593	0.7745	0.4165	0.6639
V2S5	-	0.3779	0.3779	-	0.6849	0.6849	-	0.7559	0.7559
V2S6	-	0.2887	0.2887	-	0.7276	0.7276	-	0.8400	0.8400
V2S7	0.8732	0.3458	0.5438	0.9080	0.5795	0.7386	0.7935	0.4382	0.7121
V2S8	0.8385	0.2286	0.6775	0.7420	0.3970	0.8513	0.8742	0.3206	0.8405
V2S9	0.5857	0.2264	0.5110	0.5692	0.3985	0.6670	0.5753	0.2861	0.5925
V2S10	0.8390	0.2261	0.6605	0.5965	0.5312	0.8592	0.6901	0.3897	0.8012
V2S11	-	0.4189	0.4189	-	0.6958	0.6958	-	0.6265	0.6265
Media	0.8292	0.3213	0.4612	0.7917	0.6040	0.7275	0.7575	0.5493	0.7351

Tabla 5.2: Tabla con resultados AUC para el modelo de persona de pie vs diferentes GT.

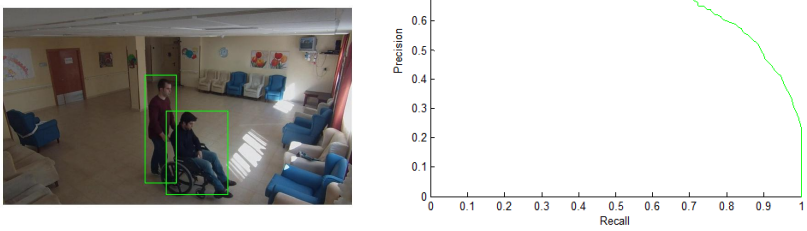


Figura 5.4: Detecciones de personas de pie para la secuencia V1S3 con el detector FASTER R-CNN. CPR frente GT de personas de pie.

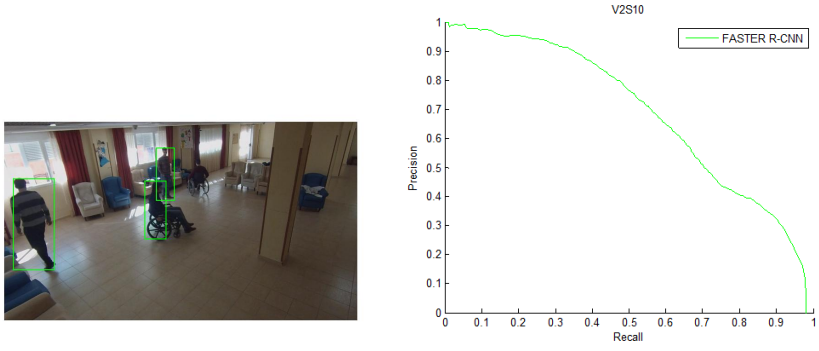


Figura 5.5: Detecciones de personas de pie para la secuencia V2S10 con el detector FASTER R-CNN. CPR frente GT de personas de pie.

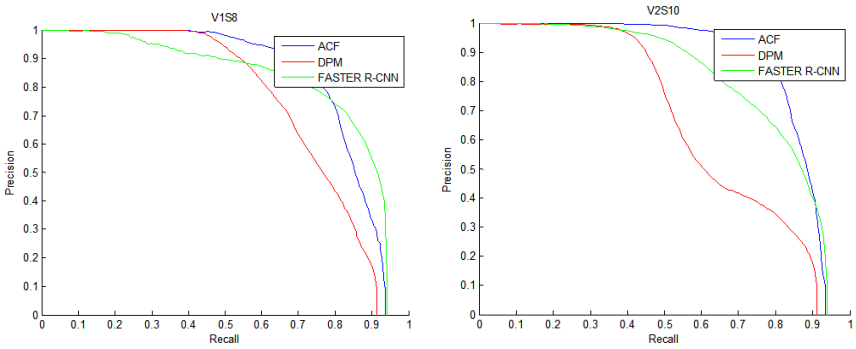


Figura 5.6: Ejemplos de curvas CPR para las detecciones de pie frente a GT total.

de pie contra el GT para todas las personas (conjunto de todas las personas independientemente de que estén de pie o en silla de ruedas) de cada secuencia, lo esperado es obtener gráficas CPR similares a las que nos da el detector DPM, figura 5.6. Es decir, curvas en las que cuando el *recall* toma valores de entre 0 y 0.5, la precisión es casi perfercta, casi 1, que indicará las detecciones de personas de pie; y cuando el *recall* tome valores mayores de 0.5, la precisión decaiga de manera significativa hasta 0, correspondiente este tramo a las personas en silla de ruedas. Por lo tanto, una vez conseguidas estas curvas CPR se espera obtener un AUC entre 0.6 y 0.7 en secuencias donde hay el mismo número de personas de pie que en silla de ruedas.

En la tabla 5.2 se puede observar que el único que cumple con esto, en general, es el detector DPM, puesto que el ACF y el FASTER R-CNN realizan, además de detecciones de personas de pie, otras detecciones, como hemos mencionado anteriormente para el caso del FASTER R-CNN, figura 5.4.

La comprobación de que los detectores además de detectar personas de pie también detectan personas en silla de ruedas se puede ver si miramos la columna en la que se comparan los resultados con el GT de personas en silla de ruedas. En vídeos donde solamente aparecen personas en silla de ruedas, para los detectores ACF y FASTER R-CNN el valor de su AUC es elevado en comparación con el valor del AUC del DPM.

5.3.2. Para modelo de persona en silla de ruedas

Los resultados en cuanto al área bajo la curva obtenidos en las detecciones conseguidas para nuestra base de datos comparándolos con los distintos GT (para personas de pie, en silla de ruedas y ambos) son los mostradas en la tabla 5.3. En este caso, el detector que mejor resultados obtiene es el FASTER R-CNN.

Algunas de las CPR que obtenemos al comparar nuestros resultados, obtenidos con los detectores seleccionados en el capítulo 3, con el GT del dataset correspondiente

Secuencias de vídeo	Detectores								
	DPM			ACF			FASTER R-CNN		
	GT de pie	GT silla	GT conjunto	GT de pie	GT silla	GT conjunto	GT de pie	GT silla	GT conjunto
V1S1	-	0.9144	0.9144	-	0.6652	0.6652	-	0.9979	0.9979
V1S2	-	0.9826	0.9826	-	0.6236	0.6236	-	0.9923	0.9923
V1S3	0.4863	0.8589	0.5314	0.2262	0.7009	0.5032	0.4935	0.9843	0.7129
V1S4	0.4732	0.7936	0.5047	0.1808	0.6054	0.3640	0.5209	0.9894	0.7083
V1S5	-	0.9135	0.9135	-	0.5207	0.5207	-	0.9532	0.9532
V1S6	-	0.9973	0.9973	-	0.7959	0.7959	-	0.9961	0.9961
V1S7	0.6332	0.4539	0.4366	0.2177	0.3144	0.2752	0.5796	0.5267	0.5877
V1S8	0.1798	0.9150	0.6069	0.0758	0.5794	0.3661	0.1581	0.9560	0.7349
V1S9	0.2561	0.8488	0.5900	0.1016	0.6585	0.3775	0.2310	0.9337	0.7087
V1S10	0.2106	0.9091	0.6121	0.0999	0.6757	0.4141	0.1937	0.9759	0.7510
V1S11	-	0.7445	0.7445	-	0.4580	0.4580	-	0.7788	0.7788
Media	0.3732	0.8483	0.7122	0.1503	0.5998	0.4876	0.3628	0.9168	0.8111

Secuencias de vídeo	Detectores								
	DPM			ACF			FASTER R-CNN		
	GT de pie	GT silla	GT conjunto	GT de pie	GT silla	GT conjunto	GT de pie	GT silla	GT conjunto
V2S1	-	0.9363	0.9363	-	0.5492	0.5492	-	0.9947	0.9947
V2S2	-	0.9865	0.9865	-	0.7775	0.7775	-	0.9990	0.9990
V2S3	0.6055	0.8395	0.4848	0.2949	0.6687	0.4838	0.4701	0.9789	0.6251
V2S4	0.5884	0.8576	0.5437	0.2216	0.6558	0.4253	0.5420	0.9866	0.6244
V2S5	-	0.7842	0.7842	-	0.6806	0.6806	-	0.9241	0.9241
V2S6	-	0.9176	0.9176	-	0.6567	0.6567	-	0.9927	0.9927
V2S7	0.4713	0.7666	0.4565	0.2888	0.7032	0.4906	0.4520	0.9196	0.6609
V2S8	0.2306	0.8817	0.6398	0.0930	0.6553	0.4279	0.2291	0.9588	0.7549
V2S9	0.1590	0.8342	0.5419	0.0630	0.7127	0.4337	0.1903	0.9649	0.6735
V2S10	0.1527	0.8304	0.5651	0.0988	0.6596	0.4601	0.1530	0.9594	0.7147
V2S11	-	0.7180	0.7180	-	0.5069	0.5069	-	0.8002	0.8002
Media	0.3679	0.8502	0.6886	0.1767	0.6569	0.5357	0.3394	0.9526	0.7967

Tabla 5.3: Tabla con resultados AUC para el modelo de persona en silla de ruedas vs diferentes GT.

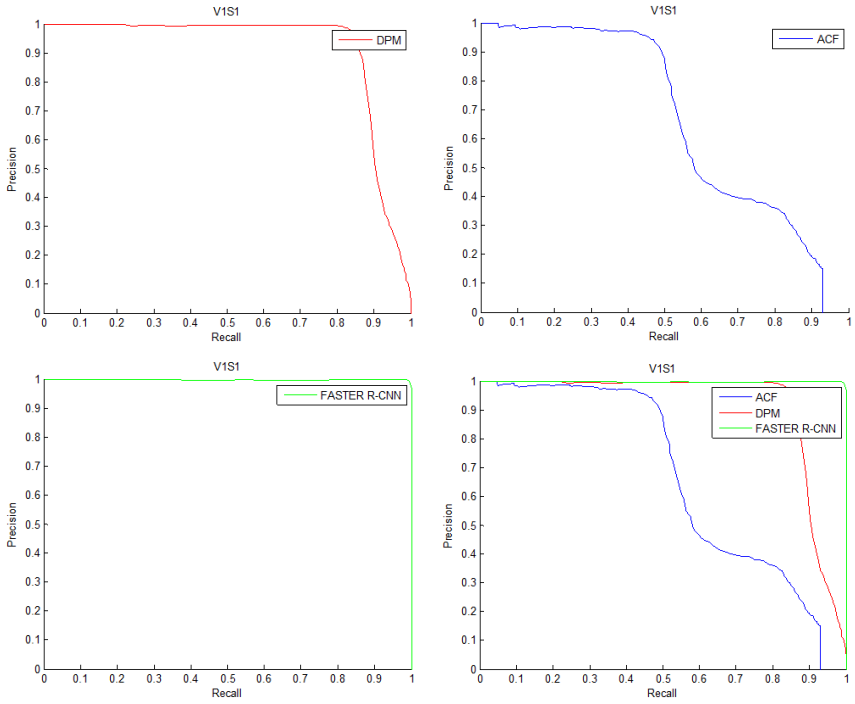


Figura 5.7: *Curve Presicion-Recall* para la secuencia V1S1 frente GT de personas en silla de ruedas.

a las personas en silla de ruedas se pueden observar en la figura 5.7.

Nuestra base de datos está formada siempre por personas en silla de ruedas, puesto que éstas aparecen en todas las secuencias mientras que las personas de pie solamente aparecen en algunas. Ésto hace que el detector FASTER R-CNN, a la hora de detectar personas en sillas de ruedas, obtenga resultados casi perfectos al compararlos con el GT de personas en sillas de ruedas.

Además, nos pasa exactamente lo mismo que para el caso anterior, 5.3.1, cuando comparamos los resultados obtenidos con los detectores con el GT de personas de pie, en secuencias donde no hay personas las curvas CPR serán inexistentes puesto que solamente se detectaran los falsos positivos.

En este caso, la influencia del número de personas y la pose en la que aparecen en las secuencias se puede ver en la columna donde comparamos los resultados con el GT total (personas de pie más personas en silla de ruedas). El valor AUC será menor en secuencias donde existen tanto personas de pie como en silla de ruedas que en secuencias en las que solamente hay personas en silla de ruedas. A la hora de ver estos resultados en las gráficas CPR lo lógico es esperar comportamientos similares al explicado en la subsección anterior. Éstos se aprecian para el detector FASTER R-CNN, y un poco para el detector DPM, en la figura 5.8.

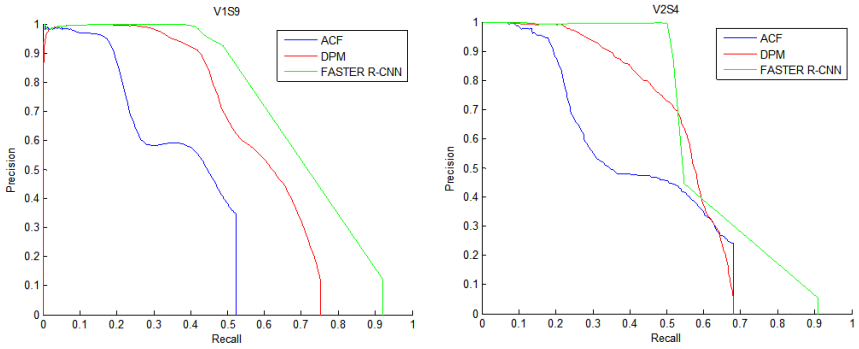


Figura 5.8: Ejemplos de curvas CPR para las detecciones en silla de ruedas frente a GT total.

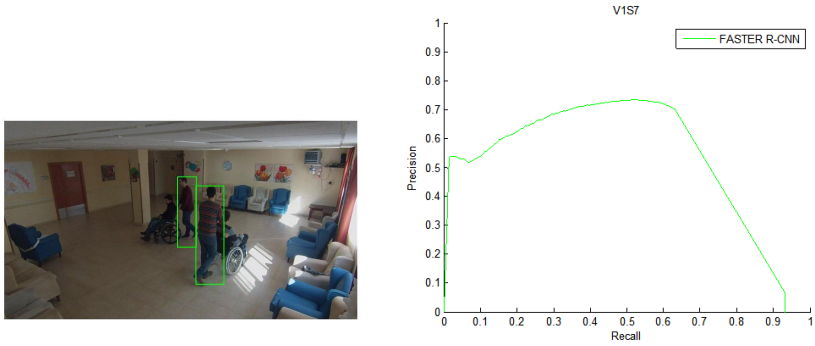


Figura 5.9: Detecciones para la secuencia V1S7 con el detector FASTER R-CNN frente a GT de personas en silla de ruedas.

Si nos fijamos en la tabla 5.3, para ambas secuencias S11 y la secuencia V1S7 se obtienen peores resultados en comparación con el resto. Si analizamos las detecciones conseguidas por nuestros algoritmos, en particular el FASTER R-CNN, puesto que es el que mejores resultados consigue en la detección de personas en silla de ruedas, éste no consigue detectar todas las personas en ambas secuencias S11 conformadas solamente por personas en silla de ruedas. La secuencia V1S7 es particular puesto que detecta personas de pie en vez de en silla de ruedas, de ahí sus malos resultados. En las figuras 5.9 y 5.10, se observan algunas frames con las detecciones que consigue el FASTER R-CNN para estas secuencias.

5.4. Conclusión

Podemos concluir este capítulo indicando que el detector que mejores resultados obtiene, en general, es el DPM, un detector "tradicional", basándonos en las gráficas obtenidas así como en las tablas 5.2 y 5.3, después de mostrar los diferentes resultados

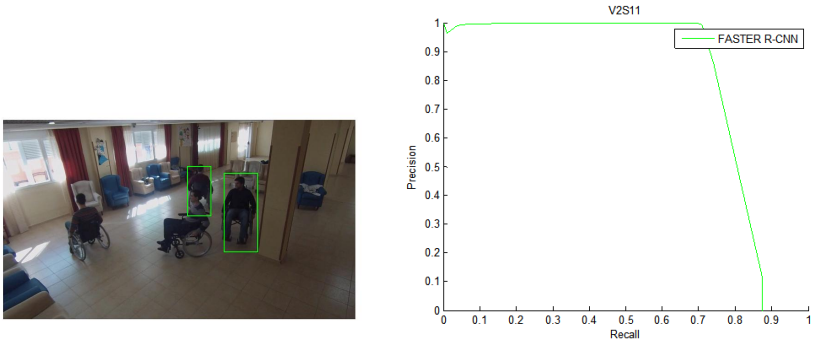


Figura 5.10: Detecciones para la secuencia V2S11 con el detector FASTER R-CNN frente GT de personas en silla de ruedas.

y factores de los que dependen los detectores como el número de personas o las diferentes poses que adoptan éstas en las secuencias.

Tras haber comparado los tres detectores en igualdad de condiciones, sobre la misma base de datos e igual métrica, a priori y según el estados del arte, capítulo 2, las redes convolucionales deberían funcionar siempre mejor, pero se observa claramente como, a pesar de ser entrenado el modelo con personas de pie, el detector basado en dichas redes convolucionales, FASTER R-CNN, detecta erroneamente personas en sillas de ruedas como personas de pie. Aunque ésto puede ser discutible, puesto que lo que está detectando son igualmente personas, aunque éstas estén sentadas en una silla de ruedas.

Capítulo 6

Conclusiones y trabajo futuro

6.1. Conclusiones

El objetivo principal de este trabajo era evaluar los resultados obtenidos, mediante el uso del detector basado en redes convolucionales, comparando dichos resultados con detectores más tradicionales, diseñando e implementando los diferentes modelos de personas necesarios así como establecer un marco de evaluación común en cuanto al *dataset* y la métrica de evaluación.

Para ello, en el capítulo 2 se ha realizado un estudio del estado del arte, basado en los artículos [9, 10, 11, 5], llegando a la conclusión de la utilización de una arquitectura en común por parte de los detectores y una clasificación de dichos algoritmos propuestos en relación a los diferentes métodos para obtener las regiones de interés así como los modelos empleados. También se introduce en este capítulo a los detectores basados en redes convolucionales a los que podemos incluir en dicha clasificación.

En el capítulo siguiente, 3, se fijan los diferentes algoritmos de detección de personas empleados en el trabajo y su funcionamiento, siendo los elegidos el detector *Deformable Parts Model*, DPM, y el detector *Aggregate Channel Features*, ACF, como los detectores "tradicionales", y el detector *Faster Region-based Convolutional Network*, FASTER R-CNN, el basado en redes convolucionales.

Los diversos modelos y parámetros para el entrenamiento de éstos se establecen en el capítulo 4.2 con el fin de que se entrenen en las mismas condiciones para poder compararlos y evaluarlos en el capítulo 5 habiendo fijado unos *dataset* comunes para el entrenamiento y otro *dataset* diferente para la evaluación. Además, estos resultados se comparan mediante la generación de las curvas *Curve Presicion-Recall*, CPR, y su área bajo la curva, *Area Under Curve*, AUC.

Éstos resultados obtenidos, las curvas y su área, se estudian teniendo en cuenta el

número de personas que aparecen en nuestro *dataset*, asi como las poses que adoptan o las visualizaciones de las distintas detecciones. Llegamos así a la conclusión de que el mejor algoritmo de detección de personas empleado está entre el DPM, un detector "tradicional", y el basado en redes convolucionales FASTER R-CNN, puesto que el primero obtiene mejores resultados en la búsqueda de personas de pie pero en la detección de personas en silla de ruedas los mejores frutos los consigue el FASTER R-CNN. Éste último no adelanta al DPM por un motivo que se puede o no tener en cuenta, puesto que a la hora de estimar o detectar a las personas de pie no solamente las llega a detectar sino que detecta también a las personas en silla de ruedas, que al fin y al cabo son personas, aunque sentadas.

6.2. Trabajo futuro

A la vista de los resultados que se han obtenido en este trabajo se propone trabajar sobre nuevos detectores así como diferentes modelos de personas, nuevas bases de datos u otras métricas de evaluación, detecciones sobre objetos, étc.

Los algoritmos "tradicionales" seleccionados han sido el detector DPM y el detector ACF, pero en trabajos futuros es conveniente la utilización de otros detectores para comparar con otro nivel de detalle el rendimiento de éstos en contraposición con los detectores basados en redes convoluciones, que no tienen que coincidir necesariamente con el elegido en este trabajo, el FASTER R-CNN. Se propone la utilización de detectores como el *Implicit Shape Model*, ISM, destinado a la detección de personas en lugares con grandes aglomeraciones de personas; o los detectores *Fusion* o *Edge* creados por el VPULab.

Por lo tanto, conviene entrenar más modelos de personas en diferentes poses como tumbadas, sentadas, agachadas, de rodillas, etc, o añadir más variaciones a los modelos empleados de persona de pie y en silla de ruedas. También sería útil comparar la eficacia y el rendimiento de los detectores no solo con modelos de personas sino con modelos de objetos.

La base de datos empleada en el trabajo para evaluar los diferentes algoritmos de detección de personas es un *dataset* compuesto principalmente por personas en silla de ruedas además de personas de pie. Sería bueno que se empleara otra base de datos compuesta por personas en toda clase de situaciones o posturas, no solamente personas de pie o en silla de ruedas, para determinar la eficacia de dichos detectores cuando hay personas de todo tipo, es decir, en escenarios mucho más reales.

Y por último, la métrica empleada es la de comparar las curvas CPR y el área bajo su curva, AUC, para valorar la efectividad de cada uno de los detectores. Ésta

puede variar comparándose así otros aspectos de los algoritmos, como por ejemplo, el coste computacional de cada uno de ellos.

Bibliografía

[1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 886–893, IEEE, 2005.

[2] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[3] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1532–1545, 2014.

[4] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," *BMVC*, vol. 2, p. 5, 2009.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

[6] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1440–1448, 2015.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, pp. 91–99, 2015.

[8] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*, pp. 818–833, Springer, 2014.

[9] Á. García-Martín and J. M. Martínez, "People detection in surveillance: classification and evaluation," *IET Computer Vision*, vol. 9, no. 5, pp. 779–788, 2015.

[10] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 4, pp. 743–761, 2012.

[11] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 12, pp. 2179–2195, 2009.

[12] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE, 2008.

[13] J. Yu, D. Farin, and B. Schiele, "Multi-target tracking in crowded scenes," in *Joint Pattern Recognition Symposium*, pp. 406–415, Springer, 2011.

[14] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *International journal of computer vision*, vol. 77, no. 1-3, pp. 259–289, 2008.

[15] P. Dollár, R. Appel, and W. Kienzle, "Crosstalk cascades for frame-rate pedestrian detection," in *Computer Vision–ECCV 2012*, pp. 645–659, Springer, 2012.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

Glosario

ACF	*Aggregate Channel Features*
AUC	*Area Under Curve*
CPR	*Curve Presicion-Recall*
DPM	*Deformable Parts Model*
FAST R-CNN	*Fast Region-based Convolutional Network*
FASTER R-CNN	*Faster Region-based Convolutional Network*
GT	*Ground Truth*
HOG	*Histogram of Oriented Gradients*
LVSM	*Latent Support Vector Machine*
R-CNN	*Region-based Convolutional Network*
ROI	*Region of Interest*
RPN	*Region Proposal Network*
SPPnet	*Spatial Pyramid Pooling Network*
SVM	*Support Vector Machine*
TFG	*Trabajo de Fin de Grado*
VPULab	*Video Processing and Understanding Lab*