UNIVERSIDAD CARLOS III DE MADRID

TESIS DOCTORAL

Detección de Peatones en el Espectro Visible e Infrarrojo para un Sistema Avanzado de Asistencia a la Conducción

Autora:
Cristina Hilario Gómez

Directores:
José María Armingol
Arturo de la Escalera

DEPARTAMENTO DE INGENIERÍA DE SISTEMAS Y AUTOMÁTICA
UNIVERSIDAD CARLOS III DE MADRID

Leganés, Octubre de 2008

TESIS DOCTORAL

Detección de Peatones en el Espectro Visible e Infrarrojo para un Sistema Avanzado de Asistencia a la Conducción

Autora: Cristina Hilario Gómez

Directores: José María Armingol

Arturo de la Escalera

Firma del Tribunal Calificador:

Firma

Presidente:

Vocal:

Vocal:

Vocal:

Secretario:

Calificación:

Leganés, de de

Índice

Capítulo 1

Introducción

Los peatones son los elementos más desprotegidos de la carretera en caso de colisión. Más de la tercera parte de los 1,2 millones de muertos y unos 10 millones de heridos anualmente en accidentes de tráfico en todo el mundo, son peatones [Wor08]. Según un informe del Consejo Europeo de Seguridad Vial (ETSC) publicado en el 2003 [ETS03], por cada kilómetro recorrido en la Unión Europea (UE) el riesgo de perder la vida en comparación con el de una persona que viaje en automóvil, era 8 veces mayor en el caso de un ciclista, 9 veces mayor en el caso de un peatón y 20 veces mayor en el caso de un motociclista. Las causas son múltiples, pero una de ellas destaca sobre las demás: al peatón no le protege ninguna coraza metálica como al automovilista, por lo que debe ser el vehículo el que proteja al peatón y no éste de los coches.

Los sistemas de protección de peatones tienen como objetivo reducir las cifras de siniestralidad a partir del desarrollo de sistemas que eviten el accidente (seguridad activa) o que minimicen las consecuencias una vez producido el siniestro (seguridad pasiva). Así como la seguridad pasiva ha ayudado enormemente a proteger a las personas cuando se encuentran dentro de un vehículo, no puede conseguirlo con igual eficacia cuando éstas están fuera. Sólo la seguridad activa será capaz de disminuir la gravedad y el número de accidentes en los que se vean involucrados peatones, ciclistas y motoristas. Las medidas que se están implementando en la actualidad, como por ejemplo capós que se elevan o airbags externos, están dirigidos a disminuir la peligrosidad del accidente, pero son medidas que se toman una vez que el accidente ha comenzado a producirse. Por ello, tanto para reducir el número de atropellos como para actuar con suficiente antelación en caso de que el accidente sea inevitable es imprescindible detectar previamente a los usuarios más vulnerables de la carretera. Como consecuencia, se está haciendo cada vez más hincapié en el concepto de seguridad activa, lo que se entiende por dotar al vehículo de sistemas inteligentes que predigan y eviten accidentes que el conductor por sí solo no puede controlar. Debido a problemas legales y psicológicos, el paso intermedio lo constituyen los Sistemas Avanzados de Asistencia a la Conducción (ADAS) que, sin llegar a tomar control del vehículo, avisen al conductor, con suficiente antelación, de un posible peligro.

Esta tesis se engloba dentro de este tipo de sistemas y su funcionamiento se especializa

dentro de entornos urbanos. Dichos entornos presentan dificultades añadidas a las carreteras y autopistas: existen otros objetos con características y formas similares a la de los peatones, además de ser frecuentes las oclusiones de éstos. El sistema propuesto detectará a los peatones que aparezcan en la escena, sin imponer ninguna restricción en cuanto a la iluminación, vestimenta o posicionamiento. Para ello, se propone construir un Sistema de Ayuda a la Conducción que vigile la parte frontal del vehículo. Se ha desarrollado un sistema estéreo en el espectro visible y un sistema estéreo infrarrojo. Ambos sistemas se han situado en dos vehículos distintos; el primero ha sido probado en el IvvI (Intelligent Vehicle based on Visual Information) desarrollado en la Universidad Carlos III de Madrid, y el segundo, hace referencia al sistema GOLD implementado por la Universidad de Parma y utilizado para el análisis del espectro infrarrojo de esta tesis. En cada sistema se han abordado distintas fases de la detección de peatones. En conjunto se ha investigado e implementado cada una de las etapas necesarias para un sistema de detección de peatones completo; determinación de las regiones de interés, detección de los candidatos potenciales, reconocimiento de los peatones e integración temporal de los resultados.

1.1. La accidentalidad de los peatones

Un estudio sobre la carga mundial de morbilidad realizado por la Organización Mundial de la Salud (OMS), la Universidad de Harvard y el Banco Mundial puso de manifiesto que en 1990 [ML96], los accidentes de tráfico eran considerados como el noveno problema de salud más importante en el mundo. El estudio preveé que para el año 2020 los accidentes de tráfico ascenderán hasta el tercer lugar en la tabla de principales causas de muerte y discapacidades a las que se enfrenta el mundo. Cabe destacar que la cantidad anual de muertos está en relación inversa con el grado de desarrollo económico y social. Los países más desarrollados, que tienen proporcionalmente una mayor cantidad de vehículos por habitante, tienen anualmente una menor cantidad de víctimas mortales. Según datos recogidos en la Asamblea General de las Naciones Unidas sobre la crisis de seguridad vial en el mundo [dSG03], en 2000 las lesiones sufridas en accidentes de tráfico mataron a más de 1 millón de personas en los países de bajos y medianos ingresos (90 % de la mortalidad mundial debida a colisiones de vehículos de motor), y a 125.000 (10 %) en los países de elevados ingresos.

Dentro de los entornos viarios, los elementos más vulnerables lo forman los peatones, ciclistas y motoristas, tanto por la fragilidad de su estructura como por la dificultad para detectarlos. Según un informe sobre prevención de los traumatismos causados por el tránsito elaborado en forma conjunta por la Organización Mundial de la Salud y el Banco Mundial [TT04], anualmente fallecen 39.000 peatones en el mundo mientras que 436.000 son heridos, lo que representa el 24 % de las víctimas mortales. En la UE ocurren anualmente 13.8 millones de accidentes, que dejan un saldo de 38.000 fallecidos y 1.7 millones de heridos. Sólo desde el punto de vista económico el coste es altísimo, casi el 2 % del Producto Nacional Bruto de los países de la Unión Europea, unos 160 mil millones de euros [ETS03]. El número de peatones y ciclistas fallecidos en ella es de casi 8.000, con unos 300.000 heridos, frente a los 3.300 muertos y 27.000 heridos en Japón. En Norte América, el número de peatones muertos son

5.000 y 85.000 heridos; y en Korea estas cifras son de 3.600 fallecidos y 90.000 heridos. A nivel mundial las estimaciones indican que los costes económicos de las lesiones causadas por accidentes de tráfico ascienden a 518 mil millones de dólares anuales (cerca de 337 mil millones de euros) [dSG03].

La información sobre la accidentalidad de los peatones muestra que la mayoría de los atropellos tienen lugar en entornos urbanos, pudiendo infringir lesiones graves incluso a velocidades bajas. Estos accidentes se producen durante el día y los grupos de edad de mayor riesgo son los niños y las personas mayores. Según la Dirección General de Tráfico (DGT), en 2004 [dT04], de las 4.741 personas que perdieron la vida en las carreteras españolas, 683 fueron por atropello. Un 40 % de los peatones fallecidos superaba los 64 años, el 26 % los 74 y el 30 % eran menores de 15 años.

En un estudio realizado en la región alemana de Hannover sobre 663 accidentes [Fue05] se vió que la mayoría ocurrieron mientras el peatón estaba en movimiento (94 %), a una velocidad del vehículo inferior a 60Km/h (96 %) y frontales (70,6 %). Esto significa que las piernas impactan con el parachoques y, entre 50 y 150 milisegundos después, el cuerpo y sobre todo, la cabeza, chocan contra el capó o el parabrisas. Cada una de estas partes del cuerpo sufren más del 30 % del total de los accidentes.

La reducción de la velocidad del vehículo es una prioridad para prevenir las colisiones en carretera. El 75 % de las lesiones ocasionadas a peatones ocurren a velocidades superiores a 40Km/h. Según varios estudios sobre las consecuencias de la velocidad en cuanto a lesiones sufridas en accidentes de tráfico, una reducción del 1 % de la velocidad disminuye las probabilidades de lesión en un 2 % a 3 %, y los casos de accidentes mortales en aproximadamente el doble. Las consecuencias en cuanto a lesiones en los peatones también se ven muy afectadas por la velocidad del vehículo: cuando la velocidad de un coche aumenta de 30 a 50 kilómetros por hora, la probabilidad de muerte de un peatón se multiplica por ocho [dSG03]. De manera que, limitando la velocidad máxima permitida en el entorno urbano hasta un máximo de 40Km/h se reducirían significativamente el número de accidentes frontales, así como la gravedad de las lesiones. En cifras, esto habría evitado el 85 % de los accidentes en los que el conductor del vehículo percibe la situación de conflicto. En el 15 % restante el conductor ni siquiera llega a percatarse de la situación de peligro, y únicamente lo hace en el preciso momento del impacto contra el peatón. Un estudio del Instituto Mapfre de Seguridad Vial realizado en Madrid entre 2000 y 2003 [dSV00] demostró que, en muchos casos, los atropellos se debían a que la fase verde del semáforo para peatones era insuficiente para atravesar la calle. Otro estudio de la revista "Consumer" [Con03], indica que cuatro de cada diez conductores españoles no respetan el paso de cebra y uno de cada cinco peatones cruza el semáforo en rojo .

En los últimos años se han registrado progresos importantes en cuanto a la protección de los ocupantes de los vehículos mediante la introducción de normas legislativas para regular el impacto de las colisiones frontales y laterales. Sin embargo, todavía no se han conseguido progresos similares con respecto a las lesiones sufridas por los peatones. Los traumas craneales causados por golpes de los parachoques y cascos son responsables del 80 % de las lesiones graves en las colisiones contra peatones. La protección de los ocupantes de los vehículos y de

los peatones pueden mejorarse aún más garantizando que los vehículos se equipen con dispositivos y mecanismos de seguridad adecuados. Se requieren leyes y medidas de imposición de la ley para garantizar unas normas mínimas de seguridad para el diseño de las partes frontales de los vehículos de motor a fin de hacerlos menos peligrosos. También se requiere esforzarse en mayor medida para promover tecnologías de seguridad que puedan contribuir a la prevención de las colisiones.

1.2. Sistemas de Protección de Peatones

Los sistemas de protección de peatones hacen referencia a la incorporación de las tecnologías más avanzadas en el vehículo, con el fin de proteger a los ocupantes de éste, así como a otros usuarios especialmente vulnerables de la vía. Un análisis sobre los principales sistemas aparecidos en los últimos años para la protección de los peatones, permite distinguir entre los antes citados sistemas de seguridad activa o primaria y sistemas de seguridad pasiva o secundaria. Como ya se ha dicho antes, la seguridad primaria hace referencia a aquellos sistemas diseñados para evitar que ocurra el accidente, mientras que la seguridad secundaria comprende a los sistemas diseñados para minimizar las consecuencias del accidente en el caso de que éste, finalmente, no pueda ser evitado.

Como principales sistemas de seguridad primaria se pueden citar los sistemas de asistencia a la frenada, la mejora de la visibilidad nocturna y la detección automática de la presencia de peatones en la escena. Entre las mejoras en los sistemas de seguridad secundaria destacan el desarrollo de nuevos materiales para el frontal de los vehículos, la propia estructura de los capós, los parachoques delanteros y, los más recientes, los capós activos y los airbags para peatones. La seguridad en los vehículos se ha enfocado tradicionalmente como una minimización de los efectos del choque. Siendo indudable su enorme impacto positivo en la disminución del número de afectados y en la severidad de las lesiones, ahora se pone el énfasis en evitar el accidente y, si éste se produce, qué medidas tienen que actuar durante los primeros milisegundos para minimizar los daños.

Si bien es importante desarrollar sistemas para la protección de peatones, no lo es menos el someter a ensayos a esos vehículos para evaluar el nivel de protección que ofrecen tanto a los ocupantes como a los peatones. Se trata de reproducir en los laboratorios de ensayo los accidentes ocurridos en la realidad, para así cuantificar tanto las medidas pasivas como las activas. Los resultados de estas pruebas pueden ayudar a reducir el riesgo del primer impacto contra el coche y del segundo contra la carretera, al permitir diseñar elementos más seguros para el peatón tras el análisis del choque.

1.2.1. Seguridad Pasiva

Dentro de los sistemas de seguridad pasiva se encuentran muchos de los dispositivos de obligado uso en la actualidad, pero cuya implantación en los vehículos ha sido progresiva, y en algunos casos, lenta. Los elementos fundamentales en relación con la seguridad pasiva en un vehículo son:

- Cinturón de seguridad: Evita que los ocupantes salgan despedidos del vehículo en caso de impacto.
- Airbag: Disminuye el contacto de los ocupantes del vehículo con los elementos del interior mediante una bolsa de aire que se infla en milésimas de segundo.
- Reposacabezas: Frena el movimiento del cuello, evitando lesiones cervicales.
- Interiores ergonómicos: Consigue que el conductor circule con mayor comodidad y esté más atento a lo que ocurre en la carretera.

Además de éstos, hay dos aspectos aún más importantes, que son la estructura o carrocería y los sistemas de deformación progresiva. Ambos afectan a la protección de los ocupantes del vehículo.

1.2.2. Seguridad Activa

La seguridad activa engloba a los dispositivos sobre los que el conductor puede actuar directamente. El objetivo es reducir el número de accidentes en la carretera gracias a un equipamiento específico que confiere estabilidad a los vehículos y disminuye el riesgo de colisión. Los dispositivos fundamentales en relación con la seguridad activa son:

- Sistema de frenado: Evita el bloqueo de las ruedas al frenar y detiene el vehículo, impidiendo que éste patine. Conocido como frenos anti-bloqueo o ABS (*Antilock Braking System*), ayuda a mantener el control de la dirección.
- Sistema de control de tracción: Garantiza la estabilidad durante la conducción, recuperando la adherencia entre neumático y el firme cuando el conductor se excede en la aceleración. Es un sistema desarrollado sobre la base del ABS y comúnmente se denomina tanto con las siglas ASR (*Automatic Stability Control*) como TCS (*Traction Control System*).
- Sistema de control de estabilidad: Evita el vuelco del vehículo y corrige automáticamente la trayectoria, impidiendo que el conductor pierda el control del vehículo. Es uno de los más revolucionarios avances en seguridad activa de los últimos años, conocido como sistema ESP (*Electronic Stability Program*).

Los sistemas ABS, los sistemas de suspensión o TCS y los sistemas de control de la estabilidad o ESP son las tres referencias claves a la hora de hablar de seguridad activa. Además de éstos, existen otros sistemas a cuyo uso estamos habituados: el sistema de dirección (que hace girar las ruedas de acuerdo al giro del volante), el sistema de climatización (que proporciona la temperatura adecuada durante la marcha), los neumáticos (cuyo dibujo es garantía de agarre incluso en situaciones climatológicas adversas), el sistema de iluminación (que permite al conductor ver y ser visto), y por último, el motor y la caja de cambios (que permiten adaptar la velocidad a las circunstancias de la carretera).

La seguridad activa está pensada para garantizar el buen funcionamiento de un vehículo en movimiento y responder a las órdenes del conductor. Según un informe del Real Automóvil Club de Cataluña (RACC) [RAC04], muchos de los accidentes ocurridos en la carreteras europeas son ocasionados por la deficiente seguridad activa de los vehículos. El estudio advierte, a modo de ejemplo ilustrativo, de la peligrosidad de un vehículo que al realizar una maniobra brusca, patina y no puede ser controlado por el conductor. Precisamente, la pericia al volante de éste y la precaución son las claves para evitar un accidente, siempre y cuando el automóvil responda como le pide el usuario.

Según datos de la DGT [dT05b], el accidente más frecuente en el 2005 fue, como en años anteriores la salida de vía del vehículo (39 %), mientras que las causas fueron, por este orden, la distracción, la velocidad y las maniobras antirreglamentarias. Un estudio estadounidense sobre el comportamiento del conductor realizado por el National Highway Traffic Safety Administration (NHTSA) [NHT97] encontró que el 99 % de los accidentes analizados eran debidos a errores humanos, siendo la causa más común debida a errores de percepción. Todos estos datos revelan que el conductor sigue siendo el principal responsable de los siniestros, y ponen de manifiesto la necesaria intervención en materia de seguridad activa para ofrecer al usuario de los mecanismos suficientes que le ayuden a reducir el riesgo de colisión.

1.2.3. Tests de Ensayo

Existen diferentes test de choque que evalúan la seguridad ofrecida por un vehículo durante el impacto con un peatón. Hay dos filosofías de test distintas:

1. Test completos (*full-scale*): En estos tests se reproduce, de un modo bastante realista, el accidente completo. En principio, sólo se sustituye al humano por un maniquí antropomórfico. Sin embargo, no se garantiza que estos tests sean capaces de reproducir la colisión. El modelo humano empleado hace que la biofidelidad sea cuestionable. Además, se precisa de una adquisición de datos compleja y la preparación de cada experimento individual consume mucho tiempo. Las simulaciones numéricas tienen el potencial de permitir evaluar la situación completa, pero aún así, resulta muy complicado obtener un modelo del peatón capaz de predecir las lesiones con precisión. (ver fig. 1.1-a).

2. Tests de componentes: Estos tests están diseñados para reproducir sólo la parte crítica de todo el accidente. Dado que las muertes pueden atribuirse casi exclusivamente al impacto con la cabeza, muchos tests se centran en analizar esa parte. Es necesario mucho conocimiento adicional para poder interpretar los resultados del modo correcto. En una situación compleja, como es el caso de un accidente, es necesario que el conjunto de parámetros del test sean dependientes de la geometría de la parte delantera del vehículo. A pesar de esto, la mayoría de los test de componentes en Europa, usan parámetros de impacto más o menos independientes de la forma del vehículo. (ver fig. 1.1-b).

   En [KFS05] proponen usar simulaciones numéricas para obtener parámetros de test dependientes de la forma del vehículo. Las simulaciones numéricas permiten obtener información sobre la cinemática de cada coche en concreto y para condiciones de choque

diversas. Para que esas simulaciones sean realistas, se emplean datos estadísticos de accidentes con peatones. Plantean un procedimiento de evaluación llamado índice VERPS (*Vehicle Related Pedestrian Safety index*), capaz de hacer frente a la mayoría de los inconvenientes un test de componentes convencional. Basándose en ese índice VERPS demuestran cómo la seguridad del peatón depende de la forma de la parte delantera del coche y cómo varía de adultos a niños. Estos tests ponen de manifiesto como ciertas medidas técnicas, como capós deformables, pueden aumentar la seguridad. De todas formas, su eficacia depende fuertemente de la geometría frontal de cada vehículo y es distinta para adultos y niños, ya que, debido a su diferencia en estatura, el vehículo choca con distintas partes de sus cuerpos. Por este motivo, una medida técnica rara vez afecta positivamente a todos los grupos de personas. Además, la comparación de distintos sistemas permite identificar las soluciones más eficientes desde un punto de vista económico.

Figura 1.1: Ejemplos de tests de impacto obtenidos de [KFS05], realizados en la *Technical University of Berlin*; (*a*) test de impacto de completo y, (*b*) test de impacto de componentes.

Estas pruebas se enmarcan en una tendencia europea, según la cuál se establece que los conductores siempre van a cometer errores. Es la llamada ”teoría del riesgo constante”, que dice que cuando el usuario percibe carreteras y vehículos más seguros, tiende a subir el nivel de riesgo y comete más errores, precisamente porque se siente más seguro. Trabajando con este planteamiento se pretende construir vehículos que transmitan seguridad, pero que el conductor perciba que sigue existiendo riesgo de accidente y por tanto, conduzca con precaución.

1.3. Nuevas Regulaciones y Pruebas de Ensayo

A partir de 1970, una serie de gobiernos europeos han estado trabajando, a través de la Comisión Experimental de Vehículos Europeos (EEVC, European Experimental Vehicles Committee), en el desarrollo de procedimientos y equipos para la evaluación de diversos aspectos de la seguridad pasiva de un vehículo. En 1995, los tests de la EEVC realizados a vehículos se generalizaron y surgió el Programa Europeo de Evaluación de Vehículos Nuevos, EuroNCAP [Eur95]. Es un programa de seguridad para automóviles apoyado por varios gobiernos europeos, fabricantes y organizaciones relacionados con el sector de la automoción a nivel mundial.

EuroNCAP realiza diversas pruebas de choque; las pruebas de seguridad pasiva miden el comportamiento del vehículo frente a distintos tipos de impacto frontal y lateral; en los últimos años se ha incorporado una prueba de medición de seguridad de niños a bordo, así como de peatones en caso de atropello. Mediante un sistema homologado de estrellas, el usuario puede hacerse una idea de la seguridad que ofrece un vehículo, siendo la máxima protección posible equivalente a cinco estrellas.

A pesar de que el EuroNCAP incluyó tests de protección a peatones, los accidentes de éstos seguían siendo elevados. En Europa el 15 % de las muertes que se producen anualmente en accidentes de tráfico son peatones. Ante esta situación, en el año 2000 [Com01a], el Parlamento Europeo aprobó un proyecto de legislación y alcanzó un acuerdo con fabricantes europeos, japoneses y coreanos para modificar los parachoques y capós de los vehículos con el fin de reducir, en un gran porcentaje, los efectos de una colisión sobre el peatón.

Un poco más tarde, en el 2003, la Unión Europea decidió adoptar normas legislativas vinculantes para que los vehículos nuevos mejorasen su seguridad con respecto a los peatones en dos etapas [Com03]. La primera etapa del Programa de Acción Europeo de Seguridad Vial comenzó el 1 de octubre de 2005, y con ella se pretende que los modelos lanzados al mercado a partir de esta fecha ofrezcan algún sistema de protección al peatón en caso de atropello. Más adelante se comentan los efectos de esta medida. La segunda fase de esta normativa para la protección de peatones se iniciará en el 2010 y es la más desafiante; a partir de esa fecha todos los coches nuevos deberán estar diseñados para conseguir dos estrellas (sobre un máximo de cuatro) en los test de choque de EuroNap y a partir de 2015, todos deberán alcanzar la máxima puntuación o no podrán ser comercializados.

Como consecuencia de la entrada en vigor de la nueva directiva y a la presión generada por el programa EuroNCAP, los fabricantes de vehículos están trabajando para ofrecer mejores sistemas de seguridad de peatones que se ajusten a las exigencias. Se están desarrollando tanto soluciones para la seguridad pasiva (airbags, capós o parachoques deformables), como para la activa (sistemas de ayuda a la conducción y sistemas automáticos), debido a que ambas están aún en su fase de desarrollo. Finalmente, se están realizando desarrollos en los sistemas post-colisión, como el aviso automático a emergencias en caso de accidente (*eCall*) [Com05], entre otros.

La mayor parte de la tecnología de seguridad desarrollada entre los años 1960 y 1990 se clasifica como la ya comentada seguridad pasiva. El desarrollo en seguridad activa comenzó hace más de 20 años con la invención de los frenos anti-bloqueo, más conocidos como ABS. A pesar de que, inicialmente se espera que el mayor crecimiento lo experimenten los sistemas de seguridad de peatones pasivos, el desarrollo de tecnologías para la percepción de peatones fiables garantizará una importante evolución de los sistemas de seguridad activos, necesaria para poder cumplir con la segunda fase de la normativa de protección de los peatones, después del 2010. Con este plan de seguridad vial, la UE espera reducir el número de fallecidos a la mitad.

En los Estados Unidos y en otros lugares, también existen Programas de Evaluación de Vehículos Nuevos (NCAP, New Car Assessment Programs) que realizan este tipo de tests y permiten que los resultados puedan ser accedidos por los usuarios. En 1979, la NHTSA (Na-

tional Highway Traffic Safety Administration) inició el NCAP, donde se realizaban pruebas de impacto frontal a una velocidad máxima de 35 mph. Mucho más tarde, se inició un programa NCAP en Australia y otro se empezó a elaborar en Japón. Si los vehículos tienen que cumplir con las recomendaciones de la EEVC, se estima una disminución en el número de peatones muertos superior al 20 % [fT05].

1.4. Situación Actual de la Seguridad en los Vehículos

Muchos de los sistemas de seguridad han sido introducidos hace relativamente poco en el mercado, y gracias a ellos, la evolución de la seguridad de los peatones muestra una tendencia positiva, aunque lamentablemente por detrás de los avances experimentados por los sistemas de protección al conductor y de los pasajeros. En este sentido, hay mucho que agradecer a los test de choque tipo EuroNCAP, ya han propiciado que algunos dispositivos antes considerados un lujo, como el airbag o el ABS, se hayan convertido en un equipamiento de serie para la mejora de la seguridad. Del mismo modo, se está ya solicitando que otros dispositivos cuya efectividad para evitar accidentes y muertes ha sido demostrada, estén también incluidos de serie en todos los vehículos. Es el caso del control de la estabilidad ESP.

Un estudio realizado por la GDV (asociación alemana de aseguradoras) demuestra que aproximadamente el 25 % de los accidentes de tráfico que provocan lesiones graves se deben a que el vehículo ha patinado. Según este estudio, este porcentaje podría reducirse considerablemente si todos los vehículos tuvieran ESP. Por ejemplo, el uso combinado del sistema de control de estabilidad (ABS) y del control de deslizamiento lateral (ESP) han permitido reducir en un 20 % los accidentes mortales en la Unión Europea, pasando de 50.000 víctimas mortales al año a 40.000 en los últimos diez años. En concreto en España, según datos de la Asociación Española de Fabricantes de Automóviles y Camiones (ANFAC) [ANF05], de los coches comercializados en el 2005, sólo un 48.7 % tienen de serie el sistema ESP (ver tabla 1.1).

Sistemas	Año 1995	Año 2005
ABS	48.5 %	100.0 %
Airbag conductor	53.6 %	100.0 %
Airbag pasajero	21.8 %	96.0 %
Airbag lateral	—	77.0 %
ESP	—	48.7 %

Tabla 1.1: Tabla que compara las tasas de los sistemas de seguridad comercializados en vehículos en 1995 y 2005. Datos obtenidos de la Asociación Española de Fabricantes de Automóviles y Camiones (ANFAC) [ANF05].

La evolución del nivel de protección a peatones queda patente a partir de las puntuaciones otorgadas por EuroNCAP a los nuevos modelos de vehículos: en los últimos cuatro años, el número medio de estrellas conseguidas por los automóviles en el aspecto relativo a la protección de peatones prácticamente se ha duplicado, pasando de 1.1 estrellas a 1.9 en la actualidad.

Los últimos test realizados son especialmente esperanzadores. Recientemente, el Citroën "C6" ha conseguido la mejor nota de protección al peatón (4 estrellas) en toda la historia de EuroNCAP [dT06]. De hecho, de los nueve modelos analizados – algunos de marcas tan prestigiosas como Audi, Mercedes o BMW –, sólo tres consiguen dos estrellas; Lexus "GS 300", Saab "9-5" y Volvo "S80". Dentro de los todoterrenos, hay que destacar las tres estrellas obtenidas únicamente por el "CR-V" de Honda. Citroën ha utilizado un sistema en su "C6", ya empleado por otras marcas como Honda (en su modelo "Legend") o Jaguar (en su modelo "XK"). Este sistema detecta el momento exacto del atropello gracias a unos sensores en un parachoques especial y eleva el capó 65 milímetros en 40 milisegundos, reduciendo los daños que produce el impacto de éste con la cabeza del peatón. En el 2002 Ford (grupo americano al que pertenece Jaguar) desarrolló un sistema parecido, que tras el impacto retrocedía el capó de forma mecánica. Al mismo tiempo, introdujeron modificaciones tanto en la estructura de sus parachoques, empleando nuevos materiales menos agresivos para el peatón, como en el diseño de sus faros, tratando de rebajar la gravedad de las lesiones sufridas por el peatón. En este sentido, en el 2005, la empresa española HBPO desarrolló un proyecto conjunto con Hella, Behr y Plastic Omnium. Han diseñado un frontal que evita daños graves, incluso a velocidades de 40km/h. Además están trabajando en parachoques deformables y también en faros, intermitentes y radiadores preparados para sufrir una colisión con un peatón [HBP05].

El diseño de la estructura de los vehículos está sufriendo cambios y se preveé que sean necesarios aún más para el cumplimiento de la normativa [Com03]. El objetivo de estas medidas es minimizar el riesgo de lesiones, diseñando las "zonas de contacto" para que cedan en la mayor medida posible. Además, se están explorando nuevos caminos, como son los faros direccionales, que amplían el campo visual del conductor o la utilización de cámaras infrarrojas colocadas en el parabrisas o debajo del parachoques, capaces de ver un peatón en la oscuridad y así evitar atropellos. En esta línea de investigación trabaja desde 2002 un grupo formado por Volkswagen, DaimlerChrysler, Siemens, VDO, CA y Faurecia (empresa del grupo PSA, formado por la alianza entre de Citroën y Peugeot).

Los resultados obtenidos de los test de EuroNCAP han ayudado a mejorar de manera considerable la seguridad de los ocupantes y de los peatones, ya que los fabricantes tratan de equipar a sus modelos cada vez con más y mejores sistemas, lo que supone una reducción de los accidentes y su gravedad. Este tipo de tests, pensados para evaluar la protección de los vehículos en caso de atropellos, son representativos de la realidad de los accidentes que ocurren a diario en las vías públicas. Aunque las pruebas de protección a peatones se realizan a una velocidad de 40km/h, sí es representativa de la realidad que se observa en las ciudades, según ha mostrado el análisis llevado a cabo por FITSA-IDIADA [FI07]. Añaden que aunque la zona de impacto suele estar bien acotada en caso de atropello, la dinámica de impacto en los casos reales es mucho más variable que en las pruebas de laboratorio. Por lo que un análisis del frontal de cada vehículo permitiría determinar para cada coche un ángulo y una velocidad más representativos de la realidad.

1.5. El Futuro de los Sistemas de Protección a Peatones

La mayoría de los fabricantes de automóviles ven el futuro de la seguridad en dispositivos y sistemas de prevención de colisiones más que de reacción ante ellos. En un futuro cercano los coches serán capaces de percibir mucho antes que nosotros la inmediatez de un accidente, para así, o bien alertar al conductor, o permitir que el vehículo tome acciones evasivas.

Sin embargo, las tendencias de evolución del sector son mucho más ambiciosas y apuntan en dirección a la creación de vehículos inteligentes [BBF99a]. Estos vehículos del futuro tendrán capacidad de decisión propia en caso de que detecten la inminencia de un accidente. Antes de que éste ocurra, su cerebro electrónico decidirá cuál es la mejor forma de reaccionar y dará las órdenes precisas para que ruedas, frenos, dirección, ayudas electrónicas, cinturones de seguridad y otras medidas empiecen a trabajar. Si, a pesar de todo, no es capaz de evitar el impacto, el coche evaluará los daños y pedirá ayuda a los servicios de emergencia. Para alcanzar ese ideal de vehículo inteligente la tendencia es transformar los coches en máquinas dotadas de algo muy parecido a un cerebro propio. En la figura 1.2 se muestran algunos prototipos de vehículos inteligentes existentes en la actualidad.

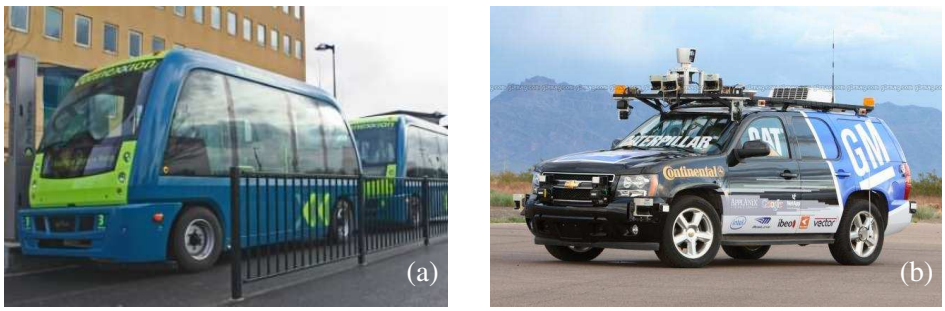


Figura 1.2: Ejemplos de vehículos del futuro; (*a*) El *ParkShuttle* en Gran Bretaña, es un sistema de transporte automatizado (imagen obtenida de http://connectedcities.eu/) y, (*b*) El vehículo que ha ganado el *DARPA Urban Challenge* en el 2007 (imagen obtenida de http://www.tartanracing.org/)

En la década pasada, la parte electrónica de los automóviles no dejó de aumentar, alcanzando en la actualidad el 30 % del coste total del vehículo. Una vez instalados los sistemas como el frenado de urgencia, la inyección electrónica o los airbags, las innovaciones siguientes se centran en ayudas a la circulación, la visión nocturna, regulación de la velocidad y la distancia (ACC, *Adaptive Cruise Control*) e incluso asistencia para el aparcamiento. Los nuevos productos que se perfilan en el horizonte son el *Stop & Go* (conducción automática en embotellamientos para evitar el estrés del conductor), la anticolisión, la adaptación inteligente de la velocidad (ISA, *Intelligent Speed Adaptation*), entre otras.

Esto que hoy en día parece ciencia ficción, puede llegar a ser común en un plazo de 10 años. De hecho, proveedores como Bosch, Autoliv, Delphi o Siemens, están comercializando y activamente investigando productos en estas líneas. Según la consultora *Frost & Sullivan* [FS06], dentro del segmento de los sistemas de seguridad pasiva, las soluciones introducidas en la estructura del vehículo tuvieron una penetración en el mercado de 72.1 % en el 2006, y

seguirá incrementándose hasta alcanzar el 100 % entre el 2011-2012. Los sistemas deformables reversibles y los no-reversibles van a sufrir un aumento desde su casi inexistencia actual, hasta alcanzar un 21 % y un 26 % respectivamente, en el 2015. En concreto, se prevee que los sistemas deformables reversibles tengan un alto grado de aceptación en los vehículos de lujo, mientras que los no-reversibles se espera que sean comunes en vehículos todoterrenos y monovolúmenes. Dentro del segmento de los sistemas de seguridad activa, cabe esperar que para el 2010 la mitad de los vehículos de lujo tengan algún tipo de sistema de aviso al conductor y para el 2015, un 25 % de los vehículos tendrán un sistema de protección de peatones que avise al conductor, mientras que menos del 2 % dispondrá de sistemas automáticos.

Finalmente, cuanto mayor sea el número de componentes de seguridad activa que se pongan en conjunto, más cerca estará la industria de poder ceder el control de un coche a un ordenador – de a bordo o no – en situaciones de emergencia, en autopistas o quizás, a tiempo completo. De momento, los fabricantes de vehículos no van a imponer al público los sistemas de seguridad activa. En lugar de esto, como ya es el caso, los componentes se van añadiendo poca a poco, pieza a pieza.

1.6. Sistemas Avanzados de Ayuda a la Conducción

Los sistemas de protección a peatones, forman parte de los Sistemas de Transporte Inteligente (ITS, *Intelligent Transportation Systems*). El interés por los ITS tuvo su inicio hace más de dos décadas, para afrontar la creciente demanda de movilidad de pasajeros y carga a nivel mundial.

Los ITS son el conjunto de soluciones innovadoras que surgen como consecuencia de la aplicación de las tecnologías de la información al sector del transporte (ver fig. 1.3). Estos sistemas integran el vehículo, la infraestructura y el conductor con el fin de conseguir condiciones de tráfico más fluidas y seguras, empleando comunicaciones electrónicas y tecnología computacional de vanguardia.

Figura 1.3: Ejemplos de la aplicación de las tecnologías de la información al sector del transporte.

Las nuevas herramientas de la tecnología de la información y de la comunicación (TIC) juegan un papel muy importante en la reducción de accidentes. Es el caso de las tecnologías encaminadas a incrementar la capacidad de control del coche, como el ABS o el ESP. Dentro de los ITS, los Sistemas Avanzados de Ayuda a la Conducción (ADAS, *Advanced Driver Assistance Systems*) suponen un paso más; predicen y evitan un accidente que el conductor

por sí solo no puede controlar. Son sistemas orientados a obtener información tanto del estado del vehículo como de la carretera, y transmitirla al conductor, para que en base a ella pueda tomar las decisiones necesarias.

A largo plazo, la pregunta sobre cuánto control dejarle al usuario del vehículo puede volverse un tema candente entre industria y gobierno. La tecnología va a estar ahí; la pregunta es si los conductores están preparados para ello. Aun cuando los sistemas de seguridad en los coches o camiones alcancen ese nivel de sofisticación, la respuesta a la pregunta será debatible. Los beneficios de una conducción completamente automática son evidentes: económicos, medioambientales, sociales y en seguridad. Sin embargo, debido a dificultades técnicas, económicas, psicológicas y legales parece más sensato, de momento, poner el énfasis en el desarrollo de sistemas avanzados de ayuda a la conducción o ADAS (Advanced Driver Assistance Systems).

El que estos sistemas de asistencia, al final sean un paso intermedio a la conducción automática o no, dependerá sobre todo de cómo se resuelvan los problemas legales y psicológicos. Los equipos deben ser los mismos, pero se relaja en una pequeña parte la exigencia de robustez del sistema (con una mayor tolerancia ante falsas alarmas) y se logra que los conductores vayan confiando en el ordenador mientras se reduce el número de accidentes.

En un futuro, las nuevas tecnologías podrían proveer vehículos con diferentes tipos y niveles de ”inteligencia” para complementar al usuario. Los sistemas de información, amplían el conocimiento del conductor en cuanto a rutas y localizaciones. Los sistemas de advertencia, como las tecnologías anti-colisión, aumentan sus habilidades para percibir el entorno. Las tecnologías de automatización y asistencia al conductor, simulan el sistema senso-motor del usuario para manejar el vehículo temporalmente en caso de emergencias o durante periodos más prolongados. Desde 1980, las principales empresas automovilísticas y otras no tan importantes, han estado desarrollando sistemas de navegación de a bordo basados en computador. En la actualidad, la mayoría de los sistemas desarrollados o en vías de desarrollo a lo largo del mundo, han incluido funciones más complejas para ayudar a conducir con mayor seguridad y eficiencia.

Sin embargo, a pesar de su potencial, la mayoría de estos sistemas todavía no se comercializan. La llegada al mercado de este tipo de vehículos centrados en el usuario dependerá de la resolución de restricciones técnicas y de costes de algunos conceptos avanzados – como los sistemas anti-colisión y de automatización–, de los intereses de los fabricantes, de los plazos de espera de producción y de la demanda del consumidor. Además, los vehículos dotados de sistemas de nueva generación se sitúan, principalmente, en el segmento de los automóviles de lujo, que representan un pequeño porcentaje del mercado. En lo que se refiere, en particular, a los sistemas para la seguridad activa, su despliegue a gran escala ha tenido que enfrentarse, en ocasiones, a numerosos problemas: principalmente barreras legales, elevado coste de los sistemas inteligentes y desconocimiento de los ciudadanos sobre estos sistemas. Además, la situación extremadamente competitiva del sector automovilístico crea condiciones poco propicias para el desarrollo de estos sistemas.

Para hacer frente a este tipo de situaciones, en el 2001 en Europa se lanzó una iniciativa relativa al vehículo inteligente que tiene como fin fomentar la utilización de las nuevas tecno-

logías para lograr vehículos más seguros, limpios y eficientes. Esta iniciativa, llamada i2010 y cuyos objetivos están plasmados en el Libro Blanco sobre transporte [Com01b], pretende ayudar a resolver los problemas derivados del tráfico vial y esperan que un enfoque global permita encontrar soluciones armonizadas. Por un lado, pretenden suprimir los obstáculos que retrasan la incorporación en el mercado de los sistemas para vehículos inteligentes. En este marco, el Foro *eSafety* desempeña un papel esencial [eSa03]. Entre sus objetivos a largo plazo está el apoyo a las actividades de investigación y desarrollo, coincidiendo con una de las prioridades que el Séptimo Programa Marco de I+D reserva a las TIC [Com07] . El último objetivo es sensibilizar acerca del potencial de los sistemas para vehículos inteligentes, contribuyendo a una mayor demanda y aceptación socioeconómica de dichos sistemas. Entre los retos de esta iniciativa, quizás el más ambicioso sea el pretender reducir a la mitad las casi 40.000 muertes de peatones registradas cada año en la red vial de la UE para el 2010. El futuro de los sistemas de seguridad se dirige no sólo a proteger al conductor y el resto de ocupantes del vehículo, sino también a proteger a peatones u otros vehículos.

1.7. Carencias, Requisitos y Sistema Sensorial para un ADAS

A medida que el automóvil avanza hacia el vehículo inteligente, aún existen carencias importantes que pueden resumirse en los siguientes puntos fundamentales:

- Sensores capaces de ofrecer una adecuada información en tiempo real en el entorno operacional y de conducción del vehículo.
- Una comprensión común acerca del comportamiento de un vehículo inteligente en diferentes situaciones operacionales.
- Actuadores y enlaces de comunicación para transferir el reconocimiento de una situación concreta a otros elementos de tráfico y así lograr una interacción inteligente.
- Un entorno político y social adecuado para soportar la introducción de vehículos inteligentes más avanzados en el mercado.

Por otro lado, cualquier sistema de a bordo para un aplicación de ITS, tiene que cumplir una serie de requisitos fundamentales:

- El sistema final que sea instalado en un vehículo comercial debe ser lo suficientemente robusto como para adaptarse a diferentes condiciones y cambios del entorno.
- Al tratarse de sistemas de seguridad críticos, deben ser altamente fiables. Por tanto, el proyecto debe cumplir todas sus fases de un modo exhaustivo y riguroso; desde los requisitos de la especificación del sistema, hasta el diseño y la implementación. Sobra decir que es de vital importancia una concienzuda fase de prueba y validación.
- Por razones de venta, el diseño de un sistema de ITS debe estar dirigido por el coste (no debería costar más del 10 % del precio del vehículo). Del mismo modo, el coste

operativo – como por ejemplo, el consumo de energía– debe ser lo más bajo posible para así no afectar al funcionamiento general del vehículo.

- El hardware del sistema no debe influir en el estilo del vehículo.
- Dado que un sistema de este tipo es controlado por el conductor, es fundamental desarrollar un interface hombre-máquina amigable.

Una vez puesto en claro qué requisitos deben cumplir este tipo de aplicaciones, cabe plantearse cómo llevar a cabo su desarrollo real. El alto nivel de automatización que se precisa en un vehículo autónomo, sólo va a poder ser alcanzado una vez que el problema más desafiante de este ámbito sea resuelto: la percepción del entorno que rodea al vehículo. Este es el centro de interés unánime para las industrias automovilísticas y los fabricantes de coches.

De entre los sensores empleados masivamente en robótica de interior, los sensores táctiles y los sensores acústicos no pueden ser usados en aplicaciones de automoción, debido a que la velocidad del vehículo inhabilita a los primeros y su reducido rango de detección, a los segundos. Los sensores basados en láser y radares detectan la distancia a la que están los objetos midiendo el tiempo de vuelo de una señal que emiten y que es reflejada por el objeto en cuestión. Por este modo de funcionamiento se clasifican como sensores activos. Sus principales inconvenientes son la baja resolución espacial y la lenta velocidad de escaneado. No obstante, los radares milímetro-ondas son más robustos ante la lluvia o la niebla que los radares basados en láser, aunque son más caros. Otro punto a favor de los radares es que no afectan al diseño de manera grave y ya se han comercializado vehículos dotados con estos sensores.

Los sensores basados en visión se definen como sensores pasivos y tienen una ventaja intrínseca sobre los sensores láser y radar: la posibilidad de adquirir datos en un modo no invasivo, y por tanto no alteran el entorno. El escaneado de la imagen es lo suficientemente veloz como para poder usarse en aplicaciones de ITS. En concreto, para algunas de estas aplicaciones la información visual desempeña un papel crucial – como es el caso de la detección de marcas de la carretera, reconocimiento de señales o identificación de obstáculos – y no requiere introducir ninguna modificación en la infraestructura vial. Desfortunadamente, los sensores de visión son menos robustos que los radares de onda milimétrica ante condiciones nocturnas, de niebla o sol directo.

Los sensores activos poseen algunas peculiaridades que en determinadas aplicaciones pueden aportar más ventajas que las cámaras de visión; pueden realizar medidas, por ejemplo del movimiento, de un modo más directo que la visión y exigiendo menos recursos de computación, ya que la cantidad de información que adquieren es bastante menor. Entre las desventajas de los sensores activos, se pueden citar el problema de la polución medioambiental, la amplia variación en ratios de reflexión motivada por distintas causas – como la forma de los obstáculos o el material – y la necesidad del máximo nivel de la señal para cumplir algunas reglas de seguridad. Aunque el mayor inconveniente son las interferencias entre otros sensores de este tipo, que puede ser un problema crítico en caso de que varios vehículos estén circulando simultáneamente en el mismo entorno.

Por tanto, preeviendo un masivo y ampliamente extendido uso de sensores autónomos, el uso de sensores pasivos, como son las cámaras, presentan ventajas clave frente al uso de

sensores activos. Una razón adicional es que, más del 90 % de los accidentes se producen por fallo humano, ya que la mayoría son de día (60 %), con buen tiempo (94 %), con vehículos en buen estado (98 %) y casi la mitad en un trayecto recto (42,8 %). Por tanto la información que darían las cámaras sería la correcta al haber buena visibilidad. No sería necesario disponer de otros sensores como radares o láseres que funcionan mejor en condiciones ambientales adversas.

Desde luego, la visión artificial no va a poder ir más allá en sus capacidades sensoriales de las del ojo humano, por ejemplo, en situaciones de mucha niebla o de noche, sin una iluminación adecuada. Pero, que duda cabe que va a poder ayudar al conductor en caso de fallo, como pueden ser situaciones de falta de concentración o de adormecimiento [BBF00, BD00, BBB+07b].

1.8. Motivación de la Tesis

La finalidad de esta tesis es desarrollar un sistema de ayuda a la conducción embarcado en un vehículo real para la protección de los peatones. A continuación se exponen las razones por las cuales se considera pertinente plantear esta tesis, junto con los objetivos iniciales marcados y las metas alcanzadas.

1.8.1. Necesidad del ADAS y Objetivo de la Tesis

Uno de los principales objetivos de los sistemas de ayuda a la conducción es evitar los accidentes de tráfico debidos a errores humanos. Las cifras de accidentes europeos reflejan un elevado número de colisiones entre peatones y vehículos. Cada año, alrededor de 475.000 peatones se ven afectados en el mundo, siendo cerca de 40.000 las víctimas mortales [TT04]. A pesar del escalofriante número de muertes, la protección de aquellos elementos más vulnerables de la circulación ha recibido escasa atención. La mayor parte de las investigaciones se han centrado en resolver tareas relacionadas con la detección de la carretera, el reconocimiento y seguimiento de vehículos, y la clasificación de objetos como señales de tráfico o paneles informativos.

Ha sido recientemente cuando la industria automovilística, los proveedores de sistemas eléctricos e investigadores han aunado fuerzas para desarrollar sistemas inteligentes para vehículos, con el objetivo de reducir las tasas de accidentes en los que se ven involucrados peatones. Varios fabricantes de coches los ofertan en sus modelos de gama alta, principalmente para la detección de vehículos, basados en radar, o el mantenimiento de la distancia lateral al carril, basados en visión por computador.

La necesidad de un sistema que advierta al conductor del posible peligro de su maniobra para peatones es manifiesta, ya que éstos suponen un elevado porcentaje del número de víctimas anuales de tráfico y no gozan de ninguna protección en caso de accidente. Estos dos hechos han motivado el desarrollo de esta tesis, cuyo objetivo final es la realización de un sistema avanzado de asistencia a la conducción (ADAS) para la detección de peatones en entornos urbanos.

1.8.2. Objetivos de la tesis

Inicialmente se plantearon una serie de hitos para la obtención de tan ambicioso objetivo. Se pueden resumir en los siguientes puntos:

1. Se pretende que el módulo funcione de forma satisfactoria en entornos exteriores complejos, entendiendo como tales, escenas no estructuradas y dinámicas. Además debe tener la capacidad de distinguir a los peatones, con independencia de su aspecto. La gran dificultad, en cuanto a la detección de peatones se refiere, estriba en la gran variedad de formas que pueden tener los peatones, así como en los cambios bruscos en sus trayectorias.

2. Cabe destacar que la información del exterior es percibida a través de una cámara estéreo instalada en el vehículo. Este hecho añade un complejidad mayor a la tarea. El objetivo es obtener un sistema lo más general posible, capaz de funcionar en un entorno exterior.

3. Después de una exhaustiva investigación de los distintos enfoques empleados por diversos grupos de investigación, el desarrollo de una metodología que integre distintos aspectos parece ser la forma más adecuada de tratar de resolver el problema de la detección. Por ello se pretende desarrollar un módulo flexible, en el que se puedan integrar de forma sencilla otras informaciones. Por ejemplo, la integración temporal de resultados o el reconocimiento de las acciones podrán mejorar el funcionamiento del sistema. El objetivo es tratar de obtener en conjunto una detección más robusta.

4. A la hora de evaluar el funcionamiento del sistema, la seguridad de los peatones será prioritario. Este aspecto está relacionado con tratar de reducir al máximo el número de falsos negativos (cuando no se detecta la persona). Así mismo, también hay que reducir la tasa de falsos positivos (cuando se detecta un peatón que no existe). Lógicamente, las consecuencias de no detectar un peatón pueden ser mucho más graves que dar una falsa alarma. Además, los falsos positivos suelen ser descartados al no mantenerse la detección en las secuencias sucesivas. Sin embargo, es bueno reducirlos por motivos de eficiencia (evitar el procesamiento innecesario). Por tanto, se medirá la robustez de nuestro sistema en función del porcentaje de errores. Se ha observado que hay sistemas muy sencillos, con muchas limitaciones que posiblemente den pocos errores. Sin embargo, el objetivo es construir un sistema lo menos limitado posible, con el menor número de falsas alarmas posibles.

5. El requisito de rapidez de respuesta, crítico en los sistemas de transporte inteligente, está englobado en el concepto de falso negativo, ya que si el módulo no proporciona la detección a tiempo, se considerará un trágico error.

6. Finalmente, el módulo de detección de peatones se integrará en la arquitectura modular del sistema de transporte inteligente IVVI (Vehículo Inteligente basado en Información Visual). Una vez desarrollado e implementado el módulo de detección de peatones, pasará a formar parte de un sistema integrado que se probará sobre un vehículo real.

Se pretende así, hacer frente a los puntos débiles de los sistemas actuales, que o bien pecan de una falta de experimentación en condiciones reales, o bien de una ausencia de integración de las diversas informaciones complementarias que se pueden obtener.

1.8.3. Estructura de la tesis

Este documento se ha estructurado de la siguiente forma:

- El capítulo 2 es un estado del arte de los sistemas de detección de peatones basados en visión artificial. Se describen todos los sistemas desarrollados hasta la fecha, poniendo de relieve los puntos débiles y fuertes de cada uno de ellos, así como las perspectivas para el futuro.
- El capítulo 3 hace referencia al sistema de percepción empleado en el desarrollo de esta tesis. Por ello, se describe tanto el sistema de visión del espectro visible – desarrollado en la Universidad Carlos III – y como el sistema de visión del espectro de infrarrojo – desarrollado en la Universidad de Parma–. Se analizan las ventajas e inconvenientes que cada uno de esos dominios puede aportar a la detección de peatones.
- El capítulo 4 se centra en las investigaciones llevadas a cabo empleando cámaras del espectro visible. Describe el sistema de detección de peatones basado en el espectro visible.
- El capítulo 5 describe los trabajos realizados empleando cámaras de infrarrojos. Describe el sistema de detección de peatones basado en el espectro infrarrojo.
- El capítulo 6 hace referencia al sistema final implementado y muestra resultados experimentales. Además, contiene las conclusiones y aportaciones de la tesis.

Capítulo 2

Detección de Peatones: Estado del Arte y Perspectivas

La detección de peatones se está convirtiendo en un foco de creciente interés dentro del ámbito de la automoción, al permitir incrementar la seguridad vial de los elementos más vulnerables de la carretera, si se dotase de esta capacidad a los vehículos. La magnitud de este problema es tal que se han adoptado diversas medidas. Por un lado, están las campañas de concienciación de la sociedad, que han ayudado a reducir el número de accidentes en carretera. Por otro lado, las empresas automovilísticas han desarrollado sistemas de seguridad pasiva, inicialmente orientados a la protección de los usuarios en el interior del vehículo – mediante p.ej. los airbags – y recientemente, orientados a la protección de los usuarios en el exterior del vehículo – mediante p.ej. los capós deformables –. Frente a estas medidas, cuyo objetivo es minimizar las consecuencias del atropello, existen otras dirigidas a tratar de evitarlo. Entre éstas, están las que se centran en ofrecer soluciones basadas en sensores, que permitan a los vehículos "ver más allá" y detectar a peatones a su alrededor. Son sistemas de seguridad activa, capaces de tomar el control del vehículo o de alertar al conductor en situaciones de riesgo.

Los sistemas de detección de peatones basados en visión, pueden jugar un papel destacado dentro del ámbito de seguridad vial de los próximos años. El elemento fundamental estos sistemas es la cámara embarcada en el propio vehículo, ya que proporciona la información del entorno en base a la cual se decide si existe o no riesgo de colisión. Esta respuesta es indudablemente muy valiosa para la creación de sistemas de seguridad para automóviles. Este tipo de aplicación es atractiva tanto para los fabricantes como para los usuarios finales. Los primeros, están interesados en vender productos con un alto valor añadido y los últimos, desean comprar vehículos más seguros.

Como consecuencia, no es de extrañar que para resolver el problema de la detección de peatones, la visión artificial haya recibido en los últimos años un creciente número de adeptos. De hecho, la detección de peatones basada en visión para aplicaciones de automoción es, en la actualidad, una de las tareas de investigación en ITS más candente. La tendencia de los últimos años así lo confirma, ya que la mayoría de los sistemas de detección de peatones hacen uso

de alguna cámara. Las razones que justifican este hecho hay que buscarlas entre las ventajas aportados por las cámaras de visión (ver capítulo 3).

Este capítulo investiga el estado del arte en este ámbito, poniendo el énfasis en los sistemas basados en el análisis de la imagen. A continuación se hace referencia a tales sistemas, que son analizados en profundidad. No sólo se describen las avances e investigaciones realizadas en favor de este tipo de sistemas a lo largo de su reciente historia, sino que además, se muestra una visión del futuro más inmediato. Además, también se va a hablar del uso combinado de la visión con otro tipo de sensores.

2.1. Descripción del Entorno de Trabajo: Características y Restricciones

Los sistemas de detección basados en visión, son difíciles de implementar debido a las características de los propios elementos que forman parte de la escena. Se pueden enumerar los siguientes como los más significativos y problemáticos:

1. **Características del entorno de trabajo:**

   - El escenario es complejo. Los objetos de interés aparecen en entornos muy saturados (con gran cantidad de elementos) y no-estructurados (los elementos en la escena no ocupan un lugar determinado en ella, aparecen de forma aleatoria, no controlada).
   - Las condiciones de iluminación son cambiantes y desconocidas, y no se pueden controlar, ya que el entorno es exterior.

2. **Rasgos de los peatones:**

   - Las personas pueden presentar una apariencia muy diversa, tanto en cuanto al tamaño y postura del cuerpo se refiere, como a la ropa. Pueden darse situaciones en las que la ropa se confunda con el fondo de la imagen o bien la indumentaria de la persona puede salirse de lo considerado como normal –como cuando se lleva sombrero– o ocultar la zona de las piernas – porque se lleva falda o abrigo–.

     Las variaciones en el tamaño se deben a que los peatones pueden aparecer a distintas distancias y a que la altura de las personas es un rasgo que varía de unos a otros y a lo largo de la etapa de crecimiento, siendo evidente esta diferencia si se compara la talla de un adulto con la de un niño.
   - Por otro lado, los movimientos de peatones resultan impredecibles. En cualquier momento pueden variar de ritmo y/o trayectoria. Además, son objetos no-rígidos. Por otro lado, el peatón puede situarse con respecto a la cámara de múltiples maneras: de frente o con un ángulo, cambiando enormemente su apariencia.
   - Además, las oclusiones y los grupos de personas añaden aún más complejidad a la tarea. También puede complicarse la tarea de detección si las personas llevan objetos – como bolsas o carritos de la compra –.

3. **Características del sistema de adquisición de datos:**

   - La cámara está en movimiento, instalada en el vehículo. El propio movimiento de la cámara (ego-movimiento) dificulta sobremanera la tarea de la detección.
   - Otro aspecto a considerar es la calidad de los sensores, que directamente influye en las imágenes a tratar. Una baja calidad de los sensores, hace que su rango dinámico sea limitado, lo que influye en su capacidad de adaptarse a cambios repentinos de iluminación y hace que aparezca una zona de la imagen más clara y otra muy oscura.

4. **Características del sistema final:**

   - El tiempo de respuesta del sistema es crítico, debiendo darse ésta lo antes posible y además, con un margen de error muy pequeño. La velocidad de los vehículos impone un límite; el sistema tiene que actuar antes de que el vehículo colisione con el peatón, bien proporcionando un cambio de trayectoria para evitar el accidente o accionando las medidas de protección de peatones como los airbags externos o el capó deformable.

Todas estas peculiaridades del entorno de trabajo no se pueden controlar. Debido a la gran dificultad que supone desarrollar un sistema que funcione eficazmente en tales condiciones, es habitual tratar de simplificar el problema, tomando una serie de supuestos. Los más habituales se pueden organizar en dos clases:

1. Supuestos sobre el movimiento. Tienen que ver con restricciones con los desplazamientos de las personas y las cámaras:

   - Movimiento de la cámara nulo o constante. En sistemas de visión de a bordo, esto es imposible. Se suele usar en sistemas con cámara fija, encargados de supervisar una zona, como las cámaras de seguridad de los bancos, del metro, etc.
   - El sujeto debe permanecer dentro de la escena. De este modo se facilita la tarea de detección. Cuando las personas entran y salen de la imagen, se corre el riesgo de cometer más errores, debido a que el algoritmo de seguimiento debe ser más robusto para poder actualizarse correctamente.
   - Sólo una persona en el entorno cada vez. Suele ser habitual en algoritmos de seguimiento. Cuando se tratan de entornos no controlados – como es el caso de los sistemas de visión embarcados en un vehículo – , esta exigencia resulta imposible de cumplir.
   - Movimientos paralelos al plano de la cámara. Reduce la dimensionalidad del problema de 3D a 2D. Se suele usar cuando se realiza un análisis del modo de andar.
   - El patrón de movimiento del sujeto es conocido. Reducen el espacio de soluciones, simplificando el seguimiento y la estimación de la postura.

- El sujeto se mueve con movimientos lentos y continuos. No se permiten movimientos bruscos. Además deben seguir una trayectoria simple y continua. Esto simplifica los cálculos de la velocidad de la cámara y el sujeto.
- El sujeto se mueve en una superficie plana. Permite calcular la distancia a la que está de la cámara.
- No admitir oclusiones. Supone que el sujeto completo es visible en todas las secuencias

2. Supuestos sobre el aspecto. Pueden hacer referencia al:

   - Posición inicial conocida. Simplifica la tarea de inicialización.
   - Conocimiento a priori del sujeto, en términos de parámetros específicos del modelo p.ej. altura, longitud de extremidades, tamaño, etc.
   - Llevan ropas con colores especiales, para facilitar la detección. En este supuesto se basa el método *"Capture Motion"* de gráficos por computador empleado en la creación de películas.

Además, hay otros dos aspectos que facilitan la tarea de la detección:

- Parámetros de la cámara conocidos. A veces, estos parámetros son necesarios para realizar cálculos; por ejemplo, para la calibración del sistema de adquisición, para la rectificación de imágenes, o para la obtención de medidas de distancias, en el caso de visión estereoscópica.
- Emplear hardware especial: integrar varias cámaras, tanto del dominio visible como del infrarrojo. Se pueden aprovechar las ventajas ofrecidas por los distintos elementos hardware [FYN+].

Conviene tener presentes los problemas y restricciones intrínsecos de estos sistemas, para poder evaluar la validez de las soluciones planteadas por otros autores.

2.2. Clasificación de los Sistemas Basados en Visión

En cuanto a detección de personas basada en visión se refiere, se han desarrollado varios sistemas tanto basados en visión monocular como estéreo. En la última década, disciplinas muy diversas han tratado de analizar el movimiento humano empleando una o dos cámaras, dando lugar, por ejemplo, a sistemas de vigilancia (para controlar la entrada en los parkings, aeropuertos, etc.), sistemas para realizar diagnósticos médicos (derivados del análisis de la forma de correr o caminar), sistemas de control (como los utilizados en las cintas de fabricación) o sistemas donde la interacción hombre-maquina es crucial (como ocurre en aplicaciones de videoconferencia) y, por supuesto, sistemas de detección de peatones.

La mayor parte de los trabajos relacionados con la detección de la forma humana en entornos saturados, vienen de estudios realizados sobre sistemas de vigilancia automáticos

[CLK99, Gav99, HTWM04]. Sin embargo, los trabajos realizados en este área, no pueden ser directamente llevados al área de la detección de peatones. La principal suposición de este campo de investigación, es el empleo de una cámara fija o cuyo movimiento sea lento. En contraposición, se puede afirmar que lo que convierte en un desafío a las aplicaciones de detección de peatones embarcadas en vehículos es, la cámara móvil, el amplio rango de apariencias que puede presentar el peatón y el entorno no-controlado involucrados en esta tarea. Por tanto, las soluciones probadas en otro tipo de condiciones, suelen fallar al tratar de aplicarlas en las propias de un entorno viario.

Independientemente del enfoque adoptado, en los trabajos realizados hasta la fecha, se suele describir la apariencia humana en términos de rasgos de bajo nivel extraídos de una región de interés. Por ello, un factor determinante para que estas técnicas alcancen el éxito, es la disponibilidad de una región correcta. Después, la detección de dichos objetos es habitualmente seguida por una etapa de reconocimiento para verificar la presencia de un peatón.

Esta tendencia generalizada permite afirmar que, el modo más habitual de enfrentarse a la detección de peatones consiste en dividir el problema en dos etapas;

- una primera, de detección de objetos, cuya finalidad es separar los objetos del frente de lo que es fondo, y
- una segunda, de reconocimiento de peatones, para distinguir a los peatones del resto de objetos.

2.3. Etapa de Detección de Objetos de Interés

Esta fase debe proporcionar zonas de la imagen susceptibles de contener a una persona: son las llamadas regiones de interés (*Regions of Interest*, ROI). Esas zonas serán las que se le proporcionen a la fase de reconocimiento. Por tanto cuanto más correctas sean éstas, más fiables serán los resultados obtenidos por la fase posterior de reconocimiento. Algunos sistemas basados en cámara estática o móvil con movimiento lento, emplean técnicas de segmentación simples para obtener regiones de interés. Entre estas técnicas, las más comunes son la sustracción de imágenes y la umbralización de ella.

Sin embargo, cuando la detección de peatones se tiene que llevar a cabo desde plataformas móviles en entornos exteriores, estas técnicas fallan. Las condiciones impuestas por los métodos anteriores, no se ajustan a las exigencias de un sistema para aplicaciones en el campo de la automoción. Por un lado, el fondo está cambiando continuamente y no existe la posibilidad razonable de poder modelarlo. Además, el hecho de que la cámara vaya instalada en el vehículo es suficiente para desechar el uso de cualquier método basado en sustracción de imágenes. Por último, la umbralización depende fuertemente de los supuestos que se impongan a la escena de partida. Por este motivo no se pueden usar estas técnicas en ambientes naturales, donde no se puede exigir, por ejemplo, que el objeto a segmentar (en este caso el peatón) lleve una ropa determinada. Además las condiciones de iluminación son cambiantes, lo que implica que la intensidad de las imágenes varía en función de la luz y por tanto es imposible diferenciar a una persona en base a su intensidad. Para el caso de aplicaciones sin restricciones, los métodos ba-

sados en estadísticos son mejores dada su adaptabilidad. En general, utilizar métodos basados en regiones tienden a ser más fiables que los basados en un sólo píxel. La parte negativa es que modelar entidades más grandes es generalmente más difícil.

Los enfoques actuales para detectar peatones usando cámaras móviles se caracterizan por realizar un análisis de más de una imagen, como en el caso del análisis del movimiento o del procesamiento de imágenes estéreo. En base a la información obtenida, segmentan aquellas zonas de la imagen que presentan un movimiento típico de un humano o que comparten unas características estéreo determinadas, como por ejemplo, la distancia. Otros sistemas, sustituyen la etapa de segmentación propiamente dicha, por un enfoque de búsqueda del foco de interés (*focus of attention*), cuya filosofía se comenta más adelante en este capítulo.

2.3.1. Segmentación basada en movimiento

El movimiento es una característica típica para detectar regiones interesantes en una escena. Hace un uso extenso de información temporal y ha demostrado ser bastante fiable si lo que se pretende es únicamente encontrar objetos en movimiento. Desafortunadamente, no detecta peatones que permanezcan estáticos y necesita analizar una secuencia breve de imágenes, antes de dar una respuesta.

El análisis del campo del flujo óptico entre una secuencia de dos o más imágenes, permite describir el movimiento coherente de puntos, rasgos o agrupaciones de píxeles en términos de velocidad. Esta técnica se ha empleado para la detección de obstáculos en sistemas instalados en vehículos [KER95, SB95]. Sin embargo, para la detección de peatones, es difícil de aplicar debido al movimiento no-rígido propio de las personas. En este caso, son escasos los trabajos que usan una detección de movimiento basado en el flujo óptico como método de segmentación. La idea básica consiste en detectar blobs con una forma o un rasgo común, como el color, que presente unos valores de flujo óptico similares y seguir sus movimientos en las imágenes subsiguientes.

[ELW03] han desarrollado un algoritmo de detección de peatones basado en visión que puede emplearse en vehículos inteligentes. Para la detección de objetos en movimiento, usan el método de estimación del flujo óptico basado en la correlación propuesto por [LRY02]. Debido a que el cálculo del flujo óptico resulta computacionalmente costoso y dado que sólo están interesados en detectar objetos en movimiento que posean un riesgo potencial de colisionar con un vehículo, se usa un esquema de localización. Este esquema busca aquellos píxeles que tienen un movimiento significativo y obtiene un conjunto de cajas o regiones de interés. Después calculan el flujo óptico de esas regiones, en términos de velocidad, que emplean para calcular el tiempo de colisión de cada píxel con un punto de referencia en la imagen. Segmentan aquellos píxeles con un bajo tiempo para colisionar, ignorando el resto y evitando subsiguientes procesamientos. El conjunto de regiones que pasan el umbral, son preprocesadas para refinar la caja que las contiene y son procesadas en la siguiente fase de clasificación. El sistema es capaz de detectar peatones en entornos saturados, pero impone varias restricciones; ni la cámara ni el coche están en movimiento y, si existe movimiento en la escena, sólo puede deberse a un único objeto en movimiento.

Las conclusiones que se pueden sacar del análisis del flujo óptico es que:

- No puede detectar objetos que se muevan rápido. Asumen que los objetos de interés se mueven en direcciones constantes, a velocidad también constante y que los objetos del fondo se mueven de forma aleatoria.
- El sistema es sensible a cambios de luz que se mantengan más de un par de segundos.
- No distingue si lo detectado es una persona, un animal o un coche, ni tampoco personas con ropas con poco contraste.
- Detecta múltiples personas, pero como una sola región.
- Si el individuo tiene un tamaño muy pequeño en la imagen, no se puede obtener un flujo correcto.
- Cuando la cámara está en movimiento resulta crítico realizar una correcta cancelación del propio movimiento de la cámara.

Otra posibilidad para segmentar el movimiento de una secuencia, se basa en la resta de imágenes consecutivas. [CD00] emplean el algoritmo de [HAD+94] para obtener una escena estabilizada de cada imagen que luego es restada de la imagen original. Así detectan regiones de movimiento a pesar de que las cámaras estén en movimiento .

2.3.2. Segmentación basada en estéreo

Su aplicación al campo de la detección de peatones comienza a finales de los 90. Puede emplearse como técnica de segmentación, al realizarse una umbralización de rango en base al análisis estéreo. La inclusión de más de una cámara aporta ventajas frente a los sistemas monoculares. En concreto, la visión estereoscópica permite:

1. Llevar a cabo un análisis de oclusiones y es robusto ante cambios de luz .
2. Obtener el tamaño real de los objetos, a partir del mapa de disparidad.
3. Se reduce el posterior tiempo de cálculo, ya que el reconocimiento tiene lugar sólo en aquellas zonas donde se han detectado objetos. Se reduce considerablemente el espacio de búsqueda.
4. Los métodos de segmentación basados en información de rango (como es el caso del mapa de disparidad) son más robustos ante cambios de la escena que aquellos métodos que segmentan basándose en información a nivel de píxel.
5. Estas técnicas son menos costosas que las técnicas basadas en la reconstrucción completa de la escena 3D, ya que se trabaja con información 2 ½ D (2D y tiempo). Por otro lado, hay que tener en cuentan que facilitan detecciones a corta y media distancia.

Por todo esto, la visión estéreo tiene unas interesantes características, de las que pueden favorecerse los sistemas de detección de peatones con independencia de que las cámaras sean estáticas o estén en movimiento.

El proyecto de asistencia a la conducción de autobuses desarrollado por Zhao y Thorpe [ZT00b] en la Universidad Carnegie Mellon, usan información de distancia con el fin de segmentar objetos. Basándose en la información del mapa de disparidades, eliminan objetos del fondo, ya que sólo están interesados en avisar cuando hay un objeto cerca del vehículo (ver fig. 2.1). Los sistemas de visión de rango se basan en la correlación entre zonas de las imágenes izquierda y derecha de la misma escena. En este tipo de sistemas, la influencia del ruido es mayor a medida que aumenta la distancia. Por tanto, de esta segmentación habitualmente no se obtiene un contorno muy preciso. Este es el mayor inconveniente de las técnicas estéreo. Zhao y Thorpe en [ZT00a] afrontan el problema mediante un algoritmo recursivo que obtiene contornos fiables de las zonas previamente segmentadas [ZT00b]. Diseñan un modelo probabilístico invariable a la translación, rotación y escalado para representar las formas de las partes del cuerpo e incluyen información contextual (relaciones de tamaño y cuerpo entre los distintos miembros). Para la detección de personas o parte de ellas, usan un enfoque Bayesiano. Después llevan a cabo una reconstrucción del contorno basado en Kalman. Es un proceso iterativo: extraen contornos, identifican partes, detectan regiones conteniendo a posibles personas, reconstruyen y vuelven a extraer contornos, hasta obtener la certeza de que existe una persona o no.

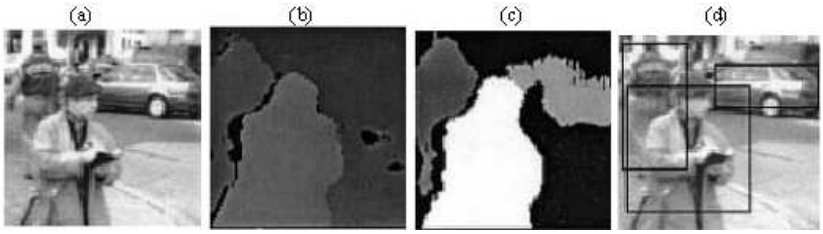


Figura 2.1: (a) imagen izda tomada con estéreo, (b) mapa de disparidad, (c) regiones y (d) resultado de la segmentación [ZT00b].

La etapa de detección da problemas, ya que las zonas segmentadas no siempre corresponden a un único objeto. Aplican un procedimiento de hipótesis y verificación para separar o unir las regiones segmentadas, que después son pasadas a un módulo de reconocimiento de peatones (ver fig. 2.2).

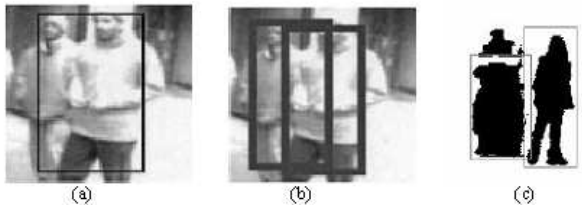


Figura 2.2: (a) y (b) muestran algunos de los problemas de las técnicas split/merge [ZT00b].

McKenna *et al.* [MJD+97] exponen la complejidad que supone el seguir a grupos de personas. Es habitual que objetos muy próximos se interpreten como una sola región o bien que un

único objeto se divida en varias regiones. Esto es debido al ruido, falta de textura y el propio límite de precisión del mapa de disparidad.

El grupo de investigadores de la Daimler-Chrysler, formado por Franke y sus colaboradores [FGG+98], implementan el módulo de detección y seguimiento de obstáculos bajo un enfoque estéreo. Para la obtención de una región de interés (ROI) eliminan los elementos de la carretera basándose en la disparidad. Para la correspondencia estereoscópica, han desarrollado una técnica basada en rasgos [FGG+98].

En la Universidad de Parma, el grupo de investigación VISLAB dirigido por Broggi, han hecho un uso extenso de la información estéreo con el fin de mejorar el funcionamiento de su sistema de asistencia a la conducción GOLD. Inicialmente, han empleado sistemas estéreo en el dominio visible [BBFS00] y en posteriores trabajos, han aplicado sus algoritmos estéreo en el dominio infrarrojo [BBF00, BBDL05, BBF+06, BBF+07a, BBC+07], calculando medidas de distancia para refinar la localización y el tamaño de las ROI. Recientemente [BBF+06], emplean un par de sistemas estereoscópicos, uno en el dominio visible y el otro en el dominio infrarrojo, beneficiándose de la ventajas de ambos espectros.

En [BRF+03] y [GZNR04], emplean la técnica basada en V-disparidad propuesta por [LAT02], que considera restricciones de la perspectiva y de la calibración para refinar la caja que contiene a cada ROI. Este método resulta adecuado para una cámara en movimiento, ya que impone pocas restricciones en cuanto a que el terreno sea plano. [BBDL05] aplican la misma idea en el dominio infrarrojo y en [BBF+06] emplean dicha técnica en ambos dominios.

En la aplicaciones de vigilancia, a veces se lleva a cabo el análisis estéreo para construir un mapa de disparidad del fondo, que se usará para la sustracción del fondo. Esto es lo que ocurre en el sistema desarrollado por el SRI International, utilizando sus cámaras estéreo integradas conocidas como *Small Vision System* (SVS) [BK99]. Los objetos segmentados se organizan en pirámides para compensar las diferencias de tamaño. Este sistema muestra una baja sensibilidad frente a elementos distractores, como son las sombras, los cambios de iluminación, objetos ocluidos o cámaras dinámicas.

2.3.3. Búsqueda de otros rasgos significativos o del *foco de atención*

Algunos sistemas evitan la etapa de segmentación, dirigiendo el foco de interés (*focus of attention*) hacia aquellas zonas de la imagen que reúnen uno o más rasgos de interés. Así, es habitual realizar una búsqueda en la imagen de simetrías, bordes o texturas, entre otros rasgos.

Se suele utilizar una ventana deslizante de tamaño variable que recorre toda la imagen en busca de esas regiones de interés. Una vez extraídas dichas regiones, se les aplican detectores de rasgos, para después emplear clasificadores de patrones estándares que determinarán la existencia de un rasgo o no en la correspondiente ventana. Generalmente estas técnicas de *"fuerza bruta"* son computacionalmente prohibitivas para ser aplicadas en aplicaciones de tiempo real. Sin embargo, recientes aportaciones, como los clasificadores en cascada de Viola y Jones [VJS05], presentan una variante del enfoque basado en ventana deslizante, que resulta viable en aplicaciones donde el tiempo respuesta es crítico.

2.3.3.1. Sistemas basados en detectores de rasgos

A la hora de detectar humanos, resulta fundamental seleccionar un conjunto de rasgos robusto, capaz de discriminar a una persona sin lugar a dudas, incluso en situaciones de fondo saturado o en difíciles condiciones de iluminación. La selección de estos rasgos no es tarea fácil. Así, hay sistemas que ponen el énfasis en tratar de definir detectores capaces de encontrar ese conjunto de rasgos.

Un detector de rasgos interesantes en la imagen, selecciona aquellas regiones más significativas del oportuno mapa o imagen de rasgos, que son interpretadas como candidatas a peatones. Algunas veces dichas regiones resultan ser invariantes al escalado y/o al cambio de punto de vista (*pointview*) de la cámara. Además, se pueden añadir invarianzas adicionales, como la rotación o la iluminación. Se han empleado detectores de rasgos de bajo nivel, tales como bordes verticales, simetrías verticales o entropía en la imagen.

En la Universidad de Parma, en el sistema GOLD [BBFS00] de detección de peatones, los rasgos que se extraen de las imágenes son las simetrías verticales, que son asociadas a peatones de pie en potencia, tanto quietos como en movimiento. Se recoge más información mediante mapas de simetría de los bordes horizontales así como del número de dichos bordes por columna. Los máximos en dicho mapa contribuyen a la identificación de las regiones de interés, que son recuadradas para ser evaluadas en la fase de reconocimiento posterior (ver fig. 2.3).

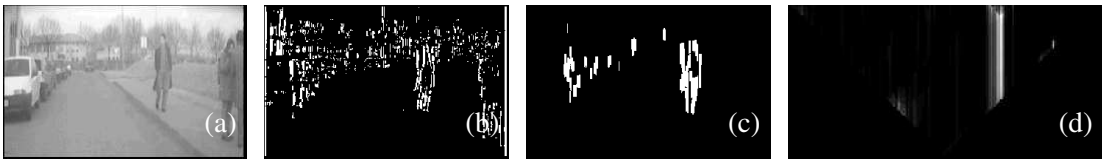


Figura 2.3: Extracción de bordes en el espectro visible [BBFS00]: (*a*) Imagen original; (*b*) Mapa de bordes verticales; (*c*) Bordes verticales después de eliminar el fondo; (*d*) Mapa de simetrías verticales.

En los trabajos posteriores [BBF+04, BCFG05], aplican estas consideraciones sobre simetrías y bordes verticales a imágenes tomadas con cámaras de infrarrojos, extrayendo los candidatos a peatones. En cambio, para la detección de peatones lejanos (entre 40 a 100 metros de distancia), realizan un análisis del contraste y brillo en la imagen capturada con una sóla cámara. Después, buscan aquellas regiones brillantes o conectadas a una región brillante [BBGM07]. El sistema de visión nocturna implementado en [THWN02] también se basa en el brillo para detectar a los posibles peatones. Además, a partir de la información de las cámaras estéreo de infrarrojo, son capaces de determinar la posición y el movimiento relativo de los peatones y avisar al conductor en caso de que exista peligro de colisión.

Recientemente, el coste de las tecnologías de infrarrojos (tanto las basadas en el infrarrojos cercano como en el lejano) ha decrecido considerablemente, favoreciendo la aplicación de dicha tecnología a la detección de peatones. En concreto, cuando el peatón emite más calor que el fondo, las cámaras de infrarrojo lejano (FIR) resultan más adecuadas que las del dominio visible para llevar a cabo la detección. POr ejemplo, [ND02] obtiene las regiones de interés

a partir de imágenes de infrarrojos monoculares, dirigiendo el foco de interés hacia aquellos píxeles con una intensidad más alta. Como método de segmentación, aplican una umbralización basada en la intensidad. Emplean un clasificador Bayesiano para determinar, a partir de un conjunto de ejemplos de peatones y de no-peatones, el umbral capaz de separar los objetos del fondo. [RWZD07] usan la idea propuesta por [ND02] para inicializar la detección.

En la Universidad de Ruhr-Universität Bochum [CEK+00] se ha desarrollado un sistema más complejo. El foco de interés es guiado por una combinación de: un mapa de la entropía local de la imagen (ver figura 2.15), un módulo de correspondencia de modelos basado en la forma de las piernas humanas y, por último, un mapeado de la perspectiva inversa (basado en visión binocular) para detecciones a corta distancia. Esta información se usa junto con un campo de activación dinámico temporal (DAF, *Dynamic Activation Field*) que resulta muy eficaz durante la etapa de reconocimiento y seguimiento.

2.3.3.2. Sistemas basados en descriptores

Un descriptor representa una región de la imagen, previamente extraída. Cada región seleccionada se caracteriza por un vector descriptor o vector de rasgos. Para ello, dichas regiones son transformadas al espacio de rasgos más adecuado, donde la información se representa de una forma más compacta y resulta más sencillo definir los correspondientes descriptores o vectores de rasgos. Son habituales descriptores basados en el análisis del componente principal (PCA, *Principal Component Analysis*), en *wavelets* o en campos receptivos locales (LRF, *Local Receptive Fields*).

Los descriptores obtenidos, son posteriormente evaluados por un clasificador que determina la presencia o ausencia de peatón. Dicho clasificador es previamente entrenado a partir de un conjunto de ejemplos, representados mediante el descriptor que se quiere clasificar. Así, podemos hablar de clasificadores wavelet, HOG (*Histograms of oriented gradient*) o SIFT (*Scale Invariant Feature Transform*), entre otros.

- **Descriptores basados en la apariencia:**

  Probablemente, el método de extracción de rasgos más conocido sea el PCA. Tiene la ventaja de reducir la dimensionalidad identificando los rasgos más característicos, es decir, los autovectores con los autovalores más altos, mientras que aquellos con autovalores bajos se considera que corresponden a ruido y se desechan. Los coeficientes PCA pueden considerarse rasgos globales, ya que cada coeficiente describe una propiedad determinada del patrón de entrada completo, y los detalles locales son suavizados debido a la reducción de la dimensionalidad.

  Franke *et al.* [FGG+98] para el proceso de búsqueda de un peatón en la imagen de entrada, trabajan con la imagen gradiente, que correlan con los primeros autovectores de una base de datos de ejemplos de personas, a los que ha aplicado PCA (ver fig. 2.4).

  En contraposición al empleo de representaciones globales, las representaciones locales de las imágenes son útiles para hacer frente a la amplia variedad de escenas conteniendo fondos saturados y objetos ocluídos. Las regiones de interés son transformadas al espacio de rasgos más adecuado. En la detección de personas, destacan las transformaciones

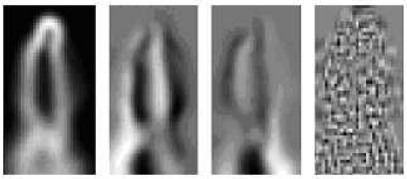


Figura 2.4: Descriptor basado en los primeros autovectores obtenidos del análisis PCA [FGG+98].

invariantes a la escala (p.ej. los descriptores SIFT o más recientemente los HOG, *Histogram of Oriented Gradient*) y las transformaciones invariantes a la iluminación (p.ej. los descriptores *wavelet de Haar*). Todos ellos son descriptores basados en la apariencia. Este tipo de descriptores, generalmente, consideran información de la intensidad (gradiente) o de la forma humana.

Entre los rasgos más comúnmente empleados para la descripción de un peatón, sobresalen los *wavelets de Haar*. La justificación del extenso uso de los *wavelets* hay que buscarlo en su capacidad para codificar rasgos locales de la imagen – como cambios de intensidades – a diferentes escalas y además permitiendo un equilibrio entre compactitud y expresividad. La base de *wavelet* más sencilla es la base de Haar y permite definir matemáticamente invariantes de la imagen. Tienen la ventaja de no verse afectados por cambios de color ni de textura y permiten definir de manera robusta clases de objetos complejos, como es el caso de las personas.

El grupo de investigadores del MIT, dirigido por Oren y Papageorgiou [OPS+97, POP98], extraen un subconjunto de coeficientes wavelet Haar de la image (ver fig. 2.5). Obtienen un representación de rasgos denso, ya que aplican *wavelets* de tres orientaciones distintas – vertical, horizontal y diagonal –, realizando además una búsqueda multiresolución en toda la imagen. [ELW03] usan el algoritmo propuesto por [OPS+97], realizando también una extracción de rasgos basada en la transformada del *wavelet de Haar* densa.

Figura 2.5: (a) Descriptor basado en Wavelet de Haar y (b) resultados del sistema de Oren y Papageorgiou [OPS+97].

El trabajo de [MPP01] es una extensión del de [POP98], donde adoptan un enfoque basado en componentes, frente al enfoque precedente basado en la detección del cuerpo entero. Como consecuencia, extraen un vector de rasgos para cada uno de los componentes del peatón, tratando así de reducir la complejidad de la descripción de la apariencia humana.

Shashua *et al.* [SGH04], también tratan de reducir la complejidad en cuanto a la apariencia de los peatones se refiere, usando un método basado en componentes. Han desarrollado un sistema de detección de peatones monocular para aplicaciones de ayuda a la conducción cuyo funcionamiento apuesta por la integración de múltiples estrategias.

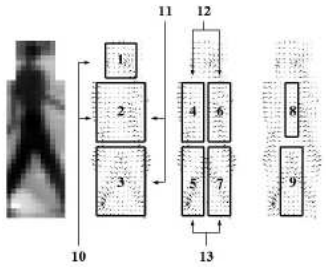


Figura 2.6: Se muestra la configuración de las 9 subregiones sobre la imagen del gradiente. La distribución de los vectores es evaluada en cada subregion. Además, se construyen 4 pares de combinaciones (regiones 10-13) [SGH04].

El sistema detecta regiones candidatas que cumplen ciertos requisitos, en cuanto a tamaño, distancia y existencia de texturas. Para obtener las medidas de distancia se basan en el reconocimiento del modo de caminar, desempeñando una segmentación basada en el movimiento. Después, cada región se divide en nueve subregiones fijas. Para cada subregión definen un vector de rasgos insensible a los desplazamientos locales de las estructuras en la imagen, mediante el uso de descriptores basados en histogramas orientados. Cada histograma de orientación se define en función de la magnitud y orientación del gradiente (HOG) (ver fig. 2.6).

En el INRIA [DT05a], han desarrollado un sistema de detección de personas que también usa descriptores basados en el histograma del gradiente orientado y obtienen mejores resultados que los basados en *wavelets* [MPP01, VJ01b, DCdB$^+$02].

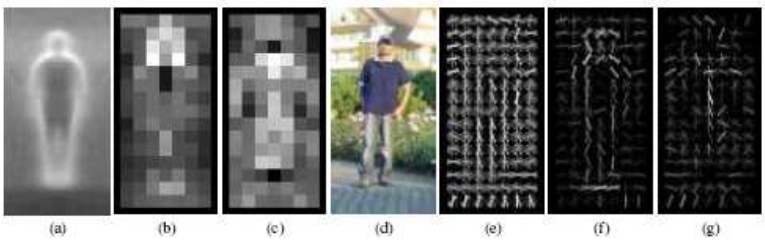


Figura 2.7: Detectores de HOG [DT05a]. (a) La media de las imágenes del gradiente obtenida del conjunto de entrenamiento. (b) Cada píxel muestra el máximo peso positivo del SVM en cada bloque. (c) Probabilidad de los pesos negativos del SVM. (d) Una imagen de prueba. (e) El descriptor R-HOG resultante. (f,g) El descriptor R-HOG pesado por los pesos positivos y negativos del SVM.

Cada imagen se divide en celdas, calculando en cada una de ellas el histograma de la

dirección del gradiente o de la orientación de los bordes. Para obtener una mejor invarianza –por ejemplo ante cambios de iluminación o ante sombras–, normalizan el contraste de dicho histograma a lo largo de un conjunto de celdas que se solapan. Como resultado, la ventana de detección consiste en una rejilla densa – ya que se solapa – de descriptores HOG. Este sistema ha sido probado con cámara estática, pero puede ser empleado en aplicaciones embarcadas en vehículos con el fin de detectar peatones. Los detectores HOG desarrollados son obtenidos a partir de características de contornos de la silueta – sobretodo, de la cabeza, hombros y piernas– (ver fig. 2.7). Si bien, el uso de histogramas basados en la orientación de los bordes era ya usado para el reconocimiento de objetos, [DT05a] introduce el concepto de histogramas densos y locales de los gradientes orientados (HOG), describiendo la imagen como un conjunto de histogramas locales.

El INSA junto con la Universidad de Parma [SARB06], proponen un sistema completo para detectar peatones aplicado a imágenes de infrarrojos. Primero, realizan un estudio de un descriptor de HOG – como el descrito en [DT05a] – aplicado a imágenes que sólo pueden contener un peatón cada vez. El sistema completo implementado permite la existencia de varios objetos, pudiendo ser o no peatones. Las ventanas potenciales son seleccionadas en función de las intensidades de los píxeles en el dominio infrarrojo y usan información estéreo para determinar la posición de los peatones. Proponen usar estas imágenes FIR durante la noche, porque entonces los peatones aparecen con una intensidad mayor que el entorno circundante.

El éxito de los *wavelet de Haar* radica en su habilidad para captar conocimiento de alto nivel, en términos de información estructural expresada como un conjunto de coeficientes *Haar*, capaces además de mostrar la relación entre la intensidad de regiones vecinas. En comparación, el uso de una representación HOG o SIFT, también captura la estructura de los bordes o del gradiente característica de la forma, siendo además sencillo controlar la invarianza a a transformaciones geométricas o fotométricas locales. Para el caso de la detección de personas, la mejor estrategia ha resultado ser un muestreo espacial basto, un muestreo de la orientación fino y una fuerte normalización fotométrica; probablemente debido a que así se permite que las distintas partes del cuerpo varíen bastante en apariencia y en movimiento, mientras la orientación se mantenga aproximadamente vertical. En concreto, los descriptores de HOG, son reminiscencias de los histogramas de bordes [FR95], los descriptores SIFT [Low04] y los *shape contexts* [SBP02].

- **Descriptores basados en el movimiento y en la apariencia:**

  Al incluir rasgos del movimiento en el detector de peatones, se aumenta el funcionamiento en una orden de magnitud, comparado con un detector estático o basados en la apariencia. El detector de peatones propuesto por Viola [VJS05], usan al mismo tiempo patrones de movimiento y de la apariencia. Otros trabajos parecidos, emplean información estática de las intensidades [NDS06], mientras que este sistema hace uso de las intensidades y movimientos obtenidos directamente de cada frame.

El detector dinámico de peatones que proponen, se basa en los filtros rectangulares propuestos para la detección de caras estáticas [VJ01a]. Para obtener la información del movimiento, calculan diferencias entre regiones en la imagen, obteniendo filtros para la dirección y la magnitud del movimiento. El detector integrado recorre toda la imagen, operando en una pirámide de imágenes y consiguiendo una detección a múltiples escalas.

Cuando se pretende hacer uso del movimiento para la detección de humanos observados con una cámara móvil, contra un fondo también dinámico, es necesario contar con rasgos que caractericen los movimientos humanos correctamente, al mismo tiempo que sean resistentes a los movimientos típicos de la cámara y del fondo. La mayor parte de los descriptores de movimiento existentes, como el de [VJS05], emplean movimientos absolutos, y por tanto, sólo funcionan bien cuando la cámara y el fondo permanecen la mayor parte del tiempo, estáticos. En este sentido, estas representaciones no poseen las propiedades ofrecidas por la familia de descriptores SIFT/HOG.

En cambio, en el INRIA [NDS06], han presentado unos descriptores que usan la diferencia del flujo, para eliminar la mayoría de las consecuencias del movimiento de la cámara, e histogramas orientados del tipo HOG, para obtener una descripción robusta. Siguiendo el trabajo anterior [DT05a], combinan el uso de esos descriptores de la apariencia basados en el gradiente, con descriptores basados en el movimiento. Calculan los histogramas orientados obtenidos del flujo óptico, y han realizado distintos experimentos considerando tanto la orientación del flujo como la orientación del gradiente del flujo. Emplean una ventana que recorre toda la imagen a diferentes escalas.

- **Descriptores adaptativos:**

  En lugar de tener que extraer el conjunto de rasgos manualmente, los perceptrones multicapa proporcionan un enfoque adaptativo, de manera que los rasgos a extraer son aprendidos a partir de los ejemplos de entrenamiento. En concreto, las redes de neuronas *feed-forward* con campos receptivos locales (LRF) son especialmente atractivos para la clasificación de imágenes 2D. Mientras que en los perceptrones multicapa (MLP, *Multi layer perceptron*) la arquitectura de la red es invariante a los datos de entrada, en las redes LRF, las neuronas de la capa oculta sólo se conectan a algunas regiones locales de la imagen de entrada. Una vez entrenada la red, se puede emplear la salida de la capa oculta como los rasgos que posteriormente se clasificarán usando algún método de clasificación – además de las redes de neuronas–.

  En [WAPF98] han presentado un algoritmo que analiza secuencias de imágenes, empleando una red neuronal con retardo (TDNN) con campos de activación locales LRF y tiempo de retardo adaptativo; al incluir los campos receptivos espacio-temporales, se tiene en cuenta la localización espacial y temporal del peatón y al aprender el retardo entre frames, se consideran rasgos de su movimiento (ver fig. 2.8).

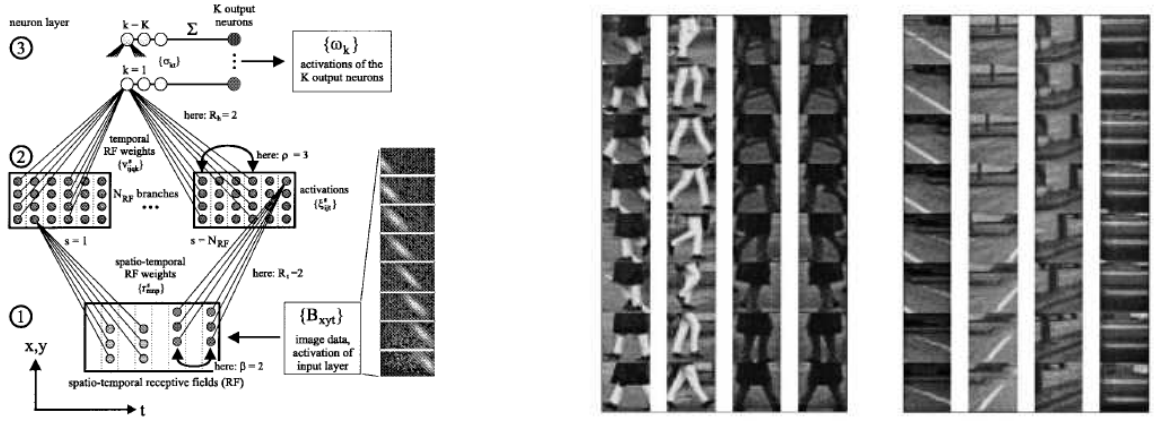


Figura 2.8: Sistema desarrollado en [WAPF98] (*a*) Arquitectura de la ATDNN con campos receptivos espacio-temporales. (*b*) Imagenes del conjunto de entrenamiento.

2.4. Etapa de Reconocimiento de Personas

El reconocimiento de objetos es uno de los retos fundamentales de la visión por computador. Las recientes técnicas de visión y de aprendizaje de máquina (*machine learning*) han aportado avances significativos. Desde un punto de vista global, se pueden catalogar los sistemas de detección de objetos – basados en visión – desarrollados hasta la fecha en tres clases principales.

1. La primera categoría corresponde a los sistemas basados en modelos, en los que se define un modelo para el objeto de interés y el sistema trata de hacer corresponder ese modelo con la imagen.

2. El segundo tipo agrupan aquellos métodos basados en invariantes de la imagen, que tratan de establecer la correspondencia a través de las relaciones existentes en patrones de la imagen –como p.ej. los niveles de intensidad.–, e idealmente, determinan unívocamente los objetos buscados. También se denominan sistemas libres de modelos (*"Model-free"*).

3. El último conjunto de sistemas de detección de objetos, se caracterizan por sus algoritmos de aprendizaje basados en ejemplos. Estos sistemas, aprenden los rasgos característicos de una clase a través de un conjunto de ejemplos etiquetados como positivos y negativos.

Los sistemas de detección de peatones basados en visión, al tratarse de un caso particular dentro de la detección de objetos, han tomado como punto de partida las técnicas de reconocimiento de patrones ya probadas en la detección de objetos genérica. De manera que, estos sistemas pueden clasificarse según el mismo criterio en:

1. Sistemas basados en modelos, donde se usa un modelo definido a mano que describe la estructura humana. Cuando interesa representar el movimiento, a veces, se incluye la

cinemática y hasta la dinámica del caminar humano. Las técnicas de reconocimiento de patrones más habitualmente empleadas en este caso, están relacionadas con las técnicas de correspondencia de patrones.

2. Sistemas libres de modelos, donde se buscan las relaciones en patrones contenidos en la imagen. La detección está dirigida por los datos, como ocurre con los métodos basados en contornos activos, que tratan de extraer la forma del peatón al verse atraídos por los rasgos de interés, o como ocurre con los métodos que tratan de reconocer el movimiento humano, basándose en la coherencia de los rasgos en el tiempo.

3. Sistemas basados en el aprendizaje. La tendencia de los últimos años ha sido adoptar un enfoque basado en aprendizaje. Los avances en teoría del aprendizaje máquina, unido a las mejoras en tecnología de computadores han favorecido, de un modo creciente, el uso de estas técnicas. Así, son habituales las técnicas de en reconocimiento de patrones basadas en *machine learning*, bien para construir un patrón o modelo a partir de ejemplos, o bien para usar clasificadores de patrones.

   Para construir un patrón o modelo, se parte de un conjunto de ejemplos positivos y negativos, y mediante un algoritmo de aprendizaje se aprenden los rasgos más representativos de cada conjunto de entrenamiento. Con este fin, son masivamente usados las técnicas de clasificación basadas en límites de decisión (p.ej. las redes de neuronas o las máquinas de vectores de apoyo, SVM, *Support Vector Machine*) y en menor medida, las técnicas basadas en la estimación de densidad (p.ej. la teoría de decisión de Bayes o el clasificador de Parzen).

   A veces, las imágenes se transforman en otro espacio con el objeto de reducir su dimensionalidad. Así, la información se representa de una forma más compacta y resulta más sencillo definir los correspondientes patrones o vectores de rasgos. Como resultado, el clasificador aprende el patrón o modelo del objeto deseado, siendo los más habituales los clasificadores los basados en PCA, *wavelets* o recientemente, HOG. Además de estos patrones basados en la apariencia, en los últimos años se han comenzado a incluir patrones de movimiento.

   Después de la fase de entrenamiento, la posterior clasificación de patrones suele afrontarse dividiendo el problema en dos partes; por un lado, se extraen los rasgos de la imagen de entrada – se construye el descriptor de la imagen – y por otro, se realiza la clasificación. Así, los descriptores de entrada al sistema de detección, son evaluados por los clasificadores convenientemente entrenados, que determinan si existe o no peatón.

Con independencia del enfoque empleado para la creación del modelo humano, se puede realizar una distinción de los sistemas de peatones desarrollados hasta la fecha en función de los indicios o pistas en que se basan. Según este criterio, se puede distinguir entre detectores de peatones basados en la apariencia y detectores de peatones basados en el movimiento. Existe la opción de combinar ambos detectores, beneficiándose de las ventajas que aportan los rasgos basados en el aspecto y en el modo de caminar, y como consecuencia, consiguiendo mejorar el rendimiento final del sistema.

La mayoría de las técnicas de reconocimiento de objetos basados en la apariencia, caracterizan los objetos por su apariencia global, considerando generalmente toda la imagen [POP98]. No son robustos ante oclusiones y sufren de una carencia de invarianza ante cambios de escala o rotación (transformaciones afines). Además, suelen ser aplicables sólo a objetos rígidos o requieren de una segmentación previa o de una evaluación en múltiples ventanas extraídas en diferentes localizaciones y a diferentes escalas. La invarianza a los cambios del punto de vista exigen el escaneado del espacio de las transformaciones afines, que es computacionalmente muy costoso.

Los detectores basados en la apariencia [DT05a, POP98], suelen recorrer toda la imagen en busca de patrones de intensidad consistentes con el objeto deseado. Los experimentos realizados, han demostrado que estos sistemas funcionan bien para la detección de caras, pero no funcionan igual de bien cuando se trata de peatones, debido a que las imágenes son mucho más variadas (debido a las posturas del cuerpo y a las ropas). En caso de que la resolución de las imágenes sea baja – p.ej. cuando el objeto en la imagen tiene alrededor de 50 píxeles –, la detección del peatón es aún más complicada. Se puede mejorar el funcionamiento de estos métodos, mejorando la intensidad en las imágenes. Otras alternativas apuntan a la inclusión de más información; de hecho, ciertos tipos de movimientos son muy característicos de los humanos, por lo que el funcionamiento del detector puede mejorar al incluir información del movimiento.

Por tanto, según este otro criterio basado en rasgos, las últimas investigaciones en materia de reconocimiento de peatones siguen tres tendencias: detección de la periodicidad típica del caminar humano en base al movimiento, el análisis de la forma o bien se basan en la integración de un conjunto de rasgos.

2.4.1. Patrones basados en el movimiento

Existe una gran tradición en visión por computador por el estudio de una secuencia de imágenes, para percibir, interpretar y describir el movimiento humano [HF82, AN89]. Sin embargo, la mayor parte de los trabajos realizados, asumían que el objeto en movimiento era ya detectado por lo que se centraban en reconocer, categorizar o analizar el patrón de movimiento a largo plazo. Uno de los pioneros en el reconocimiento de personas andando fue Hogg [Hog83]. Recientemente, el interés por tratar de etiquetar la acción que tiene lugar en la escena ha aumentado considerablemente. Este cambio ha venido impulsado, no sólo por la disponibilidad de recursos computacionales, sino por los intereses de nuevas aplicaciones, como por ejemplo, los entornos interactivos o de realidad virtual, los sistemas de vigilancia, los juegos de ordenador o las películas de cine. En el influyente artículo de Cutler y Davis [CD00] se ofrece una visión general de los trabajos relacionados en este área. En lo concerniente a la detección de peatones orientado a aplicaciones de automoción, el foco de interés está puesto en el reconocimiento del movimiento más que en una interpretación del mismo.

El patrón de movimiento humano es fácilmente distinguible de otro tipos de movimientos. Muchos artículos recientes han empleado el movimiento tanto para detectar como para reconocer personas. Es una práctica muy extendida el tratar de realizar un seguimiento de los objetos en movimiento a lo largo de varias imágenes, para luego analizar dicho movimiento buscando

indicios del ritmo al que andan las personas o bien patrones de movimiento característicos de los humanos.

Sin embargo, este esquema basado en movimiento presenta muchas limitaciones a la hora de aplicarlo a la detección de peatones:

1. Las piernas de las personas deben ser visibles.

2. El reconocimiento se lleva a cabo en base a una secuencia de imágenes. En función del número de imágenes necesarios, la identificación del movimiento puede venir retardada en el tiempo – en el caso de necesitar varias imágenes, el tiempo de procesamiento se incrementa – o la identificación puede ser inmediata – obtención del movimiento directamente de la imagen –.

3. No detecta a individuos que estén quietos o cuyos movimientos sean extraños – tales como correr, saltar, girar o vagar –. Generalmente se limitan a gente caminando a velocidad constante.

4. Los peatones son objetos no-rígidos y cada parte del cuerpo tiene un movimiento diferente. Una posible alternativa para interpretar esos movimientos complicados, es representar conocimiento sobre la forma además del movimiento.

Estos motivos hacen que a día de hoy no exista un método sencillo, eficaz y fiable de detección de peatones basado en el movimiento.

Desde un punto de vista general, los métodos empleados con el objeto de reconocer y etiquetar el movimiento humano pueden dividirse en dos grandes categorias: estructurales y no estructurales. Los métodos estructurales, usan un modelo que describe la estructura cinemática humana y a veces, hasta la dinámica. Son los llamados métodos basados en modelos. Estos modelos pueden ser creados a mano o bien aprendidos de un conjunto de entrenamiento. Los métodos no estructurales evitan ese tipo de modelos y suelen citarse como métodos basados en la apariencia o métodos libres de modelo. Los métodos libres de modelo, persiguen describir el movimiento en la imagen, de manera que, variaciones en esa descripción del movimiento indiquen la presencia humana. Estos métodos tratan de poner el énfasis en los rasgos determinantes del campo de movimiento obtenidos de una secuencia de imágenes, sin acometer una reconstrucción estructural. Se pueden obtener variaciones en el modo de caminar obtenidas de variaciones en la secuencia y, por tanto, resulta factible realizar un análisis del modo de caminar libre de modelo.

A la hora de reconocer un objeto o bien una actividad, una pista muy importante es saber que el movimiento de ese objeto es periódico, es decir, presenta un patrón que se repite en el tiempo (ver fig. 2.9).

El movimiento periódico de las personas puede usarse para reconocer individuos [RWZD07] así como para hacer un seguimiento de los mismos y también permite reconocer acciones. Además, dicha periodicidad es reconocible incluso a muy bajas resoluciones (ver fig. 2.10). Sin embargo, hay que tener en cuenta que los movimientos humanos, aun siendo repetitivos, no son regulares. Es decir, el periodo varía de un ciclo a otro, o incluso de una parte del cuerpo a otra. Esta variación puede ser pequeña, aunque habrá situaciones en las que no sea adecuado

Figura 2.9: Ejemplo del movimiento cíclico de un humano

considerar que todas las partes del cuerpo comparten el mismo periodo o que el periodo sea estático en el tiempo. Además, las condiciones del entorno – como la iluminación, las sombras o los fondos saturados – y las variaciones del peatón – tanto de posición como de forma –, corrompen la señal del movimiento periódico y afectan negativamente a la detección.

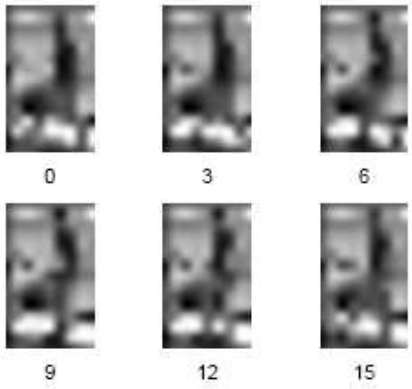


Figura 2.10: Secuencia de imágenes de baja resolución de un movimiento periódico (un peatón caminando) [CD00].

El interés por segmentar y analizar el movimiento cíclico y periódico es relativamente reciente. Durante el movimiento periódico, resulta evidente la variación de la distribución espacial del movimiento. Se puede capturar esa variación y analizar su evolución en el tiempo para obtener una descripción del movimiento periódico útil, caracterizada por una trayectoria espacio-temporal definida en algún espacio de rasgos.

Esa consistencia de la trayectoria, implica que para la escena observada, existe consistencia bien de apariencia, de flujo óptico o de la forma. Dicho de otro modo, los movimientos se pueden describir mediante su apariencia, su flujo óptico o su forma (ver fig. 2.11). La idea básica consiste en detectar aquellas regiones con una forma o una característica determinada, que comparten unos valores similares de flujo óptico, apariencia o forma, y realizar su seguimiento buscando la correspondencia en las subsiguientes imágenes.

Así, los distintos métodos no estructurales o libres de modelo existentes en la literatura, pueden agruparse como: aquellos que buscan la correspondencia de puntos, los que analizan la periodicidad de los píxeles [PN97, RWZD07], la periodicidad de movimiento de los rasgos [FL98, BHR97, RWZD07] o bien la periodicidad en función de la semejanzas entre objetos

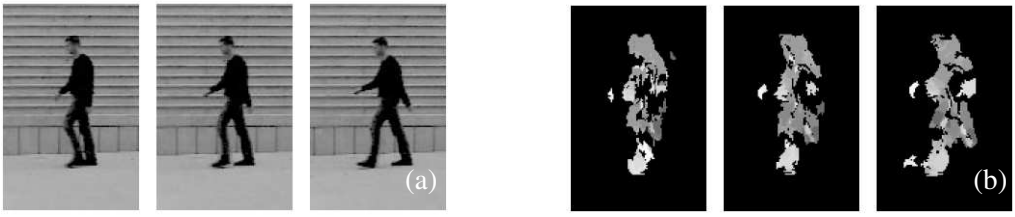


Figura 2.11: (*a*) Imágenes de una secuencia (*b*) Magnitudes del flujo óptico de la secuencia

[CD98, BH94b].

Baumberg y Hogg [BH94b] proponen un modelo que incorpora, además de la forma, la dirección del movimiento humano, permitiendo extrapolar la dirección a partir de la forma. Para ello realizan un seguimiento de regiones de una imagen a otra y describen la dirección en base a la semejanza entre regiones.

Con el fin de reconocer la periodicidad del modo de caminar humano, es habitual adoptar un análisis frecuencial de las variaciones de los patrones candidatos en el tiempo y después, seleccionar aquellos que muestran un espectro frecuencial característico del caminar humano. Se emplean métodos tradicionales, como la transformada de Fourier.

Como ejemplo, Cutler y Davis [CD00] usan una transformada de Fourier junto con una función de ventana de Hanning, para analizar las señales obtenidas mediante la correlación del patrón de los objetos detectados. Su algoritmo consta de dos partes; primero, segmentan el movimiento en la imagen y realizan el seguimiento de los objetos de interés. Después, calculan la auto-semejanza de cada objeto a medida que evolucionan en el tiempo. Si la periodicidad es estacionaria (no varía en el tiempo), el análisis de Fourier permite detectar y caracterizar dicha periodicidad. Al analizar la periodicidad de semejanzas en la imagen de regiones grandes en la imagen, en lugar de limitarse únicamente a píxeles individuales [PN97], su análisis de Fourier es mucho más sencillo.

Este trabajo es una extensión de [CD98] , permitiendo que la cámara esté en movimiento, además de realizar un análisis de la frecuencia-tiempo, que permite la detección y el análisis de periodicidades no estacionarias – es decir, que varían en el tiempo–. En este caso se emplea una función de ventana de Hanning, capaz de detectar el movimiento periódico humano (caminando o corriendo). Por otro lado, demuestran que el objeto de interés no tiene que ser segmentado del fondo – siempre que el fondo sea lo suficientemente homogéneo –. En caso contrario, el funcionamiento del sistema empeora. El sistema que describen en [CD98, CD00] es más directo que la mayoría de los sistemas basados en el análisis del movimiento de aquella época, en el sentido que trabajan directamente sobre imágenes que pueden tener una baja resolución y una pobre calidad. El sistema mide la periodicidad directamente de las imágenes seguidas y de un modo robusto. Casi todos los sistemas semejantes de aquel entonces, requerían de representaciones intermedias complejas, como puntos extraídos de los objetos seguidos o la segmentación correcta de las piernas.

[RWZD07], motivados por el trabajo de [CD00], presentan dos algoritmos para la detección de peatones probados en imágenes de infrarrojos y visibles. Obtienen mejores resultados

que [CD00], cuyo método es computacionalmente costoso y sensible a la saturación del fondo. Por otro lado, el uso de sensores del espectro infrarrojo proporciona imágenes con un nivel de ruido mayor que los sensores del espectro visible, lo que hace que los métodos basados en la similitud, como el propuesto por [CD00], fallen.

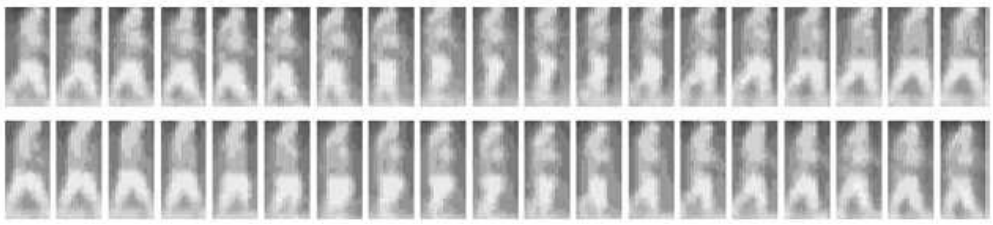


Figura 2.12: Ejemplo de la periodicidad extraída de imágenes de infrarrojos [RWZD07].

Uno de sus algoritmos se basa en un eficaz análisis frecuencial del movimiento en dos fases, que consigue filtra mejor que otros métodos del mismo tipo los píxeles no periódicos. Realiza dos test de hipótesis en cascada, comprobando primero, la periodicidad a nivel de píxel y después, lleva a cabo un análisis en conjunto de la distribución de los periodos (ver fig. 2.12).

Además de analizar la velocidad de movimientos tales como andar o correr, el modo de caminar describe el estilo o la manera en que un humano se mueve. Si bien la periodicidad permite diferenciar a un peatón de otros movimientos no periódicos —como los vehículos o las hojas de los árboles –, el modo de caminar permite diferenciar a los humanos de otros objetos que presentan un movimiento cíclico – como los animales–. Los movimientos periódicos de los peatones siguen un patrón distintivo; el balanceo de las piernas caracteriza la oscilación única de los humanos.

Este otro enfoque para el análisis del movimiento, usa modelos o patrones de la cinemática (ver fig. 2.14) y a veces, la dinámica del modo de caminar. Los métodos estructurales pueden aprender el modelo [BHR97, FGG+98, WKA00] aplicando técnicas de aprendizaje o bien, puede ser creado a mano [CEK+00].

Este enfoque clasifica un objeto en movimiento como un humano, mediante los rasgos contenidos en el patrón de movimiento cíclico. Como ya se ha comentado con anterioridad, no es sencillo decidir qué rasgos son buenos para ayudar a la extracción de patrones de movimiento cíclicos, típicos de los humanos. En principio, conviene usar rasgos globales y basados en la forma, frente a los basados en píxeles, para reducir así la sensibilidad de correspondencia de rasgos.

En varios estudios [BHR97, FGG+98, WKA00] que contribuyeron a la posterior creación del vehículo UTA (*Urban Traffic Assistant*) desarrollado por la DaimlerChrysler, se empleó un algoritmo basado en una red de neuronas con retardo (ATDNN, *Adaptable Time Delay Neural Network*) (ver fig. 2.13). Inicialmente, emplean una segmentación basada en estéreo para detectar y extraer regiones de la imagen conteniendo piernas de peatones. Después, la red ATDNN realiza un procesamiento local espacio-temporal, para detectar patrones típicos de movimiento. Así, la red es capaz de aprender los patrones de movimiento de un peatón en un ciclo completo. En la práctica se obtienen buenos resultados si las escenas duran unos segun-

Figura 2.13: Se muestra la imagen izda. estéreo. La caja negra viene determinada por el sistema estéreo. En la parte superior, se muestran las entradas a la red ATDNN; se muestra la ROI actual y las 7 precedentes. La señal de tráfico indica el paso detectado [WKA00].

dos y si el contraste entre las piernas y el fondo de la imagen es suficientemente alto.

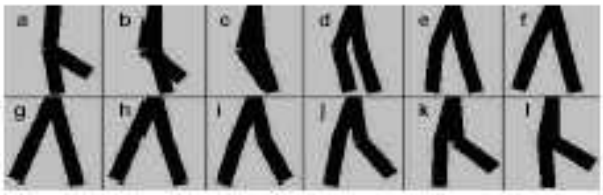


Figura 2.14: Modelo con lo pasos típicos del ciclo completo de caminar [CEK+00].

Curio *et al.* [CEK+00] han implementado un método para la detección, seguimiento y clasificación de peatones. Se hace un seguimiento del torso, de manera que se puede analizar la parte inferior del cuerpo y comprender el movimiento relativo de las piernas. La detección inicial se basa en dos tipos de datos: contorno y textura. Usan un modelo formado por varias imágenes sintéticas (ver fig. 2.14), que contienen las distintas fases que se dan durante un ciclo completo.

Los contornos los hacen corresponder con estos patrones, en un análisis multiescala. Con los resultados de la correspondencia y el análisis de la textura, detectan el movimiento periódico de las piernas, que es correlado con una curva experimental. Esta curva se obtiene de la media estadística de los periodos del caminar humano, creando un campo de activación dinámico temporal (DAF, *Dynamic activation field*). Los picos altos de la función de correlación indican la presencia de una persona (ver fig. 2.15). Por tanto, el reconocimiento final tiene lugar mediante un análisis en el tiempo del proceso de caminar en conjunto, combinando la forma del peatón y el movimiento periódico (el ritmo) de la piernas.

No obstante, obtienen errores como puede comprobarse en la figura 2.15. En principio se deben a que el algoritmo no es capaz de detectar personas que estén en zonas cercanas a la cámara, ni tampoco si el peatón presenta una forma que no está contenida el modelo (p.ej. si el sujeto lleva falda). La solución a esto último es crear un modelo para ese caso. El primer tipo de errores, tratan de resolverlos en una posterior versión optimizada de este sistema [CEK+00]. Además de la información de la textura y la correspondencia de modelos basada en contornos, añaden visión estéreo para la detección de obstáculos a distancias cortas y

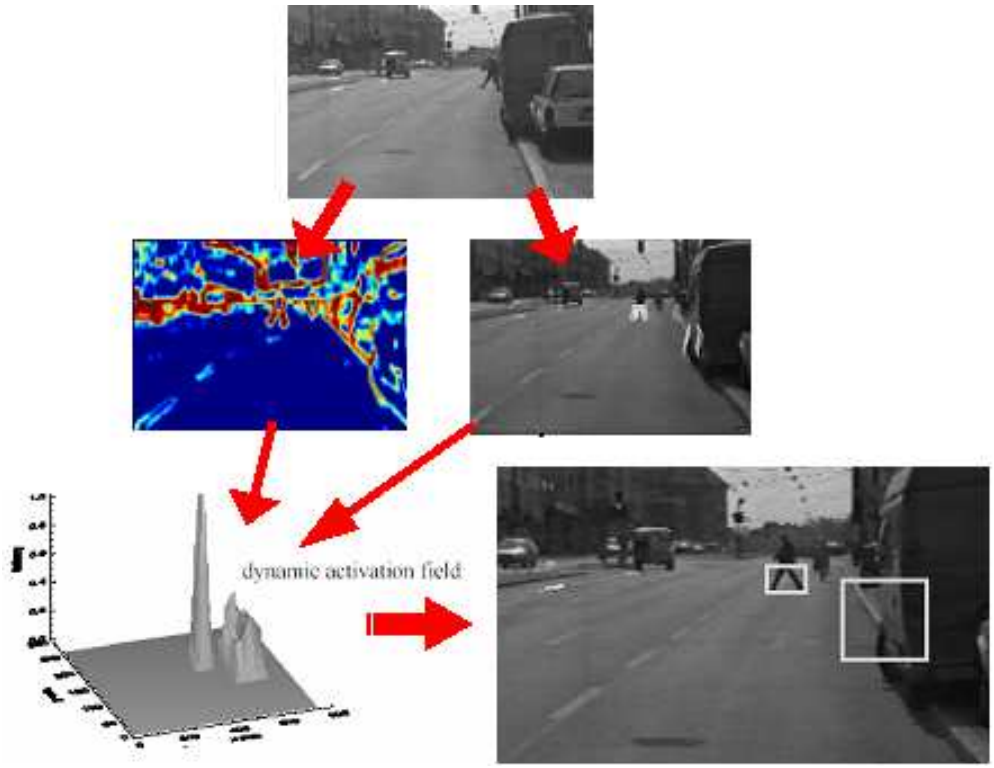


Figura 2.15: Sistema desarrollado por Curio en el que se integran varios rasgos [CEK+00].

medias. En conjunto, plantean una arquitectura flexible, que permite la integración de distintos aspectos.

2.4.2. Patrones basados en la forma

Se ha comprobado que usar un modelo basado en el movimiento es realmente complejo, sobre todo debido al movimiento de la cámara y a que los peatones siguen modelos de movimiento desconocidos o cuanto menos, impredecibles. Hay quien resuelve esto restringiendo el movimiento de la cámara o el ángulo de visión. Por tanto, aunque es un método robusto, requiere del análisis de múltiples imágenes y sólo pueden ser aplicado con sencillez en el caso de que el peatón esté cruzando la calle en la dirección del vehículo, ya que en este caso, el movimiento de las piernas es más evidente. Así es posible desglosar el modelo de movimiento en varias clases [BCZ93]. Su mayor inconveniente es su incapacidad para clasificar correctamente a personas que estén quietas o moviéndose según un patrón periódico no considerado.

Como alternativa, están los modelos basados en la forma, llamados así porque como información a priori contienen la forma aproximada del objeto. La gran ventaja de estos modelos es que permiten el reconocimiento tanto de peatones estáticos como en movimiento. La dificultad está en hallar el modo de englobar el amplio rango de situaciones bajo las que se puede presentar un peatón. Además, su funcionamiento depende de las entradas (las regiones de interés proporcionadas por la fase de detección o, en caso de ignorar esta fase, toda la imagen), por lo

que los modelos basados en la forma son más sensibles a falsos positivos que los basados en el movimiento. Por ello, para objetos que estén a corta distancia del vehículo, la información de la forma puede obtenerse con fiabilidad. Pero, para objetos a media o larga distancia no ofrece tan buenos resultados y es conveniente añadir otro tipo de información.

Los métodos de análisis de forma básicos consisten en hacer corresponder el modelo o plantilla sobre las regiones candidatas o la totalidad de la imagen. Otros métodos más sofisticados, aplican técnicas de *machine learning* para aprender el modelo y usan clasificadores para el reconocimiento final. Dentro de las técnicas convencionales de creación de modelos, el *Active Shape Model* – también conocido como PDM, *Point Distribution Model* –, ha demostrado ser un buen método para generar un modelo basado en la forma (compacto y lineal) a partir de ejemplos de entrenamiento. Pero, el PDM tradicional tiene un gran inconveniente: exige un etiquetado manual de un conjunto de puntos, llamados marcas, en cada imagen de entrenamiento, resultando un proceso muy laborioso [TCG95].

Algunos autores, en un intento por abarcar el amplio abanico de movimientos, tanto de la cámara como del peatón, así como la diversidad de formas de éste último, proponen modelos más generales. Baumberg y Hogg plantean un modelo basado en la forma flexible (*Active Shape Model flexible*) [BH94b], que se aprende automáticamente a partir de un conjunto de imágenes de entrenamiento. El método es guiado por los datos, lo que permite al modelo adaptarse a cambios, es decir, los parámetros del modelo permiten cierta variabilidad del contorno. Extraen las siluetas usando sustracción del fondo ya que trabajan con cámara estática, en un entorno interior. Rodean cada contorno con un B-*Spline* uniforme, obteniendo un vector de puntos de control. Esos puntos de control vienen a ser como las marcas en el PDM y los vectores obtenidos de los ejemplos, son alineados de un modo análogo a [TCG95]. Los autovectores con los autovalores más altos describen los modos de variación más significativos del modelo.

En [BH94b] añaden información de la variación temporal, modelando la variación de la forma como una lámina ”vibrante”. Finalmente, Baumberg [BH96], propone un método que automáticamente aprende un modelo espacio-temporal, que permite predecir el cambio esperado en la forma del objeto en el tiempo. Aumenta la credibilidad del sistema, aumentando ligeramente el tiempo de procesamiento. Este trabajo está orientado a sistemas de vigilancia y análisis del movimiento humano.

Broggi *et al.* [BBFS00] para el reconocimiento, filtran los candidatos obtenidos de la fase de detección, seleccionando aquellos objetos que muestran mayor afinidad con las características morfológicas de la forma de un humano. Para ello se basan en rasgos de la forma humana, haciendo uso de un conjunto de hipótesis: comprueban que la región candidata está dentro de unos rangos establecidos experimentalmente y descartan áreas demasiado homogéneas para contener una persona. Para refinar los límites de las ROI, buscan la cabeza del peatón en la imagen de los bordes, mediante la correspondencia con un patrón binario sencillo (ver fig. 2.16-b).

Para tomar una decisión final, comprueban si la zona en estudio puede ser asociada con una detección previa. Los trabajos posteriores de este grupo de investigación, siguen la misma estrategia basada en la aplicación características morfológicas para filtrar a los candidatos, pero empleando sensores de infrarrojos [BBF$^+$04, BCFG05, BBDL05]. Así mismo, la búsque-

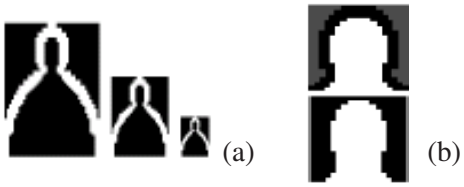


Figura 2.16: Ejemplo de los modelos de cabezas empleados en el VISLAB: (*a*) modelo para la correspondencia en el espectro visible [BBFS00] y (*b*) modelo para el espectro infrarrojo [BBC+07].

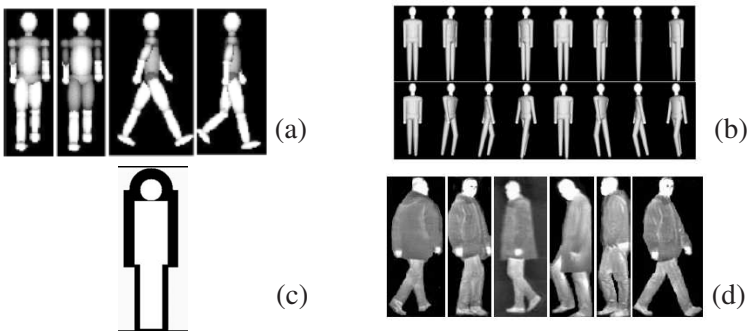


Figura 2.17: Ejemplos de los modelos de peatones empleados en el VISLAB [BBF+04] : (*a*) Modelos representando distintas vestimentas, posturas y desde distintos puntos de vista; (*b*) Ejemplos de 8 puntos de vista para peatones en pie y caminando. (*c*) El modelo simple que codifica las características morfológicas de un peatón; (*d*) Algunos ejemplos de modelos reales.

da de la cabeza también se considera para refinar las regiones de interés (ver fig. 2.16-a). En [BBF+04] emplean modelos de peatones para la correspondencia (ver fig. 2.17-b). Recientemente, han integrado varios métodos de validación de la forma humana; el detector de cabezas, un método basado en *snakes* y otro basado en modelos probabilísticos. Los *snakes* extraen la forma de los peatones en el dominio FIR [BBGD07]. A partir de la silueta extraída, utilizan una red neuronal para validar la detección. Los modelos probabilísticos, permiten evaluar la probabilidad de que en cada ROI exista un peatón [BBF+07b]. Combinando los votos de los tres validadores independientes, el sistema final determina si existe o no un peatón.

En [BCFG05], hacen un estudio entre un conjunto de modelos basados en la forma para evaluar cuál se ajusta mejor a las regiones candidatas (ver fig 2.17). Aquellos que se ajustan bien a algún modelo, son clasificados como peatones y se calculan tanto su tamaño como la distancia a la que están. Por otro lado, realizan un estudio de diferentes técnicas de validación basadas en modelos. Evalúan desde modelos muy sencillos basados en la forma, hasta modelos más complicados, que incluyen información termal. En el primer caso, filtran las regiones candidatas realizando una correspondencia de patrones empleando una sencilla máscara binaria, representación sin detalles de la forma humana. Este modelo se adaptaba bien a peatones que estén quietos, pero mostraba sus limitaciones con los peatones caminando. Además, la uniformidad del modelo no se ajusta a las texturas contenidas en los peatones reales. La co-

rrespondencia morfológica anterior fue mejorada empleando un modelo 3D de la forma humana, en escala de grises y un algoritmo de correspondencia de patrones más eficaz. El nuevo modelo comprendía diferentes posturas y actitudes humanas y podía generarse considerando distintos puntos de vista, adaptándose mejor a situaciones reales. Los modelos son precalculados, ya que requieren mucho tiempo de procesamiento como para ser calculados en tiempo real. También se consideraron diferentes intensidades de gris en el modelo, para representar las distintas temperaturas del cuerpo humana.

Se evaluaron tanto modelos sintéticos como modelos reales. Los resultados mejoran cuanto mayor es el conjunto de imágenes empleadas para la creación del modelo correspondiente, no apreciándose mejoras al incluir aspectos del sensor en el modelo –como incluir ruido en el fondo– ni al incluir información termal, ni al considerar un modelo para personas gruesas. Concluyeron que no es adecuado usar un modelo 2D o 3D con pocos detalles, ya que la correspondencia con los posibles peatones detectados en la imagen resulta inexacta. Apuntan a que un modelo que considere la información contenida en los bordes de los peatones filmados, junto con una función de correspondencia adecuada, dará mejores resultados que una correlación simple que únicamente considere los valores de los píxeles.

Los puntos débiles del sistema GOLD de Parma están relacionados con la falta de exactitud de la representación tipo caja que limita las regiones de interés. Como consecuencia, las medidas proporcionadas por la técnica estéreo empeoran dando lugar a falsas detecciones. En la fig. 2.18 se muestra como el sistema erróneamente detecta un árbol como si fuese una persona (debido a su alta simetría vertical, igual que ocurre con las personas). Como posibles soluciones, han probado a depurar el algoritmo de localización de la cabeza [BBF+04], refinar la construcción de la caja, hacer un uso mayor de la correlación temporal [BCFG05] y a la integración de sensores de infrarrojos y del espectro visible [BBF+06].

Figura 2.18: Ejemplo de errores de los modelos basados en la forma, del sistema GOLD de Parma [BBFS00].

El sistema de Philomin y Gavrila [PDD00], extrae los bordes de las imágenes y comprueba la correspondencia con un conjunto de patrones, bajo un enfoque del vecino más cercano. Philomin *et al.* [PDD00] siguiendo la vía de estudio abierta por [BH94a] y [TCG95], obtienen una técnica basada en modelos deformables optimizada (ver fig. 2.19). Por un lado, mejoran la parametrización de la curva B-spline y por otro, modifican un poco el método de alineación usado en el análisis de Procrustes, para dar más importancia a aquellos puntos que son más estables. Es decir, a la hora de alinear las formas, dan prioridad a aquellas partes que son estables. Un gran inconveniente de los modelos deformables es que dependen en exceso de su posición inicial.

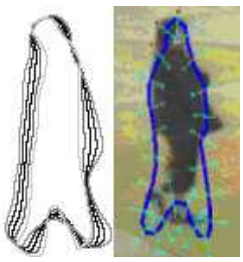


Figura 2.19: Ejemplo de los modelos flexibles usados en [PDD00].

Dentro de los métodos basados en el aprendizaje, hay que citar como indiscutible referencia el trabajo del MIT [OPS+97, POP98]. El grupo de investigadores de Oren y Papageorgiou fue uno de los precursores en la detección de peatones. Han dedicado varios años al estudio de una técnica basada en el aprendizaje de invariantes a través de ejemplos. Describen la forma humana como un subconjunto de coeficientes *wavelet* que extraen de los ejemplos de peatones o conjunto de entrenamiento. De manera que el sistema aprende lo que es un peatón a partir de ejemplos, por lo que no es necesario ningún modelo a priori. Con esos rasgos *wavelet* locales, entrenan un clasificador polinomial SVM.

El sistema detecta personas en cualquier postura, aunque sólo se ha probado con personas que están de frente o de espaldas. Básicamente se limitan a detectar peatones que entran o salen de la escena. Sin embargo, da falsos positivos si se le presentan patrones parecidos a lo que es una persona. Como algoritmo de búsqueda usan un método de ventana deslizante basado en fuerza bruta, que para obtener buenos resultados debe ser multiescala. Esto implica un coste computacional elevado. Hay que llegar a un equilibrio entre precisión en la clasificación y velocidad de procesamiento.

Los *wavelets* permiten realizar la detección a pesar de que la escena sea compleja, y sin hacer uso de información a priori, ni del entorno, ni de las personas, ni de su movimiento. Como contrapartida, requieren de muchos cálculos, resultando un sistema lento ya que tenían que realizar una búsqueda en toda la imagen a varias escalas, motivo por el cual su uso se ha restringido a detectar personas moviéndose de frente, de espaldas o de lado a la cámara. Además, la capacidad del sistema es limitado cuando se trata de reconocer personas que estén parcialmente ocluídas o con partes del cuerpo con un bajo contraste respecto al fondo.

En comparación con el trabajo de Baumberg o el de Philomin, ambos aprenden el modelo directamente de los contornos humanos previamente segmentados. En cambio en el trabajo de Oren y Papageorgiou [OPS+97] el modelo es directamente aprendido de los ejemplos y evita una segmentación explícita y el uso de detalles del movimiento.

[ELW03] usan el algoritmo propuesto por [OPS+97], realizando una extracción de rasgos basada en *wavelets* a partir de un conjunto de entrenamiento y también definen el patrón del peatón como un subconjunto de coeficientes *wavelets*. Clasifican las regiones de interés, considerando que existe peatón si el vector de rasgos *wavelet* correspondiente se aproxima suficientemente al vector de rasgos del conjunto de entrenamiento. A pesar del tiempo transcurrido entre este trabajo y en el que se ha basado, siguen sin resolver una de las mayores restricciones de la propuesta de [OPS+97]; los peatones deben aparecer de frente, nunca late-

ralmente a la cámara.

[GZNR04], también se inspiran en el trabajo de [POP98], pero en lugar de usar la transformada de *wavelet de Haar* para obtener la forma de un peatón, usan el detector de bordes horizontal y vertical de Sobel. Para la clasificación de las ROI emplean un par de SVM, uno para reconocer a peatones que están de frente o de espaldas, y el otro para reconocer posturas laterales.

[MPP01] amplían el trabajo elaborado por [POP98], al adoptar un enfoque basado en componentes, frente al enfoque precedente basado en la detección del cuerpo entero, en un intento por superar sus limitaciones. Los sistemas basados en componentes, detectan posibles partes del cuerpo humano –p.ej. la cabeza, las piernas, etc–, que luego integran para decidir si describen a la persona o no. Estos sistemas son escasos aplicados a la detección de personas, debido, por un lado, a la gran variación intraclase existente en la configuración de las distintas partes, y por otro, a la dificultad inherente a la tarea de detección de personas, cuya apariencia varía sustancialmente al estar en movimiento.

Shashua *et al.* [SGH04], también tratan de reducir la complejidad en cuanto a la apariencia de los peatones se refiere, usando un método basado en componentes. Es un sistema de detección de peatones monocular para aplicaciones de ayuda a la conducción cuyo funcionamiento apuesta por la integración de múltiples estrategias.

Relacionado con este tipo de sistemas basados en componentes, recientemente, se ha mostrado un gran interés por las estructuras de clasificación jerárquicas, consistiendo en herramientas de clasificación de datos que son combinación de otros clasificadores. Destacan los métodos de *bagging* y *boosting*.

El clasificador jerárquico propuesto en [MPP01], guía el aprendizaje basado en ejemplos a dos niveles, empleando una combinación de los clasificadores o detectores de componentes independientes de un nivel inferior, como entrada para el siguiente nivel y se ha denominado Combinación Adaptativa de Clasificadores. Emplean cuatro detectores de componentes – cabeza, piernas, brazo izquierdo y brazo derecho–, que calculan la transformada *wavelet de Haar* de cada región candidata y clasifican el vector resultante usando SVM cuadráticos. Comprueban que para cada componente, se cumplen una serie de restricciones geométricas. La mayor puntuación de las obtenidas por esos detectores de componentes, es la entrada del clasificador combinado que es un SVM lineal y determina si el patrón es una persona o no.

Shashua *et al.* [SGH04] lleva a cabo una primera clasificación basada en una imagen, pero la aprobación final la obtienen basándose en varias imágenes. En este caso, no se pueden usar SVM, ya que el tamaño de las subregiones es muy pequeño (12x36) y tienen una pobre definición, resultando complicada una localización correcta de los componentes en la imagen. Por ello, para cada subregión definen un vector de rasgos (formado por 13 elementos) insensible a los desplazamientos locales de las estructuras en la imagen, que es suministrado a una función discriminante. Los resultados obtenidos (en total son 13 elementos por 9 conjuntos de entrenamiento), son utilizados por un segundo clasificador que integra esos 117 elementos vía *Adaboost*. Cada combinación puede considerarse un detector débil (*weak learner*), obteniendo el peso más alto aquel detector que clasifique con un error menor. En conjunto, es un algoritmo de clasificación en dos fases, que obtiene en comparación, mejores resultados que el SVM

integral [OPS+97] o que el SVM en dos fases [MPP01]. Si las subregiones proporcionadas al SVM en 2 fases son muy pequeñas, su funcionamiento es peor que el del SVM integral, debido a la poca textura de las imágenes, no siendo suficientemente discriminantes (ver fig. 2.20).

Figura 2.20: Algunos errores del sistema de Shashua *et al.* [SGH04]. En la fila superior se muestran los falsos positivos, y en la inferior, los falsos negativos.

El enfoque basado en componentes maneja mejor las variaciones de iluminación y ruido en las imágenes que un detector de cuerpo entero, siendo además capaz de detectar a personas parcialmente ocluídas o ligeramente giradas, sin realizar ninguna suposición a priori sobre las imágenes a tratar. El motivo de que los sistemas basados en componentes tengan un mejor funcionamiento puede deberse a que emplean más información sobre la clase de objeto a detectar que los métodos de detección del cuerpo completo. Una representación local de la imagen divide la variabilidad de la clase en partes, cada una de las cuales tendrá su propia variación, que presumiblemente será bastante menor que la de la imagen completa. Además, al usar un conjunto de entrenamiento distinto para cada parte del cuerpo o componente, se incorpora conocimiento explícito sobre las propiedades geométricas de cada parte y se permiten variaciones en la forma, compensando los cambios de posturas o de las articulaciones. Por el contrario, usando un único conjunto de ejemplos y una representación global, es más difícil incorporar y controlar esta información.

El uso de clasificadores basados en el vecino más cercano o los clasificadores en cascada (como el SVM jerárquico) permiten integrar las distintas representaciones locales. El enfoque basado en el vecino más cercano, resulta muy eficaz cuando se tiene un clase (no un conjunto de componentes), cuando el número de descriptores es relativamente elevado (del orden de miles) y cuando loas descriptores locales se sitúan sobre regiones con mucha textura. Los SVM exigen un mapa de rasgos de orden superior (son funciones de base radial polinomiales), que se traduce en un gran número de vectores de soporte (10 % del conjunto de entrenamiento).

En el INRIA [DT05a] describen la forma humana en base a descriptores HOG. Después, este vector de rasgos es suministrado a un clasificador SVM lineal que decide si existe o no una persona. Franke *et al.* [FGG+98] Como técnica basada en el aprendizaje, proponen usar una base de datos con contornos de personas difuminados (técnica de *blurring*), a los que aplica PCA. Un problema de este método es que obtienen muchos falsos positivos.

El vehículo UTA desarrollado por la DaimlerChrysler fue concebido para la asistencia en carretera [Gav00] y aglutina varios trabajos anteriores: El sistema-Chamfer [GP99] y la jerarquía de patrones [Gav98, GG01, GG02] ha supuesto varios años de investigación. En base a los trabajos anteriores, el UTA realiza una clasificación (con redes de base radial o RBF, *Radial basis function*) de patrones basada en la textura y la forma, para verificar los candidatos obtenidos en la primera fase. A pesar de todos los esfuerzos, sigue siendo un sistema cuyos resultados dependen de un correcta segmentación del contorno. Confían en que con la integración temporal de los resultados se conseguirán reducir los errores. En la fig. 2.21 se muestran algunos ejemplos de detecciones incorrectas.

Figura 2.21: Ejemplo de errores de los modelos basados en la forma, del vehículo UTA desarrollado por la DaimlerChrysler [**?**]-b

Zhao y Torpe [Zha01] también se suman al reconocimiento basado en el aprendizaje. A partir de un conjunto de imágenes etiquetadas a mano, obtenidas del módulo de detección, una red neuronal extrae los rasgos de la forma que debe aprender. Una vez entrenada la red, clasifica las imágenes de entrada en función de esos rasgos. Como entrada a la red, sugieren usar el gradiente de la imagen en lugar de los niveles de gris. Justifican su elección argumentando que (1) la intensidad no es adecuada para representar la forma de las personas, porque éstas presentan mucha variación en color y textura y (2) son sensibles a los cambios climáticos. Para que no le afecten, habría que entrenar la red durante la ejecución.

Una posible solución invariante a cambios de color y de textura sería usar la silueta. El problema es que la segmentación basada en estéreo no siempre da buenas siluetas (si hay mucho ruido o ante oclusiones complejas). Otra solución sería usar la imagen de bordes como entrada, pero, la detección de los mismos depende en gran medida del umbral que se use. La elección del umbral tiene el inconveniente de basarse en heurísticas. Zhao *et al.* [Zha01] proponen como solución a todos los problemas citados, el uso del gradiente. Es robusto ante cambios, por tanto, la red sólo hay que entrenarla *off-line*. Presenta las mismas ventajas que la representación basada en bordes, evitando el problema de la umbralización.

El sistema da falsos positivos ante objetos cuyo contorno se asemeja al de una persona, o si el color de la persona se confunde con el fondo, o cuando dos personas están muy juntas, resultando complicado separarlas en función de rasgos de la forma o estéreo. Como posibles soluciones sugieren incluir detalles sobre el movimiento o utilizar integración temporal, para verificar la existencia o no de un peatón.

Tras un análisis comparativo realizado en la Universidad Carnegie Mellon entre su trabajo [Zha01] y el del MIT [OPS+97], los primeros consiguen una tasa de detecciones mayor siendo el número de falsas alarmas menor. El MIT obtiene mejores resultados cuando aplica la transformada *wavelet* a imágenes en RGB en lugar de aplicarlo a imágenes en niveles de gris. Al añadir color, el sistema obtiene resultados más fiables.

Entre los sistemas que emplean cámaras de infrarrojos, en [ND02] se emplea una plantilla probabilística para capturar las variaciones de la forma humana, capaz de tratar situaciones de bajo contraste o de omisión de alguna parte del cuerpo (ver fig. 2.22). La plantilla es construida a tres escalas diferentes, a partir de un conjunto de ejemplos. La función de correspondencia es ejecutada en tres diferentes mapas de probabilidad y, tras aplicar un umbral obtenido de un clasificador Bayesiano, se identifican los máximos locales como peatones. En [BBGM07] utilizan esta plantilla para realizar la clasificación de las regiones de interés. Los modelos probabilísticos también han sido empleados en el dominio visible [Hu06, SG00b]. Por otro lado, los SVM también han sido aplicados al dominio infrarrojo monocular [XF02].

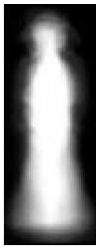


Figura 2.22: Plantilla probabilística de Nanda *et al.* [ND02]

2.4.3. Patrones basados en la forma y el movimiento

Viola *et al.* [VJS05] integran en su modelo información de la intensidad y del movimiento, pero asumen que la cámara es estática y la escena observada, sufre pocos cambios. Emplean *AdaBoost* para el aprendizaje de un conjunto de clasificadores de regiones cada vez más exigente, basados en *wavelets de Haar* y en la diferencia espacio-temporal. Cada clasificador es aprendido de un conjunto de ejemplos etiquetados como positivos y negativos, seleccionando los mejores descriptores de movimiento y de la apariencia.

Viola y Jones [VJ01b], demostraron que el uso de un sólo clasificador daría como resultado un detector lento, al requerir muchas características. Para conseguir un detector eficaz, proponen una arquitectura en cascada, donde los clasificadores más simples – con pocos rasgos – se prueban antes que los más complejos – con un gran número de rasgos–, que se sitúan al final de la cascada, procediendo en una detección que va creciendo en complejidad.

En la extensión del trabajo anterior [NDS06], el INRIA tenía como objetivo el poder detectar peatones desde un vehículo en movimiento, donde tanto la cámara como el fondo de la escena pudiesen sufrir movimientos, al menos iguales a los de los peatones filmados. Usan, una vez más, un clasificador SVM lineal para clasificar los descriptores obtenidos de las regiones resultantes. Son descriptores que combinan información de la apariencia y del movimiento.

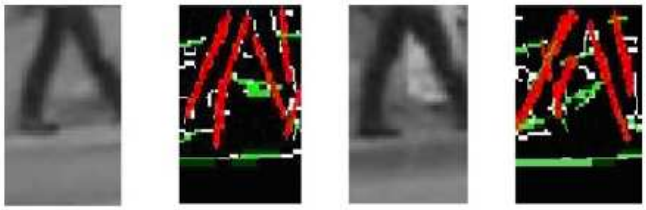


Figura 2.23: Ejemplo del método basado en correspondencia de patrones de [RWZD07]. Los píxeles en blanco corresponden a los bordes; los verdes son las líneas detectadas con la transformada de Hough; los rojos(pares de segmentos de líneas a lo largo de las piernas): líneas que forman en máximo ángulo al caminar (MPGA).

[RWZD07] también integran la forma y la apariencia con el movimiento en su segunda propuesta. Apuestan por un rasgo obtenido del contorno humano, en lugar de usar el contorno en sí mismo, ya que la segmentación de la forma puede no ser precisa. Así, se basan en el máximo ángulo formado por las piernas (MPGA, *Maximal Principal Gait angle*). Aplican un detector de bordes y aquellas regiones con un MPGA muestran un patrón basado en el ángulo destacado (ver fig. 2.12). Una vez detectado el MPGA para una imagen, evalúan la periodicidad del movimiento. Este detector en cascada, ha sido probado con imágenes de infrarrojos y con imágenes en escala de gris. El uso de sensores de infrarrojos, difuminan los objetos en la imagen, haciendo que sus contornos y apariencia se confunda con el fondo, y como consecuencia, la mayoría de los métodos basados en la forma y el movimiento fallan.

2.4.4. Integración de varias características

Como conclusión tras revisar los trabajos anteriores, se puede decir que es poco realista esperar un funcionamiento razonable a nivel de sistema basándose únicamente en la clasificación de una imagen. Sólo si se unen varias decisiones cabe esperar que el sistema segmente peatones con el suficiente nivel de fiabilidad. La clave está, por tanto, en la integración de rasgos o pistas adicionales medidas en el tiempo (p.ej. el patrón del caminar dinámico, el *motion parallax* – movimiento aparente de un objeto debido al movimiento del observador – o la estabilidad de las medidas ya detectadas), rasgos característicos de la situación (como la posición de las piernas en determinadas posturas) o mediante la inclusión de nuevas categorías de objetos, tales como vehículos o estructuras estacionarias del fondo (p.ej. farolas, árboles, marcas de la carretera, etc.). Por citar algunos autores que ha hecho uso de esta idea [BBFS00, CEK+00, HHD98, ELW03, BBF+06, BBF+07b, GM07].

Algunos sistemas más sofisticados realizan un reconocimiento de patrones usando además clasificadores [Gav00, PDD00, Zha01, Cel01] o usan en combinación un análisis de la forma junto con una detección del modo de caminar [POP98] para acometer la etapa de reconocimiento.

Shashua *et al.* [SGH04], recogen información adicional de múltiples imágenes para dar una respuesta a nivel de sistema (ver fig. 2.24). Entre las características consideradas están la periodicidad del modo de caminar, el análisis del movimiento en sí mismo y del *motion*

parallax –cuando está disponible–, la consistencia en el tiempo de los resultados obtenidos del clasificador basado en una imagen y medidas de la calidad del *tracking*.

Figura 2.24: Errores del sistema propuesto en [SGH04] (a)y (b) muestran una imagen típica de detección de peatones. (c),(d) contienen ejemplos de falsos positivos: las patas del caballo en (c) y un peatón fuera de la trayectoria a una distancia de 23m. en (d). En (e) aparece indicado con una flecha un ejemplo de un peatón no detectado.

[RWZD07] como línea de trabajo futura, proponen usar algún detector de formas [VJS05] o una jerarquía de formas [ND02], seguido de su detector de movimiento cíclico o del detector de píxeles periódicos presentados en ese artículo, para mejorar los resultados. Así no tendrían que recorrer toda la imagen, siendo desde un punto de vista computacional, más eficaz.

En [GM07] se describe el sistema para la detección y seguimiento de peatones PROTECTOR, que combina distintas pistas. La detección implica una cascada de módulos, cada uno especializado en un aspecto de la visión artificial. Así combinan algoritmos estéreo denso y no-denso, correlación basada en la forma y un clasificador basado en texturas. Como resultado, consiguen detectar correctamente entre un 62 % a un 100 % de los peatones.

2.5. Seguimiento en secuencias de imágenes

Como se ha visto, la fase de seguimiento no es una fase a la que se le haya prestado demasiado atención en los sistemas de detección de peatones. Las investigaciones han ido más en la línea de desarrollar módulos de detección robustos, tratando de reducir las falsas alarmas al máximo. Sin embargo, se ha constatado que la mayoría de sistemas proponen como futuro desarrollo el uso de información temporal. Es de esperar que hacer uso de información de secuencias previas, ayude a la toma de decisiones en el instante actual.

En esta línea, el seguimiento consiste en encontrar los objetos correspondientes entre secuencias. Las dificultades de esta tarea tienen que ver con la complejidad de la escena y de los objetos a seguir. En cuanto a los problemas relacionados con los objetos están las oclusiones y el ruido (ambos pueden hacer que los objetos se unan o dividan en nuevos objetos) y la apariencia de un objeto puede variar debido a sombras o cambios de iluminación.

El análisis de correspondencia suele llevarse a cabo usando predicción. Basándose en objetos detectados previamente y probablemente en conocimiento de alto nivel, se predice el estado de los objetos (apariencia, posición, etc.) en la siguiente secuencia y se compara (usando alguna métrica) con el estado de los objetos encontrados en la secuencia actual.

La predicción introduce una región de interés tanto en el espacio de la imagen como en el espacio de estados y así reduce la información necesaria para el procesamiento. La pre-

dicción de los diferentes parámetros de estado se basa en un modelo de cómo evolucionan en el tiempo. Puede usarse un modelo de velocidad y aceleración o modelos más avanzados de movimiento como por ejemplo caminar. Otra alternativa consiste en aprender modelos de movimiento probabilísticos.

Un método muy utilizado para predicción es el filtro de Kalman [GZNR04, BCFG05, GG02] , que además puede estimar las incertidumbres de la predicción y ser usadas para determinar las regiones de interés. Un inconveniente del filtro de Kalman es que sólo puede usarse en situaciones donde la distribución de probabilidad de los parámetros de estado es unimodal. Frente a oclusiones, fondos saturados parecidos al objeto a seguir y dinámicas complejas, la distribución es probable que sea multimodal. Como alternativa, se han desarrollado algoritmos capaces de seguir hipótesis múltiples (distribuciones multimodales). El más conocido es el algoritmo de condensación [PDD00]. Aunque como es no-paramétrico requiere una cantidad de muestras considerable. En problemas de alta dimensión, puede ser necesario usar otro método más eficaz. Otra alternativa es el filtro de Kalman extendido [YM00].

Figura 2.25: Ejemplos de sistemas de seguimiento de personas: (*a*) [CEK+00] utiliza varios rastreadores.; (*b*) Resultados de la Universidad de Reading [SM02a].

En el sistema de [CEK+00] se implementó una fase de seguimiento, analizando los resultados de distintos algoritmos clásicos. En [BvS03] realizan el seguimiento aplicando una técnica de votación a zonas simétricas. Para hacer frente a las rotaciones de la persona, usan varios rastreadores. También añaden información estéreo (ver fig. 2.25-a).

La Universidad de Leeds [PA97] fue pionera en proponer la integración de módulos especializados para el seguimiento de objetos rígidos (vehículos) y no-rígidos (personas), resolviendo posibles oclusiones. El sistema modular desarrollado esta enfocado a desempeñar tareas de vigilancia. Tienen otro subsistema independiente (el módulo de detección), que detecta objetos en movimiento e invoca al módulo de seguimiento correspondiente. Para reconocer regiones en movimiento emplean una técnica de sustracción de fondo clásica, ya que la cámara es fija.

El módulo de seguimiento de personas usa un modelo deformable 2D aprendido de ejemplos, construido con curvas B-*splines* para representar el contorno del peatón. Usan filtros de Kalman para la estimación de los parámetros de la forma. Comprueba que las regiones que le pasa el módulo de detección de movimiento cumplen una serie de requisitos morfológicos.

En la universidad de Reading, Siebel y Maybank [SM02a], tomaron como punto de partida el sistema desarrollado por Baumberg y Hogg [BH94a] para implementar el sistema de vigilancia ADVISOR (ver fig. 2.25-b). Además de seguir personas en estaciones de metro,

analizan su comportamiento. Modifican el sistema de Baumberg *et al.*, añadiendo un módulo de seguimiento de regiones y otro de detección de cabezas. El primero hace uso de la información temporal para obtener detecciones fiables. El segundo está inspirado en el algoritmo de [HHD98].

Siebel *et al.* [SM02a] tratan de solucionar los problemas que el sistema de Baumberg [BH94a] presenta ante oclusiones y en la inicialización de los modelos deformables, con sus nuevos módulos de seguimiento de regiones y detector de cabezas. Hay que recordar que estos sistemas de vigilancia tienen la ventaja de trabajar con cámara fija y generalmente en escenas interiores, permitiendo el uso de una serie de técnicas que no se pueden llevar a un sistema de detección de peatones con cámara en movimiento.

2.6. Perspectiva: Sistemas Basados en Visión Integrados en Vehículos

Este estado del arte se centra en el análisis de los sistemas de detección de peatones basados en visión, instalados en un vehículo. En la actualidad hay pocos sistemas que se ciñen a esa descripción. Uno de los primeros fue el desarrollado en el Massachussets Institute of Technology [POP98]. Tiene el inconveniente de requerir mucho tiempo de procesamiento y por ello ha limitado a la detección de peatones de frente o de espaldas.

La empresa automovilística DaimlerChrysler, junto con la Universidad de Maryland, han apostado por el desarrollo de vehículos inteligentes. Las diversas investigaciones y algoritmos desarrollados a lo largo de los años, se han consolidado en el vehículo UTA, orientado a la asistencia en la ciudad [FGG+98, GP99, Gav00]. De sus varios módulos tiene especial importancia el sistema STOP & GO, que permite el seguimiento autónomo del vehículo de delante, teniendo al mismo tiempo en cuenta otros elementos importantes de la carretera.

El Robotics Institute de la Universidad Carnegie Mellon [Zha01] han creado un módulo de detección de peatones que forma parte de un proyecto de asistencia a conductores de autobús. Su objetivo es lograr un sistema que advierta de posibles colisiones laterales, generando distintos grados de avisos, según la gravedad de cada situación. En la actualidad existen sensores para evitar colisiones laterales, pero están pensados para detectar coches o camiones, no para personas.

La Universidad de Parma, en colaboración con la de Pavia, ha desarrollado el vehículo ARGO [BBFS00], que puede incluso conducir de forma autónoma en determinadas situaciones. Consta de varios módulos orientados a crear un sistema de protección activo, que se conoce como el sistema GOLD. Entre los distintos métodos orientados a la detección de peatones, explotan las características morfológicas y la fuerte simetría de la forma de las personas. Para el reconocimiento o validación de los candidatos, han probado una amplia gama de métodos. Destacan los métodos basados en el aprendizaje, como las técnicas de reconocimiento de patrones basados en la forma o los clasificadores SVM entrenados con descriptores HOG. Recientemente [BBF+04], han integrado dos pares de sistemas de visión estéreo, combinando las ventajas ofrecidas tanto por el dominio visible como por el infrarrojo, facilitando la tarea de segmentación y de refinamiento de las regiones de interés sustancialmente.

El Institut für Neuroinformatik de la Universidad de Bochum (Alemania) pone especial énfasis en la necesidad de desarrollar arquitecturas modulares y fusionar la información de distintos sensores (radares y cámaras) para solucionar el problema de la detección. Resaltan la importancia de obtener rasgos de alto nivel que puedan ser accedidos por los distintos módulos, aumentando así la robustez y eficacia del procesamiento en conjunto. Como novedad realizan un análisis de la escena, basándose en conocimiento contextual e informan al conductor del comportamiento esperado en tales condiciones. Es un trabajo orientado a la conducción automática [**?**, BvS03].

Con el fin de mejorar la seguridad de los usuarios más vulnerables de la carretera – a saber, peatones y ciclistas –, se han lanzado algunas iniciativas. En la UE existen numerosos proyectos enfocados hacia aspectos específicos de los sistemas de seguridad para vehículos inteligentes. Algunos ejemplos son los proyectos de investigación PROTECTOR (*Preventive Safety for Unprotected Road User*, 2000-2003), SAVE-U (*Sensors and System Architecture for VulnerablE road Users protection*, 2002-2005), EDEL (*Enhanced driver perception in poor visibility*, 2002-2005), CHAMALEON (*Pre-crash application all around the vehicle*, 2000-2003) y WATCH-OVER (*Vehicle-to-Vulnerable roAd user cooperaTive communication and sensing teCHnologies to imprOVE transpoRt safety*, 2006-2008) , todos ellos financiados por la Comisión Europea.

El proyecto PROTECTOR [GGM04] y su sucesor SAVE-U [GML+03], pretenden impulsar el desarrollo de soluciones basadas en sensores para la detección de dichos usuarios, para así facilitar el uso de medidas de aviso o preventivas para evitar o mitigar el impacto de las colisiones. El proyecto PROTECTOR ha sido el primero en integrar sistemas avanzados de asistencia a la conducción (ADAS) para la detección y protección de los usuarios más vulnerables de la carretera. Por un lado, se han puesto de relieve la viabilidad técnica y los beneficios, la evaluación de riesgos y las estrategias de advertencia a perseguir en este tipo de sistemas. Por otro lado, se han identificado limitaciones. Se han empleado sensores de a bordo basados en láser, microondas y técnicas de visión artificial. El proyecto SAVE-U va a permitir la detección fiable y precisa en todo tipo de condiciones meteorológicas, uniendo la información de cámaras de infrarrojos y visible.El proyecto EDEL [Com02] está trabajando en un sistema de asistencia al conductor para visión nocturna, basado en sensores de infrarrojos cercanos y en una interface humano-máquina específica. Además, se va a llevar a cabo el seguimiento de obstáculos mediante la detección en tiempo real de la distancia, velocidad y trayectoria de los objetos basado en estéreo. El proyecto CAMALEON [Com00] ha investigado la creación de un sistema capaz de detectar una situación peligrosa, dando la posibilidad de minimizar y mitigar las consecuencias de un choque gracias a mecanismos de detección pasiva. En la actualidad, el proyecto WATCH-OVER [Com06] tiene como finalidad incrementar las posibilidades de detectar peatones, mediante el uso combinado de técnicas basadas en visión y técnicas cooperativas.

Los prometedores resultados obtenidos en las iniciales investigaciones en vehículos inteligentes, han demostrado que la completa automatización del tráfico (al menos en las autopistas o en carreteras suficientemente estructuradas), es técnicamente posible. Además de superar los problemas técnicos, deben cuidarse otros aspectos – como los legales o el impacto de una con-

ducción autónoma en los usuarios del vehículo – en el diseño de estos sistemas. En particular, la aceptación por parte de los usuarios, va a ser una pieza clave que va a afectar a la apariencia y el funcionamiento del vehículo, y el interfaz del sistema va a tener una fuerte influencia en cómo el usuario va a observar y comprender la funcionalidad del sistema.

Como consecuencia, antes de que estos sistemas estén disponibles en el mercado, deben precederlos un largo periodo de exhaustivos tests y perfeccionamiento, y un sistema de autopistas completamente automatizado con vehículos inteligentes conduciendo e intercambiando información, no será factible hasta dentro de un par de décadas. Mientras tanto, la completa automatización estará restringida a infraestructuras especiales, como aplicaciones industriales o transporte público. Después, la tecnología de vehículos automáticos será gradualmente extendida a otras áreas de transporte claves, como la movilidad de cargas. Finalmente, una vez que la tecnología se haya estabilizado y se hayan fijado las soluciones más prometedoras y los mejores algoritmos, tendrá lugar una masiva integración y expansión de esos sistemas a los vehículos privados, pero para que esto ocurra, todavía habrá que esperar otras dos o más décadas.

Capítulo 3

Especificación del Sistema de Percepción

Una vez enmarcada esta tesis dentro de los Sistemas Avanzados de Asistencia a la Conducción y revisados los trabajos precedentes, se puede afirmar que la detección de peatones es un desafiante reto y un campo abierto a la investigación. Entre las distintas soluciones que se pueden adoptar, la visión artificial ofrece una serie de interesantes cualidades, que pueden ser aprovechadas con el fin de evitar una posible colisión.

En esta tesis se ha apostado por un sistema sensorial basado en visión por computador. Con el objetivo de implementar un sistema embarcado en un vehículo para la detección de peatones, se han empleado dos sistemas de visión distintos; el primero de ellos, ha sido desarrollado en la Universidad Carlos III de Madrid por el grupo LSI (Laboratorio de Sistemas Inteligentes), mientras que el segundo, ha sido creado en la Universidad de Parma por el grupo VISLAB (*Vision and Intelligent Systems Lab*). El sistema de la Carlos III es un sistema estéreo compuesto por cámaras del espectro visible, mientras que el de la Universidad de Parma consiste en un sistema de Tetravision, formado por dos pares de sistemas estéreo: uno hace uso de cámaras del espectro visible y el otro de cámaras de infrarrojos lejano.

En este capítulo, se van a describir cada uno de los sistemas de visión empleados, abordando por un lado, aspectos técnicos relacionados con la construcción de cada sistema, y por otro, detallando los procedimientos que han sido necesarios aplicar previos a la obtención de un sistema de visión preparado para la detección de peatones. Esos procedimientos son la calibración del sistema y la rectificación de las imágenes obtenidas.

En primer lugar, se sientan las bases de las nociones matemáticas sobre proyección perspectiva necesarias para comprender la calibración y rectificación de cada sistema estéreo empleado. Después, se detalla cada uno de los sistemas de visión utilizados, tanto el desarrollado en la Universidad Carlos III, que ha sido construido por completo y ha sido necesario calibrarlo y rectificar las imágenes resultantes, como el desarrollado por la Universidad de Parma, que es un sistema que se ha utilizado ya calibrado.

3.1. Especificación del Sistema de Visión en el Dominio Visible e Infrarrojos

El motivo fundamental de analizar tanto imágenes del espectro visible como del infrarrojos, reside en la imposibilidad de resolver la detección de peatones limitándose el estudio a un sólo dominio. A continuación se describen las cualidades que aportan cada uno de esos dominios, con la idea de evidenciar tanto los puntos fuertes como los débiles. Después, se describen por separado cada uno de los sistemas de visión empleados en el desarrollo de esta tesis.

3.1.1. Propiedades de las Imágenes Visibles e Infrarrojos

Las imágenes en el dominio infrarrojo capturan un tipo de información muy diferente de las imágenes en el espectro visible. Si en el espectro visible la imagen de un objeto depende de la cantidad de luz que incide en su superficie y lo bien que la refleja, en el dominio infrarrojo, la imagen de un objeto está relacionado con su temperatura y la cantidad de calor que emite. En el espectro visible, las ventajas más destacadas son:

- Captan el color, ofreciendo una información más rica.
- Ofrecen mayor resolución que las cámaras de infrarrojos.
- Los límites de los objetos están bien definidos debido a que el contraste con el fondo es bueno.

Entre los inconvenientes;

- Las intensidades en las imágenes visibles varían de un modo significativo de unas partes del cuerpo a otras en función de la ropa.
- Muestran gran cantidad de pequeños detalles y muchas sombras.
- Son altamente sensibles a los cambios de iluminación.

El espectro infrarrojos presenta una serie de características que pueden ser interesantes a la hora de detectar humanos. Por citar las más beneficiosas;

- Al contrario que ocurre en el espectro visible, los cambios de iluminación no les afectan tanto. Los niveles de intensidad en una imagen de infrarrojos son representativas de la temperatura de las superficies de los objetos. Los peatones, generalmente, emiten más calor que los objetos estáticos, como los árboles o la carretera. Por tanto, esas regiones que contienen peatones aparecerán más brillantes que el resto – ya que toman valores altos en la escala de grises – y estarán suficientemente contrastadas con respecto a su entorno como para hacer que las imágenes de IR sean particularmente adecuadas para la localización de peatones. Obviamente, otros objetos en la escena pueden emitir calor de un modo activo, como es el caso de los vehículos o las farolas y presentar un comportamiento similar al de un peatón, imposibilitando así una segmentación basada únicamente

en valores de intensidad alta. Como posibles alternativas, se puede discriminar a los peatones en función de su forma o *aspect ratio*, aplicando técnicas de reconocimiento de patrones.

- Los humanos poseemos una temperatura similar, por tanto, el nivel de intensidad de diferentes peatones debería ser parecido independientemente del color y de las texturas de las ropas y, por tanto, la intensidad del conjunto de regiones correspondientes a una persona, deberían ser aproximadamente uniformes.

- Son menos sensibles al ruido, debido a la casi total ausencia de sombras y texturas que se originan cuando existen colores.

- Una de las principales ventajas de las imágenes de infrarrojos es que no se ven casi afectadas por los cambios de iluminación, resultando muy adecuadas para situaciones en que la luz es escasa o incluso de noche.

A pesar de estas ventajas, la detección de peatones en imágenes de infrarrojos está lejos de ser trivial. Entre sus inconvenientes, conviene destacar los siguientes:

- Las condiciones del tiempo, como por ejemplo, la lluvia o una niebla pesada, pueden modificar la firma termal de los cuerpos, limitando la efectividad de los sistemas de infrarrojos.

- En condiciones de altas temperaturas o cuando el calor del sol es intenso, la diferencia de temperatura entre peatones y el resto de objetos puede ver se disminuida. De hecho, objetos que presentan un comportamiento de radiación del calor pasivo – como las señales de tráfico, barreras, árboles, edificios o marcas de la carretera – pueden ser fuertemente calentadas por el sol, haciendo que la escena sea aún más compleja o incluso causando radiaciones o reflexiones de calor. Por añadidura, en el caso de fuerte radiación de calor externa, las ropas que llevan las personas pueden tener un comportamiento termal distinto, en función del tipo de material y del color, y por tanto, añadir texturas a la imagen.

- Por el contrario, en situaciones de temperaturas externas bajas, las ropas pueden aislar la emisión del calor de un modo significativo, y sólo algunas partes del cuerpo – como la cabeza y las manos – pueden ser apreciables. Como consecuencia, es habitual que un peatón en infrarrojos aparezca con la zona del torso más oscura que las zonas de la cabeza y las manos.

- Otro inconveniente, aunque menos grave que en el espectro visible, está relacionado con los objetos transportados por las personas.

- En comparación con las ricas y coloridas imágenes del dominio visible, las imágenes de infrarrojos resultan borrosas, tienen una peor resolución y un contraste entre fondo y objeto más bajo.

Los problemas mencionados complican aún más la detección de peatones. Pero, el dominio infrarrojos parece prometedor y justifica una profunda investigación. A menudo, en las primeras etapas del proceso de detección, se emplea la forma como un preliminar detector de patrones. En tal caso un exceso de detalles resulta una carga y complica la detección. Pero los detalles son cruciales para poder distinguir entre peatones y objetos.

En definitiva, ni las cámaras de infrarrojos ni las cámaras del espectro visible por si solas constituyen una solución completa para la tarea de detección de peatones. Se puede pensar en usar en conjunto imágenes del dominio visible y de infrarrojos con el fin de sacar provecho a los beneficios de cada tecnología.

3.1.2. Caracterización de Peatones en el Espectro Visible e Infrarrojo

El peso del funcionamiento de un detector de peatones típico, recae en el conjunto de rasgos extraídos. Si bien los rasgos habitualmente buscados en las imágenes del dominio visible lógicamente diferirán de los rasgos buscados en las imágenes de infrarrojos, existen algunas coincidencias. Los rasgos comunes empleados en ambos dominios están relacionados con rasgos únicos de los peatones, que describen las características propias de un ser humano. A continuación se detallan los rasgos que se buscan habitualmente en cada uno de los dominios.

En el espectro visible, las principales ventajas son que se puede usar el color y detectar bordes con precisión. Explotando estas dos características se pueden aplicar las siguientes técnicas. Se han dividido en dos grupos; las primeras, se usan para detectar regiones de interés y las segundas, son más apropiadas para la fase de clasificación posterior.

Para llevar a cabo la identificación de las regiones de interés, se buscan rasgos genéricos para una segmentación inicial veloz. Son comunes:

- Búsqueda de regiones con una propiedad de interés.- La simetría, tanto de los vehículos como de los peatones, y la asimetría del fondo, es un escenario de SIT típico. Por ello, se han desarrollado buscadores de simetrías para detectar regiones de interés. Es invariante a los movimientos de cabeceo del coche y a cambios de tamaño del objeto. Pero, pueden no funcionar bien para objetos en posiciones arbitrarias.

- Técnicas basadas en estéreo.- La información de la profundidad ayuda a separar los píxeles en función de la distancia, y ayudar a la detección de regiones de interés. Su bondad reside en la calidad de la calibración estéreo y la correspondencia entre puntos de las imágenes estéreo mientras que el cabeceo del vehículo durante la conducción puede originar una mala calibración, por un lado, e introducir ruido, por otro.

- Búsqueda de similitud con respecto a la secuencia del fondo.- Trata de localizar los objetos de interés, empleando técnicas de seguimiento (*tracking*) clásicas. La bondad de este método, depende de la técnica de mantenimiento del fondo, que resulta un desafío, debido a las restricciones de tiempo, luces ocasionales, movimiento de los árboles, sombras, etc. dando lugar a errores en la detección de objetos en movimiento o estáticos.

- Restricciones de forma y tamaño.- Generalmente son usadas en combinación con otros rasgos cuando por sí solas son insuficientes para diferenciar a las personas del fondo.

Para llevar a cabo la clasificación de las regiones de interés, se emplean rasgos específicos de un peatón. Estos rasgos también pueden aplicarse para la tarea de segmentación, pero generalmente, consumirían mucho tiempo. Los más usados son:

- Patrones o plantillas 2D basados en intensidad.- Permiten detectar peatones en distintas posturas Se trata de buscar regiones con mínimas medidas de distancia a la forma de los patrones. Se pueden usar técnicas basadas en bordes o contornos. A mayor número de patrones, mayor es el coste computacional. Además, resulta difícil definir una medida de correspondencia de la forma ideal.

- Búsqueda de rasgos característicos de los seres humanos.- El tono de la piel, la localización de los ojos, la cara o la cabeza, entre otros. El mayor inconveniente es que no siempre son visibles.

- Modelos geométricos 3D del cuerpo humano.- Pueden ser de ayuda, cuando múltiples modelos 2D son incapaces de detectar peatones en cualquier postura.

- Modelos basados en el movimiento humano.- Los peatones pueden ser identificados por un modelo que represente el característico modo de caminar humano. Su uso suele ser limitado – por ejemplo, aplicado a sistemas de seguimiento – , no siendo aconsejable para situaciones de conducción en el exterior. Además, la naturaleza ruidosa de la estimación del movimiento limita aún más su campo de acción.

Las aplicaciones de detección de peatones basadas en infrarrojos pueden heredar algunas de estos rasgos empleados en la detección basada en cámaras del espectro visible. De hecho, se han implementado aplicaciones basadas en cámaras estéreo de infrarrojos, bien basadas en la intensidad a nivel píxel o en la probabilidad de intensidades [ND02, SG00a]. Debido a que los peatones en las imágenes de infrarrojos contienen un mayor número de píxeles con intensidades altas que los objetos del fondo, son habituales los algoritmos basados en la búsqueda de simetrías o basados en el histograma.

Sin embargo, ciertas propiedades de las imágenes de infrarrojos, suponen un desafío para la detección de peatones y limitan la reutilización de las técnicas del espectro visible; Por un lado, debido a principios de la generación de la imagen y a la pobre resolución de las imágenes infrarrojas, es casi imposible extraer rasgos propios de un humano, como el color de la piel, ojos o cara. También resulta difícil detectar contornos bien definidos de los peatones, como por ejemplo, los extraídos en [GG02, BvS03] y no se pueden usar patrones 2D basados en contornos para detectar peatones en distintas posturas.

Además, al no existir uniformidad en la intensidad de la silueta humana, resulta inviable aplicar técnicas de correspondencia de patrones 2D uniformes o de segmentación basada en regiones (no existen razones para pensar que pueda existir correlación entre píxeles vecinos). Análogamente, rara vez se usa un modelo 3D geométrico de las partes del cuerpo humano, porque resulta arduo obtener tal cantidad de detalles descriptivos de un peatón, especialmente para situaciones reales de conducción, en las que las escenas cambian rápidamente. En segundo término, tampoco pueden aplicarse técnicas basadas en bordes, debido al bajo contraste y

falta de suficiente textura inherente a este tipo de imágenes y a los efectos ”fantasma” *ghosting effects*) que pueden aparecer con las altas temperaturas. Además, la detección de un menor número de puntos también hacen que la detección del movimiento sea menos fiable y tampoco puedan aplicarse modelos basados en el movimiento humano. Sin embargo, el dominio de infrarrojos posee rasgos diferenciadores con respecto al espectro visible y algunas de ellas aportan interesantes ventajas a la tarea de detección.

3.2. Geometría de un Sistema de Visión Estereoscópico

Aunque el modelo *pinhole* es conocido, se describe a continuación para fijar la notación que se usará en los apartados de calibración y rectificación (ver fig. 3.1).

3.2.1. Modelo de la cámara

El primer paso para desempeñar la calibración, es conocer el modelo de la cámara que se está usando. El modelo de descripción de la cámara más conocido es el modelo *pinhole*. Una cámara de *pinhole* se modela a través del centro óptico **C** y el plano de la imagen (o plano de la retina) $\Re$. Un punto **W** del mundo 3D, se proyecta en un punto **M** de la imagen 2D, como la intersección del plano $\Re$ con la línea que contiene a **W** y a **C**. La línea que contiene a **C** y que es ortogonal a $\Re$ se denomina eje óptico y su intersección con R es el punto principal. La distancia entre **C** y $\Re$ es la distancia focal.

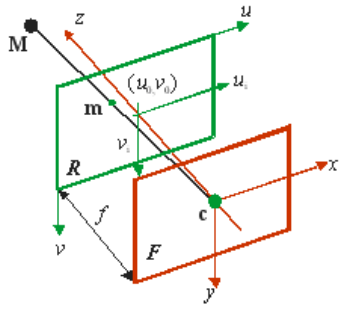


Figura 3.1: El modelo de proyección en perspectiva *pinhole*.

Sea $\mathbf{w} = \begin{bmatrix} x & y & z \end{bmatrix}^T$ las coordenadas de **W** en el sistema de referencia del mundo y $\mathbf{m} = \begin{bmatrix} u & v \end{bmatrix}^T$ las coordenadas de **M** en el plano de la imagen (en píxeles). El mapeo de las coordenadas 3D a las coordenadas 2D es la llamada proyección perspectiva, que se representa mediante una transformación lineal en coordenadas homogéneas. Sean $\tilde{m} = \begin{bmatrix} u & v & s \end{bmatrix}^T$ y $\tilde{w} = \begin{bmatrix} x & y & z & 1 \end{bmatrix}^T$ las coordenadas homogéneas de **m** y **M**, respectivamente; entonces, la transformación de la perspectiva viene dada por la matriz $\tilde{P}$:

$$\tilde{m} = \tilde{P}\tilde{w} \tag{3.1}$$

Por tanto, la cámara es modelada con la matriz de proyección perspectiva $\tilde{P}$, que puede descomponerse (aplicando la factorización QR) en el producto de otras dos matrices:

$$\tilde{m} = \tilde{P}\tilde{w} = AG\tilde{w} \tag{3.2}$$

donde la matriz A depende de los parámetros intrínsecos de la cámara y tiene la siguente forma:

$$A = \begin{bmatrix} \alpha_u & \gamma & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.3}$$

donde $\alpha_u = fk_u$, $\alpha_v = fk_v$, son las distancias focales medidos en píxeles horizontales y ver ticales, respectivamente (f es la distancia focal en milímetros, mientras que k_u y k_v son el número de píxeles por milímetro a lo largo del eje u y v), (u_0,v_0) son las coordenadas del punto principal, que viene dado por la intersección del eje óptico y el plano de la imagen, y γ es el factor de giro (*skew factor*) que modela ejes no-ortogonales u-v. Normalmente se considera que los ejes son ortogonales, siendo $\gamma = 0$. La posición y orientación de la cámara – los parámetros extrínsecos – están contenidos en la matriz de rotación **R** y el vector de translación **t**, representando la transformación rígida que convierte el sistema de coordenadas de la cámara en el sistema de coordenadas del mundo. (Representa la posición y orientación de la cámara con respecto al sistema de coordenadas del mundo)

$$G = [R|t] \tag{3.4}$$

Así, la matriz de proyección perspectiva P puede descomponerse en el producto de:

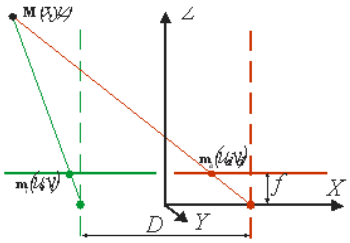
$$\tilde{P} = A[R|t] \tag{3.5}$$

Por lo que,

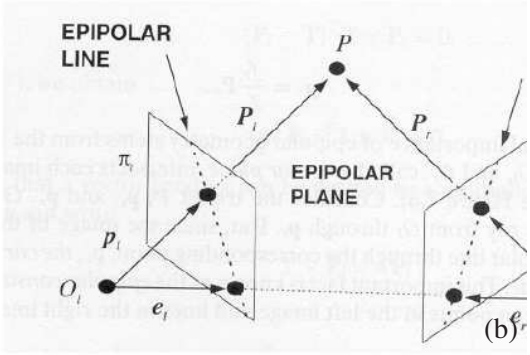
$$\tilde{m} = \begin{bmatrix} u \\ v \\ s \end{bmatrix} = \begin{bmatrix} \alpha_u & \gamma & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{00} & r_{01} & r_{02} & t_x \\ r_{10} & r_{11} & r_{12} & t_y \\ r_{20} & r_{21} & r_{22} & t_z \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = \tilde{P}\tilde{w} \tag{3.6}$$

3.2.2. Sistema estéreo con cámaras paralelas

Considérese el sistema estéreo de la Fig. 3.2-b, donde se han colocado dos cámaras exactamente iguales, separadas una distancia D (*baseline*), con ópticas de distancias focales idénticas, situadas de forma que sus ejes ópticos sean colineales. El punto del mundo 3D $\mathbf{w} = \begin{bmatrix} x & y & z \end{bmatrix}^T$ se proyecta en las dos imágenes en los píxeles $\mathbf{m_i} = \begin{bmatrix} u_i & v_i \end{bmatrix}^T$ y $\mathbf{m_d} = \begin{bmatrix} u_d & v_d \end{bmatrix}^T$



(a)



(b)

Figura 3.2: Distintas disposiciones de un sistema estéreo: (*a*) Caso "ideal" (geometría paralela); (*b*) Caso general (geometría epipolar).

En general, las coordenadas de estos dos puntos no serán las mismas, denominándose disparidad a la distancia o diferencia entre ambas. Según el modelo de óptica *pinhole* y tomando como origen de referencia el punto medio entre las cámaras, las ecuaciones para cada una de ellas son:

$$\begin{aligned} \frac{u_i}{f} &= \frac{x + D/2}{z} \\ \frac{u_d}{f} &= \frac{x - D/2}{z} \end{aligned} \tag{3.7}$$

Donde x y z son las incógnitas y el resto datos. Si se igualan las expresiones para la coordenada horizontal x:

$$x = z\frac{u_i}{f} - \frac{D}{2} = z\frac{u_d}{f} + \frac{D}{2} \tag{3.8}$$

$$z = f\frac{D}{u_i - u_d} = f\frac{D}{d} \tag{3.9}$$

Siendo d la disparidad, que en el caso particular de que ser las cámaras paralelas será siempre positiva. La coordenada y será:

$$y = \frac{v_i}{f}z = \frac{v_i}{f}\frac{Df}{d} = \frac{Dv_i}{d} \tag{3.10}$$

Si se compara con el resultado obtenido con la otra imagen:

$$\frac{v_d}{f} = \frac{y}{z} = \frac{Dv_i}{d}\frac{d}{Df} = \frac{v_i}{f} \Rightarrow v_i \equiv v_d \tag{3.11}$$

Por tanto, como $v_i \equiv v_d$, si se toma un punto de la cámara izquierda, su correspondiente en la imagen derecha estará a la misma altura. En el caso general en el que las cámaras no sean paralelas se sigue cumpliendo que dado un punto en una imagen, los posibles píxeles candidatos en la otra imagen se encuentran en una línea recta, denominada línea epipolar. Esto es así debido a la geometría de la visión estéreo, denominada geometría epipolar, que permite

restringir la búsqueda de puntos semejantes a lo largo de la línea epipolar correspondiente convirtiéndose en una búsqueda 1D horizontal.

3.2.3. El Problema de la Rectificación de Imágenes

Lo que se ha descrito hasta ahora, permite plantear uno de los problemas clave de la visión estéreo: el problema de la correspondencia. Como ya se ha comentado, si un objeto es observado desde dos puntos de vista diferentes, la posición de las proyecciones en cada imagen de un mismo punto de la escena 3D serán diferentes.

Es posible definir el problema de la correspondencia como el de la búsqueda de una función $d : (u, v) \rightarrow (d_u, d_v)$, tal que dado el punto $m = (u, v)$ de la imagen de referencia, el punto $m' = m + d(m) = (d_u, d_v)$ representa la proyección del mismo punto del espacio 3D sobre la segunda imagen. La función d es la antes mencionada disparidad y ha de ser conforme a la geometría epipolar.

En el caso de la geometría paralela descrito en el caso "ideal", $d_v = 0$ (ya que $v_i = v_d$) y por tanto, $m' = m + d(m) = (u + d, v)$. De manera que, la función de disparidad es un escalar que se calcula como la diferencia entre los valores de la coordenada horizontal de cada punto $m = (u, v)$ y su punto correspondiente $m' = (u', v)$. Es decir, $d = u' - u$.

Esta simplicidad en la formulación matemática se debe a que las líneas epipolares son paralelas y horizontales. Para que esto sea así, deben cumplirse una serie de premisas que definen la geometría paralela: que ambas cámaras sean exactamente iguales y que los ejes ópticos sean paralelos.

Sin embargo, en la práctica estas premisas no resultan nada fáciles de cumplir, como se puede confirmar tras los resultados obtenidos en la fig. 3.3 con el sistema estéreo construido en la Universidad Carlos III. En este punto cabe mencionar que, se han utilizado dos cámaras con ópticas del mismo modelo y se ha procurado que estén paralelas. Sin embargo, puede comprobarse siguiendo las líneas blancas horizontales que no lo son totalmente. Aunque pueda pensarse que las diferencias son muy pequeñas, en la fig. 3.3-c se tiene el resultado de obtener el mapa de disparidades a partir de las imágenes originales, en el que se observa que los resultados son erróneos. Dicho mapa se obtiene según el método clásico de utilización de la SAD (*Sum of Absolute Difference*) para resolver el problema de la correspondencia entre ambas imágenes. Hay que hacer notar que estos errores han sido debidos a que hemos buscado los puntos correspondientes entre las dos imágenes en las líneas epipolares que se suponían serían las líneas horizontales.

En el caso general hubiera pasado lo mismo, porque el error de fondo está en las tolerancias en la fabricación de las cámaras, las ópticas y la base mecánica que las sostiene. Por tanto, para corregir estos defectos, resulta necesario realizar un procesamiento a las imágenes estéreo conocido como rectificación.

La rectificación permite solucionar este problema, transformando la pareja de sistemas ópticos en otros dos que sean colineales y donde las líneas epipolares sean paralelas a uno de los ejes de las imágenes. En los siguientes apartados se va a ver en qué consiste y cómo puede realizarse en la práctica.

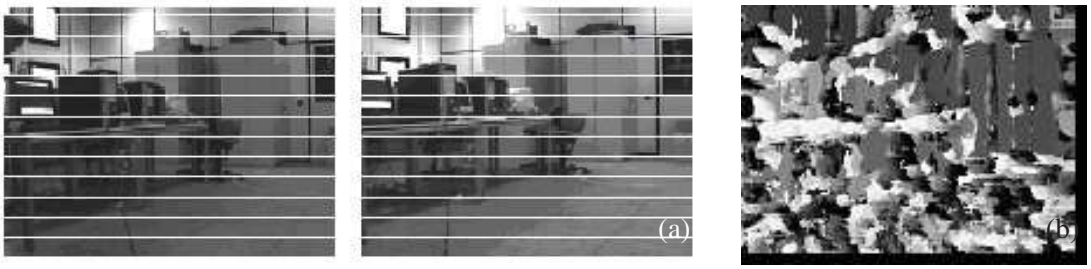


Figura 3.3: Necesidad de procesar las imágenes de un sistema estéreo: (*a*) Imágenes izquierda y derecha originales, respectivamente. Las líneas blancas horizontales reflejan que las cámaras no son totalmente paralelas; (*b*) Mapa de disparidades entre a y b.

3.3. Sistema de Visión del Espectro Visible

En este apartado se hace referencia al sistema de visión construido en la Universidad Carlos III. Se detallan aspectos técnicos y se explicitan los pasos seguidos para la calibración de dicho sistema, así como para la posterior rectificación de las imágenes obtenidas. Tanto la calibración como la rectificación, son procedimientos que forman parte de la fase inicial de puesta a punto del sistema de visión, que posteriormente será utilizado en la detección de peatones.

3.3.1. Descripción del sistema de Adquisición del Ivvl

El sistema consiste en dos cámaras con ópticas del mismo modelo que se han colocado en una estructura mecánica (ver fig.3.4) para tratar de que estén paralelas. Como ya se ha comentado en el apartado esto no es fácil debido a las tolerancias en la fabricación de las cámaras, las ópticas y a la base mecánica que las sostiene.

El tamaño de las imágenes capturadas por el sistema binocular IvvI es de 640x480 píxeles. Las cámaras se han fijado a una altura de 1180.0 milímetros del suelo y separadas una distancia (*baseline)* de 148.3 milímetros.

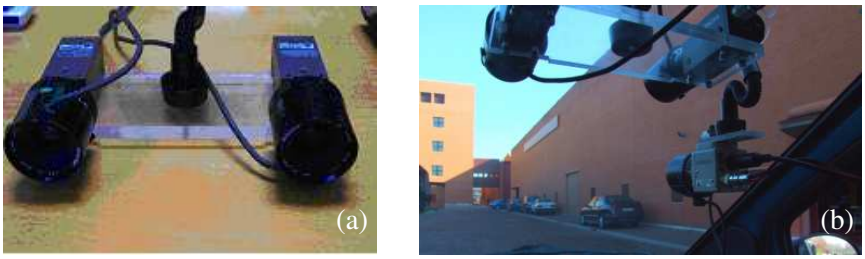


Figura 3.4: Imágenes del sistema de adquisición; (*a*) detalle del sistema estéreo durante la calibración y rectificación y, (*b*) una vez instalado en el vehículo experimental IvvI.

3.3.2. Calibración del sistema Ivvl

El método de rectificación implementado sigue el desarrollado por Fusiello [FTV00], que exige que el sistema estéreo esté calibrado, o dicho de otro modo, supone que las matrices de proyección perspectiva ($\tilde{P_{oi}} y \tilde{P_{od}}$) son conocidas. Se pueden consultar otras referencias [HG93, RBH95] para el caso de que no lo sean.

Por ello debe calibrarse el sistema estéreo y obtener los parámetros intrínsecos (distancia focal, centro de la imagen, distorsiones) y extrínsecos (matriz de rotación y vector de traslación). Con ese fin, se ha utilizado el Toolbox para Matlab desarrollado en el *California Institute of Technology, Caltech* [Bou00] que se está convirtiendo en el software estándar de calibración, aunque pueden encontrarse métodos alternativos [PD96]. Este Toolbox implementa, entre otros, el método desarrollado por Zhang [Zha99, Zha00].

El método requiere que la cámara observe un patrón plano desde unas pocas orientaciones (como mínimo dos). Tanto la cámara como el patrón plano pueden mover se libremente. No es necesario conocer el movimiento. Además, se modela tanto la distorsión radial como la tangencial de la lente.

El patrón de calibración plano empleado se muestra en la figura 3.5 así como las imágenes tomadas desde distintas posiciones de una de las cámaras (ver fig. 3.6), cuyas posiciones no tienen por qué ser conocidas. Se calibran tanto los parámetros intrínsecos como los extrínsecos relativos a las dos cámaras, gracias a la homografía (coincidencia gráfica) entre el modelo plano y su imagen. Primero se deja de lado la distorsión mientras se calcula mediante una solución analítica el resto de parámetros, para luego utilizar un método de optimización no lineal para obtenerla. Se necesita un patrón similar a un tablero de ajedrez donde el usuario señala las cuatro esquinas y el sistema de forma automática determina el resto de ellas. Las coordenadas de estas de una serie de imágenes son la entrada al algoritmo.

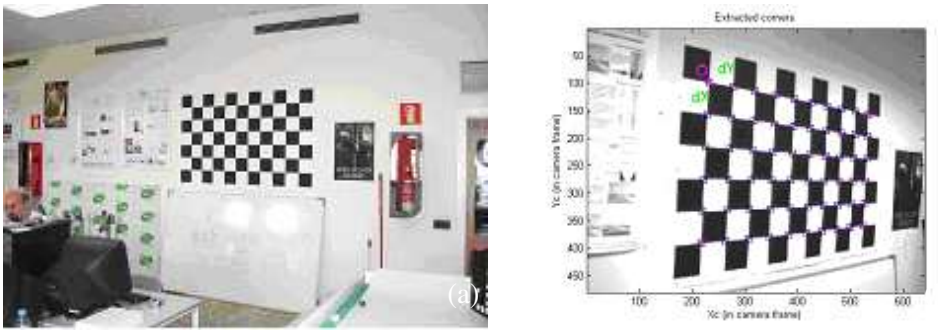


Figura 3.5: Calibración del sistema estéreo; (*a*)imagen del patrón de calibración original y (*b*) una vez analizado mediante el toolbox de calibración [Bou00].

Respecto a la distorsión (ver fig. 3.7), tiene en cuenta la radial y también la tangencial [Bro71]. Así:

$$\begin{aligned} x_r &= x_i + x_i\left(k_1 r^2 + k_2 r^4\right) + 2p_1 x_i y_i + p_2\left(r^2 + 2x_i^2\right) \\ y_r &= y_i + y_i\left(k_1 r^2 + k_2 r^4\right) + 2p_2 x_i y_i + p_1\left(r^2 + 2y_i^2\right) \end{aligned} \tag{3.12}$$

Siendo (x_r, y_r) las coordenadas reales de los píxeles, (x_i, y_i) las coordenadas ideales, k_1 y

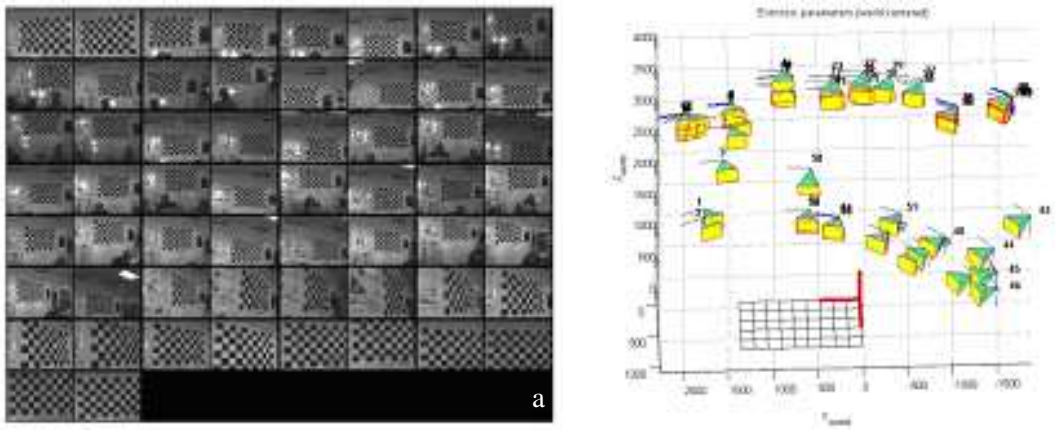


Figura 3.6: (*a*) Conjunto de imágenes tomadas con la cámara izquierda y (*b*) Distintas posiciones de la cámara respecto al patrón proporcionadas por el toolbox [Bou00].

k_2 los coeficientes de la distorsión radial,p_1 y p_2 los coeficientes de la distorsión tangencial y r la distancia al centro de la imagen.

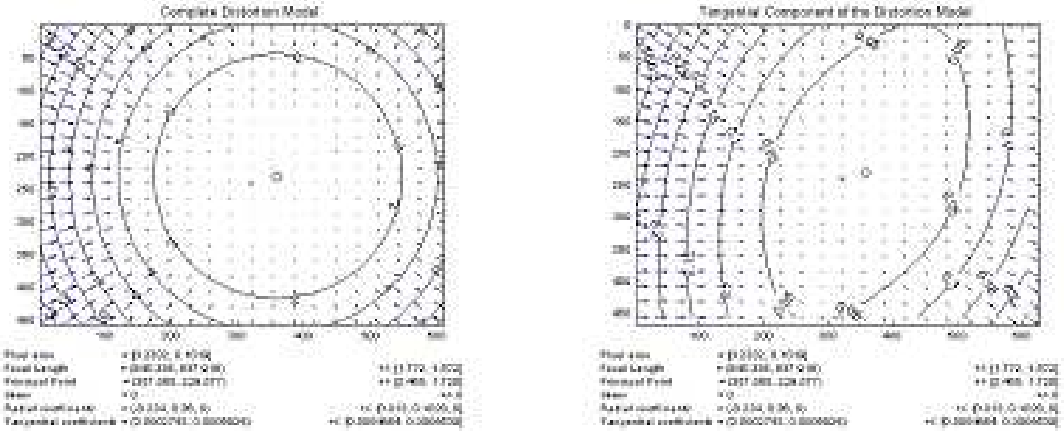


Figura 3.7: Distorsión (*a*) radial y, (*b*) tangencial.

El sistema estéreo del vehículo IVVI se colocó en 58 posiciones distintas considerando a objetos muy cercanos y lejanos (ver fig. 3.6-a). Los resultados obtenidos de la calibración aparecen resumidos en la imagen 3.9 y los errores cometidos se reflejan de un modo gráfico en la figura 3.8.

3.3.3. Rectificación de imágenes estéreo

Cualquier par de imágenes pueden ser transformadas de manera que sus líneas epipolares sean paralelas y horizontales en cada imagen. Este procedimiento se conoce con el nombre de rectificación.

Lo que se pretende con la rectificación, es definir dos nuevas matrices de proyección perspectiva ($\tilde{P_{ni}}y\tilde{P_{nd}}$), obtenidas tras rotar las cámaras originales alrededor de sus centros ópticos, hasta que sus plano focales sean coplanares y por tanto, contengan la baseline. Esto asegura que los epipolos estén en el infinito, y en tal caso, las líneas epipolares son paralelas. Para obtener líneas epipolares horizontales, la baseline debe ser paralela al nuevo eje X (eje horizontal) de ambas cámaras. Por añadidura, para conseguir una correcta rectificación, los puntos conjugados deben tener la misma coordenada ver tical. Esto se consigue, exigiendo que las nuevas cámaras tengan los mismos parámetros intrínsecos. Nótese, que si la distancia focal es la misma, los planos de la imagen serán también coplanares (ver fig. 3.10).

A continuación se describe el método de rectificación propuesto por Fusiello [FTV00], ya que sienta las bases del método de rectificación empleado en esta tesis.

3.3.3.1. Rectificación de las matrices de proyección

Resumiendo lo dicho en el punto anterior, las posiciones de las nuevas matrices de proyección perspectiva (es decir, de los centros ópticos) son las mismas que las de las cámaras originales, pero la nueva orientación (que es la misma para ambas cámaras), difiere de las originales en función de las rotaciones que hayan sido necesarias. Por otro lado, los parámetros intrínsecos son idénticos en ambas cámaras. Entonces, las nuevas matrices de proyección sólo van a diferenciarse en (la posición) su centro óptico, y pueden considerarse como una sola cámara que es trasladada a lo largo del eje X de su sistema de referencia. Así, interesa definir la posición del centro óptico, que puede obtenerse a partir de la ecuación 3.6, que se puede reescribir como:

$$\tilde{m} = \begin{bmatrix} u \\ v \\ s \end{bmatrix} = \tilde{P}\tilde{w} = \begin{bmatrix} q_1^T | q_{14} \\ q_2^T | q_{24} \\ q_3^T | q_{34} \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = [Q|\tilde{q}] \begin{bmatrix} W \\ 1 \end{bmatrix} \tag{3.13}$$

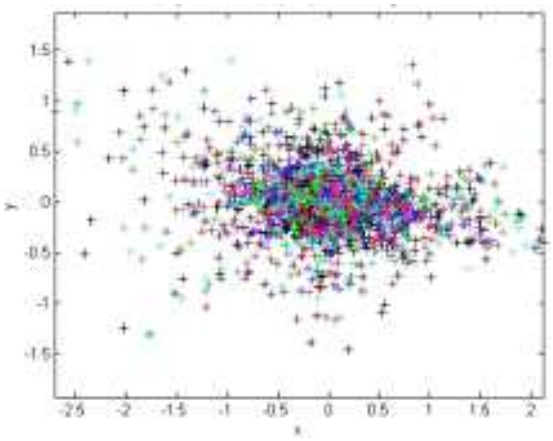


Figura 3.8: Errores cometidos en la calibración del sistema estéreo.

Tabla I. Valores de la calibración

Cámara	k_1	K_2	p_1	p_2	u_0	v_0	f_u	f_v
Izquierda	-0.24386	0.19161	0.00254	-0.00761	308.49	237.40	815.03	806.27
Derecha	-0.25275	0.32041	0.00470	-0.00905	301.74	226.90	815.40	807.86

$$\mathbf{R} = \begin{pmatrix} 1.0000 & -0.0047 & 0.0055 \\ 0.0047 & 1.0000 & -0.0025 \\ -0.0055 & 0.0025 & 1.0000 \end{pmatrix} \quad \mathbf{t} = \begin{pmatrix} -147.86 & 0.96 & 11.3 \end{pmatrix}^T$$

Figura 3.9: Resultado de la calibración del sistema estéreo, utilizando [Bou00] .

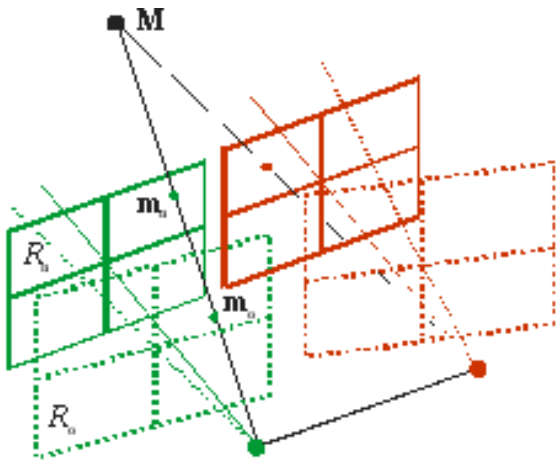


Figura 3.10: El proceso de la rectificación: en líneas discontinuos se muestra el sistema sin rectificar y en continuas, el resultado de la rectificación.

En coordenadas cartesianas, la proyección de la ecuación 3.1 se escribiría de la siguiente forma:

$$\begin{aligned} u &= \frac{q_1^T w + q_{14}}{q_3^T w + q_{34}} \\ v &= \frac{q_2^T w + q_{24}}{q_3^T w + q_{34}} \end{aligned} \tag{3.14}$$

Por otro lado, para la posterior rectificación, es necesario obtener la posición del centro óptico **C**. De la matriz 3.6 puede obtenerse esta información. El plano focal es, el plano paralelo al plano de la imagen, que contiene al centro óptico **C**. Ese plano corresponde a la fila $q_3^T w + q_{34} = 0$ de la matriz de proyección (ver fig. 3.11). Los otros dos planos, $q_1^T w + q_{14} = 0$ y $q_2^T w + q_{24} = 0$ intersectan al plano de la imagen formando el eje ver tical (u nulo) y horizontal (v nulo) respectivamente.

El centro óptico **C**, será el punto en el que intersecten los tres planos, luego debe cumplir:

$$\tilde{P} \begin{bmatrix} c \\ 1 \end{bmatrix} = 0 \Rightarrow (Q|\tilde{q}) \begin{bmatrix} c \\ 1 \end{bmatrix} = 0 \Rightarrow Qc + \tilde{q} = 0 \Rightarrow \begin{cases} c = -Q^{-1}\tilde{q} \\ \tilde{q} = -Qc \end{cases} \tag{3.15}$$

Por lo que las coordenadas c del centro óptico C vienen dadas por la ecuación

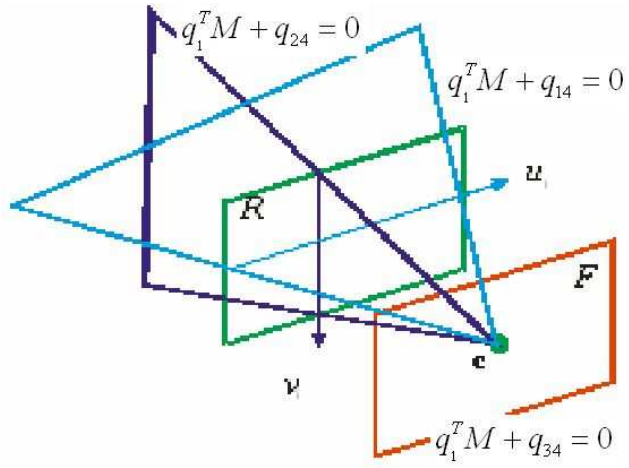


Figura 3.11: Interpretación de las filas de la matriz de proyección de perspectiva.

$$c = -Q^{-1}\tilde{q} \tag{3.16}$$

Y, además de eso, P se puede reescribir del siguiente modo:

$$\tilde{P} = [Q| - Qc] \tag{3.17}$$

Sabiendo además que, el rayo óptico asociado con un punto m de la imagen, es la línea que une ese punto con el centro óptico, es decir, es el conjunto de puntos 3D tales que $w : \tilde{m} = \tilde{P}\tilde{w}$.

$$\tilde{m} = \tilde{P}\tilde{w} = [Q|\tilde{q}]\begin{bmatrix} w \\ 1 \end{bmatrix} = [Q| - Qc]\begin{bmatrix} w \\ 1 \end{bmatrix} = Q(w - c) \tag{3.18}$$

La ecuación del rayo óptico se puede formular en forma paramétrica:

$$w = c + \lambda Q^{-1}\tilde{m}, \quad \lambda \in \mathfrak{R} \tag{3.19}$$

Las nuevas matrices de proyección se pueden formular en términos de su factorización, de las ecuaciones 3.5 y 3.17:

$$\begin{aligned} \tilde{P}_{ni} &= A\,[R| - Rc_i] \\ \tilde{P}_{nd} &= A\,[R| - Rc_d] \end{aligned} \tag{3.20}$$

3.3.3.2. La transformación de rectificación

Para rectificar, por ejemplo, la imagen de la izquierda, se necesita calcular la transformación que mapea el plano de la imagen antiguo $\tilde{P}_{oi} = [Q_{oi}| - \tilde{q}_{oi}]$ en el plano de la imagen nuevo $\tilde{P}_{ni} = [Q_{ni}| - \tilde{q}_{ni}]$. Esta transformación es la colinealidad descrita mediante la siguiente matriz:

$$T_i = \begin{bmatrix} Q_{ni} & Q_{oi}^{-1} \end{bmatrix} \tag{3.21}$$

Para la imagen de la derecha se aplica una transformación análoga. A continuación se describe el modo de calcular esta matriz. Tal y como ya se ha expresado antes, para cualquier punto w del mundo 3D, su proyección m en el plano de la imagen se puede definir con una ecuación. Ahora, el punto w se proyecta sobre dos planos de la imagen distintos ($\tilde{m}_{oi}$ y $\tilde{m}_{ni}$):

$$\begin{aligned} \tilde{m}_{oi} &\cong \tilde{P}_{oi}\tilde{w} \\ \tilde{m}_{ni} &\cong \tilde{P}_{ni}\tilde{w} \end{aligned} \tag{3.22}$$

Por otro lado, según la ecuación $w = c + \lambda Q^{-1}\tilde{m}$ (ver 3.19), las ecuaciones de los rayos ópticos referidos a cada plano de la imagen son las siguientes:

$$\begin{aligned} w &= c_1 + \lambda_0 Q_{oi}^{-1}\tilde{m}_{oi}, \quad \lambda_0 \in \Re \\ w &= c_1 + \lambda_0 Q_{ni}^{-1}\tilde{m}_{ni}, \quad \lambda_n \in \Re \end{aligned} \tag{3.23}$$

Dado que la rectificación no mueve el centro óptico de posición, sino de orientación, de las ecuaciones anteriores se puede despejar la ecuación que relaciona el plano de la imagen antiguo con el nuevo:

$$\tilde{m}_{ni} = \lambda Q_{ni} Q_{oi}^{-1}\tilde{m}_{oi}, \quad \lambda_n \in \Re \tag{3.24}$$

Si sustituimos $T_i = Q_{ni}Q_{oi}^{-1}$, obtenemos la expresión de la transformación de rectificación antes citada. De manera que, el proceso de rectificación consiste en aplicar la transformación T_i a la imagen original izquierda, obteniendo como resultado la imagen rectificada.

3.3.3.3. Descripción del proceso de rectificación aplicado

Una vez establecidos los conceptos necesarios para comprender el proceso de la rectificación, a continuación se explica el modo en el que se ha llevado a cabo dicho proceso. Después de obtener los parámetros intrínsecos y extrínsecos de cada cámara como resultado de la calibración, el método de rectificación desarrollado exige conocer las matrices de rotación y la traslación entre el sistema de coordenadas *w* del mundo y el sistema de coordenadas *m* de cada cámara:

$$\begin{aligned} m_i &= R_i w + t_i \\ m_d &= R_d w + t_d \end{aligned} \tag{3.25}$$

Si las imágenes han sido tomadas por un sistema estéreo, como es el caso, es posible obtener también la relación entre las dos cámaras según;

$$m_d = R m_i + t \tag{3.26}$$

siendo **R** una matriz de rotación y **t** el vector de traslación, según los sistema de referencia de la figura 3.12).

El Toolbox para Matlab desarrollado en Caltech [Bou00] permite obtener esta relación.

Por tanto, de las ecuaciones 3.26 y 3.25 se ver ifica que:

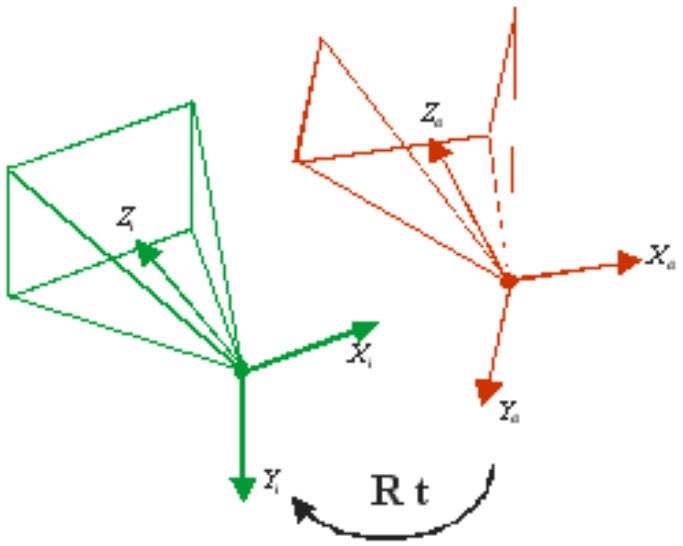


Figura 3.12: Relación entre los sistemas de coordenadas de las dos cámaras.

$$\begin{aligned} R_i &= I, t_i = 0 \\ R_d &= R, t_d = t \end{aligned} \tag{3.27}$$

Y de las ecuaciones 3.6, 3.13 y 3.27, se pueden reescribir las matrices de proyección antiguas como:

$$\begin{aligned} \tilde{P}_{oi} &= [Q_{oi}|\tilde{q}_{oi}] = A_{oi}\,[R_{oi}|t_{oi}] = A_{oi}\,[I|0] = [A_{oi}|0] \\ \tilde{P}_{od} &= [Q_{od}|\tilde{q}_{od}] = A_{od}\,[R_{od}|t_{od}] = A_{od}\,[R|t] = [A_{od}R|A_{od}t] \end{aligned} \tag{3.28}$$

Como resultado de la calibración y de la posterior rectificación, se va a definir dos nuevas matrices de perspectiva que cumplan las condiciones descritas en el apartado de rectificación de imágenes.

Se tiene entonces el plano de la imagen antiguo R_o y el nuevo R_n, que se obtendrá a través de la rectificación, por lo que a partir de las antiguas P_{oi} y P_{od} se necesita encontrar las nuevas P_{ni} y P_{nd} , que serán, tal como se vio en la ecuación 3.20, de la forma:

$$\begin{aligned} \tilde{P}_{ni} &= A_{ni}\,[R_{ni}| - R_{ni}c_i] \\ \tilde{P}_{nd} &= A_{nd}\,[R_{nd}| - R_{nd}c_d] \end{aligned} \tag{3.29}$$

La matriz de parámetros intrínsecos, $A_{ni} = A_{nd} = A_n$, tiene que ser la misma para ambas cámaras ya que deben reflejar que son cámaras idénticas. Aunque teóricamente sus valores son arbitrarios, unos deformarán más la imagen que otros y pueden dar lugar a pequeños errores. En este caso se han tomado la media de los valores obtenidos en la calibración y se ha determinado que la distancia focal para el eje ver tical y horizontal es la misma:

$$
\begin{aligned}
f_u = f_v &= \frac{f_{ui} + f_{vi} + f_{ud} + f_{vd}}{4} \\
u_o &= \frac{u_{oi} + u_{od}}{2} \\
v_o &= \frac{v_{oi} + v_{od}}{2}
\end{aligned}
\tag{3.30}
$$

Las matrices de rotación **R**, que describe cómo están colocadas las cámaras deben ser idénticas, $R_{ni} = R_{nd} = R_n$, ya que esto indica que ambos ejes ópticos están en la misma dirección. Esta matriz se va a especificar mediante vectores fila, que se corresponden con los ejes X, Y y Z respectivamente, del sistema de referencia de la cámara expresados en coordenadas del mundo:

$$
R_n = \begin{bmatrix} r_1^T \\ r_2^T \\ r_3^T \end{bmatrix}
\tag{3.31}
$$

Se pueden obtener las expresiones de los nuevos ejes, de la siguiente manera:

1.El nuevo eje X será paralelo a la baseline:

$$
r_1 = \frac{c_d - c_i}{|(c_d - c_i)|}
\tag{3.32}
$$

2.El nuevo eje Y debe ser obligatoriamente ortogonal al X y a k, que es un vector unitario arbitrario. Se toma aquel igual al vector unidad del antiguo eje Z de la imagen izquierda, siendo así el nuevo eje Y ortogonal al nuevo eje X y al antiguo eje Z:

$$
r_2 = k \wedge r_1
\tag{3.33}
$$

3.El nuevo eje Z debe ser obligatoriamente ortogonal al XY :

$$
r_3 = r_1 \wedge r_2
\tag{3.34}
$$

Finalmente, al estar los ejes ópticos en la misma dirección, el eje horizontal será aquel que contenga a los dos centros ópticos c_i y c_d, que como deben ser los mismos que en el sistema antiguo se obtienen a través de las ecuaciones 3.15 y 3.29:

$$
\begin{aligned}
c_i &= -Q_{oi}^{-1}\tilde{q}_{oi} = 0 \\
c_d &= -Q_{od}^{-1}\tilde{q}_{od} = -R^{-1}A_{od}^{-1}A_{od}t = -R^{-1}t
\end{aligned}
\tag{3.35}
$$

Sustituyendo todos estos datos, las nuevas matrices de proyección de perspectiva quedan de la siguiente forma:

$$
\begin{aligned}
\tilde{P}_{ni} &= A_n\,[R_n| - R_n c_i] = A_n\,[R_n|0] \\
\tilde{P}_{nd} &= A_n\left[R_n| - R_n^{-1}t\right]
\end{aligned}
\tag{3.36}
$$

De manera que, una vez obtenidas las matrices de perspectiva antiguas 3.28 y las nuevas 3.36 se define la transformación de rectificación a aplicar a cada imagen, que ya ha sido descrita en la ecuación 3.21.

$$\begin{aligned} T_i &= Q_{ni}Q_{oi}^{-1} = A_n R_n A_{oi}^{-1} \\ T_d &= Q_{nd}Q_{od}^{-1} = A_n R_n R^{-1} A_{od}^{-1} \end{aligned} \tag{3.37}$$

3.3.4. Resultados experimentales

Para ver la importancia de la rectificación y los resultados, se puede comparar las imágenes contenidas en la figura 3.13. Las dos primeras imágenes no están rectificadas y se aprecia que no son totalmente paralelas. El correspondiente mapa de disparidades que se obtendría a partir de esas imágenes se muestra en la figura 3.13-e. Por el contrario, si la imágenes son rectificadas (ver fig. 3.13-c y -d), la mejora en el mapa de disparidad obtenido es notable (ver fig. 3.13-f).

Cabe destacar que, si bien las imágenes capturadas por el sistema estéreo son de 640x480 píxeles, el procesamiento de las mismas se realiza una vez que las imágenes han sido reducidas a 320x240 píxeles. El motivo reside en la necesidad de acelerar los cálculos lo máximo posible, para que el algoritmo final pueda ser utilizado en un vehículo inteligente.

3.4. Sistema de Visión del Espectro Infrarrojo. Sistema TETRAVISION

El sistema de infrarrojo empleado en esta tesis, ha sido construido en la Universidad de Parma y forma parte de un sistema más complejo llamado Tetravision. Recibe este nombre porque está compuesto de cuatro cámaras, organizadas en dos sistemas estéreo que, además, trabajan en dos dominios distintos: el infrarrojo y el visible.

El algoritmo de reconocimiento de peatones que se ha implementado, sólo emplea imágenes tomadas por las cámaras de infrarrojos lejano. A pesar de que dicho sistema se ha utilizado ya calibrado y rectificado, se considera oportuno explicar brevemente la calibración y rectificación del sistema Tetravision, para tener un conocimiento del sistema en conjunto.

3.4.1. Descripción del Sistema de Adquisición Tetravision

En este apartado se mencionan los aspectos más importantes del sistema de adquisición de vídeo. A fin de asegurar que el sistema de visión se mueva lo menos posible, se ha construido una estructura metálica que se coloca sobre el techo del vehículo. En esta estructura se han montado cuatro soportes para cámaras con 3-ejes motorizados; dos son para las cámaras de infrarrojos lejano y las otras dos, para las del espectro visible. Se ha puesto especial atención para evitar que las cámaras se muevan debido a vibraciones, incluso en el caso de soportar grandes pesos, como es el caso de las cámaras de infrarrojo lejano utilizadas (ver fig.3.14). La tabla 3.15 refleja los parámetros principales de ambos sistemas.

Cada cámara de infrarrojo lejano está conectada a un sistema basado en bt878 de adquisición de imágenes dedicado, mientras que las imágenes del dominio visible son adquiridas a través de un bus IEE1394. Ni las cámaras de infrarrojos ni las del espectro visible disponen de un sistema de sincronización externo; por tanto, se han experimentado pequeñas diferencias en los tiempos de adquisición de las cuatro cámaras.

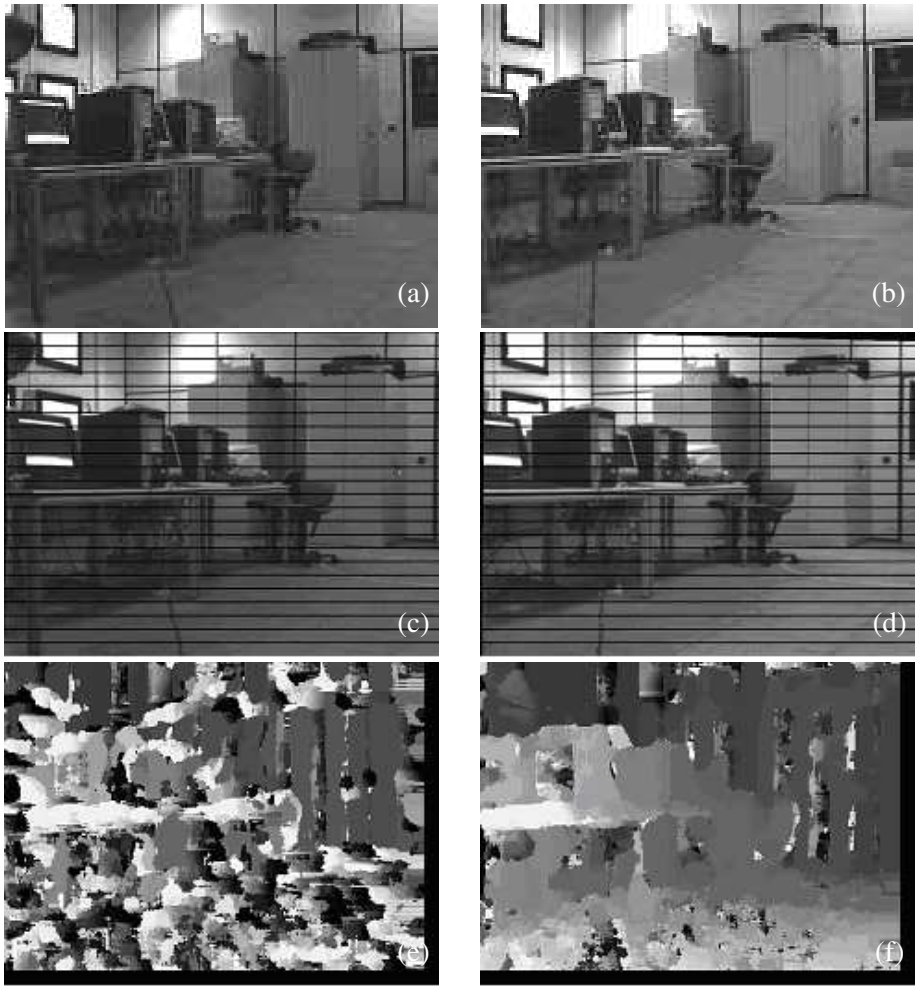


Figura 3.13: Comparación de resultados entre imágenes rectificadas y sin rectificar; (*a*) y (*b*) Imágenes izquierda y derecha sin rectificar, (*c*) y (*d*) Imágenes izquierda y derecha rectificadas, (*e*) Mapa de disparidades entre a y b (*f*) Mapa de disparidades entre c y d.

Figura 3.14: Sistema Tetravision. Las 4 cámaras utilizadas para la adquisición de imágeners se instalan en el techo del vehículo [BBF+07a].

Characteristic	FIR	Visible
Wavelength	7-14 μm	0.4-0.7 μm
Sensor type	Uncooled FPA	CCD
Sensor size	320×240 pixel	640×480 pixel
Image depth	8 bit	8 bit
Horizontal FOV	0.1535 rad	0.2216 rad
Vertical FOV	0.1182 rad	0.1680 rad
Baseline	0.500 m	1.000 m
Standard	NTSC	IIDC

Figura 3.15: Tabla que muestra los parámetros principales del sistema Tetravision [BBF+07a].

Por otro lado, la resolución de ambos sistemas es distinto. Las cámaras del dominio visible capturan imágenes de 640x480 píxeles, mientras que las del dominio infrarrojo son de 320x240 píxeles. Este hecho añade complejidad a la hora de hacer corresponder las regiones de interés halladas en cada uno de los dominios. La figura 3.16 muestra un ejemplo de un escenario típico tomado por el sistema de infrarrojo y del dominio visible, donde se puede apreciar esta diferencia.

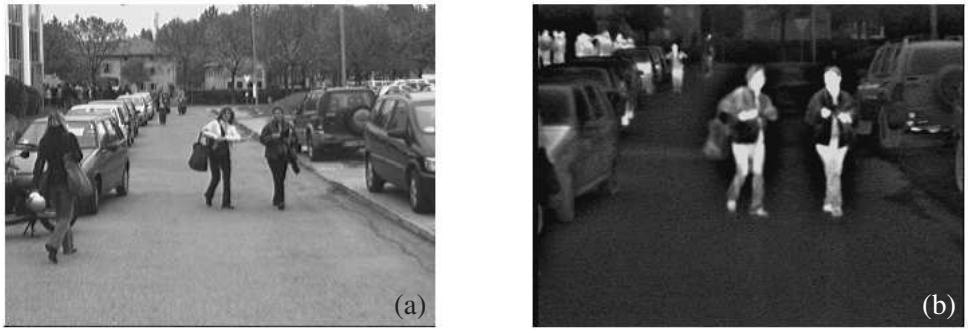


Figura 3.16: Ejemplo de un escenario típico tomado por el sistema Tetravision [BBF+07b]; (*a*) imágen obtenida a partir del sistema estéreo visible e (*b*) infrarrojo.

3.4.2. Calibración de las Cámaras del Sistema Tetravision

El procedimiento llevado a cabo para la calibración de las imágenes y para la orientación de las cámaras se compone de dos fases. Durante la primera fase, la calibración se realiza de un modo manual, operando directamente sobre el hardware. El vehículo se posiciona en un punto preciso de una rejilla de calibración (ver fig. 3.17-a), que consiste en un trozo de carretera plano, donde se han situado un conjunto de marcas a distancias conocidas, hasta cubrir una distancia máxima de 40 metros. Cada par de cámaras (del dominio visible e infrarrojo),

enfocan en mayor o menor medida la misma escena. Debido a que cada pareja emplea distintas ópticas, para poder realizar una correspondencia de coordenadas entre ambos dominios, hay que redimensionar las imágenes capturadas. Para cada cámara, el *roll* debe ser nulo, de tal manera que, se puede calcular la disparidad a partir de objetos que estén en una misma línea horizontal en ambas imágenes. Una vez obtenida la disparidad, se ajusta el *yaw*, que debe ser tal que permita encuadrar el mismo objeto con ambas cámaras. Finalmente, el *pitch* se ajusta en función de la distancia máxima y mínima que se quiere considerar.

Debido a que el hardware no es capaz de obtener la precisión requerida, es necesaria una segunda fase, que se lleva a cabo sobre imágenes filmadas. Una herramienta software desarrollada en el Vislab para el sistema GOLD (ver fig. 3.17-b) [BBG+03], permite alinear las imágenes, ayudando a refinar las posiciones de las cámaras (el *yaw, pitch,roll*). Por otro lado, el mismo interfaz muestra histogramas del color y luminancia de las imágenes adquiridas, permitiendo hacer un ajuste de las aperturas y ganancias de las cámaras. Para saber qué punto del mundo real le corresponde a cada píxel en la imagen, se basan en el conocimiento de la posición exacta de las cámaras, que obtienen respecto a un punto fijo (que es el origen de la rejilla de calibración).

Figura 3.17: Sistema de calibración GOLD [BBG+03] (*a*) Detalle de la rejilla de calibración y, (*b*) Ordenador que ejecuta la herramienta software para refinar la calibración.

3.5. Conclusiones

A partir de ahora, se va a diferenciar el trabajo realizado en dos partes. Primero, se va a detallar la investigación llevada a cabo usando el sistema de visión construido en la Universidad Carlos III. Después [en el capítulo], se expondrá el algoritmo desarrollado usando el sistema Tetravision de Parma. Finalmente, se pondrán en conjunto los resultados obtenidos por ambas partes.

Capítulo 4

Detección de Peatones en el Dominio Visible

Este capítulo se centra en la detección de peatones en el espectro visible. La metodología propuesta combina el uso de varios enfoques, divididos en las etapas: la primera, orientada a la detección de obstáculos, proporciona una lista de peatones potenciales, que la posterior fase de reconocimiento debe filtrar mediante un enfoque basado en la forma. Por último, se extraen las siluetas de las regiones clasificadas como peatones. Esta última etapa permitirá en un futuro el análisis de la posición de las piernas, tal y como se ha realizado en el dominio infrarrojo.

El detector de obstáculos sienta sus bases en características simétricas y morfológicas de los seres humanos. Sin embargo, debido a que estos rasgos no son exclusivos de las personas, es necesario el uso de otras técnicas que ayuden a eliminar falsos positivos. Pero se asegura que aunque lo detectado no sea un peatón, sigue siendo un obstáculo.

Por ello, la clasificación se realiza en base a la forma humana, mediante una técnica basada en el aprendizaje supervisado: el análisis del componente principal (PCA), que aplicado a los bordes del contorno en lugar de considerar las intensidades, permite ser usado en entornos donde la iluminación no puede ser controlada. Finalmente, la extracción de la silueta también se realiza mediante una técnica basada en la forma. Los modelos deformables, y en concreto, los contornos activos se deforman para tratar de adaptarse a la forma del peatón. A continuación se expone cada uno de los módulos que integran en sistema de detección final, así como los resultados experimentales obtenidos.

4.1. Descripción del Sistema Visible

El sistema de protección a peatones del dominio visible se ha construido sobre dos bases: la visión estereoscópica y la apariencia humana. Se han desarrollado varios módulos aprovechando los puntos fuertes de cada una de ellas. La detección de regiones de interés se realiza empleando técnicas estéreo no-denso, la clasificación se basa en la apariencia humana y la

verificación final combina el uso de técnicas estéreo denso con el análisis basado en la forma. La figura fig. 4.1 muestra el flujo del algoritmo completo.

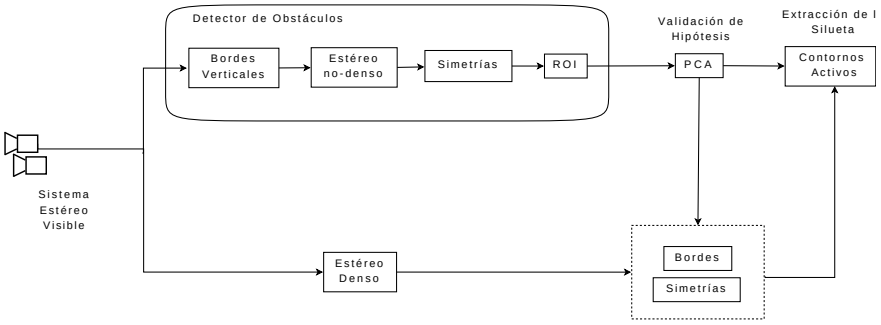


Figura 4.1: Diagrama de flujo del sistema global.

1. **Detección de obstáculos:**

   Para obtener una lista de regiones en la imagen conteniendo a peatones potenciales, se explotan los rasgos propios de los humanos: en concreto, los bordes verticales y la simetría vertical que presenta su silueta.

   - **Detección de bordes verticales:**

     Este enfoque se basa en el hecho que el contorno de un peatón presenta más bordes verticales que horizontales. El uso de cámaras del espectro visible, permite capturar mayor nivel de detalles que las cámaras de infrarrojos. Esto implica que las imágenes contendrán más bordes, pero también mayor cantidad de ruido, siendo necesario realizar un preprocesamiento previo a la extracción de bordes. Para ello se ha aplicado una amplitud de escala a las imágenes filmadas.

   - **Cálculo del mapa de disparidad no-denso:**

     Las regiones de interés se generan a partir del mapa de disparidad. Como paso previo, se rectifican las imágenes de la cámara izquierda y derecha. Se ha implementado un algoritmo estéreo no-denso, basado en la correspondencia de los bordes verticales extraídos en la etapa precedente.

   - **Simetrías:**

     El mapa de disparidad no-denso se filtra en base a medidas de distancia. Así se obtienen un conjunto de imágenes, cada una de ellas conteniendo únicamente aquellos píxeles que se encuentren en un rango de profundidad. Después, se realiza la búsqueda de simetrías verticales en cada una de esas imágenes. Se recorre toda la imagen, pero el número de puntos que se deben considerar son los pertenecientes a los bordes verticales, resultando rápido. El tamaño de la ventana de búsqueda queda establecido en función del rango de disparidades contenido en cada imagen. De este modo, se obliga a que la búsqueda de simetrías cumpla unas restricciones en cuanto al tamaño estimado del peatón.

- **Generación de Regiones de Interés:**

  A partir de la búsqueda de simetrías se determinan las regiones de interés. Para tomar esta decisión, sólo se consideran aquellas ventanas que cumplen las condiciones geométricas antes citadas. Pero, debido a que en el entorno puede haber otros objetos con una simetría vertical fuerte, es necesario filtrar las regiones candidatas.

  Se han impuesto restricciones en cuanto al ancho mínimo que deben tener los bordes considerados para obtener el eje de simetría. De este modo se eliminan objetos como postes, farolas o árboles, que son elementos habituales en entornos urbanos y que ocasionan multitud de falsos positivos. Además, se exige que el número de votos de cada ventana esté por encima de un umbral, que varía en función del tamaño de la ventana en evaluación. Así, para la selección del mejor candidato se considera la relación existente entre, el número de puntos de borde simétricos y el tamaño de la ROI que los contiene.

2. **Validación de hipótesis basado en PCA:**

   Las hipótesis que no han sido eliminadas, son evaluadas mediante la técnica del análisis del componente principal (PCA). La finalidad de esta fase es realizar un reconocimiento basado en la forma humana. Debido a que el PCA es muy sensible a los cambios de intensidad, se ha optado por emplear imágenes de bordes y distancias, en lugar de las imágenes de peatones sin procesar, ya que son más robustas ante dichos cambios. Una vez que el algoritmo ha aprendido, es capaz de clasificar el contenido de las ROIs como peatón o no peatón.

3. **Detección de la forma basado en *Snakes*:**

   Finalmente, los contornos activos o *snakes* se emplean como método de segmentación de peatones. Uno de los mayores inconvenientes de este método es la inicialización, que se ha resuelto al realizar la siembra dentro de las ROIs proporcionadas después del análisis del PCA. Sólo se consideran las que han sido clasificadas como peatones, ya que únicamente interesa detectar la forma humana, no la del resto de los obstáculos detectados. Como resultado, el *snake* extrae el mejor contorno posible.

- **Cálculo del mapa de disparidad denso:**

  La lista de ROIs que sobreviven al análisis del PCA, son analizados en mayor profundidad durante esta fase. El objetivo es ayudar a filtrar las falsas detecciones. Para ello se utiliza la técnica de contornos activos que segmentan la forma del objeto encerrado en cada ROI.

  Este análisis de la forma se efectúa en un rango de disparidad concreto, establecido en función de la distancia a la que esté cada ROI o peatón potencial. De este modo, se consiguen eliminar los elementos del fondo, que podrían confundir al contorno activo, generando una segmentación errónea. Para la obtención del mapa de disparidades, se ha desarrollado un algoritmo estéreo basado en regiones, ya

que permite obtener unos bordes mejor definidos que el estéreo basado en rasgos. Esta decisión se debe a que los contornos activos extraen mejor la siluetas de los objetos cuanto más precisos sean los límites de los mismos.

- **Simetrías y bordes:**
  Como paso previo a la ejecución de los contornos activos, se calculan las energías externas en cada ROI, que son las responsables de atraerlos hacia los rasgos de interés. En este caso, se vuelven a emplear los bordes (tanto verticales como horizontales) y las simetrías verticales. Como se ha dicho, estas energías se calculan en un rango de disparidad concreto. Para ello se filtran las imágenes empleando como máscara aquel rango de disparidad que sea coherente con la distancia estimada de cada ROI. Esta medida se obtiene bajo la suposición del *mundo plano*.

El objetivo es obtener la forma del peatón contenido en cada ROI, para su posterior evaluación. Esta última fase de reconocimiento está sin implementar en el dominio visible, pero su finalidad sería reconocer la posición de las piernas. Una posible propuesta para realizar esta tarea ha sido implementada en el dominio infrarrojo y se trata en el capítulo **??**.

4.2. Visión estéreo para la obtención de mapas de disparidad

La visión estéreo es una interesante técnica sensorial pasiva en la que se basan los ADAS más recientes [FGaG+01, GM07, BBB+07a]. Los proyectos orientados a la comprensión de escenas de tráfico urbanas lo han venido empleando para la detección de obstáculos arbitrarios [BBFS00] y así como para la estimación de su movimiento relativo [FGG+98].

El hecho de que resulte factible la obtención de medidas 3D a partir de las imágenes, las convierte en técnicas muy atractivas para el cálculo de mapas de profundidad o de disparidad. Sin embargo, para obtener el mapa de disparidad es necesario resolver el problema de la correspondencia enunciado en el capítulo 3 [Zha00], que se define como la búsqueda del punto correspondiente en una imagen para cada punto característico de la otra imagen. Este problema no es exclusivo de la visión estéreo; aparece en otros ámbitos de la visión por computador, como son, el reconocimiento de objetos, el análisis de secuencias de imágenes y el análisis de imágenes tridimensionales.

En lo referente a las técnicas de visión estéreo, los algoritmos de búsqueda de la correspondencia se clasifican en dos categorías:

1. Métodos basados en áreas o superficies: Explotan la resolución fotométrica (intensidad) de los píxeles por medio de áreas o ventanas sobre la imagen. Obtiene buenos resultados sobre imágenes con textura importante. Permiten crear mapas densos de disparidad siendo posible incluso obtener precisión subpíxel. Además, son fáciles de paralelizar.

   Como inconveniente está el hecho de que asumen una superficie continua, presentando problemas en presencia de discontinuidades. Además, son muy sensibles a variaciones fotométricas, requieren de un proceso posterior de eliminación de falsas correspondencias y tienen problemas con las oclusiones.

2. Métodos basados en características o rasgos: Explotan las primitivas de alto nivel obtenidas de la imagen, que comprenden un conjunto de características invariantes a la proyección en mayor o menor medida; pueden ser puntos, líneas o curvas de los contornos, o bien las regiones, entre otros. De estas primitivas se seleccionan los rasgos distintivos (como por ejemplo, la posición, orientación o curvatura) que permita establecer la correspondencia.

   Proporcionan una información más dispersa pero más robusta cuanto más significativa sea la primitiva. La información que contienen es más rica que los niveles de intensidad y permiten utilizar restricciones geométricas entre las primitivas.

En general, los algoritmos estereoscópicos emplean como entrada las primitivas extraídas de la imagen y devuelven como resultado un mapa de disparidades o directamente la posición tridimensional de esas primitivas. Existen distintas técnicas, caracterizadas por la elección de las primitivas, las restricciones que les son asociadas y la estrategia de la correspondencia.

En esta tesis, como pasos previos al cálculo de la correspondencia, se ha realizado la calibración del sistema estéreo y la rectificación de las imágenes, con el fin de facilitar la obtención del mapa de disparidad. Para la calibración se ha empleado el método de Bouguet [Bou00], explicado en el capítulo 3. De ese modo se consigue mejorar el alineamiento epipolar y se reduce el efecto de la distorsión de la lente en aquellos píxeles apartados del centro focal.Sin embargo, hay sistemas que realizan la correspondencia sin rectificar las imágenes previamente [RD96]. En la figura 4.2 se muestran los resultados obtenidos después de rectificar la imágenes de ambas cámaras.

Con el fin de aprovechar las ventajas ofrecidas por los distintos enfoques estéreo, se han desarrollado dos módulos; uno basado en rasgos y el otro basado en regiones. Ambos tienen en común que no exigen un hardware específico, pero son capaces de ser ejecutados en tiempo real, y serán empleados cuando sea más útil detectar simetrías verticales (estéreo basado en rasgos) o aplicar contornos activos (estéreo basado en regiones).

4.2.1. Limitaciones de la segmentación basada en estéreo

Algunos autores [SGH04], se cuestionan los beneficios de basarse en la disparidad para segmentar peatones en entornos urbanos debido a que la cantidad de disparidad defectuosa es considerable. De hecho, algunas técnicas básicamente *"bottom-up"*, como las que tratan de agrupar la información de profundidad no-densa [ZT00a] o aquellas basadas en la detección de obstáculos empleando la técnica *V-disparity* [BRF+03, GZNR04], tratan de enfrentarse a esos escenarios saturados, bien cuando los peatones no están bien diferenciados del resto de objetos en el entorno o cuando la superficie de la carretera está considerablemente tapada por otros obstáculos. En tales casos, los rasgos de los peatones aparecen frecuentemente fusionados con los de otros objetos del entorno y el subsiguiente clasificador tiene problemas para tratar el mapa de disparidades resultante.

En el caso en estudio, la segmentación basada en mapas de disparidad ha presentado esta limitación. Por ello, se ha desarrollado otro método que en lugar de confiar en una segmentación basada en la disparidad, se apoya en la información de la disparidad para localizar las

Figura 4.2: Resultado de la rectificación aplicado a; (*a*) la imagen izquierda y (*b*) derecha tomadas por el sistema estéreo y (*c*) la imagen de la izquierda y (*d*) de la derecha obtenidas como resultado de la rectificación.

regiones de interés en la imagen. Este enfoque se enmarca dentro de los basados en la búsqueda del foco de interés, que precisamente sustituyen la segmentación de objetos en sí misma, por una detección de regiones caracterizadas por unos rasgos de interés. Más adelante se explica la metodología empleada para alcanzar este objetivo.

4.3. Rango de la detección

Los parámetros intrínsecos de la cámara son utilizados para establecer las relaciones entre las coordenadas $3D$ del mundo y los píxeles de la imagen, suponiendo que la carretera sea plana delante del sistema de visión y suponiendo que el cabeceo del vehículo se puede ignorar. De hecho, estas estrictas suposiciones se pueden considerar válidas en las proximidades del vehículo (hasta un máximo de 20 m.), incluso cuando exista cierta pendiente en la carretera. Por el contrario, en el área lejana (a más de 20 m. de distancia), los resultados pueden ser menos fiables.

Bajo la hipótesis de ”mundo plano”, la calibración se emplea para ajustar las correspondencias entre:

1. Distancias en el mundo 3D y líneas en la imagen,

2. Tamaño de los objetos de interés 3D y el tamaño de las ROIs

4.3.1. Distancias en el mundo 3D y líneas en la imagen

Siendo conocidos los parámetros intrínsecos de la cámara, la relación entre la disparidad de los objetos en la imagen y la distancia Z a la que se encuentran en el mundo real con respecto a la cámara es directa. Como se observa en la fórmula siguiente, una es la inversa de la otra.

$$Z = \frac{f \cdot baseline}{disparidad} \tag{4.1}$$

donde, Z es la profundidad o distancia a la que están los objetos en la imagen, f es la distancia focal y *baseline* es la distancia entre las cámaras. La distancia focal y la separación existente entre las cámaras son datos aportados por el módulo de calibración, explicado en el capítulo anterior.

A modo de ejemplo, en 4.3 se dibuja la posición en la imagen de objetos que están a unas distancias entre 3 y 30 m (siendo el rango de disparidad correspondiente entre 4 y 40).

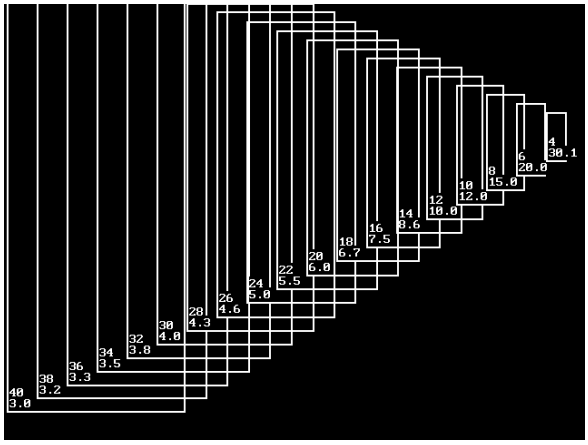


Figura 4.3: Tamaño de las regiones de interés correspondientes a un peatón de 175 cm de alto y 90 cm de ancho que esté a diferentes distancias. Los cálculos se han realizado en base a los parámetros intrínsecos de las cámaras que componen el sistema estéreo del dominio visible. El rango de detección va de 3 a 30 metros, siendo la disparidad de entre 4 a 40 píxeles y el tamaño de la imagen es de 640x480.

4.3.2. Tamaño de los objetos de interés 3D y de las ROIs en la imagen.

El sistema global propuesto tiene como finalidad la detección de adultos que puedan aparecer en la escena. Como referencia, la imagen 4.3 muestra las regiones de interés (ROIs) correspondientes a un peatón de 190*cm*. de alto y 75*cm*. de ancho a diferentes distancias (a mayor distancia, menor tamaño). A partir de las restricciones geométricas del sistema de visión empleado, se puede determinar el tamaño del peatón (o de la ROI que lo contiene) en la imagen.

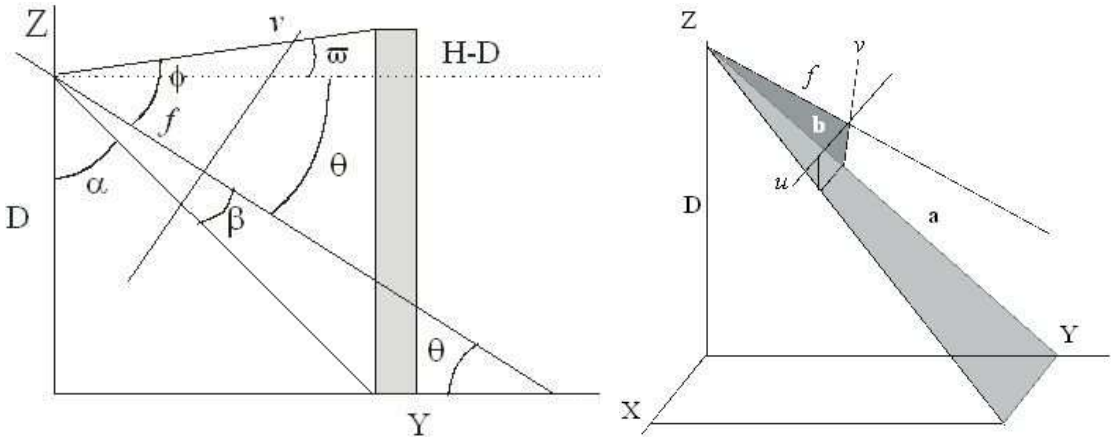


Figura 4.4: Esquema del sistema de visión estéreo empleado; (*a*) Vista de perfil y (*b*) 3D.

Si se observa el sistema 4.4 de perfil, se tiene que para los puntos situados sobre el suelo:

$$\tan\alpha = \frac{Y}{D} = \tan(\pi/2 - \theta + \beta) = \frac{1}{-\tan(-\theta+\beta)} = -\frac{1+\frac{v}{f}\tan\theta}{\frac{v}{f}-\tan\theta} = \frac{f+v\tan\theta}{f\tan\theta - v} \tag{4.2}$$

Luego la relación entre la coordenada Y del mundo y su proyección en la imagen v vendrá dada por las ecuaciones:

$$Y = D\frac{f+v\tan\theta}{f\tan\theta - v} \qquad v = f\frac{Y\tan\theta - D}{Y + D\tan\theta} \tag{4.3}$$

Para un punto que se encuentre a esa distancia Y pero a una altura H se tendrá:

$$\tan\phi = \frac{v}{f} = \tan(\theta+\varpi) = \frac{\tan\theta + \frac{H-D}{Y}}{1-\tan\theta\frac{H-D}{Y}} = \frac{Y\tan\theta + H - D}{Y - \tan\theta(H-D)} \tag{4.4}$$

por lo que:

$$Y = (H-D)\frac{f+v\tan\theta}{v - f\tan\theta} \qquad v = f\frac{Y\tan\theta + H - D}{Y - \tan\theta(H-D)} \tag{4.5}$$

Una vez conocida Y puede obtenerse X ya que como:

$$\frac{u}{b} = \frac{X}{a} = \frac{u}{\sqrt{f^2+v^2}} = \frac{X}{\sqrt{Y^2+D^2}} \tag{4.6}$$

Despejando X y sustituyendo el valor de Y obtenido con anterioridad:

$$X = \frac{u\sqrt{Y^2+D^2}}{\sqrt{f^2+y^2}} = \frac{Du}{f\,sen\theta - v\cos\theta} \tag{4.7}$$

Aunque como lo que interesa es el incremento de anchura:

$$\Delta X = D\frac{\Delta u}{f\,sen\theta - v\cos\theta} \qquad \Delta u = \frac{f\,sen\theta - v\cos\theta}{D}\Delta X \tag{4.8}$$

De este modo, la coordenada v de las ecuaciones 4.3 y 4.4 establecen la base y la altura de la ROI, mientras que la coordenada u de 4.8 impone el ancho. La figura 4.3 muestra los cálculos gráficamente.

Con esto se sabe para cada distancia el área máxima que puede ocupar un peatón (se establece el tamaño de cada ROI). Queda por determinar la posición de cada peatón en la imagen. Dicho de otro modo, falta por encontrar la distancia real a la que se encuentra (fila en la imagen) y la columna en la imagen en la que está. Lo primero se obtendrá mediante el estéreo, y lo segundo, a través de la simetría de los bordes verticales.

4.4. Generación de hipótesis basada en estéreo disperso

La fase inicial del sistema consiste en la detección de obstáculos, entendiendo como tales, aquellos elementos de la carretera que pueden correr peligro frente a un vehículo. El detector de obstáculos se basa en el estéreo disperso para la generación de hipótesis (posibles peatones). Como resultado final, proporciona una lista de ROIs que es suministrada a la subsiguiente fase de clasificación. A continuación se explica en detalle todo el proceso.

4.4.1. Preprocesamiento de las imágenes de entrada

El procesamiento se inicia con el cálculo de la amplitud de escala para independizar un poco la imagen de la iluminación y aumentar el contraste de las imágenes capturadas. Este paso es importante, ya que las intensidades de las imágenes estéreo pueden sufrir variaciones, dificultando la búsqueda de la correspondencia. Esto puede deberse a desiguales características del elemento sensor de la cámara derecha e izquierda, como ocurre con el brillo o el contraste. Otro motivo para que las intensidades sean distintas puede hallarse en las diferencias condiciones de iluminación para cada posición de las cámaras. Si se van a usar las imágenes de entrada sin preprocesar, será necesario emplear una medida de similitud invariante a la iluminación. La solución más comúnmente adoptada consiste en preprocesar las imágenes de entrada. Las técnicas más extendidas suelen basarse en la substracción de la media o en la convolución mediante la laplaciana de la gaussiana.

Al emplear una técnica estéreo basada en rasgos, mediante la amplitud de escala se pretende realzar el contraste de los bordes, para favorecer su detección. Para ello se obtiene el histograma y se ve en nivel de gris que tiene más píxeles en el entorno $5 - 250$ (así evitamos píxeles muy oscuros o claros). Después se busca el primer nivel de gris que tiene un tanto por ciento de píxeles igual o superior al máximo. De igual forma se busca el último que cumple esta condición. Ambos valores son el nuevo 0 y 255 y se reescala la imagen.

Además de estas dos tareas, una orientada a la mejora del contraste y la otra, a la reducción de las diferencias de iluminación, el preprocesamiento de las imágenes pueden perseguir la extracción de información adicional que ayude en la búsqueda de la disparidad. El algoritmo

Figura 4.5: Ejemplo de la amplitud de escala aplicado a; (*a*) una de las imagénes estéreo obtenida con la cámara visible y (*b*) el resultado obtenido al aplicar la amplitud de escala. Las imágenes han sido redimensionadas a 320x240 píxeles.

estéreo no-denso implementado en esta tesis, realiza una segmentación basada en bordes para dirigir la búsqueda de regiones con similitudes fotométricas. A partir de los bordes verticales extraídos de la imagen de la izquierda, se buscan los puntos correspondientes en la imagen de la derecha. De este modo, el algoritmo busca puntos de interés en los que realizar la correlación para reducir el tiempo de cálculo. En este caso, tomando como base las características antropomórficas de un peatón, se tendrán en cuenta aquellos puntos con una inclinación superior a 60 grados respecto al eje horizontal. Las regiones uniformes y los bordes que no alcancen ese grado de inclinación no son considerados en el cálculo de la correspondencia. De entre los posibles bordes verticales contenidos en la ventana de correlación, se seleccionan aquellos cuyo gradiente sea máximo (*Non-Maximum Supression*).

4.4.2. Medidas de similitud

Las medidas de similitud más sencillas se basan en la diferencia de intensidad a nivel de píxel, como es el caso del cálculo de la diferencia absoluta (*Absolute Difference*, AD) o de la diferencia al cuadrado (*Squared Difference*, SD). Los algoritmos que usan esta única medida de intensidad para comparar puntos se conocen como algoritmos *píxel-a-píxel*. Desafortunadamente, las imágenes discretas tienen un número de distintas intensidades de gris limitado, no resultando suficientemente discriminantes cuando las intensidades de píxeles que no se corresponden es la misma o están corrompidos por el ruido. Por otro lado, las intensidades a nivel de píxel pueden verse afectadas al muestrear la imagen. Por tanto, resulta necesario emplear una medida menos sensible ante cambios de tamaño de la imagen. Para aumentar la capacidad discriminante, se suele emplear la suma de la diferencia de los píxeles en una ventana alrededor del píxel de interés, como por ejemplo la suma de diferencias absolutas (*Sum of Absolute Differences*, SAD) o la suma de diferencia al cuadrado (*Sum of Squared Differences*, SSD).

En esta tesis se han evaluado distintas medidas de similitud;

- Suma de diferencias absolutas (*Sum of Absolute Differences*, SAD):

$$SAD = \sum_{i,j} |I_{izda}(i,j) - I_{dcha}(x+i, y+j)| \tag{4.9}$$

- Correlación cruzada normalizada (*Normalized Cross Correlation*, NCC):

$$NCC = \frac{\sum_{i,j} (I_{izda}(i,j) \cdot I_{dcha}(x+i, y+j))}{\sqrt{\sum_{i,j} I_{izda}(i,j)^2 \sum_{i,j} I_{dcha}(x+i, y+j)^2}} \tag{4.10}$$

- Suma de diferencias absolutas con media cero (*Zero mean Sum of Absolute Differences*, ZSAD):

$$ZSAD = \sum_{i,j} \left| (I_{izda}(i,j) - \overline{I_{izda}}) - (I_{dcha}(x+i, y+j) - \overline{I_{dcha}}) \right| \tag{4.11}$$

La suma de diferencias absolutas se ve muy influenciada por la ganancia de las cámaras, siendo necesario preprocesar previamente las imágenes para reducir este problema. Un preprocesamiento muy extendido consiste en la aplicación de la Laplaciana.

Otro modo de reducir las diferencias en intensidades entre las imágenes capturadas por las dos cámaras consiste en utilizar la resta de la media de la ventana que rodea al píxel de interés. A esta clase pertenecen la correlación normalizada y la suma de diferencias absolutas con media cero. La ventaja del segundo método frente al primero es que la correlación normalizada exige mayor tiempo de cálculo. Por este motivo, los resultados mostrados en este capítulo hacen referencia al uso de la suma de diferencias absolutas con media cero (ZSAD) (ver fig. 4.6-a y b).

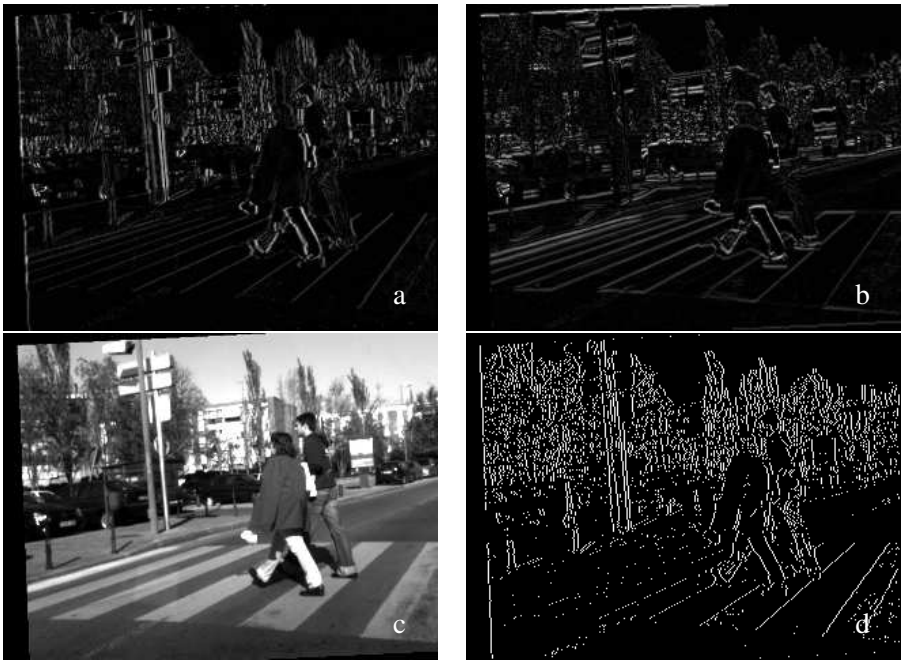


Figura 4.6: El cálculo del ZSAD se realiza considerando los bordes con una pendiente superior a 60 grados obtenidos a partir de; (*a*) los bordes verticales y, (*b*) los bordes horizontales (*c*) extraídos de la imagen original izquierda. (*d*) Muestra los bordes seleccionados como resultado del proceso. Serán los que se utilicen en el cálculo de la correspondencia.

Cuanto mayor sea el tamaño de la ventana, más robusto es el algoritmo frente al ruido. Sin embargo, ventanas más grandes con un tamaño fijo calculadas para el píxel central darán como resultado unas estimaciones de la disparidad menos precisas [KER06]. Esto se debe a que distintas áreas de una superficie van a corresponderse únicamente con una proyección en la otra imagen. Por otro lado, las oclusiones cerca de los bordes de los objetos pueden ocasionar problemas. Cuando una ventana grande esté centrada en un punto del fondo cercano a un borde de un objeto, casi con toda seguridad contendrá parte del objeto de interés. Pero, en el caso de una oclusión, una gran cantidad de píxeles del fondo contenidos en la ventana, no van a aparecer en la otra imagen. La medida de similitud, en este caso, va a considerar erróneamente disparidades pertenecientes a objetos en lugar del fondo.

En la literatura pueden encontrarse varias maneras de mejorar la correspondencia basada en ventanas. Kanade y Okutomi [KO94] propusieron una ventana de tamaño adaptativo; dada una estimación inicial de la disparidad, usan una técnica estadística para calcular el tamaño óptimo de cada ventana de correspondencia. A pesar de todo, la estimación del tamaño óptimo de ventana para cada punto es computacionalmente costoso.

Otras alternativas consideran técnicas de multiescalado, caracterizadas por el uso de distintos tamaños de ventanas, donde las ventanas más grandes proporcionan robustez, mientras que las más pequeñas añaden precisión a la correspondencia. El inconveniente es que los errores cometidos en un nivel más basto pueden propagarse a escalas más finas. Un enfoque parecido se adopta al permitir que la localización del píxel central (aquel píxel sobre el que recae el resultado de la correspondencia) varíe [BS99, FTV00, HIG02]. En esta tesis se ha utilizado una ventana de tamaño fijo de 7x7, para imágenes de 320x240 y una disparidad máxima de 20 píxeles, ya que los resultados obtenidos permitían alcanzar un equilibrio entre precisión y tiempo de procesamiento, pudiéndose ejecutar en tiempo real (ver fig. 4.7). Al realizar la correlación considerando únicamente los puntos de la imagen con una pendiente máxima de 60 grados respecto a la horizontal, se aceleran considerablemente los tiempos de cómputo al tener en cuenta un menor número de puntos (ver fig. 4.6-d).

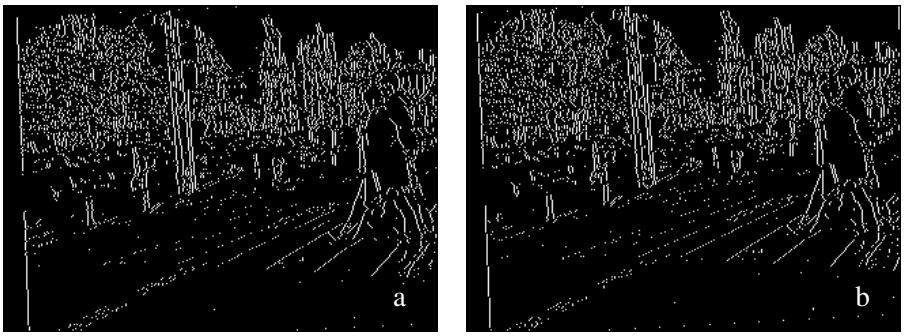


Figura 4.7: El cálculo de la correspondencia se realiza considerando las intensidades en la vecindad de los puntos de bordes verticales, (*a*) de las imágenes izquierda (*b*) y derecha respectivamente, que se muestran binarizados.

4.4.3. Búsqueda de la disparidad

El objetivo de las estimación de la disparidad estéreo es encontrar los puntos correspondientes en ambas imágenes. Si la geometría de las cámaras es conocida, la búsqueda se reduce a la línea epipolar. En el caso de que las imágenes se rectifiquen (ver el capítulo 3), los puntos correspondientes estarán en la misma fila en las dos imágenes.

El espacio de disparidad contiene todas las posibles correspondencias para una misma fila en la imagen izquierda y derecha. El algoritmo desarrollado sólo realiza la búsqueda de izquierda a derecha, limitando el espacio de búsqueda entre una disparidad mínima de 4 píxeles y un máximo de 20. De este modo se cubre un rango de distancias entre 3 y 15 metros aproximadamente (ver tabla 4.2). Una vez que se han calculado los valores de similitud en el espacio de búsqueda de la disparidad, se pueden buscar las correspondencias correctas. Un modo directo de conseguirlo consiste en buscar la disparidad que proporciona un valor máximo. Este enfoque se conoce con el nombre de *"the Winner Takes All"* (WTA).

Entre los mayores inconvenientes del enfoque WTA, destacan dos; por una lado, no resuelve el problema de las oclusiones y por otro, es muy sensible a la medida de similitud empleada, ya que su búsqueda se basa en la optimización de esa medida. En caso de regiones con poca textura, puede no existir un valor de similitud máximo o en el otro extremo, en el caso de estructuras repetitivas, pueden existir varios máximos ambiguos. En estas situaciones el WTA es propenso a fallos.

La secuencia de imágenes en la figura 4.8 muestra los mapas de disparidades obtenidos para una secuencia de imágenes adquiridas con el sistema binocular. Se han diferenciado con colores aquellos píxeles que comparten el mismo rango de disparidad. Los colores son más oscuros a distancias mayores.

4.4.4. Postprocesamiento: Corrección de errores.

El conjunto de correspondencias obtenidas por un algoritmo basado en áreas pueden ser escasamente fiables y por lo tanto se debe incorporar un paso de validación para incrementar la robustez. Una fuente principal de errores son los puntos ocluidos, es decir, puntos que debido a las discontinuidades de profundidad pueden ser vistas sólo en una de las dos imágenes. Una propuesta efectiva en la detección y rechazo de puntos ocluidos [Fua91, SM00, SMMN03], valida sólo aquellas correspondencias coherentes en el emparejamiento de izquierda a derecha (correspondencias directa) y de derecha a izquierda (correspondencias inversa). Otras importantes fuentes de errores son las regiones de baja textura de la imagen y con patrones repetitivos. Algunos autores han tratado de resolver el problema considerando los dos [HIG02] o incluso los tres [MDM02] mejores valores de correspondencia del espacio de búsqueda de disparidad.

El método empleado para obtener las disparidades, como ya se ha explicado, sólo considera los píxeles de bordes verticales. Convendría por tanto, obtener un contorno del peatón lo más preciso posible. Esto resulta complicado para el caso que nos ocupa: los bordes se extraen de imágenes exteriores, cuyas condiciones no pueden ser controladas, por lo que rara vez se obtiene un contorno exento de discontinuidades. Además, las cámaras del espectro visible

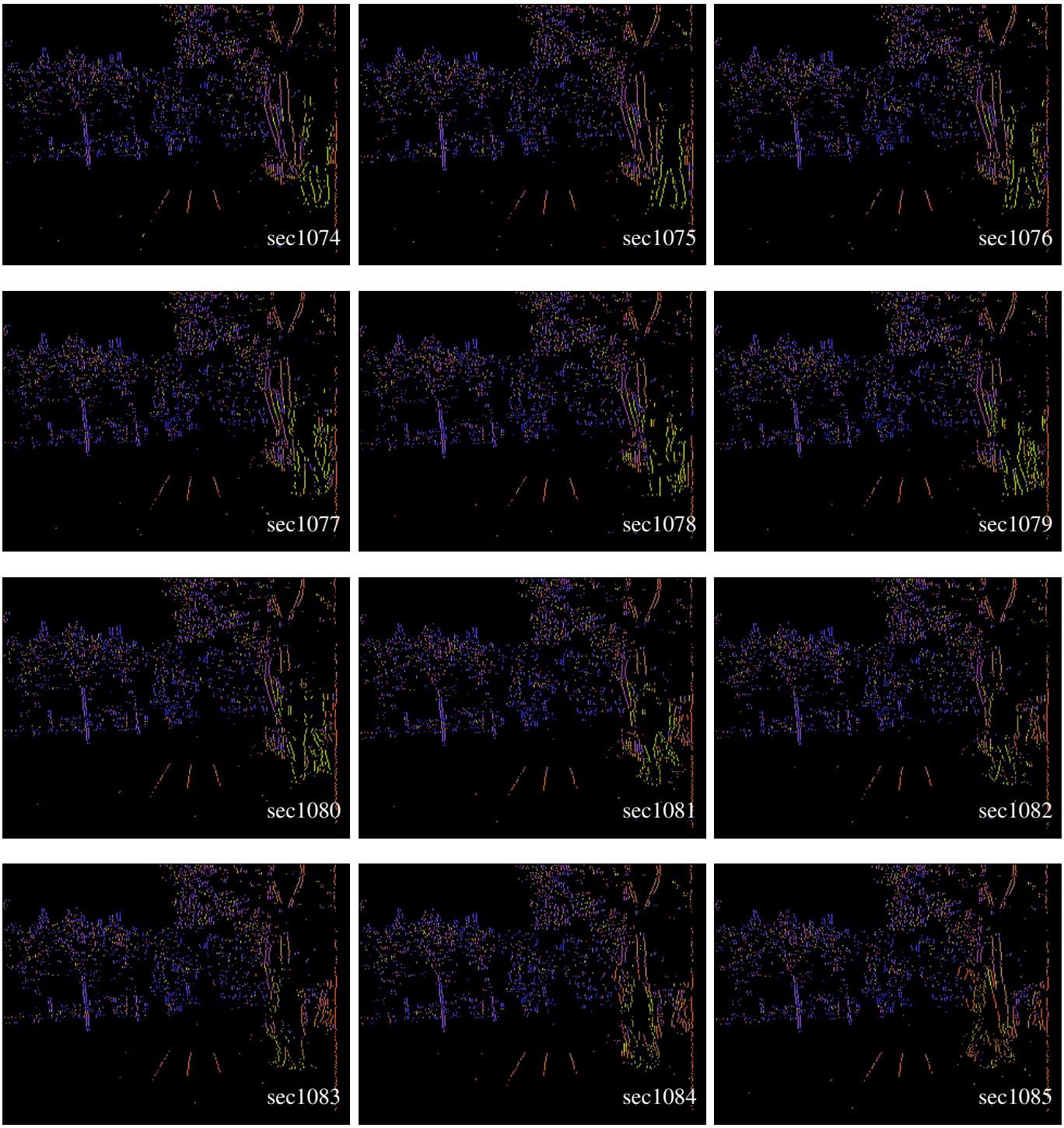


Figura 4.8: Mapas de disparidades obtenidos para una secuencia de imágenes. Los colores cálidos (amarillos y naranjas) corresponden a los objetos cercanos (entre 2-3 metros de distancia, siendo la disparidad entre 20-30), mientras que los objetos lejanos aparecen en gamas de azul (entre 4-15 metros o disparidades entre 4-14).

tienen la capacidad de captar gran cantidad de detalles, mostrando además de los bordes del peatón, otros muchos bordes pertenecientes a texturas en la ropa, elementos en la carretera (p.ej. líneas) u otros objetos del entorno (p.ej. las ramas y troncos de los árboles).

Como consecuencia de lo anterior, el mapa de disparidades generado posee discontinuidades y es propenso a fallos en la estimación de la disparidad. Para la corrección de esos errores se realizan operaciones morfológicas, con el fin de propagar la información de la disparidad disponible de un modo sencillo y rápido. Se han considerado las disparidades vecinas a cada píxel del mapa. Así, siendo la ventana de búsqueda de 3x3, si dos píxeles vecinos tienen disparidades iguales y el píxel central no, se corrige este último. Se sigue un algoritmo *"hit-or-miss"*, empleando los elementos estructurales mostrados en el esquema 4.1. De este modo se consigue rellenar los huecos y corregir los errores cometidos en la imagen estéreo basada en bordes verticales (ver fig. 4.9).

X			X			X				X			X			X	
O			O				O				X		O			O	
X				X				X				X			X		X

Tabla 4.1: Elementos estructurales usados para corregir los errores en el mapa de disparidades

4.4.5. Filtrado basado en la disparidad

A partir del mapa de disparidades corregido, se generan n imágenes de disparidad obtenidas al dividir el intervalo de disparidad global en n agrupaciones distintas. Como resultado de este proceso, se obtiene un conjunto de imágenes, cada una de ellas conteniendo únicamente los píxeles correspondientes a su grupo de disparidad. Se persigue así que cada nueva imagen contenga aquellos objetos que se encuentren en un rango de disparidad concreto.

Para generar los distintas agrupaciones se consideran dos factores: el número de disparidades que se quieren agrupar en cada nueva imagen y el número de disparidades que se van a solapar o reconsiderar entre la imagen de disparidad de un grupo c_i y del siguiente grupo c_{i+1}. El determinar estos parámetros depende de la bondad de la imagen de disparidades previamente obtenida. Los resultados que se van a mostrar (ver fig. 4.10), se han obtenido considerando un caso arbitrario; a saber, las disparidades se han agrupado de tres en tres, reconsiderando para cada imagen de un nivel $i + 1$ los dos mayores valores de disparidad del nivel anterior i, siendo por tanto el solape entre agrupaciones de dos disparidades.

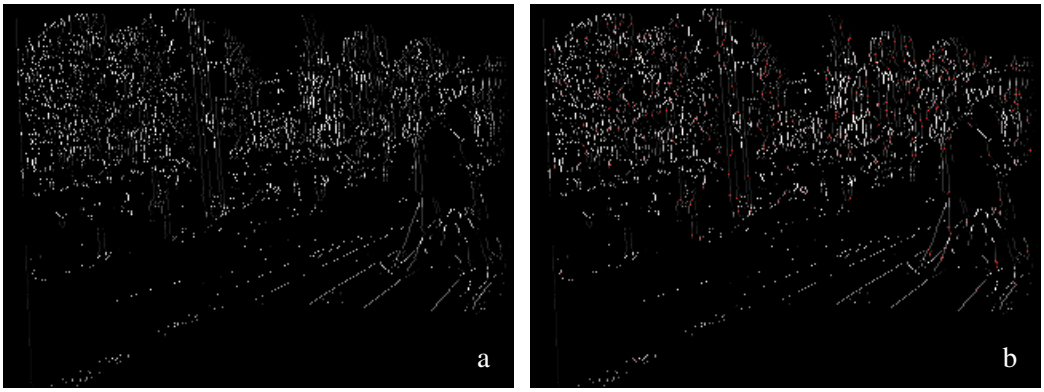


Figura 4.9: Resultado de las aplicar los elementos estructurales de la figura 4.1 (*a*) a el mapa de disparidades. (*b*) Se analiza la imagen estéreo y si dos píxeles tienen disparidades iguales y el de en medio no se corrige este último. En rojo aparecen los píxeles modificados.

Haciendo uso de la fórmula 4.1, se calcula el rango de distancias considerado en cada mapa de disparidad de nivel *i*. Para las agrupaciones de disparidades ya comentados, se obtienen los intervalos de distancias mostrados en la tabla 4.2. Se puede apreciar como a medida que la disparidad disminuye, el intervalo de distancias abarcado en cada mapa aumenta exponencialmente. Sin embargo, este hecho puede afectar a la precisión de la detección de objetos que se encuentren a más de 10 metros. A mayores distancias, el hecho de agrupar en el mismo mapa a objetos que a una distancia de más de 2 metros entre sí no impide que sea detectado. Si el objetivo es separar a los objetos con una mayor precisión, se puede generar el conjunto de imágenes *i* considerando medidas de distancias en lugar de las disparidades a la hora de umbralizar el mapa de disparidades original.

Una vez que se umbraliza la imagen de disparidades original, se hace crecer los puntos extremos de cada una de las nuevas imágenes, para intentar corregir pequeños errores. Para ello se emplean los elementos estructurales que se describen en el esquema 4.3. La imagen 4.11 muestra el resultado de estas transformaciones morfológicas aplicadas a una de las imágenes de disparidades.

4.4.6. Localización del foco de interés basado en simetrías

La detección de los peatones se apoya en los resultados obtenidos por el algoritmo estéreo no-denso. La búsqueda del *foco de interés* va a estar guiado por los bordes del mapa de disparidades. Cada hipótesis o ROI generada va a ser evaluada en base a medidas de simetrías. Al final de este proceso, el sistema toma la decisión de desechar o mantener cada hipótesis. Serán los subsiguientes módulos quienes verifiquen la existencia o no de un peatón.

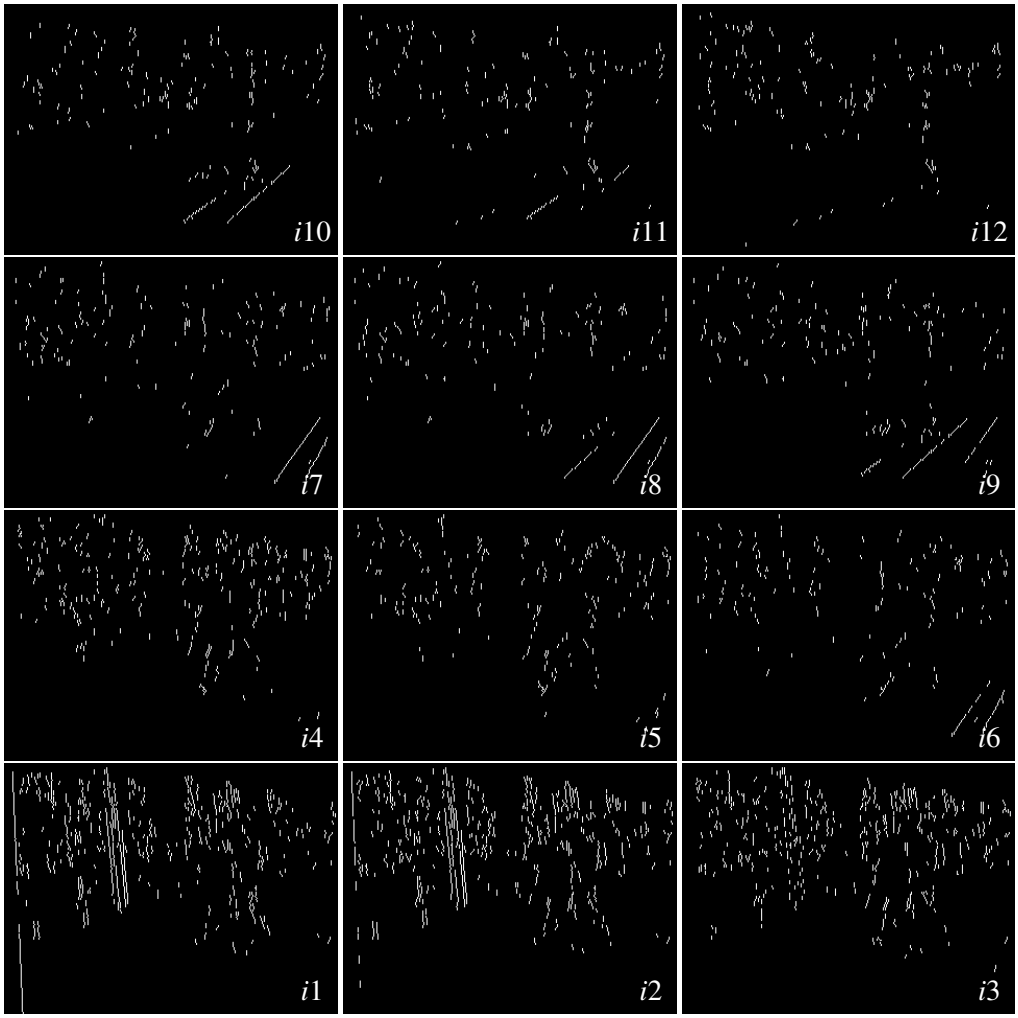


Figura 4.10: El mapa de disparidades previamente corregido, se filtra, obteniendo 15 imágenes nuevas, cada una de las cuales contiene las disparidades en un rango, tal y como muestra la tabla 4.2. En este ejemplo, se ha considerado un rango de disparidad máximo de 20 y cada imagen agrupa tres disparidades distintas. De izda. a dcha. se muestran 12 de las imágenes segmentadas de menor a mayor distancia.

Nivel i	Agrupación c de disparidad	Distancias (m.)
15	4-6	15.0 -10.0
14	5-7	12.0 - 8.6
13	6-8	10.0 - 7.5
12	7-9	8.6 - 6.7
11	8-10	7.5 - 6.0
10	9-11	6.7 - 5.5
9	10-12	6.0 - 5.0
8	11-13	5.5 - 4.6
7	12-14	5.0 - 4.3
6	13-15	4.6 - 4.0
5	14-16	4.3 - 3.7
4	15-17	4.0 - 3.5
3	16-18	3.7 - 3.3
2	17-19	3.5 - 3.2
1	18-20	3.3 - 3.0

Tabla 4.2: Relación entre los intervalos de disparidad de cada agrupación y los correspondientes rangos de distancias

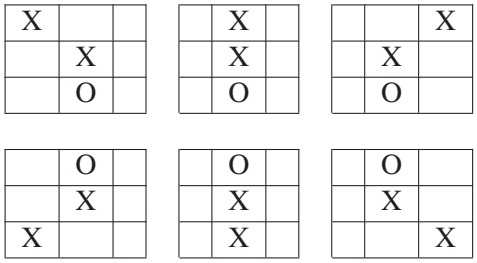


Tabla 4.3: Elementos estructurales usados para aplicar transformaciones morfológicas a las disparidades segmentadas

4.4.6.1. Definición de las regiones de interés

Para definir los objetos de interés se usa un tamaño y una relación de aspecto concretos. El tamaño de un peatón se ha determinado mediante una altura de 190 centímetros y una anchura de entre 15 a 75 centímetros. El motivo de considerar una anchura variable es doble. Por una parte, durante el ciclo de caminar el ser humano va cambiando de posición las piernas y los brazos, siendo, como consecuencia, el ancho que ocupa variable. Por otro lado, existen otros objetos en entornos urbanos con una fuerte simetría vertical. Es el caso de los árboles, postes y farolas, que son los responsables de generar multitud de falsos positivos en los sistemas de detección de peatones. Con el fin de filtrar estos casos, se ha impuesto como límite inferior una anchura de 15 centímetros. Por tanto, sólo se evaluarán aquellos obstáculos en la escena con un relación de aspecto en el rango de 2,5 a 12,6. Además, esto evita perder el tiempo analizando regiones que no cumplen las condiciones indicadas.

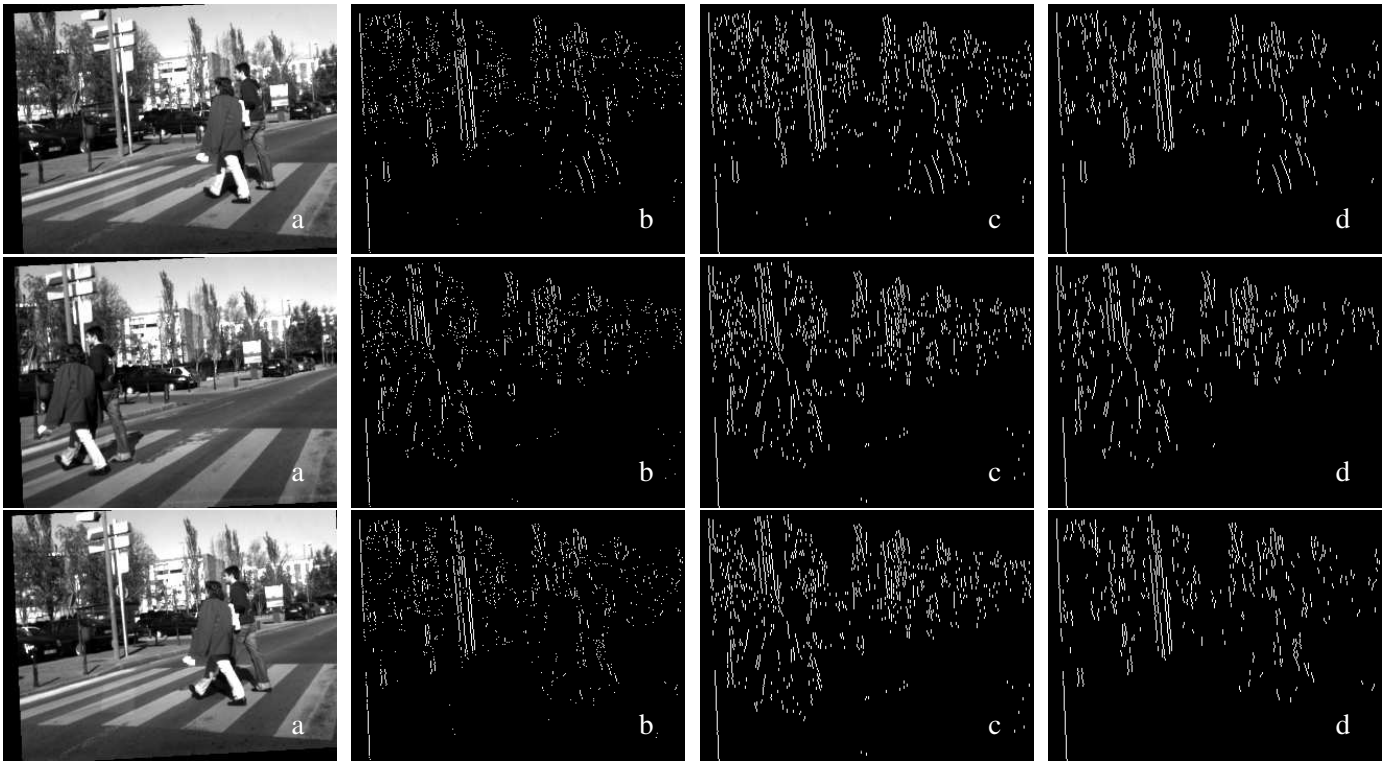


Figura 4.11: Fase de validación del mapa de disparidades. Se muestra el proceso para distintos mapas filtrados por disparidad. (*a*) Es la imagen de la cámara izquierda rectificada y mejorado el contraste, (*b*) Los mapas sin corregir se han filtrado por disparidad, mostrando una de las imágenes de disparidad. (*c*) El resultado de aplicar las transformaciones morfológicas y (*d*) al eliminar los píxeles aislados y los bordes de escasa longitud.

La búsqueda de peatones se realiza en cada mapa de disparidades, comenzando por el nivel actual hasta llegar al nivel máximo. Para cada uno de esos mapas, se sabe de antemano las dimensiones de los objetos de interés. Esto se debe a varios factores; primero, el sistema está calibrado y segundo, se conoce el tamaño de los objetos deseados en el mundo real ($190cm.$ de alto y $15-75cm.$ de ancho). Así, se puede calcular *off-line* los tamaños de las ROIs y almacenarlos en una tabla, que es accedida durante la ejecución en tiempo real, acelerando su ejecución. Todos los cálculos realizados para la obtención de las dimensiones de las ROIs se han detallado en el apartado 4.3. Para cada mapa se almacena el alto, ancho mínimo y ancho máximo de la región, cuyo contenido se evaluará después.

La separación de los obstáculos en varias imágenes en función de su disparidad, permite realizar la búsqueda empleando una ventana que se ajusta a la distancia que se espera encontrar el objeto. Se puede considerar que no son ventanas de tamaño fijo, ya que se admite cierta variabilidad con respecto al ancho.

4.4.6.2. Extracción de regiones de interés basado en simetrías

La búsqueda de los peatones del sistema implementado se limita a distancias entre 3 y 15 metros. El correspondiente rango de disparidades va de 4 a 20 píxeles. Este rango es suficiente como para cubrir las necesidades de un ADAS orientado a evitar colisiones.

Explotando la fuerte simetría vertical que muestran las personas, se estudia la existencia de un eje de simetría vertical en cada una de las posibles ROI. Para cada mapa de un nivel concreto, se recorren todas las posiciones de la línea en la que tiene su base la correspondiente ROI, realizando el análisis de simetría en función de los píxeles que encierra. Debido a que los peatones no son los únicos elementos de la carretera que pueden presentar esta simetría, se impone a los puntos que su anchura esté entre el mínimo y el máximo esperado. Así se consigue evitar falsas detecciones, como en el caso típico de los árboles y las farolas.

Finalmente, se establece un sistema de votación, ya que dentro de una misma ROI pueden existir varios ejes de simetría. El candidato con un mayor número de votos es seleccionado como la mejor posición del eje. Aquellas ROI que no superen un umbral de votos, serán desechadas. Como resultado se obtiene una lista de ROIs que serán proporcionadas al siguiente módulo, encargado de confirmar cada una de las hipótesis.

4.5. Resultados Experimentales del Detector de Obstáculos

El detector de obstáculos ha sido integrado en el vehículo experimental IvvI y ejecutado en entornos urbanos. El sistema es capaz de dar una respuesta en tiempo real. En la figura 4.12 se muestran algunos resultados proporcionados durante la conducción del vehículo. Se pueden apreciar falsos positivos, que serán eliminados en la siguiente fase de clasificación. Estas falsas detecciones son debidos a objetos con una fuerte simetría vertical, como es el caso de las vallas a lo largo de la acera o las señales de tráfico. En ocasiones se dan algunos falsos negativos, pero son excepciones que no se mantienen en la secuencia, recuperándose la correcta detección del peatón en las siguientes imágenes.

4.6. Verificación de Hipótesis basado en PCA

Como resultado de la fase de detección de obstáculos el sistema obtiene un conjunto de hipótesis aún por validar. En la figura 4.13 se muestran algunos ejemplos. La información de las simetrías verticales no es un rasgo suficientemente discriminante como para identificar a los peatones de entre el conjunto de posibles objetos. De ahí que resulte imprescindible una fase posterior que estudie el contenido de las ROIs para poder emitir un juicio más fiable acerca de su contenido. Es necesario volver a insistir en la dificultad que entraña esta tarea; la creación de un algoritmo capaz de reconocer personas es complicado, ya que es una tarea de alto nivel, difícilmente reproducible mediante un modelo computacional. A esta limitación cabe añadir las restricciones propias de un ADAS. Destacan la necesidad de dar una respuesta en tiempo real y la fiabilidad de esa respuesta.

El objetivo que se persigue es la creación de un algoritmo de reconocimiento de peatones rápido, razonablemente sencillo y preciso. El propuesto confía en su capacidad para reconocer

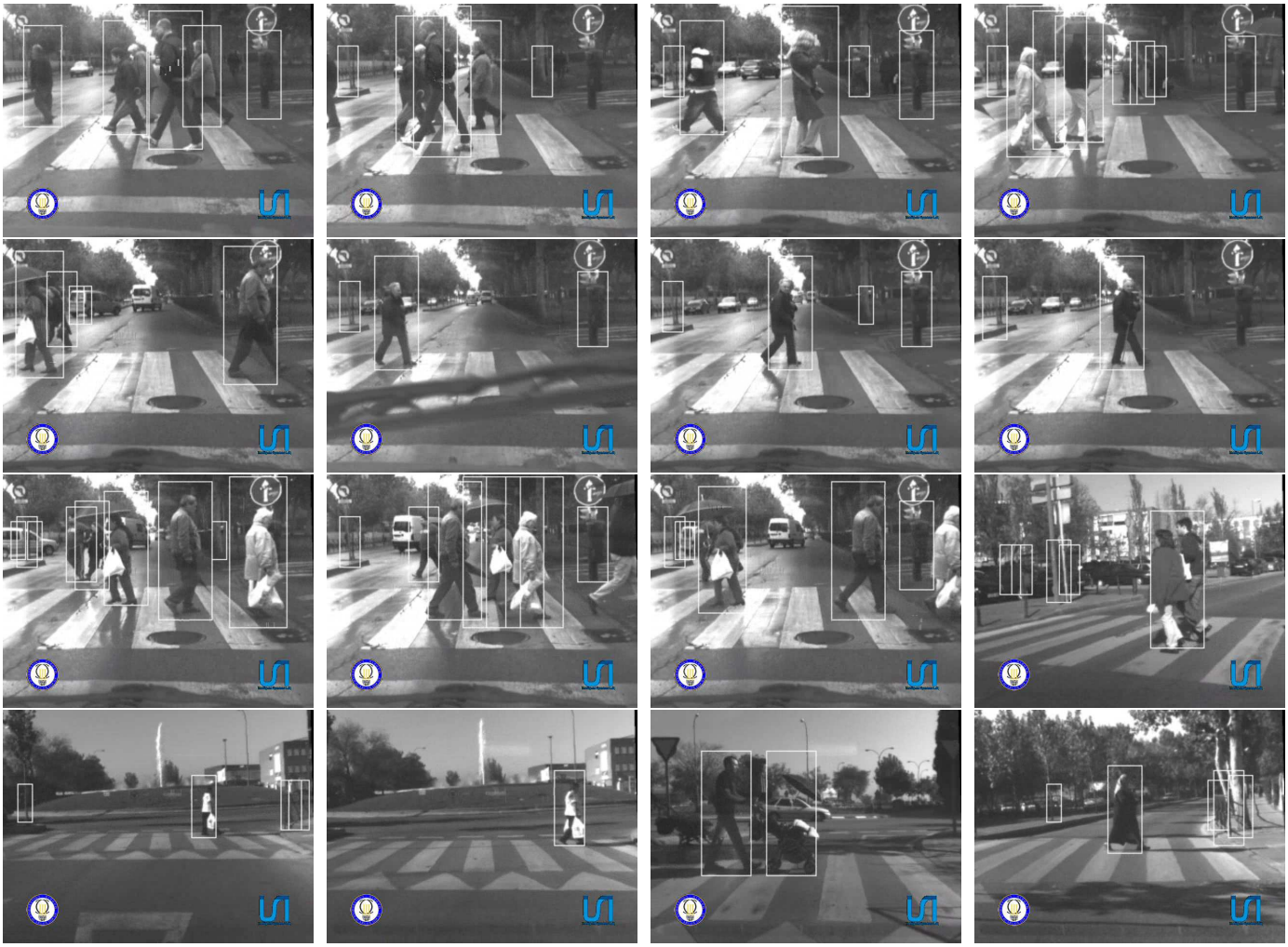


Figura 4.12: Imágenes que muestran los resultados del detector de obstáculos una vez integrado en el vehículo experimental IvvI. Los obstáculos candidatos a peatones son recuadrados para ser evaluados durante la fase de clasificación.

patrones, sin tener una dependencia de modelos 3D ni de una detallada geometría del cuerpo humano.

A pesar de que el reconocimiento de personas es un proceso de alto nivel, existe una cierta estructura implícita en la tarea. Turk y Pentland [TP91], sacaron provecho a parte de esa estructura, planteando un esquema para el reconocimiento de caras basado en un enfoque de la teoría de la información; a partir de un conjunto de caras, tratan de codificar la información más relevante capaz de distinguir a cada una de ellas. Se basa en la búsqueda del componente principal de la distribución de las caras. El análisis del componente principal (PCA) también se conoce como la transformada (discreta) de Karhunen-Loève (KLT) o la transformada de Hotelling (HT). Recientemente, se ha utilizado mucho en el ámbito de la visión por computador y en el reconocimiento de patrones. En concreto, es una técnica muy extendida para parametrizar la forma, la apariencia y el movimiento [Edw98, MA95, MN95, TP91].

Figura 4.13: Conjunto de imágenes conteniendo a peatones y a no-peatones obtenidas durante la fase de detección de obstáculos y utilizadas en el entrenamiento basado en el PCA.

La detección de peatones entraña una mayor complejidad que la de muchos otros objetos, debido a que las personas pueden mostrar una basta variación en apariencia, siempre que las extremidades (piernas y brazos) aparezcan en diferentes posiciones. A esto hay que añadir los diferentes tipos de ropas y colores que pueden llevar. Estas características de los peatones, hacen imprescindible el uso de un método robusto, capaz de aprender la alta variabilidad presente en la clase peatón.

Se propone un sistema de clasificación de peatones basado en PCA, cuyo objetivo es eliminar las ROIs suministradas por la fase previa que tengan una probabilidad muy baja de contener a un peatón. Para ello se ha creado un clasificador basado en la reconstrucción de imágenes mediante PCA y posterior comparación de la imagen reconstruida con la original. Este enfoque, basado en la teoría de la información de codificar y decodificar las imágenes de los peatones, facilita una comprensión de su contenido resaltando los rasgos locales y globales más significativos. Estos rasgos pueden o no tener relación con los rasgos de un peatón que un humano extraería de un modo intuitivo, como por ejemplo la cabeza, las piernas y los brazos.

4.6.1. EigenPedestrians para el reconocimiento

El enfoque basado en PCA transforma las imágenes de los peatones en un pequeño conjunto de imágenes de rasgos característicos, que se han denominado *eigenpedestrians* y que

son los componentes principales del conjunto de imágenes de peatones de entrenamiento. El reconocimiento tiene lugar proyectando una nueva imagen en el subespacio creado por los *eigenpedestrians* (constituyen el espacio de los peatones).

Originariamente, esta técnica la desarrollaron Sirovich y Kirby [SK87] para representar con precisión imágenes de caras. Posteriormente, Turk y Pentland [TP91] aplicaron esta idea para el aprendizaje y reconocimiento de caras, dando lugar a las conocidas *autocaras* o (*eigenfaces*).

4.6.2. Cálculo de los Eigenpedestrian

Las imágenes de peatones contenidas en el conjunto de entrenamiento son de dimensiones NxM igual a $30x69$, resultando un vector de 2070 dimensiones para cada imagen.

El PCA es una técnica utilizada para reducir la dimensionalidad de un conjunto de datos, siempre que esos datos no estén distribuidos de una forma aleatoria y se puedan describir en función de un subespacio de menor dimensión. Esta técnica trata de encontrar aquellos vectores que mejor agrupen la distribución de imágenes de peatones, definiendo el subespacio denominado espacio de peatones. Para ello se realiza una transformación lineal que escoge un nuevo sistema de coordenadas para el conjunto original de datos, en el cual la mayor varianza del conjunto de datos es capturada en el primer eje (llamado el Primer Componente Principal), la segunda varianza más grande es el segundo eje, y así sucesivamente. Las componentes principales o vectores del nuevo sistema, son los autovectores de la matriz de covarianza del conjunto de imágenes. Estos autovectores pueden interpretarse como un conjunto de rasgos, que combinados, caracterizan la variación existente en la colección de imágenes de peatones. Cada autovector van a ser de longitud 2070, describiendo a una imagen de NxM.

Sea un conjunto de entrenamiento formado por $I_1, I_2, I_3...I_M$ imágenes. La media de ese conjunto se define como $\psi = \frac{1}{M}\sum_{n=1}^{M} I_n$. Cada imagen se diferencia de la media un vector $\phi_i = I_i - \psi$

El análisis del componente principal trata de encontrar el conjunto de M vectores ortogonales (los autovectores) u_n y sus autovalores asociados λ_k que mejor describen la distribución de datos, obtenidos de la siguiente matriz de covarianza:

$$\begin{aligned} C &= \frac{1}{M}\sum_{n=1}^{M} \phi_n \phi_n^T \\ &= AA^T \end{aligned} \tag{4.12}$$

donde la matriz $A = [\phi_1 \phi_2 ... \phi_M]$. Esta matriz es de dimensión 2070^2, siendo intratable determinar 2070 autovectores y autovalores. Se puede resolver este problema pasando del orden del número de píxeles en la imagen (2070), al orden del número de imágenes M del conjunto de entrenamiento [TP91]. Se ha empleado una colección de $M = 84$ peatones, constituyendo el límite superior de posibles *eigenpedestrians* que se podrían obtener.

Cada una de las imágenes del conjunto de entrenamiento se puede representar con precisión como combinación lineal de los *eigenpedestrians*. Sin embargo, los peatones se pueden aproximar utilizando sólo los mejores *eigenpedestrians*, que son aquellos con los mayores

autovectores, y que por tanto retienen aquellas características del conjunto de datos que contribuyen más a su varianza dentro del conjunto de imágenes de peatones.

Uno de los motivos para usar un número menor de *eigenpedestrians* tiene que ver con la eficiencia computacional. Los mejores M' *eigenpedestrians* abarcan un subespacio M' dimensional de todas las posibles imágenes. La idea que subyace es que las imágenes de los peatones se pueden representar de un modo más económico mediante su proyección en un número reducido de imágenes base obtenidas mediante los autovectores más significativos de la matriz de covarianza.

4.6.3. Rasgos característicos: Bordes verticales y distancias a bordes

Debido a que los peatones aparecen con multitud de colores y diferentes texturas, no es recomendable usar rasgos basados en estas características para llevar a cabo el reconocimiento. Por este motivo, se ha decidido usar la imagen de los bordes verticales, ya que el contorno de un peatón muestra un predominio de éstos, y así se elimina una gran cantidad de información poco útil para el clasificador. Se han implementado y comparado dos clasificadores distintos; uno se construye a partir de la información de los bordes verticales y el otro, en función de las distancias a esos bordes.

Como paso previo al procesamiento de las imágenes, ha sido necesario aumentar el contraste de las mismas, para así mitigar los efectos de los cambios de iluminación. Después, se obtienen los bordes verticales y se calculan las distancias a los mismos. La figura 4.14 contiene una representación del tipo de imágenes empleadas durante el entrenamiento de ambos clasificadores.

4.6.4. Reconstrucción de la imagen usando PCA

El siguiente paso consiste en reconstruir la imagen original. Para ello, se proyecta la imagen en los componentes principales u_k^T(constituyen el espacio de peatones), y a partir de esa proyección w_k se va a tratar de reconstruir la imagen original, mediante la siguiente operación;

$$w_k = u_k^T(I - \psi) \tag{4.13}$$

para $k = 1...M'$, donde u_k^T son los autovectores, I es la imagen y ψ es la media. El número de autovectores más significativos M' seleccionados heurísticamente en función de los autovalores. En la mayoría de aplicaciones es habitual usar un número de autovectores que abarcan una varianza de entre el 65 %-90 %. En la gráfica 4.15 se observa el rango de autovalores y el porcentaje de varianza que los primeros n vectores cubren. En los experimentos realizados se han comparando los resultados obtenidos usando distintos números de autovectores, siendo $M' = [5, 10, 15, 10, 25, 79]$.

Las proyecciones o pesos forman un vector $\Omega^T = [w_1 w_2 ... w'_M]$ que describe la contribución de cada *eigenpedestrian* a la hora de representar la imagen original, considerándolos como un conjunto de bases para las imágenes de peatones. Este vector se utiliza para encontrar la distancia ϵ entre la imagen original y la reconstruida.

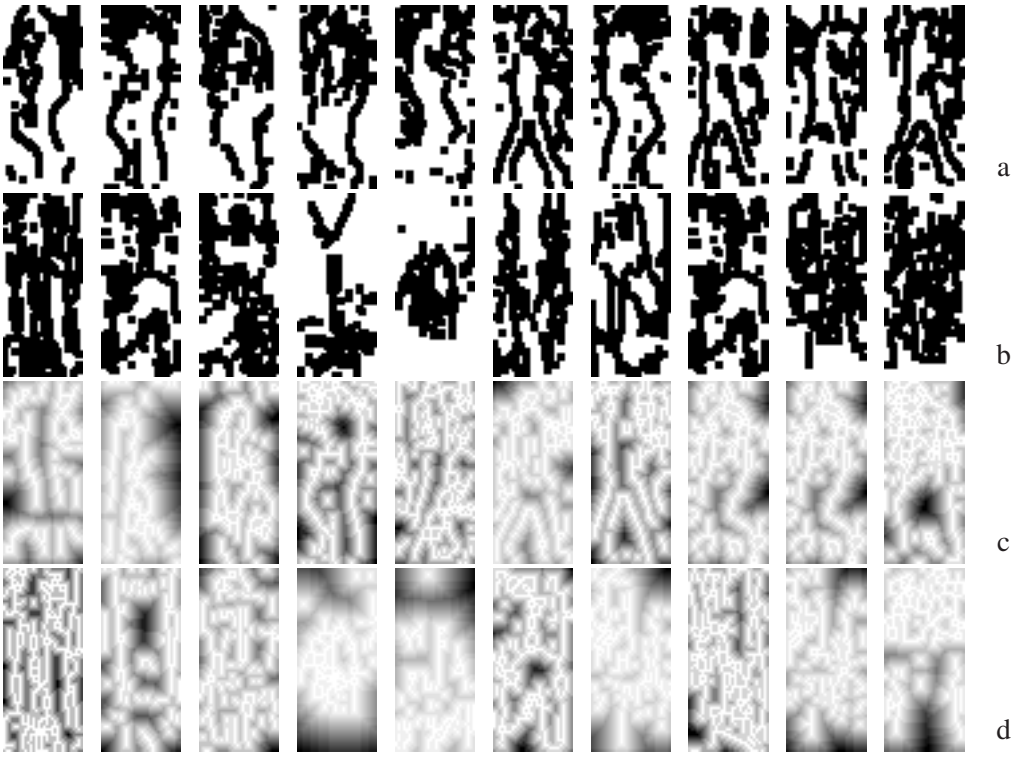


Figura 4.14: Algunos ejemplos del tipo de imágenes utilizado durante el entrenamiento del PCA basado en bordes y del basado en distancias, (*a*) bordes a peatones, (*b*) bordes a no-peatones, (*c*) distancias a peatones y (*d*) distancias a no-peatones.

Para calcular la imagen I'_k reconstruída a partir de la proyección w_k, se realizan las siguientes operaciones:

$$I'_k = \sum_{i=1}^{M'} u_k.w_k + \psi \tag{4.14}$$

A mayor número de autovectores que se usen para obtener la proyección, menor será la pérdida de información que se sufrirá, y por tanto la reconstrucción de la imagen será más precisa. En la figura 4.16 se visualizan como imágenes los 20 primeros autovectores obtenidos a partir de los bordes verticales y de las distancias a los mismos.

4.6.5. Clasificación en función de la reconstrucción

Se utilizan los *eigenpedestrians* o el espacio de los peatones para detectar a los peatones. La distancia ϵ entre la imagen original y su reconstrucción, no es más que la distancia entre la imagen normalizada (restándole la media) $\phi = I - \psi$ y su proyección en el espacio de los peatones $\phi_f = \sum_{i=1}^{M'} u_k.w_k$.

Como se observa en las figuras 4.17 y 4.18, las imágenes de peatones (en la primera columna) no cambian drásticamente al proyectarlos en el espacio de peatones (en la segunda columna), mientras que la proyección de imágenes de no-peatones cambian bastante. La tercera

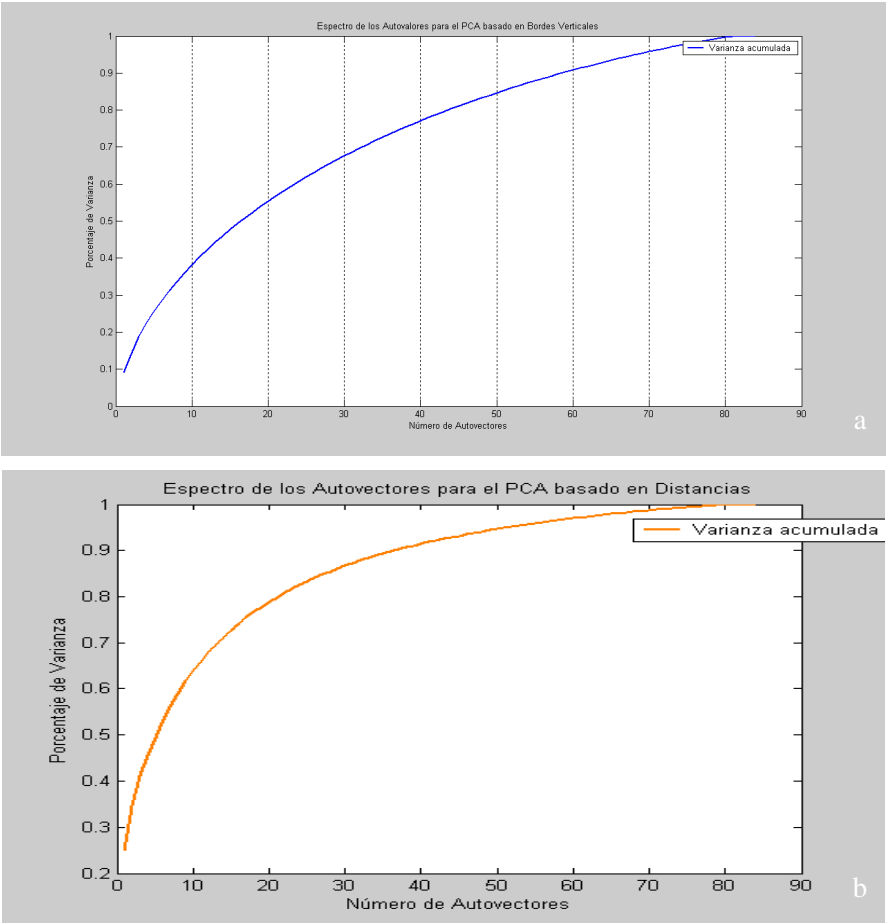


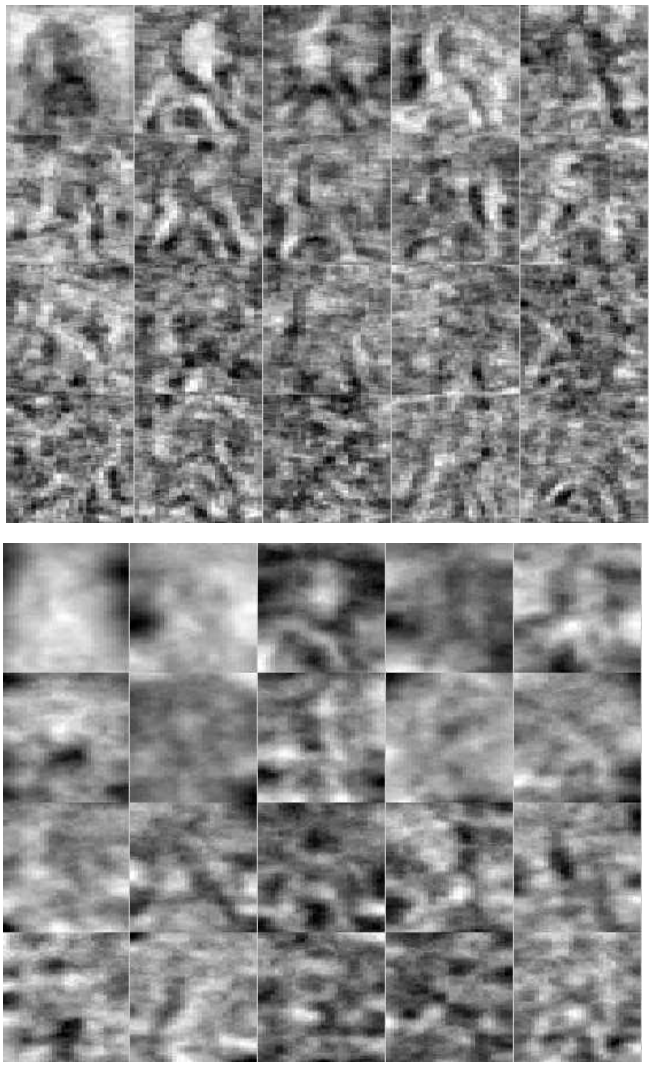
Figura 4.15: Gráfico que representa la varianza acumulada para los primeros n autovectores referidos a; (a) los bordes verticales y (b) las distancias a los mismos.

columna contiene el peatón tal y como es adquirido por la cámara. Para detectar los peatones de la escena, se va a calcular la distancia entre la subimagen contenida en la ROI y el espacio de los peatones. Esta distancia se usa como una medida del parecido de la subimagen con un peatón. Cuanto menor sea esa distancia, más parecidas serán ambas imágenes, existiendo un umbral θ tal que, las proyecciones de las imágenes correspondientes a peatones deberían estar por debajo de ese umbral $\theta < \epsilon$. El clasificador toma una decisión en función del siguiente criterio:

4.7. Resultados Experimentales del PCA

Ambos clasificadores se han comparado para determinar cuál ofrece mejores resultados. Se han obtenido un máximo de 79 autovectores, pero para medir su funcionamiento, se han considerado distinto número de autovectores. Las curvas ROC de la figura 4.19 permiten comparar cómo afecta este parámetro a los resultados.

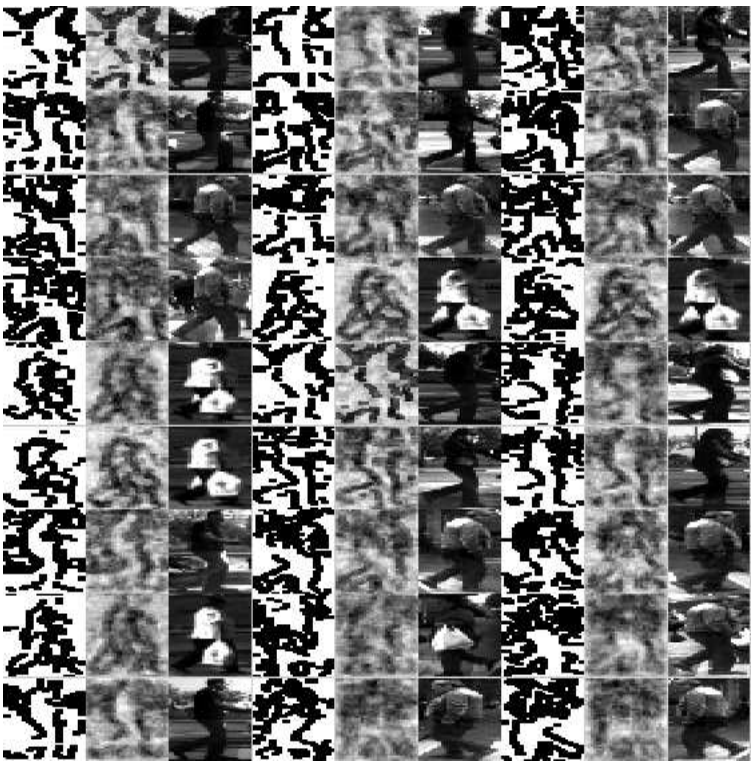
Un análisis del área debajo de cada una de las curvas ROC, ha demostrado que el PCA



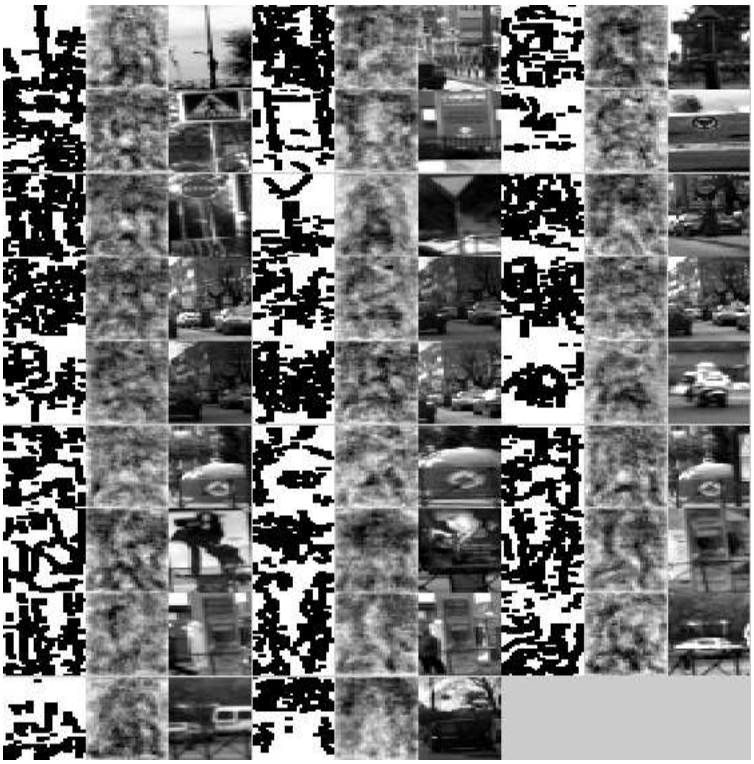
a

b

Figura 4.16: Los eigenpedestrians se pueden visualizar como una imagen. En estos ejemplos se han empleado los 20 primeros autovectores obtenidos de los (*a*) bordes verticales y (*b*) las distancias a bordes.

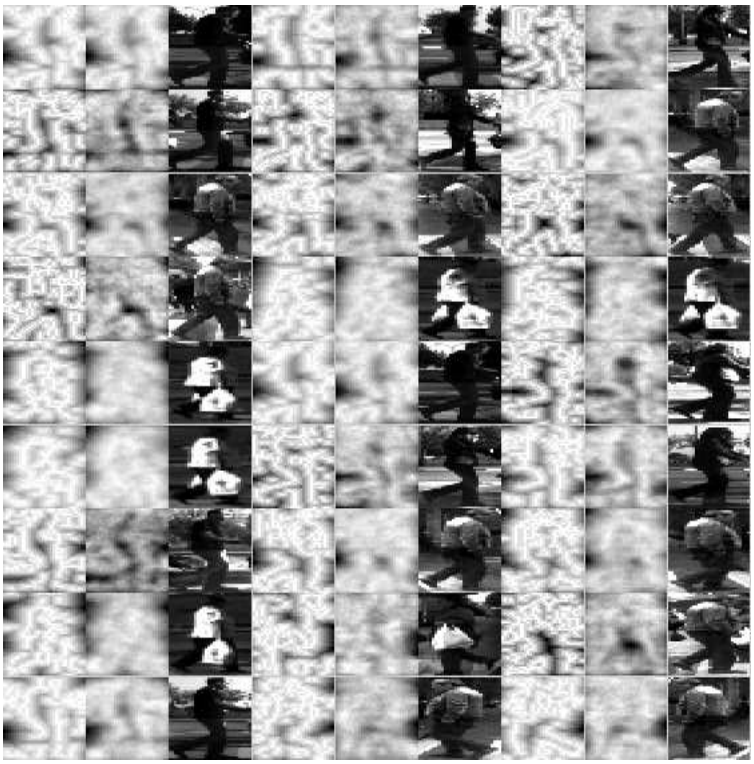


a

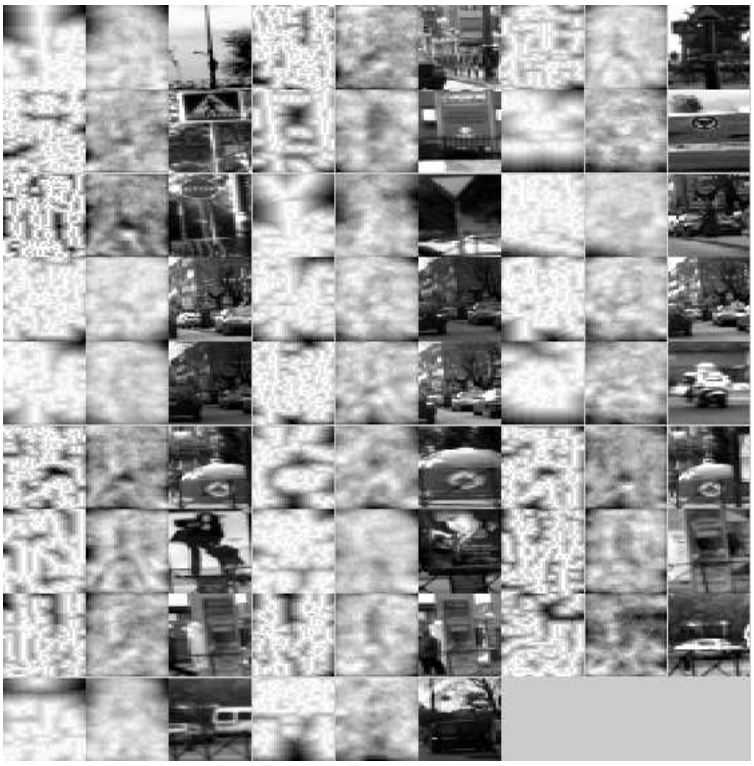


b

Figura 4.17: Resultados obtenidos después de la reconstrucción: la primera columna es la imagen de bordes, la segunda columna representa la imagen reconstruída y la tercera, la imagen original. (*a*) Contiene el proceso realizado a ROIs conteniendo a peatones y (*b*) a no peatones. Se observa que la reconstrucción trata de aproximar el resultado a un peatón, aunque el objeto detectado sea un no-peatón.

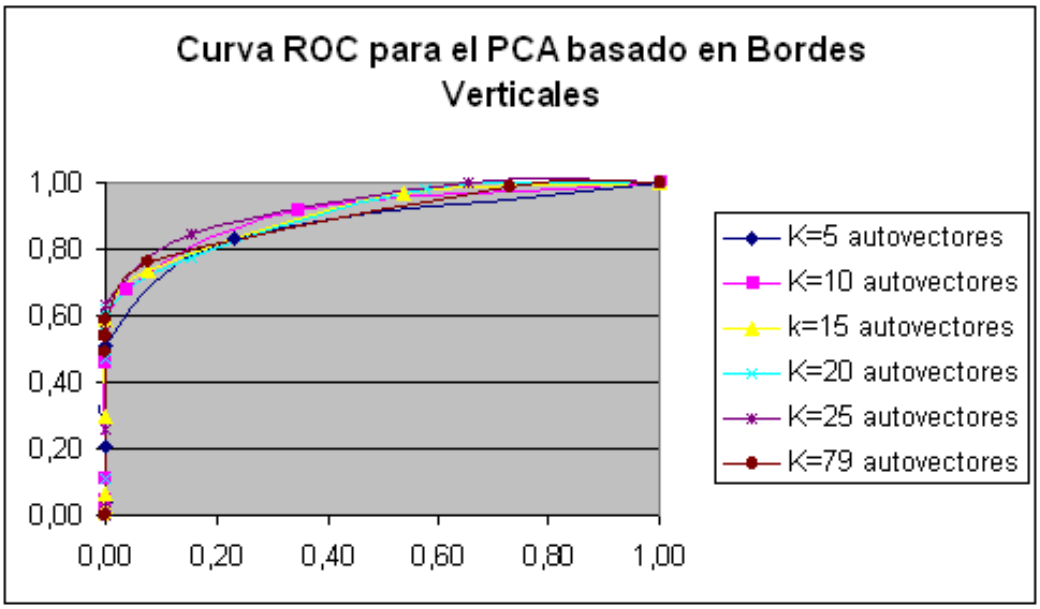


a

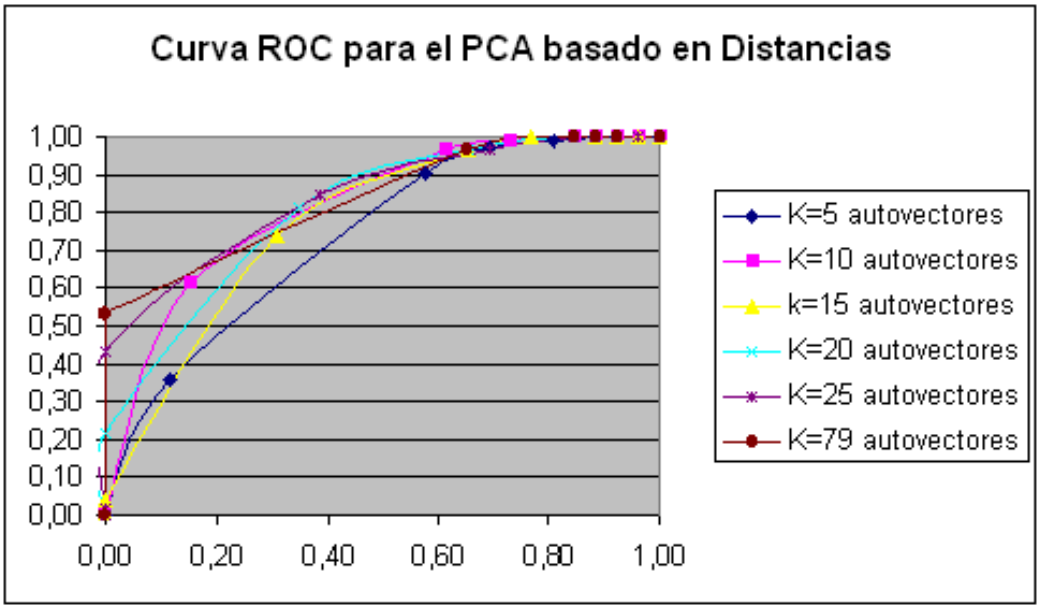


b

Figura 4.18: Resultados obtenidos después de la reconstrucción: la primera columna es la imagen de distancias, la segunda columna es la imagen reconstruída y la tercera, la imagen original. (*a*) Contiene el proceso realizado a ROIs conteniendo a peatones y (*b*) a no peatones.



a



b

Figura 4.19: Curvas ROC calculadas variando el número de autovectores considerados. Se han tomado valores para $M' = [5, 10, 15, 10, 25, 79]$. (*a*) Resultados obtenidos usando los bordes verticales como rasgos y (*b*) las distancias a los mismos.

basado en los bordes verticales alcanza unos porcentajes mayores (85 % aprox.) que el PCA basado en distancias (80 % aprox.), obteniendo un 92 % utilizando 25 autovectores. A pesar de decidir utilizar el PCA basado en bordes, cualquiera de los dos consigue una tasa alta de clasificaciones correctas.

Cabe resaltar que las imágenes almacenadas en el conjunto de entrenamiento no han sido seleccionadas, sino que el número de imágenes de peatones disponibles se ha dividido en dos conjuntos de manera arbitraria; un 54 % de las imágenes se han usado para el conjunto de entrenamiento, y el 46 % restante se ha empleado como conjunto de test, estando formado por 39 imágenes de cada fase del ciclo de caminar. En total se han evaluado 156 imágenes de peatones y 110 de no peatones.

Por último, se pueden mejorar los resultados considerando en el entrenamiento únicamente aquellas imágenes de peatones con unos bordes claros y precisos. Como consecuencia los componentes principales obtenidos representarían con mayor precisión a los peatones. En la figura 4.14 se observa como en algunos casos resulta difícil saber si la imagen contiene un peatón o no. A pesar de esas condiciones, los resultados obtenidos permite clasificar correctamente un porcentaje elevado (entorno al 80 %) de las ROIs, sin imponer ninguna restricción a la iluminación, posición o tamaño de los peatones.

4.8. Detección de la Forma Humana Basada en Contornos Activos

Los contornos activos son un método muy utilizado en el reconocimiento de patrones para extraer la forma de un objeto, aún cuando la misma no sea muy precisa. Su éxito radica en la capacidad para integrar conocimiento físico y topológico en el proceso de segmentación, pudiendo llevar a cabo una correcta interpretación de la imagen aunque la información de la escena sea insuficiente. Esta técnica de segmentación surgen de la mano de Kass, Witkin y Terzopoulos [KWT88] a finales de los 80. Su objetivo era desarrollar un método capaz de detectar contornos relevantes en la imagen mediante la minimización de la energía de una curva. A diferencia de la mayoría de las técnicas tradicionales del momento, su propuesta resultó ser innovadora al tratarse de un modelo activo, en el sentido de que la curva varía de forma por sí misma. Debido al modo en que evolucionan estas curvas son más conocidas como *snakes*. Un *snake* es una curva continua y elástica, que a partir de una posición inicial se deforma hasta ajustarse al rasgo deseado, influenciado por fuerzas externas e internas. En los últimos años, los contornos activos han sido extensamente empleados para llevar a cabo la segmentación de imágenes.

Las fases previas del sistema propuesto en esta tesis, proporcionan una lista de ROIs que encierran peatones potenciales. Durante la fase de reconocimiento, se han eliminado aquellas detecciones con una probabilidad baja de contener una forma humana. La lista de ROIs proporcionada a los *snakes* puede contener falsos positivos. Mediante el uso de los contornos activos, se pretende extraer la silueta o contorno del peatón. Si se analiza la forma extraída, se pueden filtrar esos falsos positivos, así como reconocer la actividad del peatón en base a la posición de las piernas. El sistema implementado en esta tesis no realiza el reconocimiento de

la postura en el dominio visible. Cumple su cometido de detectar peatones una vez ha obtenido la silueta humana.

A pesar de que los *snakes* son modelos flexibles que pueden resultar muy eficaces para obtener el contorno de un objeto no rígido, poseen dos inconvenientes importantes: primero, los resultados dependen de la posición inicial del modelo. Como se ha comentado, se debe tener una idea aproximada de la forma del objeto. El sistema aquí descrito, propone el uso de la visión estéreo como técnica de segmentación de objetos que proporcione esa forma aproximada al contorno activo. La solución adoptada consiste en situar un *snake* en el interior de cada ROI obtenida de la etapa anterior, habiendo segmentado previamente su contenido mediante el uso de mapas densos de disparidad. El segundo problema a resolver tiene que ver con la correcta selección de las energías, ya que de su formulación va a depender la manera en que se deforma el contorno.

4.8.1. Segmentación Basada en el Mapa de Disparidad Denso

Los sistemas estéreo en tiempo real aplicados a vehículos inteligentes [LAT02, FH02] han utilizado mayoritariamente, técnicas de visión estéreo no-densas. Se han desarrollado numerosos algoritmos basados en rasgos que buscan un subconjunto de puntos de interés en los que realizar la correspondencia para así alcanzar los requisitos de procesamiento en tiempo real. Pero, empleando únicamente la información de profundidad dispersa, la fase de segmentación de objetos resulta más complicada. Por ejemplo,si se extraen los bordes verticales de los objetos en la escena, resulta muy difícil determinar qué bordes pertenecen a un mismo objeto. Este hecho complica el uso de otras fases de procesamiento, como por ejemplo la clasificación o el seguimiento, ya que requieren de una cierta segmentación de la imagen. Por este motivo, resulta tan atractivo el uso de la visión estéreo densa, que trata de estimar la disparidad para todos los puntos en la imagen.

La visión estéreo densa ha sido aplicada a otros campos de investigación; la robótica móvil, los sistemas de vigilancia, la videoconferencia, el modelado 3D o la realidad virtual se han beneficiado de las potentes capacidades del estéreo denso para afrontar complicados problemas de la percepción basada en visión. La aplicación del estéreo denso a los vehículos inteligentes es sólo posible si el mapa de disparidad puede ser generado en tiempo real [KER06]. La técnica *una instrucción, múltiples datos* SIMD (*Single Instruction Multiple Data*) ofrece la posibilidad de acelerar el rendimiento del software, empleando una única instrucción para procesar múltiples datos simultáneamente. Debido a que el paralelismo sólo es respecto a los datos, se pueden evitar problemas más complicados como es la sincronización de procesos. En los últimos años, los procesadores de propósito general han sido dotados con capacidades SIMD en respuesta a las exigencias de las aplicaciones multimedia. Hoy en día, el conjunto de intrucciones SSE2 es una característica estándar disponible en la mayoría de las CPUs de propósito general, y es el que se ha utilizado en esta tesis. De hecho, se han evaluado varias implementaciones SIMD, cuyos resultados y conclusiones se exponen en este capítulo.

El algoritmo estéreo denso propuesto se basa en el desarrollado por la Universidad de Bolonia dentro del proyecto VIdeo DEcoder by Touch (VIDET) [SM02b]. Conocido como Videt Stereo Algorithm (VSA), consiste en un algoritmo basado en áreas, que permite generar

datos de profundidad en tiempo real gracias al desarrollo de estrategias de optimización que se van a comentar más adelante. Para más información ir al anexo A.

Las imágenes de 640x480 píxeles, adquiridas por el sistema binocular son preprocesadas como paso previo a la obtención del mapa de profundidad. Para compensar las diferencias de iluminación se han evaluado dos técnicas: el fitlro LOG (Laplaciana de Gaussiana) y la substracción en cada píxel del valor medio de intensidad calculado alrededor de cada píxel. Con la introducción de estos dos preprocesamientos se consigue eliminar la problemática de la luminosidad con lo que esto conlleva, pero se incorpora el hecho de que el valor de los píxeles puede ser tanto positivo como negativo. Esto sólo repercutirá en el tipo de dato a usar dentro del código pero, en principio, conduce al uso de tipos de datos de mayor tamaño, pudiendo ralentizar el proceso.

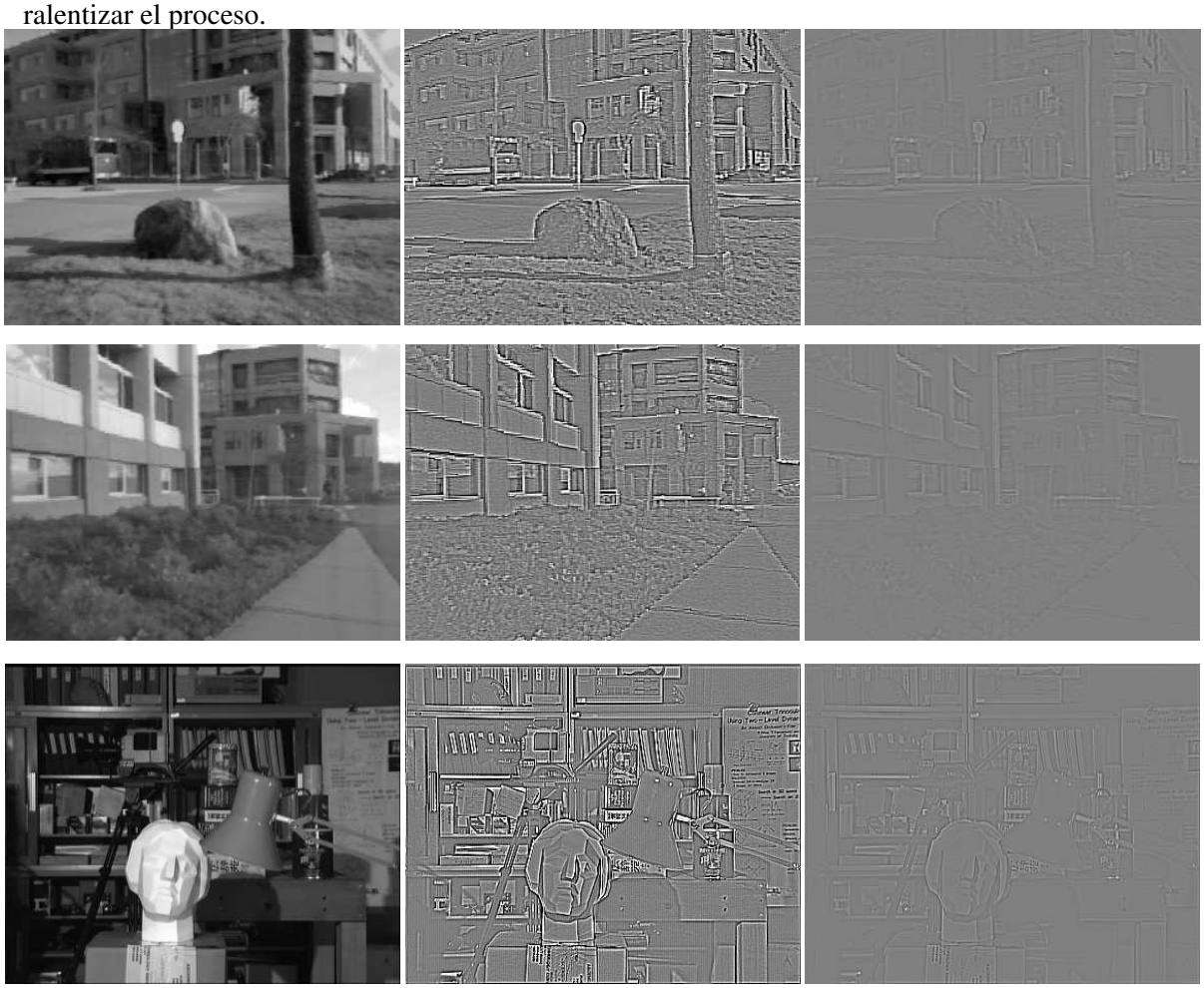


Figura 4.20: Preprocesamiento efectuados a las imágenes capturadas por el sistema binocular. La columna de la izquierda corresponde a la imagen izquierda, y las dos siguientes son el resultado de aplicar el filtro LOG y la resta de la media respectivamente.

En la figura 4.20 se muestran algunas imágenes tomadas por la cámara izquierda con sus correspondientes preprocesados, el método de la Laplaciana de la Gaussiana así como el

método de la resta de la media de la ventana. Todas ellas han sido procesadas utilizando una ventana de 8x8 y una disparidad máxima de 30 píxeles.

Como se puede observar el preprocesamiento a través del método de la Laplaciana de la Gaussiana proporciona imágenes con una mayor claridad que las del método de la resta de la media de la ventana. En un principio se podría pensar que esto conduciría a la obtención de mejores mapas de disparidad. Por contra el mapa de disparidad de los dos tipos de procesados viene a ser muy similar, con lo cualquiera de los dos métodos resultaría válido para llevar a cabo el tratamiento previo de las imágenes. Esto va a proporcionar la ventaja de poder elegir entre los dos métodos según resulte más cómodo o preferentemente más rápido en su ejecución.

4.8.1.1. Medidas de similitud

Al igual que para el cálculo del mapa no-denso, es importante seleccionar la medida de similitud que proporcione unos resultados buenos y al mismo tiempo sea el más apropiado desde el punto de vista de la implementación. Este último punto debe cuidarse en especial ahora que el mapa denso va a obtenerse empleando SIMD.

De entre los posibles métodos matemáticos, los que menos operaciones requieren son: el de la suma de las diferencias absolutas – tanto (SAD) (ver la ecuación 4.9) como el cálculo correspondiente con media cero (ZSAD) (ver la ecuación 4.11) – y el de la suma de las diferencias cuadradas que se define de forma parecida mediante las siguientes ecuaciones:

- Suma de diferencias cuadradas (*Sum of Squared Differences*, SSD):

$$SSD = \sum_{i,j} (I_{izda}(i,j) - I_{drcha}(x+i, y+j))^2 \tag{4.15}$$

- Suma de diferencias cuadradas con media cero (*Zero Mean Sum of Squared Differences*, ZSSD):

$$ZSSD = \sum_{i,j} ((I_{izda}(i,j) - \overline{I_{izda}}) - (I_{drcha}(x+i, y+j) - \overline{I_{drcha}}))^2 \tag{4.16}$$

Por una parte el método de la suma de las diferencias absolutas (SAD) se puede implementar con sumas, restas y comparaciones, mientras que para el método de la suma de las diferencias cuadradas se necesita además la realización de multiplicaciones (se puede observar que los métodos de la suma de diferencias absolutas con media cero y para el de la suma de diferencias cuadradas con media cero sólo difieren de los anteriores en la adicción de operaciones de resta, por lo que el concepto es el mismo). Debido a esto, se ha decidido utilizar como medida de correspondencia la suma de diferencias absolutas (SAD y ZSAD en cada caso) en vez de la suma de diferencias cuadradas (SSD y ZSSD), ya que se implementa con operaciones más sencillas y por lo tanto más rentables en cuanto a coste computacional.

4.8.1.2. Cálculo del SAD

La correspondencia basada en áreas se ha implementado según el cálculo SAD del algoritmo VSA [SM00]. Se trata de un esquema distinto a los trabajos anteriores del INRIA

[FhM+93] y del CMU [KYO+96], y permite llevar a cabo el consistente control izquierda-derecha sin ningún cálculo adicional de la función de correspondencia estéreo. Se alcanza una optimización operativa mediante el uso de cálculos incrementales dirigidos a los cómputos redundantes, minimizando así la cantidad de nuevos cálculos a ejecutar para cada valor de disparidad.

Partiendo de imágenes con niveles de luminosidad iguales o que hayan sufrido un preprocesamiento se puede directamente realizar dicho cálculo, representado mediante la ecuación $SAD(x, y, d) = \sum_{i,j=-n}^{n} |I_{izda}(x + j, y + i) - I_{dcha}(x + d + j, y + i)|$ del anexo (ver A.1). Para realizar esta operación es aconsejable utilizar funciones cuanto más simples mejor para que su coste operacional sea mínimo, repercutiendo en unos tiempos de operación más reducidos. Esto se traduce en realizar simples operaciones de sumas y restas así como de comparaciones. El valor absoluto demandado se podría implementar entonces a través de la siguiente fórmula:

$$|a - b| = máximo(a - b, b - a) \tag{4.17}$$

Se podría pensar aquí que el uso de valores con signo seguiría siendo necesario para el cálculo de las restas ya que éstas darán como resultado valores negativos. Por el contrario, si se observa lo que se pretende conseguir, se ve que no es así. Si se trabaja con valores sólo positivos al hacer la resta y obtener un valor negativo éste, por desbordamiento, se ajustaría al valor mínimo de un valor positivo, el cero. Esto podría presentar un problema a la hora de hacer la comparación, pero como lo que se busca es el valor máximo entre las dos restas, una de las dos restas será siempre positiva, o si a y b tiene valores idénticos, valdrá cero. Así, en el caso de obtener el valor negativo, éste se ajustará a cero y será el mínimo de los dos valores de todas formas. Por ello se podría trabajar sólo con valores positivos ya que no presentaría ningún problema.

4.8.1.3. Búsqueda de la disparidad

El espacio de disparidad contiene todas las posibles correspondencias para una misma fila en la imagen izquierda y derecha. Las posibles correspondencias para un punto en la fila izquierda forman una columna en este espacio, mientras que las posibles correspondencias para un punto en la fila de la derecha forman una fila (ver imagen 4.22). El algoritmo desarrollado sólo realiza la búsqueda de izquierda a derecha, limitando el espacio de búsqueda entre una disparidad mínima de 4 píxeles y un máximo de 20. De este modo, como ya se ha comentado antes, se cubre un rango de distancias entre 3 y 15 metros aproximadamente.

Una vez calculado el valor de la suma de las diferencias absolutas de cada uno de los píxeles de las imágenes para el rango de 4 a 20 disparidades, se busca el valor óptimo bajo un enfoque WTA; se selecciona la disparidad para cada píxel cuyo valor de la suma de las diferencias absolutas se hace mínimo dentro del rango. Como consecuencia se obtiene el mapa denso de disparidad entre las dos imágenes de partida. Este modo de operar es el mismo que se utilizó para la obtención del mapa no-denso.

A la hora de realizar la implementación del método del SAD, éste podría llevarse a cabo a través de sumas, restas y comparaciones en código C, recorriendo todos los píxeles de las

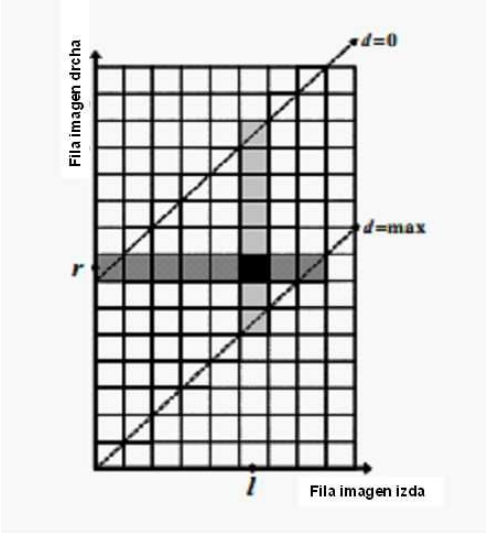


Figura 4.21: Espacio de búsqueda de la disparidad [KER06]. La columna gris clara corresponde al área de búsqueda de izquierda-a-derecha, mientras que la fila gris oscura corresponde al área de búsqueda de derecha-a-izquierda.

imágenes en búsqueda de la disparidad mínima entre estos. Sin embargo, los métodos de fuerza bruta son computacionalmente prohibitivos y desde luego, inviables para su ejecución en tiempo real. El algoritmo propuesto ha sido optimizado aprovechando las instrucciones paralelas orientadas a multimedia del estilo SIMD proporcionadas por la tecnología MMX. Más concretamente, y debido al estado actual de los microprocesadores, se han utilizado las ampliaciones de flujo SSE2 incorporadas en el Pentium 4. Por otra parte estas ampliaciones incorporan intrínsecos, que son funciones predefinidas que trabajan sobre código fuente (están escritas en C) y facilitan la utilización de las ampliaciones de flujo SSE2 al liberar a los programadores de tener que programar en lenguaje ensamblador y de controlar los registros. Además, el compilador optimiza la programación de las instrucciones por lo que los ejecutables funcionan más rápido. ver el anexo A) para más detalles.

4.8.1.4. Optimizaciones al cálculo de la correspondencia

Inicialmente, se ha implementado la estrategia de correspondencia estéreo íntegramente en C, aplicando las optimizaciones para el cálculo del SAD propuestos en el algoritmo VSA. Este método resulta fácilmente implementable pero da unos resultados de bajo rendimiento. Por ello, se introdujo la primera mejora al código a través del uso de *intrínsecos*.

- **Detalles de implementación:**

  Al implementar el cálculo del SAD utilizando los *intrínsecos* de las ampliaciones de flujo SSE2 se observa que éstos utilizan datos de 128 bits. Por otra parte las imágenes tratadas presentan los valores de los píxeles con 256 niveles de gris, lo que se traduce en formato tipo *char*, de 8 bits (1 byte). Con esto se llega a la conclusión de que se

podría trabajar con 16 píxeles en paralelo. Por el contrario, recordando la necesidad de trabajar con datos con signo se hace necesario utilizar datos tipo *short* (2 bytes), lo que reduce a 8 los valores con los que se puede llegar a trabajar en paralelo. Partiendo de estas consideraciones previas se realizó en un inicio, el cálculo del mapa de disparidad utilizando una ventana de 8x8 píxeles. Para el cálculo de los valores SAD de una fila de la ventana, se cargan los 8 píxeles de la imagen de la derecha y después, los 8 de la imagen izquierda, utilizando el intrínseco $\_mm_{l}oadu_{s}i128$ en cada caso. Son únicamente necesarias dos restas $a - b$ y $b - a$ y calcular el valor máximo de los resultados para calcular 8 diferencias absolutas. La suma de los valores absolutos requiere ir sumando todas las diferencias absolutas calculadas (ver el anexo A.2.1.1 para más detalles).

A pesar del cambio, los tiempos de ejecución obtenidos son prácticamente similares a la implementación en C, lo que obliga a la introducción de la segunda optimización mediante el uso de *intrínsecos* de más alto nivel. En concreto, se usa el $_{m}m_{s}ad_{e}pu8$, que a partir de dos grupos de datos de 16 valores cada uno, devuelve dos datos distintos de 64 bits cada uno (ver el anexo A.2.1.2 para más detalles). El problema reside en que nuestros datos son de 16 bits y este intrínseco trabaja con datos de 8 bits.

La utilización de datos de 16 bits se debía a la necesidad de trabajar con valores positivos y negativos, ocasionado como resultado del preprocesamiento aplicado a las imágenes. Sumando a todos los píxeles de la imagen +127 se consigue que todos tengan un valor comprendido entre 0 y 256. Partiendo de la ecuación A.1 se ve claramente que esto no entraña ningún problema a los cálculos de la disparidad ya que el valor de SAD seguirá siendo el mismo.

$$\begin{aligned} SAD\,(x,y,d) &= \sum_{i,j=-n}^{n} |I_{izda}\,(x+j,y+i) - I_{dcha}\,(x+d+j,y+i)| = \\ &= \sum_{i,j=-n}^{n} |(I_{izda}\,(x+j,y+i)+127) - (I_{dcha}\,(x+d+j,y+i)+127)| \end{aligned} \tag{4.18}$$

De esta forma se consigue trabajar con datos de 8 bits permitiendo la utilización del intrínseco antes mencionado. Para el cálculo de los valores SAD de una fila de la ventana, se cargan los 8 píxeles de la imagen de la derecha y después, los 8 de la imagen izquierda, igual que antes. Pero, el cálculo del SAD se realiza con una única instrucción; a saber, el intrínseco $_{m}m_{s}ad_{e}pu8$.

- **Tamaño de la ventana de búsqueda:**

  Los tipos de datos utilizados por los intrínsecos tienen un tamaño de 128 bits, por lo que se podría llegar a realizar operaciones con 16 datos de 8 bits en paralelo. Con la necesidad de utilizar datos de 16 bits esta posibilidad quedaba anulada, pero si se tratan las imágenes después del preprocesamiento mediante 4.18 se logra trabajar con valores positivos de 8 bits, permitiendo ampliar la ventana al tamaño de $16x16$ píxeles. Así se consigue que el cálculo del mapa de disparidad sea más denso, ya que se opera con más información en los alrededores del píxel sobre el que se realiza el cálculo.

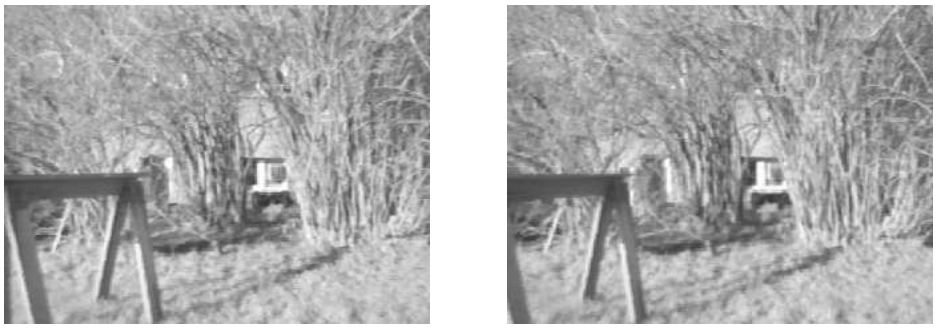


Figura 4.22: Imagen izquierda y derecha sobre la que se han realizado cálculos del SAD (ver fig 4.23 para distintos tamaños de ventanas.

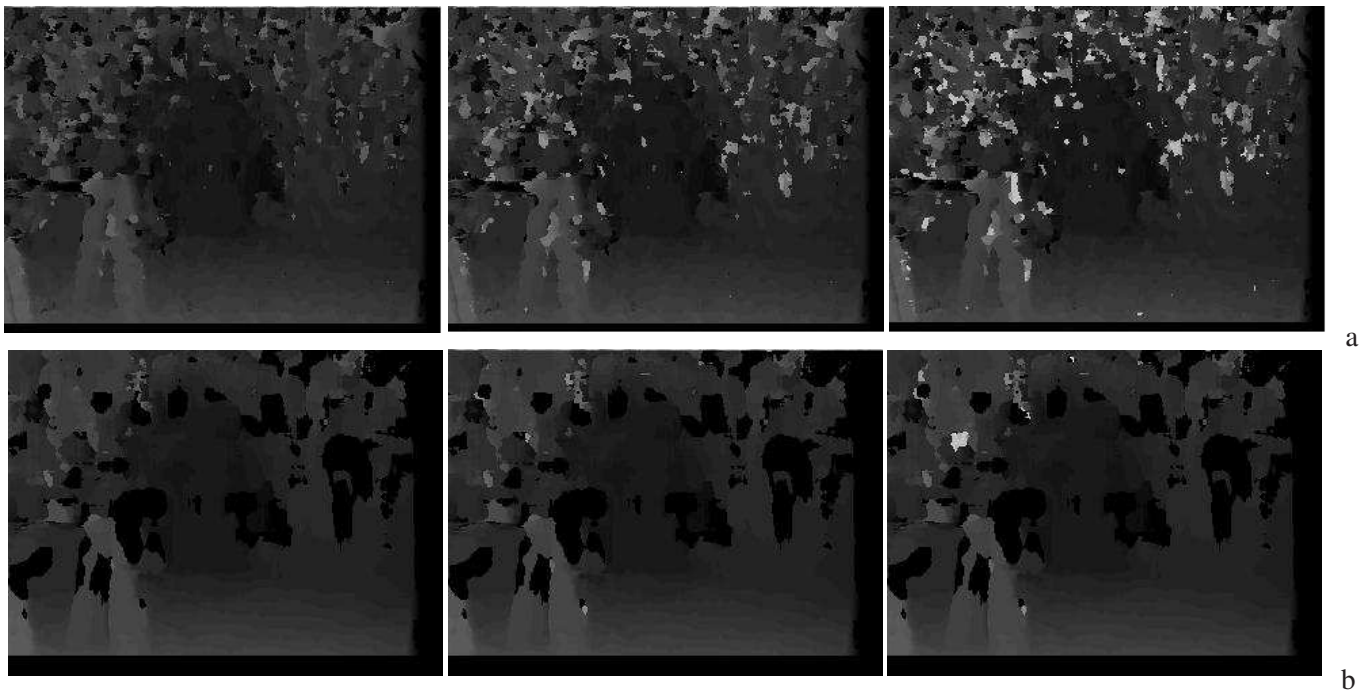


Figura 4.23: Mapas de disparidad obtenidos para distintos tamaños de ventanas, siendo la disparidad máxima de cada columna 20, 30 y 40 píxeles respectivamente; (*a*) Resultados para la implementación con ventana 8x8 y (*b*) con ventana 16x16.

Partiendo de la posibilidad de trabajar con valores positivos, se logra operar con 16 datos de 8 bits en paralelo. En un primer momento, esto permite el uso de una ventana de 16x16 pero por otra parte, esta posibilidad puede ser también aprovechada para ventanas dobles de 8x8 píxeles, lo cuál puede ser interesante si no se pretende conseguir un mapa de disparidad tan denso. En la figura 4.24 se aprecian las dos ventanas de 8x8, una en gris claro que va de $x-n$ a $x+n$ y otra en color más oscuro que va de $x-n+8$ a $x+n+8$,

que se implementa utilizando 8 columnas, cada una con 16 valores.

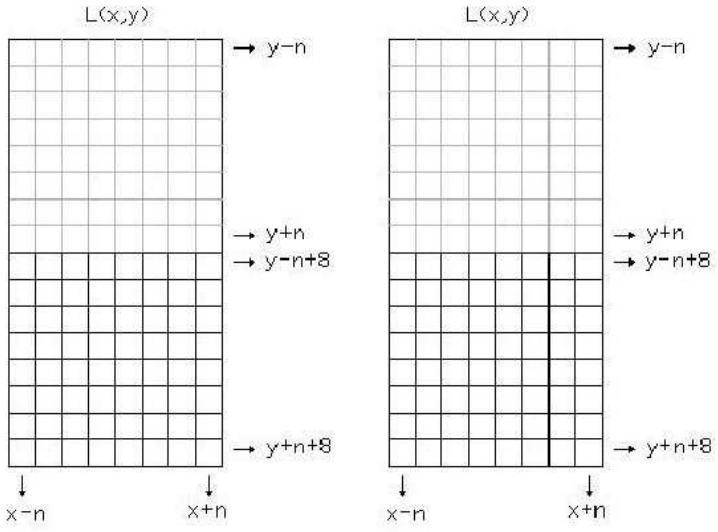


Figura 4.24: Método de dos ventanas de 8x8

Cuando se trabaja con una ventana de 8x8 al cargar los datos se almacenan 16 valores, 8 valores que se utilizan y otros 8 que no. Esto representa una falta de eficiencia a la que se suma el hecho de que el intrínseco que nos permite realizar el cálculo del valor de SAD, $\_mm_sad_epu$, opera esos 8 valores no utilizados devolviendo la suma de las diferencias absolutas correspondientes. Ante este caso cabe la posibilidad de intentar aprovechar el resultado de estas operaciones. Esto puede llevarse a cabo utilizando dos ventanas en paralelo de 8x8. Si la carga de datos se realiza por columnas se tendrá dos ventanas de 8x8 una encima de la otra, tal y como se observa en la figura 4.24.

Si se aplica el método de recorrido por la matriz hasta ahora utilizado, estas ventanas se irán desplazando en horizontal hasta llegar a la $x_{máxima}$, para posteriormente pasar a la posición $x = 0, y = y + 1, d = 0$. La diferencia en el método se dará cuando se pase a la posición $y =$ (8 y sus múltiplos). En este caso el cálculo de valor de SAD ya habrá sido calculado por la segunda ventana así como los cálculos de valores SAD asociados a los siguientes 8 valores de y. Con esto se consigue reducir a la mitad el número de operaciones necesarias, decreciendo considerablemente el tiempo de procesado de las imágenes.

4.8.1.5. Análisis de los resultados estéreo-denso

En este apartado se analizan las distintas implementaciones del cálculo del mapa de disparidad anteriormente expuestas, con el objetivo de seleccionar la que ofrece mejores ventajas en cuanto a la velocidad de procesamiento se refiere. Se ha realizado una comparativa entre los diferentes métodos de correlación con respecto al tamaño de ventana que utilizan y en cuanto al valor de disparidad máxima utilizado. Se ha utilizado un procesador Intel Pentium 4 a 3000

MHz con 1 Gbyte de memoria. El tamaño de las ventanas consideradas es de 8x8 y de 16x16 y los valores de disparidad máxima varian entre 20, 30 y 40 píxeles. Los resultados mostrados en la tabla 4.4 se han obtenido procesando imágenes de 320x240 píxeles.

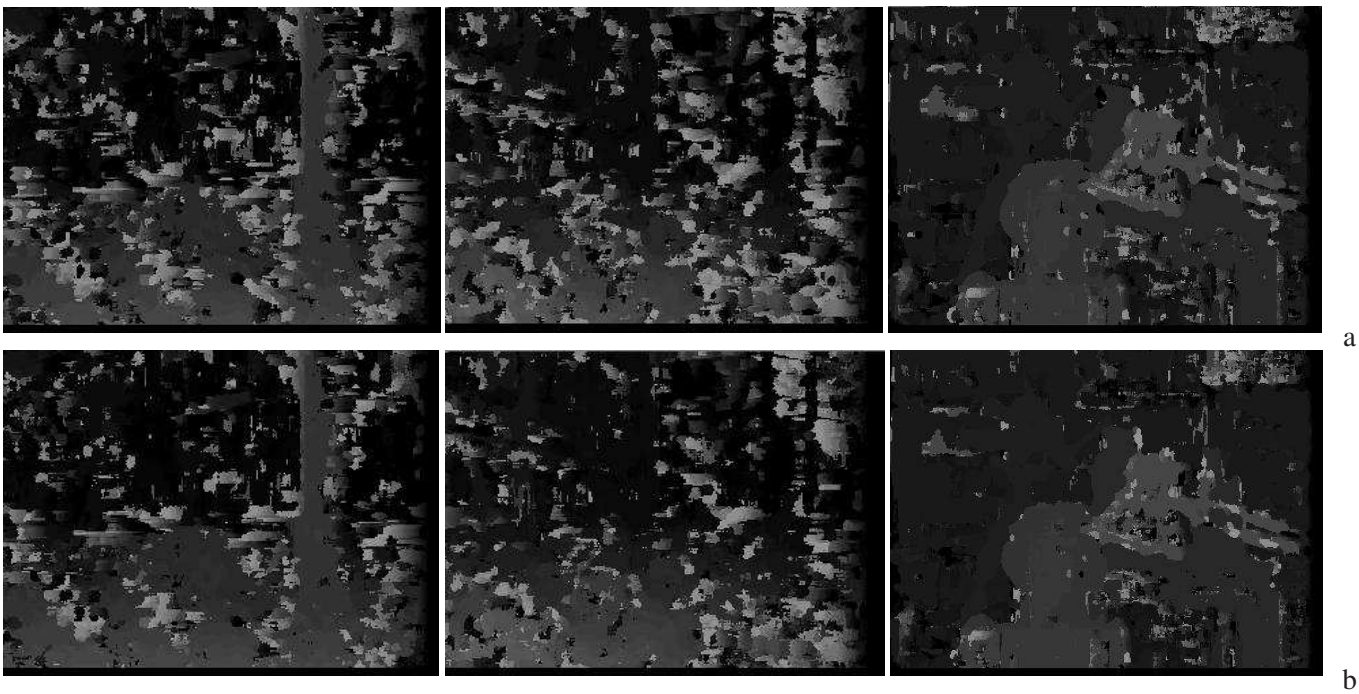


Figura 4.25: Mapas de disparidad obtenidos a partir de las imágenes preprocesadas (*a*) con el filtro de la Laplaciana y (*b*) mediante la resta de la media. Las imágenes originales son las mostradas en la figura 4.20.

Método utilizado Disparidad máxima	20	30	40
Sin intrínsecos 8x8	94 ms	125 ms	157 ms
Con intrínsecos 8x8 sin la función ${}_m m_s ad_e pu8$	86 ms	110 ms	156 ms
Con intrínsecos 8x8 con la función ${}_m m_s ad_e pu8$	31 ms	38 ms	47 ms
Con intrínsecos doble 8x8 con la función ${}_m m_s ad_e pu8$	16 ms	24 ms	31 ms
Sin intrínsecos 16x16	141 ms	218 ms	250 ms
Con intrínsecos 16x16 con la función ${}_m m_s ad_e pu8$	109 ms	156 ms	218 ms

Tabla 4.4: Tabla de tiempos de procesamiento del SAD (en ms.)

Un análisis de los resultados permite concluir que las implementaciones que usan intrínsecos son considerablemente más rápidas que el resto. Además, con cada sucesiva mejora introducida se ha conseguido reducir el tiempo de operación, demostrando que las mejoras son tales. Se podría remarcar que al aumentar el valor de disparidad máxima aumenta también el tiempo de procesado, lo que es fácilmente explicable ya que un mayor valor de la disparidad máxima conlleva más cálculos en procesado. Además, los métodos que usan ventanas de 16x16 píxeles son más lentos. Esto es debido a que necesitan un mayor un número de cálculos

iniciales, así como durante el proceso, ya que el tamaño de la ventana es mayor. En cuanto al método más rápido es el último método desarrollado, a saber, la implementación que usa una doble ventana de 8x8 es la más rápida. Como se ha explicado anteriormente, optimiza el cálculo de valores SAD siendo capaz de reducir prácticamente a la mitad los tiempos de operación. Éstos entran dentro del marco de tiempo real, por lo que se ha decidido utilizar este método para el cálculo de los mapas densos.

En cuanto a los preprocesamientos utilizados se llega a la conclusión de que cualquiera de los dos puede ser usado ya que proporciona unos resultados similares en cuanto a la calidad del mapa de disparidad. Entonces sólo se deberá tener en cuenta cuál de los preprocesamientos resulta más fácil de implementar, proporcionando unos tiempos de preprocesamiento menores. Teniendo en cuenta la diferenciación por tamaño de ventana se observa que, el uso de ventanas de mayor tamaño va a producir unos mapas de disparidad más densos ya que recogen más información del entorno. Por contra, esto va a conducir a unos tiempos de cómputo mayores, por lo que habrá que trabajar con el equilibrio calidad-tiempo. Dentro del ámbito de disparidad máxima permitida es de subrayar que este valor dependerá de las imágenes con que se trate, o lo que viene a ser lo mismo, del entorno en que trabaje el sistema. Por ello este valor ha de ser elegido a través de una serie de pruebas para ver cuál es el más conveniente en cada caso. Llega un momento en que aunque se de una valor de disparidad máxima permitida mayor no se va a conseguir mejores mapas de disparidad, sino que sólo conducirá a una ralentización del procesado.

De todo esto se obtiene como conclusión final la importancia de los ajustes de los valores, a través de un estudio previo del ambiente en que va a estar situado el sistema para conseguir optimizar su rendimiento. Como consecuencia, se ha establecido una disparidad máxima de 20, tal y como se había establecido para la obtención del mapa no-denso. Este rango de disparidad ha demostrado ser suficiente para la detección de peatones que se encuentren a una distancia máxima de unos 15 metros. En cuanto al preprocesamiento de las imágenes adquiridas por el sistema binocular, se ha optado por realizar la resta de las medias, siendo también la misma solución que la adoptada para la compensación de la iluminación del mapa no-denso.

4.8.1.6. Postprocesamiento basado en *snakes*

Una vez que se ha realizado el cálculo de la disparidad, es necesario llevar a cabo un post-procesamiento. Esto se debe a que los algoritmos estéreo son propensos a errores cuando hay poca textura o patrones repetitivos en la imagen, fallan ante oclusiones y la estimación de la disparidad es limitada. En otros trabajos, se afrontan estos problemas mediante una detección de errores, interpolación subpíxel y eliminación de oclusiones [FTV00, HIG02].

Sin embargo, en esta tesis no se ha optado por ninguna de ellas. Esto es debido a que el resultado del algoritmo estéreo, en sí mismo, no es importante para el sistema de ayuda a la conducción que se persigue construir. Es la subsiguiente fase de segmentación la que puede aportar información interesante sobre la presencia de un peatón. El estéreo-denso se utiliza como método para guiar la segmentación basada en contornos activos.

4.8.2. Extracción de la silueta mediante Contornos activos o *Snakes*

En este trabajo se ha seguido el enfoque clásico propuesto por Kass et al. [KWT88] empleando una representación paramétrica de los contornos, describiendo cada curva a través de la ecuación $v(s) = (x(s), y(s))$, donde s es la longitud normalizada en el rango $[0, 1]$. Se ha seleccionado esta representación por su sencillez y facilidad a la hora de interpretar la deformación que sufre el modelo, ya que sólo hay que analizar los cambios sufridos por un conjunto discreto de puntos.

Los puntos de la curva sufren los efectos de un campo generado por un conjunto de fuerzas, siendo éstas el único control que se tiene sobre el movimiento de los puntos; por este motivo, la capacidad del *snake* de encontrar el contorno deseado en la imagen sólo va a depender de la cuidadosa selección de dichas fuerzas. Cada punto del *snake* se va a mover hasta alcanzar una posición donde exista un mínimo de energía asociado con ese campo.

Las fuerzas y sus energías asociadas, pueden dividirse en dos categorías distintas: internas y externas.

4.8.2.1. Formulación de la energía interna

Son las responsables de la topología del modelo. El cometido de estas energías es tratar de evitar que se produzcan discontinuidades y deformaciones en la forma del modelo, es decir, tratan de imponer cierta condición de suavidad y continuidad a la curva.

Cabe decir que, en una implementación clásica (paramétrica) de contorno activo, la forma y el muestreo de los vértices están interrelacionados (ver fig. B.1 del anexo B). De manera que, cambios en la parametrización del modelo afectan a su forma. Por tanto, es fundamental controlar el muestreo de los puntos de la curva. La energía del contorno propuesta se calcula separando la componente tangencial (controla el muestreo) y la componente normal (controla la forma) de la fuerza interna.

- Componente tangencial de $f_{interna}$:

  Los puntos del contorno se van a mover en la dirección de un vector local tangente $\overrightarrow{t_i}$ cuya dirección es la línea que une dos puntos vecinos, tratando de que la separación entre los puntos se aproxime a la distancia media entre los puntos del contorno (ver fig. 4.26). Esta idea fue propuesta por Williams y Shah, quienes modificaron ligeramente la aproximación por diferencias finitas (ver anexo B.4) empleada por Kass et al., a fin de evitar la prematura contracción del contorno.

  Para mantener una parametrización uniforme, se ha considerado el vector tangente al vecino anterior y posterior de cada vértice:

$$\begin{aligned} \overrightarrow{t_i} &= \overline{dist} - |v_i - v_{i-1}| \\ \overrightarrow{t_{i+1}} &= \overline{dist} - |v_{i+1} - v_i| \end{aligned} \tag{4.19}$$

  De este modo se trata de mantener una separación uniforme entre los puntos. Por otro lado, además de controlar la separación entre los puntos, se puede actualizar el número

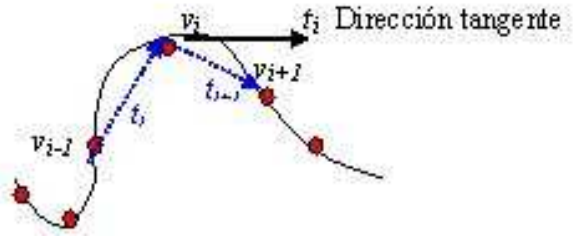


Figura 4.26: Modo en el que se calcula la tangente al vecino anterior y posterior del vértice actual i.

de puntos del contorno, insertando puntos si la separación entre 2 vértices vecinos es superior a un umbral o bien, eliminando puntos en caso contrario.

- Componente normal de $f_{interna}$:

  La componente normal está relacionada con la estimación de la curvatura propuesta por [WS92] y [KWT88], representada mediante la siguiente fórmula:

$$curvatura = k_i \approx |v_{i-1} - 2v_i + v_{i+1}|^2 \tag{4.20}$$

  Es importante destacar cómo las variaciones en la parametrización afectan de un modo importante a aquellas zonas donde la curvatura es alta (ver fig. 4.27. Pero si los puntos están distribuidos de manera uniforme, la aproximación de Williams y Shah [WS92] de la curvatura da buenas estimaciones.

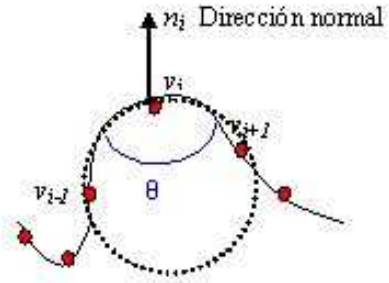


Figura 4.27: Detalle del cálculo de la curvatura. Es aproximada con el cálculo de la segunda derivada a lo largo del contorno.

Resumiendo, la evolución de un contorno discreto puede escribirse descomponiendo la fuerza interna en su componente tangencial y normal, como sigue:

$$\begin{aligned}(f_{interna})_i &= \left(f_{tangencial}\right)_i + (f_{normal})_i \\ &= \left|\overline{dist} - |v_i - v_{i-1}|\right| + \left|\overline{dist} - |v_{i+1} - v_i|\right| + |v_{i-1} - 2v_i + v_{i+1}|^2 \\ &= \left|\overrightarrow{t_i}\right| + \left|\overrightarrow{t_{i+1}}\right| + k_i \cdot \overrightarrow{n_i}\end{aligned} \tag{4.21}$$

El contorno alcanza el equilibrio, cuando ambas componentes se hacen cero. Recordando que se emplea un enfoque de minimización de la energía de la curva, a la hora de implementar el *snake*, se busca minimizar la energía de cada componente.

La componente tangencial aporta continuidad de tipo 0 a la curva, que simplemente supone la unión de 2 segmentos de la curva. El gradiente de esos segmentos en el punto de unión así como su curvatura puede ser diferente. Se minimiza la 1ª derivada a lo largo del contorno.

La componente normal, mediante el cálculo de la curvatura proporciona continuidad de tipo 1, ya que garantiza que los gradientes en el punto de unión de dos segmentos de curva son idénticos. Es decir, la tangente es ese punto es la misma. Esto se consigue minimizando la 2ª derivada (la 1ª derivada o gradiente es constante, por tanto).

A la hora de controlar la forma del modelo, la componente normal de la energía interna es la que más peso tiene. Sin embargo, hay que cuidar su formulación para evitar efectos secundarios no deseados. En este trabajo, se busca minimizar la 2ª derivada a través de la fórmula 4.20, o lo que es lo mismo, se está aplicando un suavizado Laplaciano a la curva (ver el apartado B.2 del anexo B). Como consecuencia, se está minimizando la energía elástica del contorno, por lo que el contorno tiende a reducir su capacidad para estirarse y se encoge.

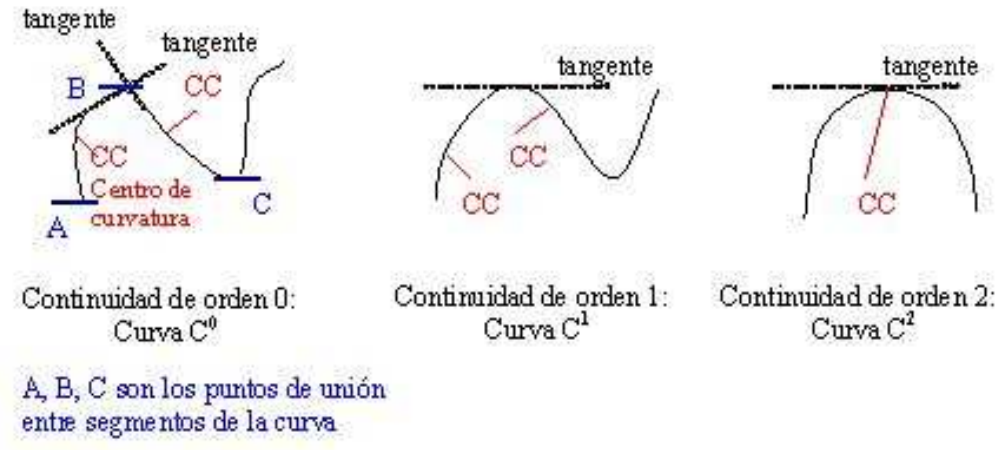


Figura 4.28: Tipos de continuidad.

Hasta aquí, se persigue obtener un espaciado uniforme de los puntos, restringido a la dirección tangente y cambios suaves de curvatura, en la dirección normal. Sin embargo, considerando el caso a tratar, un peatón muestra ciertos rasgos en su forma que conviene explotar. Hay zonas de su contorno que presentan la misma curvatura, como pueden ser la cabeza y los pies. Así, se propone imponer continuidad de tipo 2 a la curva, que garantiza precisamente continuidad de curvatura; esto es, tanto el gradiente (o la tangente) como el centro de curvatura es la misma (ver fig. 4.28).

La nueva fuerza de regularización se obtiene de la siguiente forma:

$$concavidad = c_i \approx |v_{i-2} + 4v_{i-1} + 6v_i - 4v_{i+1} + v_{i+2}|^2 \tag{4.22}$$

Como ventaja añadida, al calcularse la 4ª derivada a lo largo del contorno, esta fuerza ocasiona un menor encogimiento que el suavizado Laplaciano anterior. A efectos prácticos, para evitar calcular esta energía en todos los puntos, su cálculo se restringe a aquellos segmentos de la curva donde existe cierta suavidad de curvatura, o lo que es lo mismo, donde existe continuidad de tipo 1.

La formulación de la energía interna empleada queda de la siguiente forma:

$$\begin{aligned}(f_{interna})_i &= \left(f_{tangencial}\right)_i + (f_{normal})_i + \left(f^*_{normal}\right)_i \\ &= \left|\overrightarrow{t_i}\right| + \left|\overrightarrow{t_{i+1}}\right| + k_i.\overrightarrow{n_i} + c_i.\overrightarrow{n_i}\end{aligned} \tag{4.23}$$

4.8.2.2. Formulación de la energía externa

Son las responsables de empujar al contorno hacia las zonas deseadas de la imagen. La imagen es interpretada como un campo de potenciales, siendo los mínimos locales los que atraen al *snake*. Según como esté definida la función del potencial, el contorno será atraído hacia ciertos rasgos.

No obstante, un gran inconveniente de los potenciales es que su efecto tiene carácter local. Sin embargo, si una parte del *snake* encuentra algún rasgo de la imagen con la suficiente baja energía, arrastra al resto del *snake* hacia esa zona. Este efecto puede aumentarse aplicando un filtro de suavizado a la imagen, de manera que el potencial obtenido de su procesamiento tenga un carácter menos local. Otra forma de conseguir que el *snake* vaya a un mínimo local, consiste en realizar la búsqueda a varios niveles de resolución, empezando por una imagen de baja resolución, donde el *snake* se aproxima a una zona con un mínimo local, para luego, refinar el ajuste del contorno a una resolución mayor.

En este trabajo se ha aplicado un filtro gaussiano a cada una de las ROIs heredadas de la clasificación basada en PCA, antes de calcular el campo potencial generado por los rasgos contenidos en su interior. A continuación se exponen cuáles son los rasgos de cada ROI hacia los que interesa que el *snake* sea atraído:

- Bordes de la imagen:

  Sin duda, el uso más común que se ha dado a los contornos activos ha sido el de detector de bordes. Para ello, se puede definir un potencial basado en el gradiente:

  $$P_{bordes}(x, y) = |\nabla (G(x, y) * I(x, y))| \tag{4.24}$$

  donde se aplica una Gaussiana a la imagen para detectar los cambios de intensidad grandes y el campo de potenciales se construye en términos de variaciones de intensidad en valor absoluto (ver figura 4.31). Así se consigue mayor robustez ante ruido y cambios de iluminación.

  Sin embargo, guiarse por la energía de los bordes da muchos problemas; el *snake* puede ir hacia bordes débiles o verse distraído de los rasgos de interés por otros rasgos menos interesantes, pero más fuertes. Es necesario incluir algún término más en la especificación de las fuerzas externas. En los siguientes apartados se detallan los 2 términos que se han añadido en este caso.

Figura 4.29: Pasos para la obtención del potencial debido a los bordes en la imagen ; (*a*) Efecto de la gaussiana aplicada a la imagen original y (*b*) Gradiente de los bordes obtenidos de la imagen suavizada y filtrada mediante las disparidades del mapa estéreo.

- Distancias a bordes verticales:

  Para evitar que el *snake* se vea atraído por bordes que no interesan, hay que definir una formulación más exacta y más restrictiva. A la hora de detectar peatones, uno de los rasgos más característicos es el predominio de bordes verticales frente a los horizontales en su silueta. Por tanto, parece obvio que hay que incluir esa cualidad en la definición de la energía.

  A fin de evitar el carácter extremadamente local de los campos de potenciales tradicionales, se define un potencial que se extiende de manera suave a lo largo de una larga distancia. Para ello se emplean mapas de distancias a los rasgos que nos interesan. Para los bordes verticales, interesa saber la distancia a la que está en *snake* de un borde de ese tipo. Así, el *snake* se ve afectado no sólo por los rasgos que le rodean.

  El campo de potencial basado en el mapa de distancias a los bordes verticales se obtiene de:

  $$P_{dist\_bordes\_vert}(x,y) = dist_{CHAMFER}((x,y), P_{borde\_vert}) \tag{4.25}$$

  Este mapa muestra para un píxel (x, y) de la imagen, la distancia al borde vertical más cercano. La métrica empleada es la distancia chamfer.

- Distancias a los ejes de simetría:

  Con los 2 potenciales anteriores obligamos a que el contorno se vea atraído por todo tipo de bordes, pero exigimos que cuanto más cerca esté de un borde vertical, mayor sea la fuerza de atracción. Además de esta condición, se impone otra, que restringe aún más la acción de las fuerzas externas; se explota el hecho de que los peatones presentan una considerable simetría vertical. Ya se hizo uso de esta heurística en una fase previa, para decidir dónde sembrar los *snakes*. Por tanto, se genera un mapa de distancias a partir de los ejes de simetría obtenidos por el módulo estéreo.

  La fórmula de este nuevo potencial es la siguiente:

  $$P_{dist\_sim\_vert}(x,y) = dist_{CHAMFER}((x,y), P_{ejes\_vert}) \tag{4.26}$$

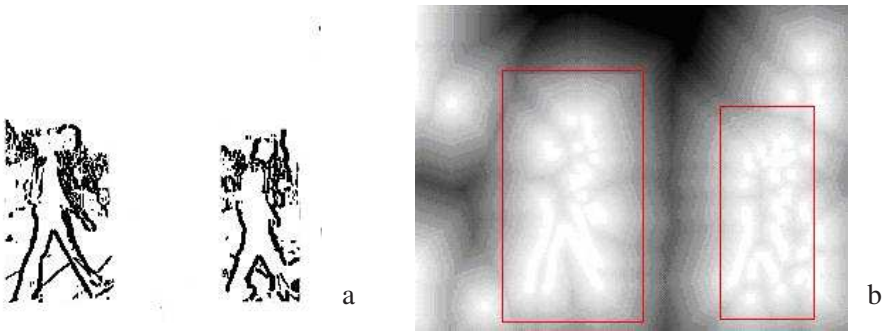


Figura 4.30: Pasos para la obtención del potencial debido a las distancias a los bordes verticales (*a*) Gradiente de los bordes verticales en la imagen sin filtrar (*b*) Mapa de distancias a esos bordes.

La interpretación es análoga a la fórmula (25). El mapa de distancias muestra para cada píxel (x, y) la distancia el eje más cercano.

- Distancias a bordes:

  Si no se le impone alguna restricción más al modelo, puede deformarse en exceso. Para impedirlo, se permite su deformación hasta que alcance un borde. Para ello se construye un mapa de distancias que combina información sobre la distancia a la que están los ejes verticales con información sobre la distancia a la que están los bordes, obtenida mediante una expresión análoga a 4.25:

$$P_{dist\_bordes}(x,y) = dist_{CHAMFER}((x,y), P_{borde}) \tag{4.27}$$

  Se persigue así favorecer aquellos píxeles que perteneciendo a un borde están próximos a su eje de simetría vertical.

La formula de la energía de la imagen propuesta queda de la siguiente forma:

$$(f_{externas})_i^* = P_{bordes}(x,y) + P_{dist\_bordes\_vert}(x,y) + P_{dist\_sim\_vert}(x,y) \cdot P_{dist\_bordes}(x,y) \tag{4.28}$$

Cada uno de los mapas de distancias representan una función logarítmica de la imagen, ya que interesa que la energía sea mayor cuanto más cerca se esté de los bordes o del eje de simetría, pero forzando a que su fuerza disminuya logarítmicamente con la distancia.

4.8.2.3. Descripción de la deformación del contorno

Bajo una formulación de Lagrange (ver anexo B), la curva se deforma hasta que alcanza una situación de equilibrio. Esto ocurre cuando las fuerzas que le afectan se hacen cero y la curva alcanza el reposo:

$$(f_{internas})_i - (f_{externas})_i = 0 \tag{4.29}$$

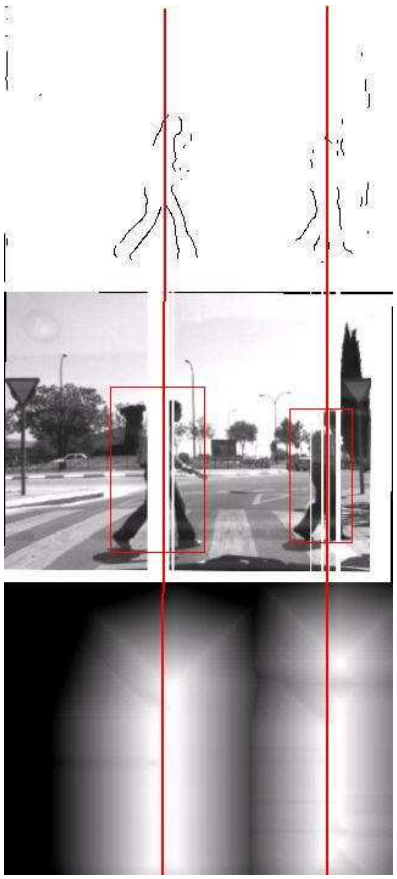


Figura 4.31: Pasos para la obtención del potencial debido a las distancias a los ejes de simetría vertical ; [Arriba] Imagen de bordes verticales filtrada con el mapa-denso estéreo. [Centro] En blanco se han dibujado los ejes de simetrías verticales. [Abajo] Campo de potenciales obtenido de las distancias a esos ejes. En rojo se muestra el eje de simetría cuya energía es máxima.

Para tratar de encontrar ese estado, se ha empleado un enfoque de minimización de energía. A cada curva se le asocia una energía, cuya estimación varía a medida que sus puntos se mueven. La solución viene dada por un mínimo de la energía total de la curva, definiéndose ésta en los términos de las fuerzas internas y externas expuestas (ver las ecuaciones 4.23 y 4.28):

$$E_{snake} = (f_{internas})_i + (f_{externas})_i \tag{4.30}$$

La estimación de la energía proporcionada por cada las fuerzas internas era la siguiente:

$$(f_{internas})_i = \underbrace{\left(f_{tangencial}\right)_i} + \underbrace{\left(f_{normal}\right)_i} + \underbrace{\left(f^*_{normal}\right)_i} \tag{4.31}$$

donde cada término de energía es aproximado mediante:

$$\begin{aligned} E_{continuidad} &= \left|\frac{dv_i}{ds}\right|^2 \approx \left|\overline{dist} - \left|\overrightarrow{v_i}\right|\right| + \left|\overline{dist} - \left|\overrightarrow{v_{i+1}}\right|\right| \\ E_{curvatura} &= \left|\frac{d^2v_i}{ds^2}\right|^2 \approx \left|\overrightarrow{v_i} - \overrightarrow{v_{i+1}}\right|^2 \\ E_{concavidad} &= \left|\frac{d^4v_i}{ds^4}\right|^2 \approx \left|\overrightarrow{v_{i-1}} - \overrightarrow{v_{i+2}}\right|^2 \end{aligned} \tag{4.32}$$

Minimizando la energía de continuidad, se tratan de evitar discontinuidades en la curva así como que se estire. De ese modo se controla la elasticidad del contorno. Por otro lado, al minimizar la energía de curvatura, se evita flexiones en la curva. Así se controla su capacidad para doblarse o lo que es lo mismo, proporciona una estimación de la rigidez del contorno. Por último, minimizando la energía de concavidad se evita que la curva se contraiga al favorecer que no haya cambios en el centro de curvatura de cada segmento. A cada uno de esos componentes se les da un peso local dependiendo de la relevancia que se le quiera dar a su valor de energía en cada vértice i del contorno:

$$E_{internas} = \alpha_i E_{continuidad} + \beta_i E_{curvatura} + \theta_i E_{concavidad} \tag{4.33}$$

Valores altos de α hacen que la continuidad de tipo 0 sea más importante que el resto de términos, con lo que se tendería a una curva con los puntos distribuidos de manera uniforme. Llevado al extremo, se obtendría un círculo. Por el contrario, si a es 0 en un punto significa que puede darse una discontinuidad en dicho punto. O lo que es lo mismo, la curva admite una parametrización no uniforme. Del mismo modo, valores altos de β pondrían el énfasis en minimizar la segunda derivada con lo que la curva tendería a ser una recta. Si es 0, se permite discontinuidad de tipo 1, formando una esquina en ese punto. Finalmente, valores altos de θ hacen que la continuidad de tipo 2 cobre más relevancia, imponiendo que el centro de curvatura a lo largo de la curva no varíe y así, la curva tiende también a un círculo. Si es 0, no se consideran los valores de concavidad. Si no existen otras fuerzas que compensen las internas, el contorno activo termina por convertirse en un punto.

La energía proporcionada por los campos de potenciales en la imagen se ha estimado mediante:

$$(f_{externas})_i^* = \underbrace{P_{bordes}(x,y)} + \underbrace{P_{dist\_bordes\_vert}(x,y)} + \underbrace{P_{dist\_sim\_vert}(x,y) \times P_{dist\_bordes}(x,y)} \tag{4.34}$$

Igual que antes, a cada término de energía se le asigna un peso con carácter local, o lo que es lo mismo, personalizado a cada vértice i del contorno:

$$E_{externas} = \gamma_i \cdot E_{gradiente} + kd_i \cdot E_{dist\_bordes\_vert} + ks_i \cdot E_{dist\_sim\_vert} \tag{4.35}$$

Una de las tareas más importantes a la hora de usar contornos activos está relacionado con la correcta selección de las energías asociadas con la imagen, ya que éstas son las que determinan cómo va a ser el *snake* atraído hacia los rasgos deseados. En concreto, la energía de la imagen debe decrecer en las regiones donde se quiere que el *snake* sea atraído. Cuanto

más brillante sea un píxel en la imagen, más baja es su energía asociada. Por tanto, los puntos de la curva se verán atraídas hacia regiones brillantes, con rasgos (bordes y distancias a ejes verticales) fuertes, en la imagen.

En conjunto, la energía total del *snake* se calcula minimizando ambos términos de energía:

$$\begin{aligned} E_{snake} = E_{internas} + E_{externas} = \alpha_i E_{continuidad} + \beta_i E_{curvatura} + \theta_i E_{concavidad} \\ -(\gamma_i \cdot E_{gradiente} - kd_i \cdot E_{dist\_bordes\_vert} - ks_i \cdot E_{dist\_sim\_vert}) \end{aligned} \tag{4.36}$$

Un último apunte respecto a esta fórmula; El signo del potencial debido a los bordes es negativo, ya que interesan las zonas con gradiente alto. Para el resto de términos de la energía externa, a menor distancia del correspondiente rasgo, menor energía. Así, se premian aquellas zonas de la imagen con un valor alto del gradiente y que estén lo más cerca posible de un borde vertical y de un eje de simetría vertical. Por otro lado, la energía interna busca valores pequeños de continuidad, curvatura y concavidad. Se favorecen, por tanto, deformaciones suaves en la forma del *snake*. Los parámetros $\alpha_i, \beta_i, \theta_i, \gamma_i, kd_i$ y ks_i indican el grado de influencia de cada uno de los términos, en cada punto i del contorno.

4.8.2.4. Detección de la forma humana

La técnica adoptada para la obtención de los contornos de los peatones, consta de varias fases.

1. Inicialización:

   Es crucial seleccionar un modo para situar al *snake* en su posición inicial. El algoritmo de inicialización utilizado integra la información obtenida de dos fuentes; por un lado, las fases previas de detección y de reconocimiento basado en el PCA, proporcionan una lista de ROIs encerrando objetos. Por otro lado, el mapa de disparidad denso obtenido a partir del sistema estéreo, proporciona la forma aproximada de esos objetos, requisito fundamental, como se ha comentado, para la extracción de los contornos mediante *snakes*.

   El procedimiento consiste en colocar el *snake* en el centro de cada ROI, una vez que se han eliminado los objetos del fondo. Después de utilizar distinto modelos, el que ofrece mejores resultados combina una forma triangular en la zona de las piernas y otra elíptica en el torso y la cabeza (ver fig. 4.32). Los resultados experimentales (ver fig. 4.35 han demostrado que el *snake* se ajusta mejor a la parte superior de un peatón, siendo propenso a fallos en la parte inferior (ver fig. 4.36). Esto es debido a la dificultad de los *snakes* para extraer la forma interior de las piernas. Por ello, se ha decidido usar la forma triangular original durante el *tracking* de cada ROI. Por el contrario, una vez inicializada la parte superior de la misma, la deformación alcanzada en la imagen precedente se utiliza como punto de partida en la siguiente.

   Siendo conocidas la distancia a las que está cada ROI, resulta sencillo emplear esa medida de distancia para segmentar el mapa de disparidades calculado antes (ver fig. 4.31), eliminando así los objetos que no interesan.

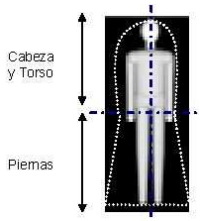


Figura 4.32: El modelo de *snake* utilizado durante la fase de inicialización. Se ha combinado una forma triangular para la detección de las piernas, con una forma elíptica para el torso y cabeza.

2. Búsqueda del mínimo local de energía:

   Una vez situado el *snake*, su energía se minimiza moviendo iterativamente cada punto de la curva, tratando de hallar un mínimo local. Existen varias soluciones en la literatura al problema de la minimización del *snake*[WS92][XuPrince98][Kass88]. En esta propuesta se ha extendido el algoritmo voraz [WS92] por ser rápido y eficaz. Un inconveniente es que hay que mantener los puntos del contorno uniformemente distribuidos para que la estimación de la curvatura y de la concavidad sean correctos. No obstante, si se consigue que la parametrización del contorno sea uniforme, Williams y Shah obtenían una mejor aproximación de la curvatura que otros autores, como Kass *et al.*.

   En esencia, su propuesta calculaba varios términos de energía (continuidad, curvatura e imagen) y mueve el punto actual al vecino más cercano donde la suma de esos términos de energía sea minimizado. El proceso se repite de forma iterativa y se dice que el algoritmo converge cuando el número de puntos que no han alcanzado un equilibrio, en el sentido de que aún es posible moverlos, está por debajo de un umbral. En esta implementación, se sigue esa misma idea siendo los términos de energía calculados los de la fórmula 4.36.

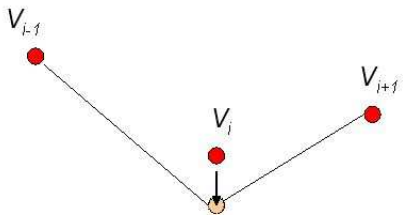


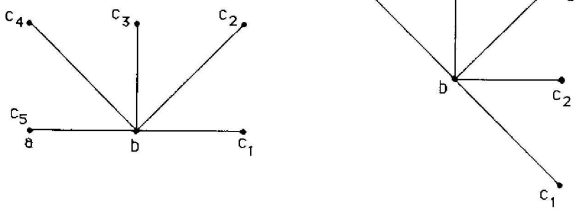
Figura 4.33: El algoritmo de Williams et al. [WS92] calcula en cada iteración, la energía del punto y de cada uno de sus vecinos (la vecindad aparece representada por la rejilla). Los puntos anterior v_{i-1} y posterior v_{i+1} del contorno se usan para calcular los términos de la función de energía. El punto se mueve a aquel punto de la vecindad con menor energía.

   Al final de cada iteración se ha incluido un paso que funciona como un proceso de alto nivel, ya que ofrece *feedback* al proceso de minimización de la energía. En función de la curvatura calculada mediante 4.37 en cada punto *i* del nuevo contorno, se va a controlar la flexibilidad y rigidez de ese punto para la próxima iteración.

$$curvatura^* = \left[\frac{\Delta x_i}{\Delta s_i} - \frac{\Delta x_{i+1}}{\Delta s_{i+1}}\right]^2 + \left[\frac{\Delta y_i}{\Delta s_i} - \frac{\Delta y_{i+1}}{\Delta s_{i+1}}\right]^2 \tag{4.37}$$

Se modifican los pesos de continuidad α_i y curvatura β_i de aquellos puntos que tienen un valor de curvatura mayor que sus vecinos anterior y posterior, permitiendo que en la siguiente iteración esos puntos tengan la flexibilidad y rigidez deseada. En la figura 4.34 se han dibujado dos distribuciones de puntos. Interpretando los valores de curvatura obtenidos mediante la fórmula 4.37, resulta sencillo establecer los umbrales máximos según el nivel de curvatura que interese. Se han considerado como umbrales los valores mostrados en la última columna de la figura 4.34.

Esta idea es herencia de [WS92], que se ha extendido para llevar a cabo un análisis de la concavidad de ciertos puntos de interés. Aquellos puntos por encima de un umbral de curvatura, que suele ser un valor próximo a cero, se considerarán puntos candidatos a que su centro de curvatura no varíe y por tanto se incrementa el peso de la concavidadθ_i en dichos puntos. Se pretende favorecer aquellas zonas donde el centro de curvatura no varía, ya que se espera que correspondan a la cabeza y pies de un posible peatón.



c	$(d\theta/ds)^2$	κ^2	$\lvert\mathbf{v}_{ss}\rvert^2$	$\lvert\vec{\mathbf{u}}_i - \vec{\mathbf{u}}_{i+1}\rvert^2$	$\left\lvert\frac{\vec{\mathbf{u}}_i}{\lvert\vec{\mathbf{u}}_i\rvert} - \frac{\vec{\mathbf{u}}_{i+1}}{\lvert\vec{\mathbf{u}}_{i+1}\rvert}\right\rvert^2$
			Horizontal case		
1	0.0	0.0	0.0	0.0	0.0
2	0.42	0.512	0.40	1.0	0.59
3	2.47	8.0	2.0	2.0	2.0
4	3.80	64.0	2.34	5.0	3.41
5	9.87	∞	4.0	4.0	4.0
			Diagonal case		
1	0.0	0.0	0.0	0.0	0.0
2	0.42	0.512	0.40	1.0	0.59
3	1.23	4.0	1.0	4.0	2.0
4	3.80	64.0	2.34	5.0	3.41
5	4.93	∞	2.0	8.0	4.0

Figura 4.34: Resultado del análisis de la curvatura realizado por [WS92] ; (*a*) Se muestran dos distribuciones de puntos a, b, c. La localización del punto c puede ser cualquiera de las posiciones C1 , C2 , C3 o C4. A la izda está el caso horizontal y a la derecha el diagonal que se emplean para realizar cálculos de curvatura. (*b*) Comparación de la estimación de la curvatura con distintos métodos realizado por Williams y Shah. Se han recuadrado los valores de los umbrales empleados para modificar los pesos α_i y β_i.

3. Análisis de la forma:

   El sistema de detección de peatones obtiene como resultado del módulo de los contornos activos, la silueta de los obstáculos reconocidos como peatones. Como ya se ha explicado, este proceso consta de varias fases: la validación de hipótesis basada en el PCA, tomaba la lista de ROIs proporcionada por la fase previa de detección de obstáculos. Después de evaluar cada ROI, aquellas con una probabilidad de contener un peatón baja eran desechadas. Después, la lista de ROIs clasificadas como peatones es proporcionada al módulo de *snakes*, encargado de extraer la forma humana.

4.9. Resultados Experimentales de los Snakes

Los *snakes* presentan una serie de inconvenientes; Debido a la tensión del *snake*, éste tiene capacidades elásticas, pero tiende a encogerse hasta que alguna otra energía fuerza un comportamiento opuesto. En general, el *snake* va a disminuir su longitud a medida que se va adaptando a la forma del objeto. Esto no tiene porqué ocurrir de un modo uniforme; cuando el objeto detectado es un peatón el *snake* se contrae mucho por la zona de la cabeza y menos a lo largo del cuerpo, siendo la concentración de puntos distinta a lo largo del contorno. Para evitar esta situación, el *snake* se vuelve a muestrear periódicamente a medida que se va contrayendo, manteniendo controlada la distancia media entre los puntos. De este modo se mantienen las características mecánicas del contorno activo, aunque éste cambie de longitud. Aunque, en determinadas situaciones, ha conducido a tener que emplear un gran número de puntos. Por ejemplo, para detectar personas que estén a unos 3 metros, en una imagen de 640x480 píxeles, fueron necesarios unos 70 puntos aproximadamente.

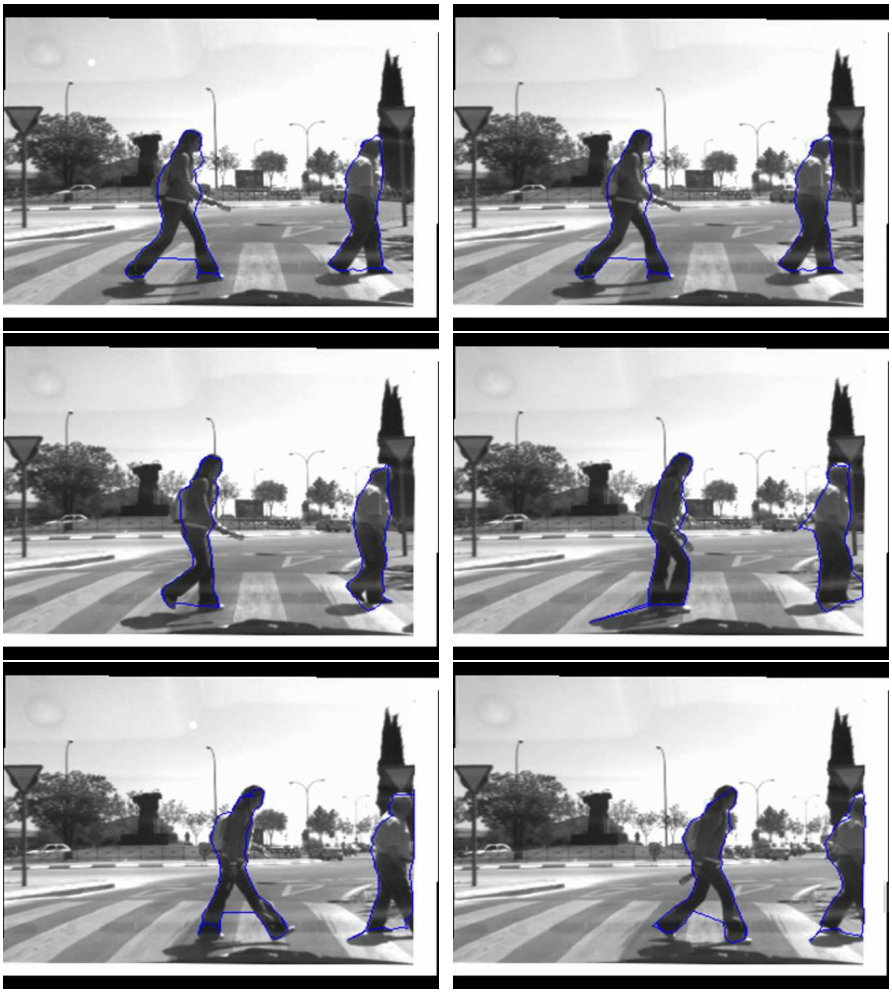


Figura 4.35: Imágenes que muestran los resultados de la extracción de la forma de los peatones mediante *snakes*.

Una importante limitación de esta técnica, es que sólo considera rasgos locales. La fórmula de la energía se evalúa únicamente en los píxeles vecinos a cada punto del modelo. Por ese motivo, se puede quedar atrapado en un mínimo local. Las 2 medidas empleadas, por un lado el suavizado aplicado a las imágenes a varios niveles, y por otro, los mapas de disparidad consiguen mitigar esos errores al incluir información global en la vecindad de cada punto.

Si se aumenta el tamaño de la vecindad, cabe esperar un menor número de iteraciones para llegar a la convergencia y una menor tendencia a caer en mínimos locales. Sin embargo, en imágenes reales esto no siempre es así, ya que cuanto mayor sea el tamaño de la ventana, más probabilidades habrá de que incluya información de otras fuerzas que no interesan. Además, hay que llegar a un compromiso entre el tiempo de cómputo y el tamaño de la vecindad. En nuestro caso, se emplearon ventanas de 7x7, 9x9 y 11x11 dependiendo de las distancias de los objetos a detectar.

Finalmente, hay que hacer especial mención al modo en que los valores de los parámetros se han obtenido. Se ha estudiado la sensibilidad de cada parámetro, ya que no hay una ciencia que diga cuáles con los parámetros adecuados, sino más bien se trata de llegar a un equilibrio entre todos y hay combinaciones que llevan a los mismos resultados. Debido a el modo en que se ha definido la función de la energía, es fundamental una distribución uniforme de los puntos del contorno. Por ello, se comienza con un valor muy alto de α para todos los puntos, a fin de asegurar el cumplimientos de esa condición. El resto de pesos de las energías internas son muy bajos en la primera iteración. A medida que se van moviendo los puntos del contorno en cada iteración, los pesos de los mismos varían, dependiendo de sus valores de curvatura y concavidad. En cuanto a los pesos de las energías externas, es crucial elegirlos del modo correcto, ya que estás energías son las que estabilizan a las energías internas. Se comienza con unos pesos que en conjunto sumen un valor ligeramente inferior al peso asignado a a. En este caso se han dividido los pesos casi equitativamente entre las 3 fuerzas externas propuestas. Finalmente, dependiendo del problema en estudio, habrá que enfatizar algún término más que otro o modificar el número de puntos.

Figura 4.36: Ejemplos de los problemas típicos de los *snakes*. En rojo aparece la forma alcanzada en la secuencia anterior, y en azul, la deformación alcanzada en la secuencia actual a partir de la forma (roja) previa. Se puede apreciar como para la zona de las piernas se utiliza la forma triangular original; (*a*) Puede darse el caso de que algún punto quede atrapado en un mínimo local y (*b*) los snakes tienen dificultades para adaptarse a la parte interior de las piernas.

Capítulo 5

Detección de Peatones en el Dominio Infrarrojo Lejano

El sistema de detección de la Universidad de Parma es fruto de varios años de trabajo y esfuerzo. El grupo de investigación VISLAB, dirigido por los profesores A. Broggi y M. Bertozzi, comenzó su particular conquista hacia la obtención de un vehículo inteligente en el año 90. Su prototipo ARGO se hizo famoso a nivel mundial, al desempeñar una conducción automática durante casi 2000 km a lo largo de Italia, durante el *MilleMiglia in Automatico Tour* celebrado en 1998 [BBFC99, BBF99b].

Por ello, las incursiones iniciales del VISLAB en el mundo de la protección de peatones, seguían el enfoque empleado hasta la fecha en otras tareas relacionadas con la conducción automática [BBCF01]; usaban las cámaras del dominio visible que ya tenían instaladas en el vehículo.

Los primeros experimentos [BBFS00, BRF+03] demostraron que el dominio visible aplicado a la detección de peatones tenía carencias; las imágenes se ven afectadas por cambios de iluminación y además, muestran un excesivo número de detalles y sombras, que aun siendo de gran ayuda en la fase de clasificación, resulta un inconveniente en las primeras etapas del proceso de detección.

Las cámaras de infrarrojo (lejano) precisamente, carecen de esos problemas; las imágenes no se ven afectadas por los cambios de iluminación y generalmente son menos ruidosas, debido a su relativa ausencia de sombras y texturas originadas por los colores. Sin embargo, pueden aparecer texturas debidas otras causas, como son los cambios en las temperaturas y las distintas ropas que llevan los peatones.

A continuación, se explica el algoritmo de detección de peatones desarrollado en Parma, ya que el algoritmo basado en infrarrojos implementado en esta tesis forma parte del sistema final. Dicho sistema final se conoce con el nombre GOLD *General Obstacle and Lane Detection* y fue el que originariamente guió al vehículo experimental ARGO [BBFC99, BBB+07a] y sobre el que se han ido añadiendo mejoras. Para la adquisición de imágenes se ha utilizado el sistema de adquisición Tetravision, del que se han explicado sus elementos en el capítulo 3.

5.1. Descripción del Sistema GOLD

El sistema GOLD se basa en el uso simultáneo de dos parejas de cámaras estéreo del dominio visible y del dominio infrarrojo lejano (*Far Infrared*, FIR), respectivamente. La idea principal es beneficiarse de las ventajas ofrecidas tanto por las cámaras de infrarrojos lejano como por las del espectro visible, enfrentándose al mismo tiempo a las deficiencias de cada uno de esos sistemas.

Para la detección de peatones, siguen un enfoque basado en la búsqueda del foco de interés en ambos dominios y llevan a cabo la detección de áreas cálidas, la detección de bordes verticales y el cálculo simultáneo en los dos dominios del mapa de disparidad. Como resultado obtienen una lista de áreas de la imagen que potencialmente contienen un peatón y que es filtrada mediante un análisis de simetrías. Finalmente, el reconocimiento de los peatones se basa en la forma humana. Se integran las decisiones de tres métodos; dos de ellos se basan en el uso de modelos probabilísticos para detectar la cabeza y el cuerpo, y el tercero usa modelos deformables para detectar el contorno del peatón. La figura Fig. 5.1 muestra el flujo del algoritmo completo. El módulo etiquetado como *Probabilistic Model* es el resultado de las investigaciones de la autora de esta tesis en la Universidad de Parma [BBF+07b, BBF+07a]. El resto del sistema GOLD ha sido íntegramente desarrollado por el VISLAB.

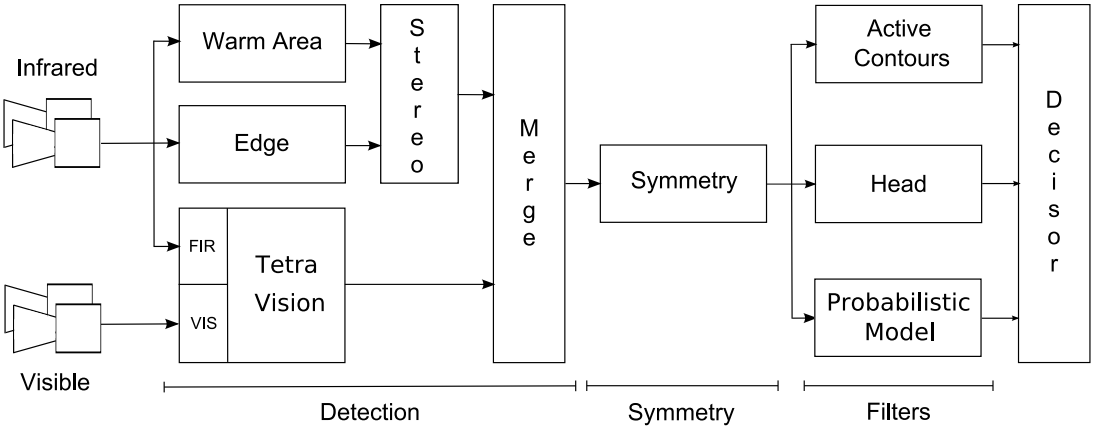


Figura 5.1: Diagrama de flujo del sistema global .

1. **Etapa de Detección:**

   Para la detección de regiones de interés, combinan los resultados obtenidos de varios procesamientos. Por un lado, mediante el cálculo de la pendiente de la carretera se obtiene la inclinación de las cámaras, siendo un dato útil para posteriores fases de la detección. Por otro lado, las imágenes FIR se prestan a un procesamiento concreto basado en sus cualidades. Además, las cámaras estéreo que forman el sistema Tetravision permiten obtener los obstáculos de ambos dominios.

   a) **Detección de la pendiente de la carretera:**

Son muchos los sistemas basados en visión para vehículos inteligentes que se apoyan en el supuesto de estar moviéndose en un terreno plano o –más realistamente– en un terreno con una ligera pendiente variable. Sin embargo, esta suposición puede llevar a errores (relativos a los parámetros de la escena o en las detecciones) no sólo cuando la carretera no sea realmente plana, sino también debido a que la cámara está instalada en un vehículo en movimiento y por tanto, continuamente varía su inclinación. Por ello, se usa un enfoque de V-disparidad [LAT03, BCFG05, BBF+07a], que aprovecha la existencia de los dos sistemas estéreo para calcular la pendiente real de la carretera.

Después de aplicar el filtro de Sobel, con el fin de realzar los rasgos – en especial, los bordes – en la imagen, se calcula la correlación para diferentes disparidades, para cada par (izquierda y derecha) de filas de las imágenes de Sobel. Debido a que los pares estéreo han sido calibrados y rectificados convenientemente, las mismas filas en las imágenes estéreo son líneas epipolares.

Como resultado se obtiene una nueva imagen: la imagen V-disparidad, que contiene los valores de correlación. A mayor intensidad del píxel, mayor es la correspondencia. Este proceso se ha realizado tanto para imágenes del dominio visible como del infrarrojo, pero debido a que los elementos del terreno son escasos en las imágenes de infrarrojos, la imagen V-disparidad obtenida no permite recuperar información de la carretera. Por el contrario, se puede apreciar cómo en las imágenes del dominio visible, los bordes del terreno producen una línea inclinada que ofrece información sobre la sección trasversal de la carretera. Por ello, sólo se usa el dominio visible para calcular la pendiente del terreno.

b) **Procesamiento realizado únicamente en el dominio infrarrojo lejano:**

Con el fin de obtener una lista de regiones en la imagen conteniendo a peatones en potencia, se realizan dos procesamientos sobre las imágenes de infrarrojo lejano: la detección de regiones cálidas y un filtrado basado en bordes.

- **Detección de regiones cálidas:**
  Este enfoque se basa en el hecho que los peatones generalmente emiten más calor que el fondo, apareciendo como regiones más brillantes en la imagen de infrarrojo lejano. La búsqueda en la imagen se dirige hacia zonas contiguas con una intensidad alta. Para refinar las cajas que contienen a cada región, se calculan los histogramas verticales y horizontales en cada caja, hasta que su tamaño no se puede reducir más. Las cajas demasiado pequeñas para contener a un peatón, son eliminadas.
- **Detección de bordes:**
  Una de las características intrínsecas de las imágenes de infrarrojos lejano, es su ausencia de texturas. Esto se debe a que los patrones termales presentan una homogeneidad mayor que las rasgos basados en el color. Como consecuencia, las imágenes de infrarrojo contienen menos bordes que las del espectro visible; además, estos bordes serán con mayor probabilidad, bordes de objetos. El

problema del ruido en los enfoques basados en la detección de bordes, queda así reducido.

Normalmente, la forma humana se caracteriza por tener un mayor número de bordes verticales que el fondo o que otros objetos. Por ello, este enfoque se basa en la búsqueda de regiones en la imagen que contienen un número elevado de bordes verticales. Para filtrar otros objetos que también contienen bordes verticales – como los árboles, coches o farolas –, se eliminan aquellos bordes verticales regulares y más largos que un umbral dado. Los píxeles conectados se agrupan en cajas, que una vez filtradas para eliminar aquellas que son muy pequeñas o que contienen regiones frías, pasan a formar una lista de cajas.

- **Correspondencia estéreo:**
  Para calcular la posición y el tamaño real de las cajas detectadas en los dos pasos anteriores, se busca sus homólogas en la otra imagen. Debido a que se conocen los parámetros de calibración del sistema, se puede estimar una región de búsqueda en la otra imagen para cada caja, y se selecciona la mejor correspondencia. Después, basándose en triangulación, se estima la distancia entre los objetos en la imagen y el sistema de visión.

c) **Detección de obstáculos Tetravision independiente:**

El hecho de tratar con dos sistemas de visión estéreo distintos simultáneamente, entraña ya cierta dificultad, debido a que cada sistema estéreo caracteriza distintos puntos de vista, distintas aperturas angulares y, con frecuencia, distintas velocidades de adquisición y resoluciones. Además, no se puede realizar una búsqueda de puntos homólogos, ya que es muy difícil establecer una correspondencia de rasgos entre dominios espectrales distintos.

El sistema GOLD, procesa por separado cada flujo estéreo para detectar obstáculos, para más tarde fusionar los resultados de cada dominio distinto.

Se calcula una imagen de disparidades (*Disparity Space Image*, DSI), subdividiendo la imagen derecha en regiones de 3x8 – motivado porque un peatón se caracteriza por rasgos verticales – y realizando una búsqueda de regiones homólogas en la imagen izquierda. Una vez más, debido a que los ejes ópticos de los dos sistemas estéreo son paralelos, dos líneas correspondientes en ambas imágenes serán líneas epipolares. Además, la información de la calibración permite reducir aún más el espacio de búsqueda. Para cada región, se considera el mejor valor de correspondencia y se calcula la disparidad. Aquellas regiones grandes que compartan una disparidad similar son marcados como obstáculos. Las porciones de cada columna de la imagen DSI, que presentan una disparidad variable, son eliminados al considerarse elementos del fondo.

d) **Etapa de unión y filtrado:**

Como ya se ha comentado, cada una de las regiones obtenidas como resultado del proceso de segmentación precedente, se delimitan mediante cajas. Sin embargo,

debido a que se aplican distintas técnicas para procesar la misma escena, con frecuencia, distintas cajas pertenecen al mismo obstáculo; es necesario un proceso de unión, que fusione cajas que estén lo suficientemente próximas en coordenadas del mundo real. Al basarse en las coordenadas del mundo, si los obstáculos se solapan parcialmente, el sistema es capaz de detectarlo y no se unen. Después de la unión, se filtran aquellas cajas que no cumplen unas condiciones de tamaño adecuadas.

Finalmente, se refina el tamaño de las cajas resultantes. De hecho, se usa el conocimiento que se tiene de la pendiente de la carretera para calcular el punto de contacto entre cada objeto contenido en una caja y el terreno, conseguiéndose así, hacer coincidir las bases de las cajas, con la carretera.

2. **Filtrado basado en Simetrías:**

   El detector de obstáculos descrito opera a bajo nivel y genera una lista de cajas. Por ello se ha incluido un proceso que elimina los obstáculos que no son peatones y además, hace que cada caja contenga un peatón. Emplean un procesamiento basado en simetrías para dividir, refinar y realizar una preliminar validación de cada caja.

   Esta fase se realiza sólo en el dominio infrarrojo, ya que los resultados experimentales han demostrado que este tipo de imágenes contienen un menor número de pequeños detalles que las del dominio visible, siendo más adecuadas para el cálculo de la simetría. El cómputo de la simetría vertical de cada caja considera varios aspectos: los niveles de gris, los bordes verticales y la densidad de esos bordes.

3. **Fase de Reconocimiento:**

   La etapa de validación consta de tres procesos distintos, que evalúan cada caja para determinar si existe una forma humana. Uno de los procesos se basa en la búsqueda de la cabeza y los otros dos, en la forma humana completa mediante contornos activos y modelos probabilísticos.

   - **Detección de la cabeza:**

     Para cada peatón potencial, se evalúa si está presente la cabeza. Sólo se realiza en el dominio infrarrojo, ya que es el rasgo más evidente en dicho dominio. El detector de cabezas combina tres técnicas distintas: dos basadas en correspondencia de modelos y una, en la búsqueda de regiones cálidas.

     El primer método de correspondencia de modelos, se basa en una técnica de correspondencia de patrones, combinando el uso de dos modelos de cabeza –uno de ellos binarizado–. El segundo método de correspondencia se basa en un enfoque probabilístico. A partir de un conjunto de entrenamiento, se genera un modelo probabilístico de la cabeza, que se hace corresponder con el contenido en la imagen. Ambos métodos buscan el valor más alto de correlación. Por último, se realiza un análisis de regiones cálidas contiguas que satisfagan unos criterios de similitud relativos a una cabeza.

- **Modelos basados en contornos activos *snakes*:**

  Con el fin de extraer el contorno del objeto contenido en cada caja, sitúan un *snake* alrededor de los límites de cada caja. Debido a la diferencia de temperatura entre una persona y el fondo, se ha considerado más conveniente realizar este proceso de extracción de contornos sobre las imágenes de infrarrojos.

  El resultado obtenido va a ser evaluado, en la fase de clasificación, por una red neuronal que dará un voto, determinando si existe o no un peatón.

- **Modelos probabilísticos:**

  Este enfoque se ha empleado para validar la existencia de un peatón en cada caja (*bounding box*). Se crea un modelo probabilístico de la forma humana, que es usado para dar un voto a cada candidato potencial.

  Este algoritmo forma parte de la presente tesis y a continuación va a ser explicado en profundidad. Se ha considerado conveniente exponer el sistema global dentro del cual ha sido integrado este método, para tener una mayor comprensión del flujo del sistema.

4. **Validación final:**

   Se ha incluido esta etapa para filtrar los falsos positivos. Los votos de cada evaluador se combina linealmente para proporcionar una decisión final.

5.2. Descripción del módulo de detección de peatones probabilístico

Un problema fundamental de la visión por computador consiste en detectar y reconocer objetos en entornos saturados. La tarea puede ser especialmente complicada cuando los objetos no pueden ser fácilmente extraídos porque sus bordes se confunden con el fondo, y sus texturas y colores no pueden ser usados para la segmentación. Esto ocurre con las imágenes FIR, pero resultan adecuadas para la detección de peatones cuando el fondo emite menos calor que las personas, durante la noche o en condiciones de baja iluminación. Además, suelen contener menos ruido (y por tanto, menos detalles), facilitando los pasos iniciales de la detección. Por estos motivos, este módulo se basa en las imágenes capturadas con las cámaras estéreo de infrarrojo.

Una vez implementado, ha sido integrado en el sistema GOLD y se han realizado pruebas con secuencias de imágenes adquiridas con el sistema Tetravision. Los objetivos inicialmente planteados eran conseguir detectar las piernas del peatón y realizar el posterior reconocimiento de la postura de las mismas.

5.2.1. Dificultades para el reconocimiento de peatones en imágenes FIR

En el capítulo 3 (ver el apartado 3.1.1) se expuso la dificultad que conlleva la detección de peatones en el dominio infrarrojo. El método propuesto ha tenido que tratar, en concreto, con los siguientes inconvenientes:

- Rara vez son completamente visibles todas las partes del cuerpo de una persona. Esto puede ser debido a oclusiones (ocasionadas por otros objetos en la escena o por auto-oclusiones con otras partes del cuerpo), o a errores en el proceso de segmentación.

- Las ropas afectan la huella termal de la forma humana. Es habitual que un peatón en infrarrojos aparezca con la zona del torso más oscura que las zonas de la cabeza y las manos, haciendo la detección aún más complicada.

- Las imágenes de infrarrojos facilitan la tarea de detección de objetos que son o más cálidos o más fríos que el fondo, siendo así suficientemente diferenciables del fondo. Pero, múltiples factores – ambientales o la emisión de calor por parte de otros objetos – pueden afectar la respuesta termal de la imagen, complicando sobremanera la detección de objetos.

Como consecuencia, no son factibles ni técnicas de correspondencia basados en patrones uniformes, ni técnicas de segmentación basadas en regiones. Tampoco dan buenos resultados los enfoques tradicionales basados en bordes, debido al bajo contraste y la pobre resolución de las imágenes de infrarrojos. En este contexto, los métodos de reconocimiento de objetos más exitosos se basan en la integración de medidas locales (como p.ej. la intensidad, la textura o el color) de la imagen junto con conocimiento global de la forma del objeto. Es habitual representar esos dos conocimientos mediante modelos. Esta idea, llevada al caso de la detección de peatones, se puede concretar en un conjunto de posibles tareas:

1. Modelar la complejidad de la apariencia humana:
   Integrar múltiples medidas de la imagen (p.ej. contornos, intensidad, siluetas, etc.) y aprender sus valores óptimos para reducir la incertidumbre.

2. Modelar las complejas restricciones estructurales del cuerpo humano:
   Aprender las proporciones del cuerpo, así como las representaciones estructurales que describan el movimiento típico entre las distintas partes del cuerpo.

3. Representar y propagar la incertidumbre:

   - Beneficiarse de la estructura del problema (p.ej. simetrías, localidad) durante la etapa de búsqueda.
   - Integrar la información a lo largo del tiempo, de una forma compacta y eficiente.

Por último, los métodos utilizados para la detección de peatones deben depender del menor número de restricciones posibles siendo al mismo tiempo sencillos, rápidos y eficaces.

Para la detección de peatones en imágenes FIR se ha optado por modelar la apariencia humana, siendo en este caso el reto más importante, poder abarcar la gran diversidad en cuanto a formas, tamaños, posturas y texturas que pueden presentar tanto las personas, como el entorno que les rodea. Además de esto, la apariencia de una persona depende de su postura; es decir, su posición y orientación con respecto a la cámara. De manera que, una visión lateral de un peatón difiere mucho de una visión frontal del mismo.

Un detector de peatones potente debe poder adaptarse a estas variaciones y, al mismo tiempo, ser capaz de distinguir al peatón del resto de objetos que puedan aparecer en la escena. Con este fin, se ha adoptado un enfoque probabilístico, inspirado por el modelo propuesto por Nanda y Davis [ND02] en el 2002.

5.2.2. Esquema general del módulo

El módulo desarrollado sigue un enfoque basado en el aprendizaje de modelos y consta de dos fases; la primera fase, de entrenamiento, y la segunda, de test. Durante la fase de entrenamiento, el sistema debe aprender un límite de decisión que separe un objeto peatón del resto de objetos en la imagen. Se pueden adoptar dos enfoques: *bottom-up* o *top-down*. Bajo el enfoque *bottom-up* o *feed-forward*, se parte de la información en la imagen para tratar de saber si existe o no peatón. Suelen desarrollarse modelos discriminadores o condicionales, basados por ejemplo en SVM, RVM o *AdaBoost*. Se caracterizan por imponer condiciones a las observaciones (rasgos extraídos de la imagen) para establecer la clase a la que pertenecen. Bajo el enfoque *top-down*, se parte del modelo que puede haber generado la observación para tratar de clasificar el contenido en la imagen. Se caracterizan por modelar las observaciones, asociando predicciones a los rasgos extraídos de la imagen. Por tanto, se puede formalizar un proceso generativo como una distribución de probabilidad. El clasificador de Bayes es un modelo generativo.

En método que se propone se clasifica dentro de estos últimos. La figura 5.2 muestra el flujo del algoritmo completo.

1. **Fase de Aprendizaje:**

   El aprendizaje se realiza sobre un conjunto de imágenes de peatones y de no-peatones tomadas con el sistema FIR estéreo construido en la universidad de Parma. El proceso se realiza a mano, seleccionando y etiquetando las correspondientes imágenes de forma supervisada por una persona.

   - Extracción de la silueta:

     A partir de imágenes de entrenamiento, que contienen a un único peatón en diferentes posturas, se determina un umbral capaz de segmentar su silueta.

   - Creación de los modelos probabilísticos:

     El conjunto de siluetas extraídas en el paso anterior pueden caracterizarse mediante una distribución de probabilidad binomial o de *Bernoulli*.

2. **Fase de Test:**

   Durante la fase de test, se evalúa el funcionamiento del sistema sobre una secuencia de imágenes de FIR en tiempo real.

   - Módulo probabilístico:

     El reconocimiento de los peatones se basa en la correlación de cada modelo con las ROIs obtenidas de las fases precedentes. Sin embargo, este método no es un

simple *pattern matching*, ya que cada píxel del modelo contiene información sobre su probabilidad de ocurrencia.

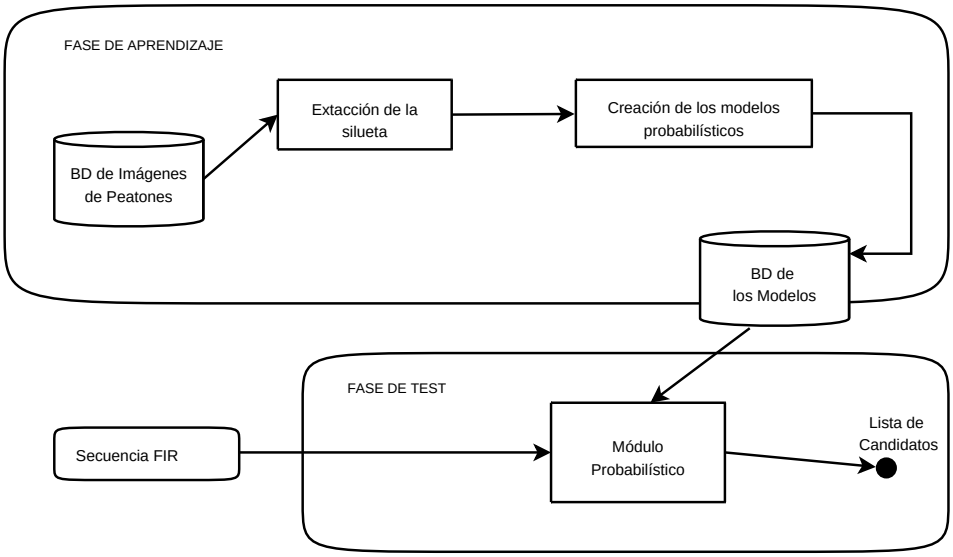


Figura 5.2: Esquema general del módulo probabilístico, compuesto por la fase de aprendizaje y la fase de test.

5.3. Enfoque probabilístico: Modelos basados en la apariencia

Nanda y Davis ([ND02]), desarrollaron en el 2002 un sistema de detección de peatones en tiempo real, capaz de reconocer peatones a partir de imágenes FIR de baja calidad, bajo contraste y ante formas de peatones incompletas. Proponen utilizar modelos probabilísticos para abarcar la gran diversidad de formas humanas que pueden darse. Por un lado, las imágenes FIR proporcionan las ROIs y, por otro, los modelos probabilísticos ayudan al reconocimiento de los peatones. Sin embargo, determinar un modelo humano no resulta sencillo en el dominio infrarrojo, ya que todas las partes del cuerpo humano rara vez son visibles. Por ello, resulta crucial el decidir qué rasgos se van a usar para representar a los peatones.

5.3.1. Extracción de rasgos: Siluetas

Las siluetas capturan la información esencial de la forma. Además son insensibles a los atributos de las superficies: textura, color (intensidad) y vestimenta. En el aspecto negativo, los detalles internos de la silueta pueden ser espurios debido a la ropa y la segmentación puede ser pobre. Sin embargo, en el dominio infrarrojo las siluetas aparecen más homogéneas que en el dominio visible, viéndose menos afectados por esos problemas.

Por otro lado, la naturaleza cíclica de la acción de caminar, asegura que la silueta de cada individuo será repetida a intervalos regulares. Además, se pueden apreciar similitudes entre

las siluetas de un conjunto de peatones. Debido a sus características, se ha decidido describir la apariencia en base a la silueta humana.

Inicialmente, se siguió la propuesta de Nanda y Davis,que se va a explicar a continuación, para la extracción de los rasgos de interés. Sin embargo, resultó no ser adecuada para el tipo de imágenes FIR en estudio, planteándose una solución alternativa.

1. **Distribución Gaussiana para las clases peatón y no-peatón:**

   La dificultad está en determinar el valor que permite distinguir un peatón del fondo. En [ND02], este valor se calcula a partir de un conjunto de entrenamiento formado por 1000 cajas conteniendo peatones. Calculan la media y desviación estándard para los píxeles pertenecientes a los peatones $N_1(\sigma_1, \mu_1)$ y para los píxeles pertenecientes al fondo $N_2(\sigma_2, \mu_2)$. Por tanto, consideran que las distribuciones de los datos son gaussianas. Suponen que las prioridades para las clases de peatones y de no-peatones son equiprobables, pudiendo emplear la clasificación Bayesiana para calcular el umbral con la siguiente fórmula:

$$umbral_1 = \frac{\sigma_1\sigma_2}{\sigma_1 + \sigma_2} \ln\left(\frac{\sigma_1}{\sigma_2}\right) + \frac{\sigma_1\mu_2 + \sigma_2\mu_1}{\sigma_1 + \sigma_2} \tag{5.1}$$

   Para el caso en estudio, inicialmente se consideró esta fórmula. A partir de un conjunto de entrenamiento formado por imágenes de peatones en distintas fases del caminar, así como de imágenes de no-peatones, se calculó el umbral. La especificación del conjunto de imágenes de entrenamiento empleados se muestran en la tabla 5.1.

   El valor del umbral obtenido fue $umbral_1 = 56$. Un análisis de las intensidades de los píxeles FIR demostró la dificultad de separar a los peatones del fondo con este umbral, ya que muchos píxeles del fondo tienen un valor superior a ese umbral (ver figura 5.3-b). Por otro lado, los peatones del conjunto de entrenamiento no podían aproximarse mediante gaussianas, como puede apreciarse en la figura 5.4-b.

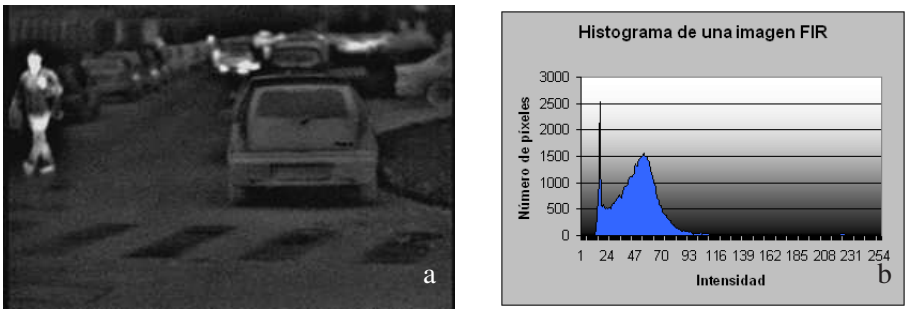


Figura 5.3: Ejemplo de una de las imágenes FIR empleada en el análisis de intensidades; (*a*) imagen típica FIR y (*b*) muestra los valores de las intensidades. El valor máximo corresponde a la intensidad 226, teniendo el 98.36 % del total de los píxeles una intensidad menor que 127.

2. **Distribución Gaussiana para la clase no-peatón:**

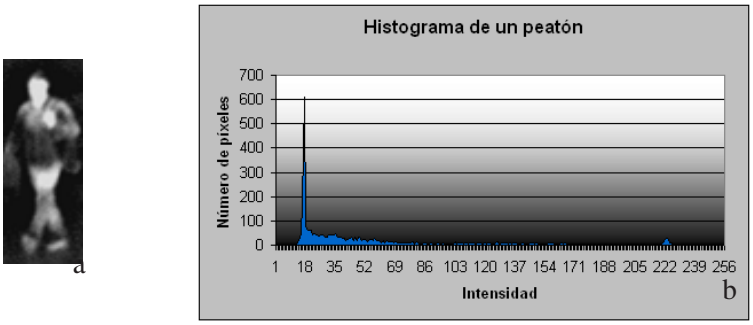


Figura 5.4: Ejemplo de una de las imágenes de peatones empleada en el análisis de intensidades; (*a*) imagen que contiene un peatón en infrarrojos; (*b*) histograma que muestra los valores de las intensidades.

Con el fin de obtener una extracción de la silueta lo más precisa posible, se ha realizado un análisis de la distribución de intensidades del fondo de un conjunto de imágenes FIR como el que se muestra en la figura 5.5. El conjunto de entrenamiento empleado está formado por 240 imágenes de fondo, como se indica en la tabla 5.2.

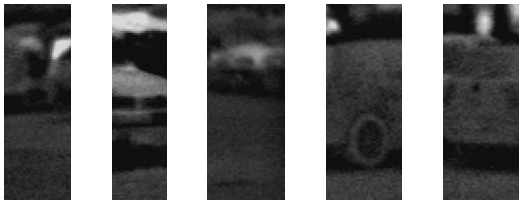


Figura 5.5: Ejemplos de imágenes de fondo FIR tomadas del conjunto de entrenamiento.

La distribución de los datos sigue una gaussiana como la de la figura 5.6, de media y desviación típica $N(\mu = 51.13, \sigma=16.63)$.

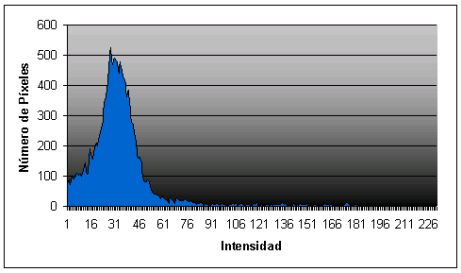


Figura 5.6: Distribución gaussiana que caracteriza los píxeles del fondo FIR.

En una distribución normal o gaussiana se cumple que aproximadamente la totalidad de los datos se encuentran en ± 3 desviaciones típicas σ de la media μ. Dado que la clase de los no-peatones sigue una distribución gaussiana, el umbral que diferencia a peatones de los no-peatones puede calcularse basándose en el siguiente supuesto;

$$T = \mu + 3\sigma \tag{5.2}$$

Considerando que los píxeles pertenecientes a un peatón se espera que sean más brillantes que los del fondo, sólo se va a tener el cuenta el límite superior representado en la ecuación 5.2 que, en esta situación, asegura que más del 99 % de los datos del fondo están por debajo de ese umbral. Según los cálculos realizados para el conjunto de entrenamiento de imágenes del fondo, el valor obtenido para dicho umbral ha sido;

$$T = 101{,}026 \tag{5.3}$$

Usando este valor se pueden umbralizar las imágenes FIR para extraer la silueta de los peatones.

5.3.2. Distribución de los rasgos: binomial o de *Bernoulli*

El siguiente paso consiste en modelar los rasgos extraídos de la imagen (las siluetas), que permitan determinar la probabilidad de que en dicha imagen exista un peatón;

$$P(imagen|modelopeaton) \tag{5.4}$$

Sin embargo, no resulta sencillo modelar la definición 5.4, ya que no se conocen los rasgos característicos que definen la apariencia humana. Por ejemplo, no se sabe si las distribuciones reales son Gaussianas, multimodales o de Poisson. Estas propiedades son desconocidas ya que resulta imposible analizar los estadísticos combinados de un gran número de píxeles. Dado que no se conoce la estructura real de la distribución de peatones, una posible solución es seleccionar modelos lo suficientemente flexibles como para adaptarse a un amplio rango de estructuras.

Para el caso en estudio, dado que los rasgos son siluetas binarias (ya que han sido previamente umbralizadas aplicando la ecuación 5.3), se pueden modelar mediante una distribución binomial o de *Bernoulli*. Esto es así porque cada píxel de la silueta (o variable aleatoria) sólo puede tomar valores $\{0, 1\}$. Por tanto, se puede representar la probabilidad de que cada píxel individual de la imagen esté activo (a 1), como una variable aleatoria independiente de *Bernoulli*, y las probabilidades para todos los píxeles que forman una silueta, como una distribución binomial o de *Bernoulli*. Debido a que los datos son binarios, esta distribución resulta una suposición paramétrica razonable. En definitiva, se obtiene una representación probabilística de la silueta que va a ser utilizada como modelo.

Esta solución se basa en el modelo propuesto en [LDT03]. Se trata de un método que de manera automática extrae la silueta de los peatones y aprende el correspondiente modelo. Además, utilizan dicho modelo para mejorar la silueta extraída, ya que eliminan ruido y completan partes eliminadas durante la segmentación.

5.3.3. Creación de los modelos

Se considera que los peatones caminan en una dirección frontal o lateral con respecto al eje de la cámara. Los casos estudiados comprenden a personas caminando en un plano básicamente paralelo al plano de la imagen y siempre en la misma dirección. En la figura 5.7 se muestran algunos ejemplos de entrenamiento. En estas condiciones, la naturaleza cíclica de la apariencia basada en la silueta es claramente apreciable. La misma fase de un ciclo del caminar humano, aparecerá repetidamente en una secuencia. Como consecuencia, se obtendrá una mejor estimación de la silueta si se emplean todas las siluetas que corresponden a la misma fase.

Figura 5.7: Algunos ejemplos de las imágenes de entrenamiento conteniendo peatones

Estas consideraciones han conducido al siguiente método para la obtención de un modelo del peatón a partir de un conjunto de entrenamiento:

1. El conjunto de entrenamiento está formado por imágenes de peatones tomadas con las cámaras FIR del sistema GOLD. Esas imágenes son separadas manualmente en cuatro clases, dependiendo de la fase en que se encuentra el peatón. Para ello, se analiza la posición de las piernas clasificando cada imagen como *piernas abiertas*, *piernas semiabiertas*, *piernas cerradas* o *piernas semicerradas*.

   Cada subconjunto de entrenamiento va a estar formado por distintas imágenes de peatones, exigiendo que todas tengan la misma altura. En cambio, no se imponen restricciones a la anchura, ya que su dimensión varía mucho durante el ciclo del caminar, afectado tanto por la posición de las piernas como de las manos. En las tablas 5.2, 5.3 y 5.4 se detallan las características de los conjuntos de entrenamiento utilizados.

2. Para extraer la silueta de los peatones, se umbralizan las imágenes de entrenamiento aplicando el umbral obtenido mediante la aproximación de una distribución gaussiana para los píxeles del fondo (ver ec. 5.3). Las siluetas correspondientes a cada fase del ciclo del caminar (agrupadas en las cuatro clases ya citadas), se alinean asumiendo un periodo del caminar constante. Para ello, todas las imágenes de entrenamiento son transladadas para que el centro de la imagen esté en todas ellas en la misma posición.

3. Después, se calcula la distribución de probabilidad de las siluetas binarias asignadas a cada fase o clase. El modelo probabilístico va a representar la frecuencia con que cada rasgo o píxel i se ha observado a 1 en el conjunto de siluetas de entrenamiento.

5.3.3.1. Conjuntos de entrenamiento y modelos generados

Para la generación de los modelos probabilísticos, se han probado varios conjuntos de entrenamiento distintos. Todos ellos contienen imágenes de peatones del dominio infrarrojo y han sido divididos en clases, en función del número de modelos distintos que se quieran generar. Conviene que cada clase contenga el mismo número de imágenes de peatones, para no favorecer unos modelos probabilísticos frente a otros (ver tabla 5.1).

El conjunto de entrenamiento inicial separaba las imágenes de los peatones en dos clases; *piernas abiertas* y *piernas semicerradas*. En la figura 5.8 se muestran los dos modelos generados.

Tipo de imágenes	Número de imágenes
Fondo de la imagen	400
Piernas abiertas	95
Piernas semicerradas	81
Piernas cerradas	388
Todo tipo de peatones	516

Tabla 5.1: Conjunto de entrenamiento 1. Sólo se consideraron las clases piernas abiertas y semicerradas

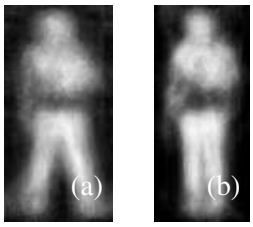


Figura 5.8: Modelos obtenidos después del aprendizaje, considerando distintas fases del caminar humano; (*a*) piernas abiertas y (*b*) semicerradas.

Sin embargo, los resultados experimentales han demostrado la necesidad de crear nuevos modelos capaces de reconocer otras posiciones intermedias de las piernas. Por ello, los nuevos conjuntos de entrenamiento utilizados consideran dos nuevas posiciones de las piernas; *piernas semiabiertas* y *piernas cerradas*. El primero de ellos está formado por un total de 240 imágenes, agrupadas en cuatro clases de 60 imágenes (ver tabla 5.2). Los modelos aprendidos a partir de esos datos se muestran en la imagen 5.9.

Pero las clases de este conjunto de entrenamiento agrupan a los peatones en función de la posición de las piernas, sin considerar su orientación respecto a la cámara. Se ha considerado oportuno separar las imágenes de los peatones considerando si se mueven frontal o lateralmente con respecto al eje de la cámara. Así se obtiene una mejor estimación de las probabilidades, al corresponderse todos los ejemplos a la misma orientación y fase del proceso de caminar.

Por ello, finalmente, se han usado dos nuevos conjuntos de entrenamiento; uno, para el caso del movimiento lateral y el otro, para el caso del movimiento frontal. Cada uno de los conjuntos es posteriormente separado en clases. El número de imágenes de estos dos conjuntos

Tipo de imágenes	Número de imágenes
Fondo de la imagen	240
Piernas abiertas	60
Piernas semiabiertas	60
Piernas semicerradas	60
Piernas cerradas	60

Tabla 5.2: Conjunto de entrenamiento 2:Piernas abiertas,semiabiertas,semicerradas y cerradas

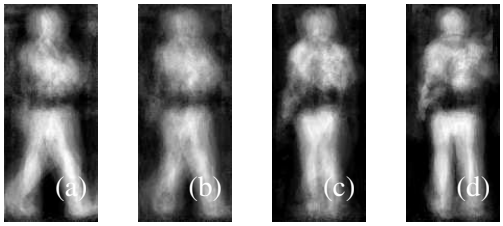


Figura 5.9: Modelos obtenidos después del aprendizaje, considerando distintas fases del caminar humano; (*a*) piernas abiertas, (*b*) semiabiertas, (*c*) cerradas y (*d*) semicerradas.

es menor que los anteriores, ya que en las secuencias FIR utilizadas para extraer los ejemplos, el número de casos que cumplen parecidas condiciones de orientación, posición de las piernas y tamaño, son limitados (ver tablas 5.3 y 5.4).

Para el caso del movimiento lateral, se consideran las cuatro clases ya citadas. En la imagen 5.10 se muestran los modelos probabilísticos obtenidos. Así mismo, para el movimiento frontal también se consideran cuatro clases; *piernas abiertas* ,*piernas semiabiertas*,*piernas cerradas* y *piernas semicerradas*. La imagen 5.11 muestra los modelos aprendidos.

Tipo de imágenes	Número de imágenes
Fondo de la imagen	120
Piernas abiertas	30
Piernas semiabiertas	30
Piernas semicerradas	30
Piernas cerradas	30

Tabla 5.3: Conjunto de entrenamiento 3, utilizado para aprender el movimiento lateral

En resumen, el método implementado se basa en el aprendizaje de una distribución de probabilidades de modelos de siluetas de peatones, a diferentes fases del ciclo de caminar. A continuación se explica el algoritmo que ha hecho uso de estos modelos.

5.4. Descripción del algoritmo

El algoritmo implementado toma como punto de partida el método para la extracción de regiones de interés (ROI) en imágenes FIR propuesto por [ND02]. Un profundo análisis y eva-

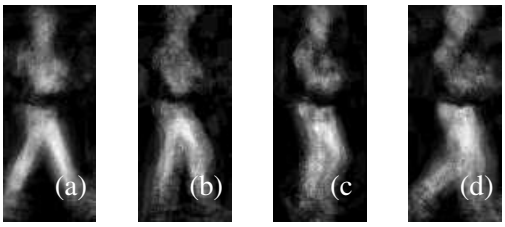


Figura 5.10: Modelos aprendidos para el movimiento lateral; (*a*) piernas abiertas, (*b*) semiabiertas, (*c*) cerradas y (*d*) semicerradas.

Tipo de imágenes	Número de imágenes
Fondo de la imagen	120
Piernas abiertas	30
Piernas semiabiertas	30
Piernas semicerradas	30
Piernas cerradas	30

Tabla 5.4: Conjunto de entrenamiento 4, utilizado para aprender el movimiento frontal

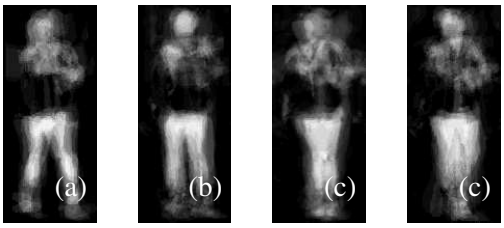


Figura 5.11: Modelos aprendidos para el movimiento frontal; (*a*) piernas abiertas, (*b*) semiabiertas, (*b*) cerradas y (*b*) semicerradas,

luación de la idea original ha permitido descubrir las debilidades de sus métodos, planteándose como una de las aportaciones de esta tesis, nuevas técnicas más robustas.

La detección de peatones se ha dividido en varias tareas, mostradas en la figura 5.12, que pueden ser agrupadas en las dos fases típicas con que se suele abordar el reconocimiento de objetos: la fase de detección de objetos y la posterior fase de clasificación.

5.4.1. Fase de Detección de Objetos de Interés

El objetivo de esta fase es la extracción de las regiones de interés de la imagen. Esta tarea es crucial, ya que la bondad de la clasificación posterior depende de la precisión de las regiones extraídas. Sin embargo, el algoritmo desarrollado confía en la lista de cajas (*bounding boxes*) proporcionadas por las etapas previas del sistema GOLD. De manera que el foco de interés de este algoritmo son las regiones contenidas en esas cajas.

A pesar de que el funcionamiento final del algoritmo concierne sólo al dominio infrarrojo, se han integrado las regiones de interés extraídas de ambos dominios, visible e infrarrojo.

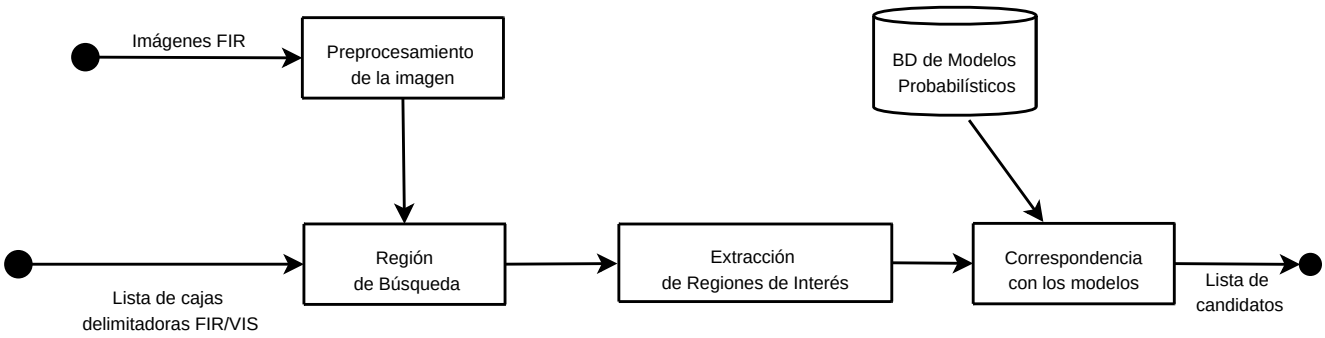


Figura 5.12: Diagrama de flujo del módulo probabilístico.

5.4.1.1. Preprocesamiento de las imágenes de entrada

La información de entrada al módulo probabilístico consiste en: la secuencia de imágenes FIR tomadas con el sistema estéreo Tetravision y una lista combinada de cajas extraídas de los dominios visible e infrarrojo.

No se ha impuesto restricciones a las escenas a tratar; ni en cuanto a las condiciones de iluminación, ni en lo referente al número de objetos en la escena o en movimiento. Teniendo en cuenta estas condiciones, se puede pensar en aplicar un preprocesamiento a las imágenes FIR, con el fin de facilitar la tarea de segmentación de las siluetas. Se han evaluado dos tratamientos distintos (ver fig. 5.13). La binarización de la imagen de entrada hacía que un gran número de píxeles del fondo fuesen considerados como objetos de interés, y a la hora de hacer corresponder el modelo del peatón, se observó un incremento de falsos positivos. En cambio, la umbralización simple modificaba los píxeles por debajo de un umbral a 0, consiguiendo así clasificarlos como fondo y dejaba intactos el resto de píxeles por encima del umbral, dejando la clasificación de los mismos para el enfoque probabilístico.

Figura 5.13: Ejemplos de los distintos tipos de preprocesamiento evaluados (*a*) Binarización y (*b*) umbralización simple de la imagen FIR de entrada.

5.4.1.2. Determinación de la región de búsqueda

En lugar de recorrer toda la imagen, la detección de los posibles peatones se limita a ciertas regiones. Las cajas proporcionadas por las etapas previas del sistema GOLD, dirigen el foco de interés hacia zonas que comparten una serie de características en cuanto a simetría, disparidad, intensidad y bordes verticales [BBF+07a].

Si bien se han tomado esas cajas como punto de partida para la búsqueda, sus estimaciones no siempre han resultado correctas. Los resultados experimentales han demostrado que los problemas más críticos del módulo probabilístico propuesto están relacionados con el *aspect-ratio* y la inexactitud de esas cajas. Por ello, ha sido necesario reestimar cada una de ellas.

1. Ajustar cada observación considerando el "*aspect-ratio*".

   Es sabido que una persona mantiene unas proporciones de tamaño. Haciendo uso de este conocimiento, se fuerza a que las observaciones o regiones encerradas en las cajas (*Bounding boxes*) tengan unas dimensiones adecuadas para contener a una persona. En caso de que las nuevas observaciones sean demasiado pequeñas o demasiado grandes, se eliminan de la lista de cajas, al asumir que no pueden contener una forma humana. El tamaño de las cajas es coherente con el de los modelos probabilísticos existentes, cubriendo un rango de dimensiones entre 210x90 a 29x12 píxeles.

   Desde luego, existen otros objetos en la escena que pueden producir una caja de dimensiones compatible con un peatón. Este hecho puede ocasionar un falso positivo si la fase de clasificación no es capaz de filtrarlo.

   A pesar de que el *aspect-ratio* no es un buen criterio para filtrar los resultados finales, si es una forma adecuada de generar hipótesis o candidatos potenciales. La decisión final se delega a la fase de clasificación.

2. Determinación de la zona de búsqueda.

   A pesar de haber redefinido el tamaño de cada caja, el proceso anterior no asegura que sus límites sean correctos. Puede ocurrir que encierren sólo parte de un peatón, perjudicando o incluso impidiendo la posterior detección.

   La solución adoptada consiste en definir una región de búsqueda alrededor de cada caja, permitiendo recuperar aquellos píxeles mal extraídos en etapas anteriores; cada caja se va a desplazar dentro de su región de búsqueda, hasta localizar la mejor observación. Este desplazamiento va a ser guiado por el criterio establecido en la clasificación.

Como resultado, cada caja reestimada tendrá asociada una región de búsqueda (ver fig. 5.14). De este modo, se permite que cada caja tenga una información más global del entorno que le rodea, pudiendo correguir errores de segmentación previos. Como consecuencia, se consigue disminuir el número de falsos positivos e incrementar la confianza de las detecciones.

5.4.1.3. Extracción de las regiones de interés de la secuencia FIR

Una vez establecida la región de búsqueda, el siguiente paso consiste en la segmentación de las ROI limitada a esa región.

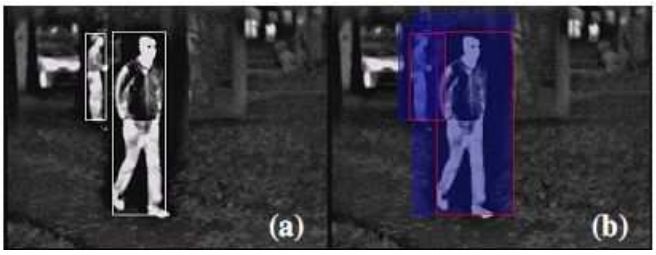


Figura 5.14: Determinación de la región de búsqueda (*a*) *Bounding Boxes* originales, proporcionadas por las etapas previas de GOLD, (*b*) Zona de búsqueda (azul) y los rectángulos (rojo) que corresponden a los *Bounding Boxes* redefinidos.

Para la extracción de las siluetas a partir de la secuencia de imágenes FIR, se han probado diversas técnicas:

- **Segmentación basada en la disparidad:**
  Se pueden reutilizar las regiones segmentadas en la etapa de detección de obstáculos FIR por el sistema GOLD. Sin embargo, con frecuencia había que unir y/o filtrar las regiones ofrecidas por el algoritmo estéreo. Por ello, se decidió desestimar esta opción.

- **Segmentación basada en intensidad:**
  Otra opción consiste en umbralizar las regiones. En el trabajo de [ND02], los objetos son extraídos del vídeo empleando la umbralización basada en intensidad, expresada mediante la siguiente ecuación;

$$th(x,y) = \begin{cases} I(x,y) & \text{if } I(x,y) \geq umbral_1 \\ 0 & \text{if } I(x,y) < umbral_1 \end{cases} \tag{5.5}$$

  donde $th(x,y)$ hace referencia al valor asignado al píxel en la posición (x,y) tras aplicar el $umbral_1$, obtenido mediante la ecuación 5.1.

  Como ya se ha comentado antes, resulta complicado separar a los peatones del fondo con este umbral y la clasificación bayesiana empleada por Nanda *et al.* supone que los píxeles correspondientes a los peatones siguen una gaussiana.

  Impulsados por los problemas obtenidos tras la aplicación de ese $umbral_1$, se ha seleccionado como valor el $umbral_2$ obtenido aplicando la ecuación 5.2, que consideraba como gaussianas la clase de no-peatones.

  Se propone como alternativa a la umbralización 5.5, la siguiente;

$$th(x,y) = \begin{cases} 255 & \text{if } I(x,y) \geq umbral_2 \\ 0 & \text{if } I(x,y) < umbral_2 \end{cases} \tag{5.6}$$

  donde los píxeles que estén por debajo del umbral (se espera que pertenezcan al fondo) son puestos a 0, mientras que los que estén por encima (se espera que correspondan a la silueta), se activan.

En la figura 5.15 se puede comparar los resultados obtenidos aplicando cada uno de esos dos umbrales.

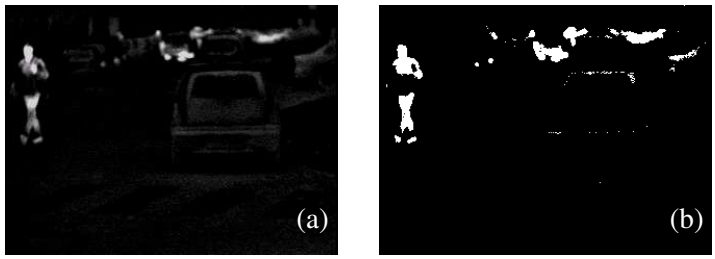


Figura 5.15: Segmentación basada en intensidad; (*a*) Resultado obtenido al aplicar la ecuación 5.5 de [ND02] y, (*b*) mediante la ecuación 5.6 que se propone como alternativa.

Cabría decir que las imágenes FIR utilizadas por Nanda *et al.*, deben de capturar peatones con unos valores de intensidad muy altos, ya que el umbral que usan es 190, y aseguran que obtenían buenos resultados aplicando valores en un rango de 170 a 205, y por otro lado, los peatones deben ser muy homogéneos para poder representar su distribución mediante gaussianas.

En las imágenes capturadas por el sistema Tetravision, los valores máximos de la intensidad capturados son alrededor de 220, estando la mayoría ($> 98\,\%$) de los píxeles por debajo de 127 (ver fig. 5.3).

- **Segmentación basada en modelos:**
  Si se emplea un simple umbralización, hay ocasiones en que los píxeles del frente de la imagen no se distinguen del fondo. Al basarse únicamente en los valores de la intensidad, si ésta no supera el umbral, se considerará como perteneciente al fondo.

  Una representación del peatón basado en modelos incorpora expectativas a la estructura de los peatones, y en base al nivel de confianza asociado a las expectativas, se puede ignorar el ruido contenido en las imágenes y rellenar la estructura esperada allí donde falta información.

  1. Aprendizaje del modelo global:

     Con el objeto de eliminar el ruido sistemático que suele aparecer en las secuencias de vídeo, se aprende un modelo que representa la apariencia de todos los peatones y al que vamos a llamar modelo global.

     Cabe esperar que los errores cometidos en la segmentación del fondo, tengan una probabilidad muy baja de ocurrir en la misma posición para una población de peatones. Para ello se construye un modelo basado en la silueta, a partir de un conjunto de entrenamiento de siluetas de peatones genéricos. El procedimiento es idéntico al explicado en el apartado 5.3.3, pero considerando todas las imágenes contenidas en el conjunto de entrenamiento. Es decir, las 240 imágenes contenidas en el conjunto de entrenamiento 5.1. La figura 5.16-a muestra el modelo global.

2. Segmentación basada en el modelo:

   Se puede utilizar el modelo global para eliminar el ruido y rellenar los agujeros en cada imagen, refinando la extracción de la silueta. Dado que el modelo se obtiene de la media de un conjunto de siluetas, se puede interpretar esa media como la estimación de máxima probabilidad para los parámetros de un proceso generativo de una población de siluetas. Es decir, la posición L de cada píxel es un proceso independiente de *Bernoulli*, con parámetro $\theta_L = p(L = 1)$.

   El objetivo es determinar, para cada peatón en la secuencia FIR, el valor binario de cada uno de los píxeles de su silueta. Para ello, se puede obtener la distribución a posteriori de θ_L dada la secuencia y establecer una probabilidad a priori en función de los parámetros de la población. En principio, se podría usar como umbral el valor máximo a posteriori de θ_L. Sin embargo, debido a que la probabilidad a priori del modelo sólo es válida para formas binarias estáticas, sólo se puede umbralizar con seguridad aquellos píxeles para los que la forma es estática en el tiempo, correspondiendo a una baja varianza para los procesos de *Bernoulli*. Empíricamente, se ha restringido la validez de la probabilidad a priori al rango $\theta_L \geq 0{,}75$ y $\theta_L \leq 0{,}25$. El resto de los píxeles no se modifican. Aplicando este conjunto de umbrales, se obtiene la máscara mostrada en la figura 5.16-b.

   Se propone usar la siguiente umbralización, que activa (pone a 255) aquellos píxeles que se corresponden con el torso y la cabeza del peatón, mientras que desactiva (pone a 0), aquellos píxeles que se encuentran lejos de los bordes de la silueta. Los píxeles que se dejan sin modificar corresponden a las piernas y los bordes de la silueta.

$$th(x,y) = \begin{cases} 255 & \text{if } p(x,y) \geq 0{,}75 \text{ AND } I(x,y) \geq umbral_2 \\ 0 & \text{if } p(x,y) \leq 0{,}20 \text{ AND } I(x,y) < umbral_2 \\ \text{I(x,y)} & \text{en el resto de los casos} \end{cases} \tag{5.7}$$

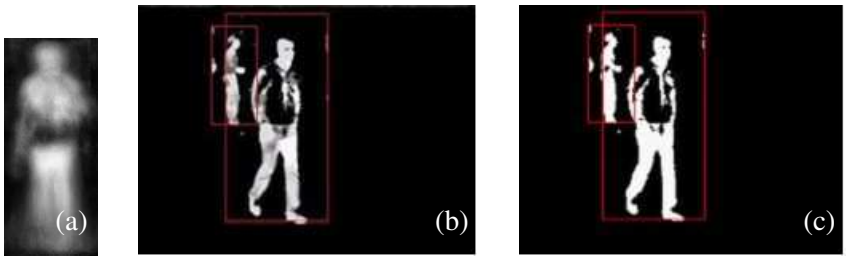


Figura 5.16: Comparación de los distintos métodos de segmentación; (*a*) Modelo global obtenido a partir de 5.1 (*b*) Resultado de la segmentación basada en el modelo (ec. 5.7) y, (*c*) de la segmentación basada en intensidad (ec. 5.6).

5.4.2. Fase de Clasificación de los peatones

Una vez extraídas las regiones de interés, tiene lugar la fase de reconocimiento de los peatones. Se hace uso de los valores de intensidad de cada píxel para clasificar los objetos de interés como peatones o no-peatones, bajo un enfoque probabilístico.

5.4.2.1. Detección basada en modelos probabilísticos

La lista de regiones proporcionada por la fase de detección anterior, contendrá peatones potenciales. La decisión final se toma en base a la probabilidad de que en cada región exista una persona.

Dado un modelo probabilístico y una caja a evaluar, se han probado dos fórmulas distintas para calcular la probabilidad combinada de que contenga un peatón. La primera de ellas se basa en la fórmula planteada en [ND02], mientras la segunda se propone en este algoritmo, en un intento por hacer frente a las limitaciones de la primera.

- **Probabilidad combinada original:**

  Para cada píxel (x, y), la probabilidad $C_1(x, y)$ de que la caja de dimensiones $m \times n$ que rodea a ese píxel contenga un peatón, se calcula como la suma de las contribuciones de los píxeles pertenecientes a la imagen th_{xy}. En este caso, th_{xy} se obtiene tras aplicar el umbral basado en modelos definido en la fórmula 5.7 a la imagen FIR original.

  La contribución de un píxel (i, j) en la región segmentada th_{xy}, viene dada por la expresión $(th\_xy(i, j) - 127) * (p(i, j) - 0{,}5)$, donde p es el modelo probabilístico. Un píxel con valor para $th_{xy}(i, j)$ superior a 127 (y por tanto más próximo a blanco que a negro), tendrá una contribución positiva proporcional al producto de su proximidad al blanco (es decir, $\left(th_{xy}(i, j) - 127\right)$) y su medida de confianza de que sea realmente blanco (i.e. $(p(i, j) - 0{,}5)$).

  Empleando $(t - 0{,}5)$, se asegura que la contribución es positiva si $(t \geq 0{,}5)$ (se espera que un píxel sea blanco y realmente lo es), y negativo si $(t < 0{,}5)$ (se espera que el píxel sea negro y lo es).

  La misma fórmula funciona para aquellos píxeles con una intensidad menor que 127. La contribución será positiva cuando el píxel observado tenga un valor próximo al valor estimado del píxel correspondiente en el modelo probabilístico (es decir, observamos un píxel negro en la imagen $th_{xy}(i, j) < 127$) y esperamos un píxel negro, ya que $p < 0{,}5$).

  Esto se expresa mediante la siguiente convolución:

$$C_1(x, y) = \sum_{i=1}^{m} \sum_{j=1}^{n} \left(th_{xy}(i, j) - 127\right) * (p(i, j) - 0{,}5) \tag{5.8}$$

- **Mejora a la fórmula original:**

  Sin embargo, la fórmula original 5.8 ha resultado inadecuada al aplicarla a las imágenes FIR en estudio. Los resultados experimentales han demostrado que la intensidad de la

mayoría de los píxeles está por debajo del umbral dado, por lo que la contribución a $C_1(x, y)$ para los píxeles oscuros es mayor que la obtenida a partir de los píxeles *claros*. Como consecuencia, la mayoría de los píxeles aportan una contribución negativa.

Otro inconveniente que se suma al anterior, tiene que ver con la dificultad de evaluar las probabilidades obtenidas. Aunque esos valores de correlación estén dentro de un rango, los límites del intervalo son desconocidos, como se constata de los valores mostrados en la tabla 5.17.

DIMENSIÓN / MODELO	168x72	135x58	108x46	87x37	70x30	56x24	45x19	36x15	29x12
ABIERTO	167655	104071	73046	47858	29560	21070	12716	7734	4473
SEMI-CERRADO	207936	143317	89212	59729	42550	25416	14137	9154	6047

Figura 5.17: Resultados de las probabilidades obtenidas aplicando la fórmula original de Nanda *et al.* (ec. 5.8)

Por estos motivos, se ha realizado una mejora al método original, proponiendo una nueva fórmula de correlación para tratar de resolver esos inconvenientes. La correlación se realiza sobre las regiones extraídas con el método de umbralización simple descrito en 5.6.

Esta vez no se usa el modelo general como máscara para la segmentación. El objetivo de usar ese modelo, era incrementar la intensidad de los píxeles del frente de la imagen. Como resultado, el contraste con el fondo era mayor y la tarea de segmentación se facilitaba.

En cambio, usando la ecuacion 5.6 para extraer los objetos de interés, no hay necesidad de aumentar la intensidad de la imagen ya que sus valores son, o 0 (corresponde al fondo) o 255 (frente de la imagen).

La nueva fórmula de correlación se define como:

$$C_2(x, y) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(th_{xy}(i, j) - 127)) \times (p(i, j) - 0{,}5)}{\sum_{i=1}^{m}\sum_{j=1}^{n}|p(i, j) - 0{,}5|} \tag{5.9}$$

donde $th_{xy}(i, j)$ es la imagen de entrada umbralizada después de aplicar el filtro descrito en la ecuacion 5.6. Así, la contribución de los píxeles es proporcional a los valores correspondientes en la plantilla o modelo probabilístico; únicamente considerando el signo de los píxeles de entrada – positivo si pertenece al frente y negativo en caso contrario– , en lugar de considerar los valores de la intensidad, la correlación resultante es más precisa, ya que no favorece la presencia de algunos píxeles frente a otros.

Además, la probabilidad es calculada de manera separada para los píxeles del fondo y del frente, en un intento por dar el mismo peso a la contribución de cada región. Como consecuencia, el valor de probabilidad conjunta se obtiene sumando los dos términos normalizados,

$$C_2(x,y) = \frac{1}{2}\left(C_{2\ fondo}(x,y) + C_{2\ frente}(x,y)\right) \tag{5.10}$$

Por último, la fórmula propuesta 5.10 está normalizada entre [-1, 1], permitiendo una forma más sencilla de comparar los valores de correlación obtenidos que mediante la fórmula anterior 5.8.

5.4.2.2. Selección del mejor candidato. Máximo a posteriori

Cada uno de los modelos probabilísticos generados, son escalados para ajustarse a la caja o cajas (en caso de que la búsqueda sea multidimensional) contenidas en cada región. En cada caja se calcula la probabilidad combinada de que exista un peatón empleando la fórmula 5.8 o la 5.10. Considerando cada caja o observación x como una hipótesis, se asignará a la clase $C = \{abiertas, semiabiertas, cerradas, semicerradas\}$ de máxima probabilidad. Es el enfoque de hipótesis maximum a posteriori o MAP:

$$\hat{C}_{MAP} = \underset{C_i}{\text{argmax}}\ P(C|x) \tag{5.11}$$

El cálculo de la probabilidad de que dada una hipótesis o posible peatón, verdaderamente lo sea es complicada. Esa probabilidad $P(C|x)$ se denomina probabilidad a posteriori ya que estima la probabilidad una vez se tiene una observación. Esta probabilidad debe responder a la pregunta: si se está esperando un peatón de la clase *piernas abiertas*, cuál es la probabilidad de que esa clase sea lo observado en la imagen. El método empleado para este cálculo consiste en la correspondencia de patrones probabilísticos, asignando a la observación la etiqueta de mayor probabilidad. Es un enfoque basado en la máxima verosimilitud. En el siguiente capítulo se explica en detalle la fórmula 5.11, así como el concepto de máxima verosimilitud.

Si la probabilidad de que en una caja exista un peatón está por debajo de un umbral, se considera que no existe peatón. Este umbral lo establece el usuario del sistema. Se han realizado pruebas aplicando distintos umbrales y por defecto se emplea un valor de 0,55. El valor de la probabilidad es un indicativo de la confianza de que exista un peatón en la caja.

5.5. Rango de la detección

A la hora de diseñar el algoritmo resulta fundamental definir cuál va a ser el objeto deseado, es decir, establecer el rango del alto y ancho de los peatones.

5.5.1. Definición de los objetos de interés

Hay que recordar que el algoritmo hereda una lista de regiones de interés. El sistema GOLD define los objetos de interés en base a un tamaño y un *aspect-ratio* específicos. El tamaño de los peatones queda definido por:

- Altura: $160cm - 200cm$
- Ancho: $40cm - 80cm$

La gran tolerancia del ancho considera distintas posturas del peatón (por ejemplo, cuando un peatón se cruza en la trayectoria del observador). De hecho, sólo se consideran como válidas las combinaciones de alto y ancho que satisfacen unos límites concretos del *aspect-ratio* (se considera un rango de 2,4 a 4,0 para el ratio $alto/ancho$.

5.5.2. Especificación del rango de la detección

La presencia de un peatón se comprueba en cajas de distintos tamaños, situadas en distintas posiciones de la imagen. Sin embargo no todas las cajas deben ser consideradas. Es por tanto fundamental determinar un rango de cajas razonables, que permita obtener unos resultados suficientemente precisos. Se toma como base los tamaños de las *bounding boxes* heredadas de GOLD:

- La caja más pequeña es de 28x7 píxeles
- La caja más grande es de 100x40 píxeles.

Los límites del alto de la caja (28 y 100 píxeles) fueron determinados experimentalmente, mientras que los límites del ancho (7 y 40 píxeles) fueron calculados en función de los valores definidos para el alto y ancho deseados. Es decir; $7 = 28/(160/40), 40 = 100/(200/80)$.

Sin embargo, esta elección limita el área de detección frontal del coche. Con la esa especificación de las cajas se puede detectar un peatón de 170 centímetros de alto que se encuentre en un rango de distancias entre 13 a 46 metros [BBF+04]. En realidad, el rango de detección se reduce aún más, debido a que el rango de alturas permitidas se restringe a $(160cm - 200cm)$. De hecho, para que el sistema sea capaz de detectar, a todos los peatones (con una altura comprendida dentro del rango permitido) que se encuentran a una distancia, el área de trabajo se reduce aún más, cubriendo entre $15m - 43{,}5m$ (ver fig. 5.18-a).

Como solución se adopta un enfoque multiresolución; se reduce el tamaño de la imagen original para ampliar el rango de detección e incluir a peatones que se encuentren más cerca de la cámara. Para los peatones lejanos, no se puede superar el límite superior de $43{,}5m$ ya que no se puede compensar la baja resolución de la información contenida en cajas muy distantes. Como resultado de la reducción de la imagen, el tamaño de las *bounding boxes* de peatones cercanos cumplen los límites impuestos por el sistema en cuanto al máximo tamaño permitido.

La figura 5.18-b se muestra el nuevo rango de detección, aplicando un submuestreo de $1 : 2{,}15$ a la imagen original, cubriendo un espectro de $7m - 20m$. El límite inferior no puede ser más bajo, ya que viene impuesto por el sistema de visión. En conjunto, la búsqueda de

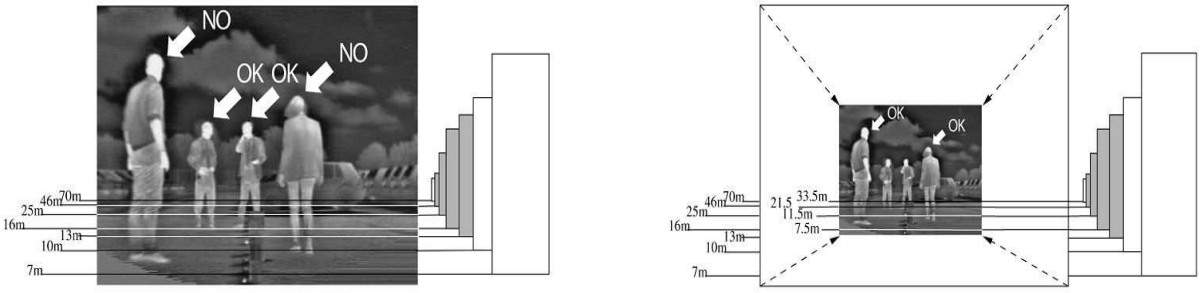


Figura 5.18: Rango de detección; (*a*) Peatones de diferentes alturas, situados a distintas distancias. La *bounding box* contiene a un peatón de 170 cm. de alto a diferentes distancias. En blanco, el rango de detecciones factible para un peatón de 170 cm. de altura, (*b*) Después del submuestreo, los peatones próximos a la cámara caen en el rango de detección.

peatones se va a realizar en un rango de distancias entre $7m$ y $43{,}5m$. El submuestreo $1:2{,}15$ equivale a usar un rango de *bounding boxes* entre $2{,}15xBB_{hmin} - 2{,}15xBB_{hmax}$ para al alto de las cajas y $2{,}15xBB_{wmin} - 2{,}15xBB_{wmax}$. Esto equivale a definir,

- La caja más pequeña es de 60x15 píxeles
- La caja más grande es de 215x86 píxeles.

5.5.3. Definición de los modelos. Enfoque multiresolución

En cuanto a la especificación de los modelos probabilísticos,se ha tenido en cuenta las dimensiones empleadas en la descripción de las cajas, ya que posteriormente serán evaluadas por esos modelos. Al mismo tiempo, se asegura que la evaluación de las cajas sea coherente con el módulo previo, considerando todos los posibles peatones que se encuentren en un margen entre 7m y 43.5m. Se han llevado a cabo varias pruebas modificando el tamaño de los modelos, hasta determinar experimentalmente unas dimensiones de 210x90 píxeles, manteniendo un *aspect-ratio* de 2,3, cercano al 2,5 de las cajas.

Sin embargo, un sólo modelo no cubre todos los posibles tamaños que pueden presentar las personas en las imágenes, resulta necesario adoptar un enfoque multiresolución. Pero, debido a que un modelo probabilístico no se puede muestrear por debajo del 75 % de su tamaño original sin que pierda su eficacia, se ha decidido aplicar un factor de reducción del 0.80 % aplicado al modelo más grande. Como resultado se obtienen 10 modelos, cuyas dimensiones van de 210x90 hasta $29x12$ (ver fig. 5.19).

A pesar del elevado número de modelos, no añade carga computacional al sistema, ya que son creados durante la fase de aprendizaje previa.

5.6. Resultados experimentales

El módulo desarrollado ha sido probado en diferentes situaciones, utilizando un vehículo experimental equipado con el sistema Tetravision. El algoritmo propuesto ha sido ejecutado en

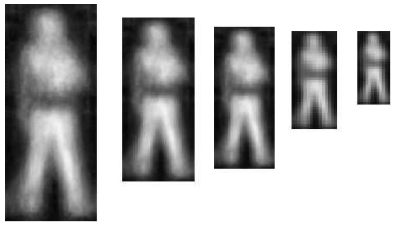


Figura 5.19: Cada modelo original de 210x90 se ha reducido hasta un tamaño mínimo de 29x12. En total se obtienen 10 modelos de tamaños 210x90, 168x72, 135x58, 108x46, 87x37, 70x30, 56x24, 45x19, 36x15, 29x12. En la figura se muestran algunos ejemplos del modelo de piernas abiertas.

tiempo real en varias secuencias FIR. La figura 5.31 muestra algunos resultados de la correspondencia con los distintos modelos probabilísticos. Si la confianza en la detección es superior a un umbral dado, la *bounding box* se dibuja en color rojo, queriendo decir que la región tiene una alta probabilidad de contener una forma humana, o en color azul, si la confianza no es tan elevada. Se puede observar en los resultados que el algoritmo es capaz de detectar uno o más peatones, aún estando próximos, parcialmente ocluidos o en presencia de fondos complejos.

No obstante, un enfoque de correlación basado en patrones (*template matching*) falla cuando la forma humana en la imagen es muy distinta al conjunto de entrenamiento. Por ejemplo, en la figura 5.21-b, el peatón que está sentado en la motocicleta no es reconocido, debido a que los modelos probabilisticos creados no codifican a humanos que estén sentados.

Uno de los problemas más críticos tiene que ver con el *aspect ratio* y las *bounding boxes* imprecisas. La figura 5.21-c y d muestran como algunos objetos en la escena pueden producir una *bounding box* que es compatible con un peatón. Esto puede hacer que la correlación con el modelo humano de lugar a falsos positivos. Otro problema tiene que ver con *bounding boxes* conteniendo partes de un peatón. Esto puede generar falsos negativos, ya que la forma del peatón es incompleta y la correlación falla. Algo parecido ocurre ante las oclusiones, aunque no es un problema que se mantenga durante varias imágenes, por lo que el *tracking* podría resolver este error.

En el caso de cajas conteniendo a peatones muy lejanos (a una distancia de más de 30m), caracterizadas por encerrar una información de muy baja resolución, pueden ser fácilmente confundidas con otros objetos en la escena que presenten unas características termales similares (como ocurre habitualmente con los vehículos). En la figura 5.21-a se observa que el sistema es capaz de detectar a peatones que estén a más de 30 metros.

El funcionamiento del algoritmo ha sido evaluado a través de las curvas ROC (*Relative Operating Characteristic*) mostradas en las gráficas 5.22 y en función del umbral de correlación. La primera curva (5.22-a) ha sido obtenida al ejecutar el sistema empleando la ecuación de correlación 5.8. Utilizando este método, el porcentaje de detecciones correctas crece de modo suave. La segunda curva (5.22-b) muestra el resultado de aplicar la ecuación 5.10, sólo en aquellas regiones que han sido segmentadas mediante la ecuación 5.6. Los datos utilizados para generar las curvas ROC se componen de 100 *bounding boxes* conteniendo a peatones y no-peatones.

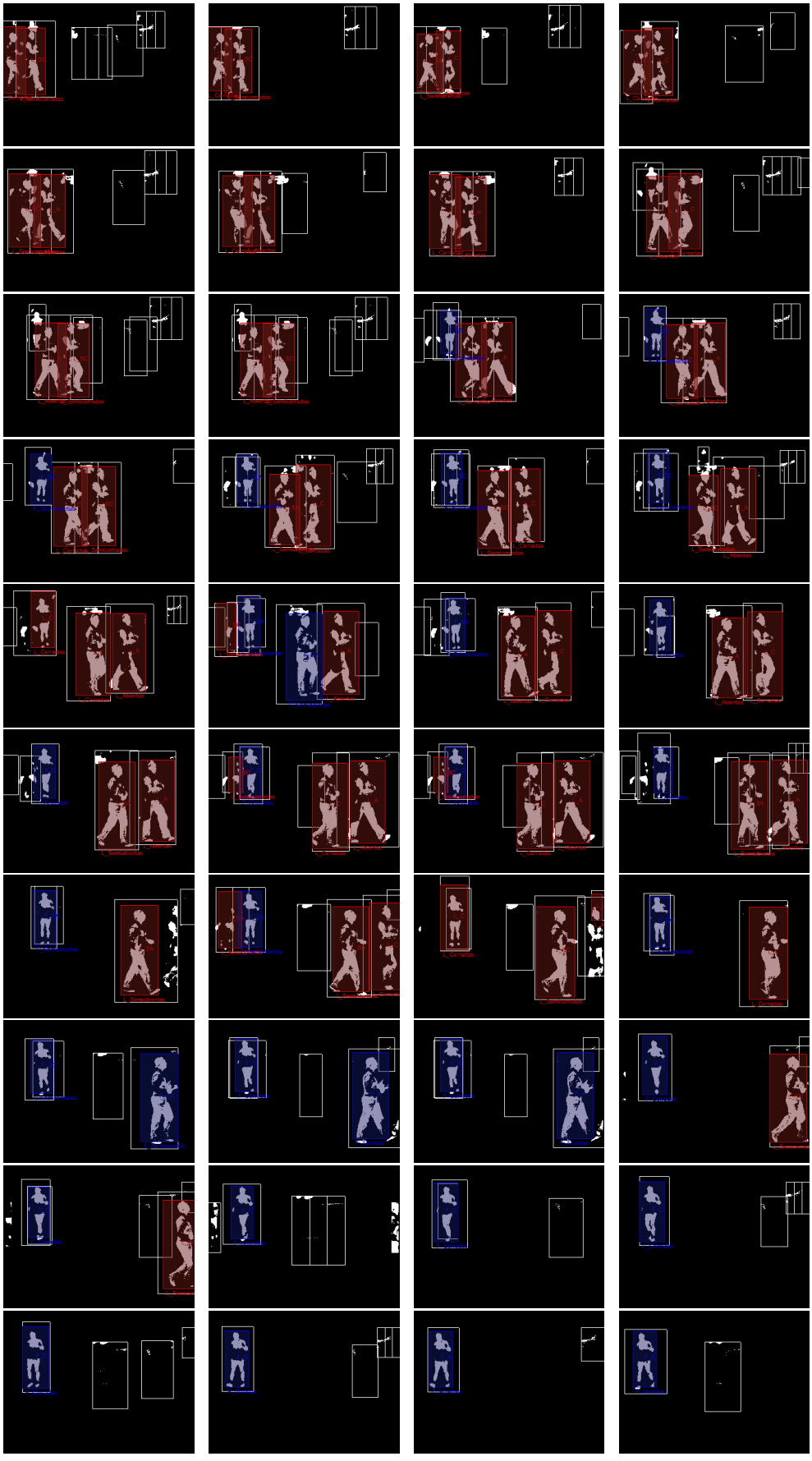


Figura 5.20: Secuencia de imágenes FIR con las etiquetas asignadas por el matching probabilístico, que son la entrada a los modelos ocultos de Markov. En rojo se dibujan las detecciones con una alta probabilidad de ser peatones, mientras que el azul corresponde a detecciones con una menor confianza.

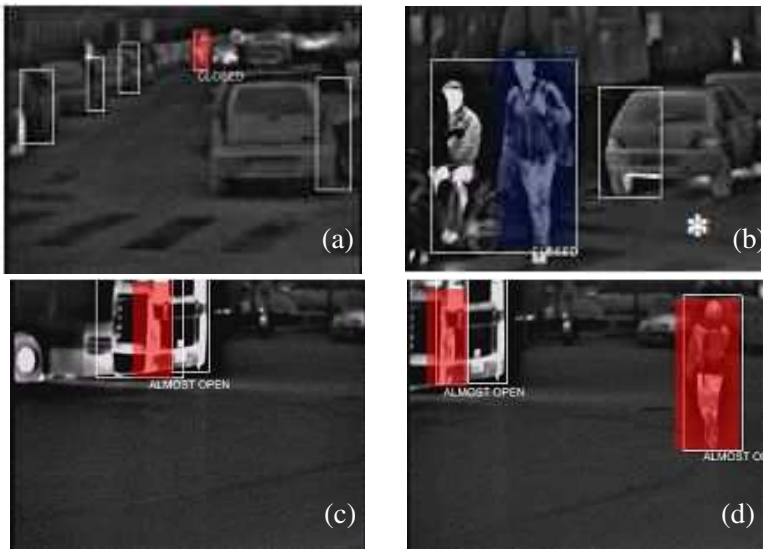


Figura 5.21: Resultados erróneos y detecciones lejanas; (*a*) Ejemplo de un peatón muy lejano (a una distancia de más de 30m), detectado por el sistema. (*b*) Errores debidos a la inexistencia de un modelo probabilístico que reconozca peatones sentados y (*c*) y(*d*) errores debidos al filtrado basado en el *aspect ratio*; objetos con un tamaño similar al de un peatón pueden dar lugar a falsas detecciones.

Se puede observar que en estas condiciones el sistema es capaz de detectar con éxito más del 85 % de los peatones en la escena, con un número muy bajo de falsas detecciones por imagen. Por tanto, la mejora de la robustez del algoritmo se debe a las modificaciones realizadas a la implementación original.

5.7. Integración temporal de los peatones detectados

En este capítulo se ha hablado del enfoque probabilístico como metodología adoptada para el reconocimiento de peatones. Este enfoque se ha extendido al reconocimiento de los peatones en el tiempo, como se va a explicar más adelante.

Cuando se dispone de suficiente información sobre el entorno (por tanto, es un proceso determinista), el enfoque de la lógica permite tomar decisiones correctas. Desafortunadamente, casi nunca se dispone de toda la verdad sobre el entorno (es un proceso estocástico), debiendo desarrollar sistemas capaces de comportarse bajo incertidumbre. En el caso de la detección de peatones a partir de imágenes, sólo se dispone de información local y generalmente, incompleta. En estas condiciones resulta un reto tratar de descubrir si existe una persona en la imagen.

La teoría de la probabilidad proporciona un marco completo y formal para la toma de decisiones realizadas bajo incertidumbre. El razonamiento (incierto o no) forma parte de unos de los enfoques de la inteligencia artificial (IA), basado en la aplicación de las matemáticas a la toma de decisiones. Bajo el enfoque racional, se trata de alcanzar el mejor resultado, o

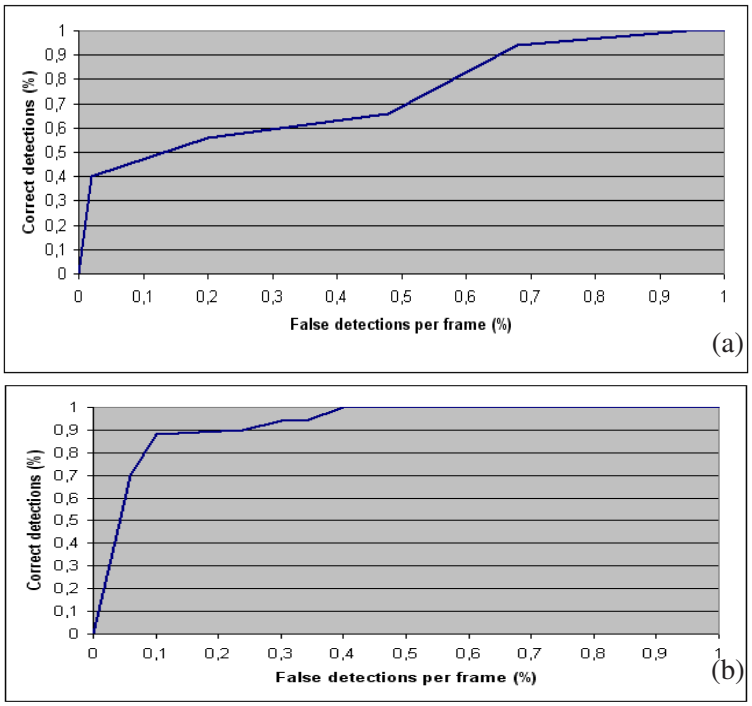


Figura 5.22: Evaluación del algoritmo mediante curvas ROC en base al umbral de correlación; (*a*) curva obtenida al aplicar la ec. 5.8 (basada en la idea original de Nanda) y (*b*) al aplicar la ec. 5.9. Se observa el incremento en el número de detecciones correctas al utilizar la modificación propuesta.

cuando hay incertidumbre, el mejor resultado esperado.

La regla de Bayes y el área resultante llamado análisis Bayesiano conforman la base de las propuestas más modernas que abordan el razonamiento incierto en los sistemas de IA [RN03]. Un buen modelo de la tendencia actual es el campo del reconocimiento del habla, donde las aproximaciones basadas en los modelos ocultos de Markov (*Hidden Markov Models*, HMM), han pasado a dominar el área recientemente. Esto es debido a que; primero, se basan en una rigurosa teoría matemática y en segundo lugar, los modelos se generan mediante un proceso de aprendizaje a través de conjuntos de datos reales, lo cual garantiza una funcionalidad robusta.

5.8. Reconocimiento del caminar humano mediante modelos ocultos de Markov

La motivación de usar modelos ocultos de Markov surge de la necesidad de corregir los errores cometidos por el módulo probabilístico previo. La finalidad de dicho modulo es etiquetar las siluetas extraídas de las imágenes FIR, llevando a cabo una clasificación de la posición de las piernas. Sin embargo, son varias las causas que pueden conducir a un etiquetado erróneo. En líneas generales:

- Errores de etiquetado: Existe un peatón, pero no se etiqueta adecuadamente. Una seg-

mentación deficiente ocasiona este tipo de errores. En cualquier caso, el peatón es detectado, pero equivoca el etiquetado.

- Falsos positivos: No existe un peatón, pero se etiqueta como si existiese Los falsos negativos se originan por la segmentación errónea de zonas que emiten calor en la imagen.
- Falsos negativos: Existe un peatón, pero el sistema no lo detecta. Las consecuencias de éstos son mucho más graves que los dos casos anteriores, al no poder avisar al conductor de la existencia del peatón.

La finalidad de usar HMMs es corregir los errores de etiquetado y reconocer el tipo de desplazamiento que está efectuando el peatón, a saber, si se está moviendo frontalmente o lateralmente a la cámara. La inferencia bayesiana permite modelar información ruidosa y proporcionan un modo de efectuar correcciones.

5.8.1. Especificación del modelo

La idea de modelar un canal de información ruidoso se ajusta al caso en estudio. Se puede considerar que las siluetas de peatones son defectuosas, debido a que las siluetas correctas han sufrido alteraciones al pasar por un canal de comunicaciones ruidoso. Este canal introduce ruido eliminando píxeles de la silueta del peatón o mediante alguna otra alteración, que dificulta la tarea de reconocer la ”verdadera” silueta. Por tanto, el objetivo es modelar ese canal. Dado ese modelo, se encontrará la silueta verdadera (identificada por una etiqueta), pasando un conjunto de siluetas correctas por ese modelo del canal ruidoso y luego comprobando cuál es la que más se parece a la silueta defectuosa. Los efectos del canal ruidoso, generalmente pueden reflejarse mediante un modelo de observación.

Este modelo es un caso especial de inferencia bayesiana; se observa una observación o_i – una silueta defectuosa a la que se ha asignado una etiqueta – y hay que encontrar la etiqueta q_i correcta que identifique la posición de las piernas. De entre todas las posibles etiquetas se quiere encontrar aquella que maximiza la siguiente expresión, obtenida aplicando la regla de Bayes (ver ec. C.3):

$$\hat{q}_{MAP} = \underset{q_i \in Q}{\operatorname{argmax}} \; P(O|Q)P(Q) \tag{5.12}$$

La probabilidad de que el modelo genere una secuencia de observaciones O concreta viene definido por $P(O|Q)$. Mientras que la probabilidad de distribución sobre las posibles etiquetas ocultas es modelado por $P(Q)$. La etiqueta $\hat{q}$ más probable, dado que se ha observado una etiqueta incorrecta o_i, puede calcularse mediante el producto de la probabilidad a priori de la etiqueta $P(q_i)$ y la verosimilitud de la observación $P(o_i|q_i)$, seleccionando la etiqueta que de el mayor producto.

Para resolver la fórmula 5.12 se ha considerado un modelo bigrama o de primer orden (ver Anexo C.12). Según la condición de Markov de primer orden, la probabilidad de un estado q_i sólo va a depender del estado anterior q_{i-1}. Debido a que se está analizando una secuencia, existe una dependencia entre la etiqueta asignada en el instante t_{i-1} y la asignada en el instante

siguiente t_i. Si se analiza el movimiento de las piernas de un humano durante un ciclo, se puede apreciar esa relación; si en una imagen el individuo tiene las piernas abiertas, en el instante siguiente la posición de las mismas habrá sufrido una variación respecto al estado anterior (semicerradas o semiabiertas).

La decisión de usar un modelo bigrama se debe a varios motivos; en principio, se consideró usar un modelo bigrama por sencillez. Si se considera un modelo de Markov de un orden superior, los cálculos de la probabilidad se complican. Por otro lado, para predecir el movimiento de las piernas en un estado la mayor influencia la tiene el estado predecesor. El hecho de incluir más estados supone complicar el modelo, no estando claro que aporte para este caso una mayor robustez.

Por último, las personas pueden seguir distintas trayectorias. Las dinámicas que describen cada dominio estarán contenidas en su correspondiente modelo de evolución. Como ya se expuso en el capítulo anterior, se considera tanto a peatones que caminan de frente como de lado con respecto a la cámara. Por tanto debemos tener dos modelos distintos para modelar cada una de esas trayectorias.

El algoritmo final debe ser capaz de corregir los errores de etiquetado cometidos por el módulo probabilístico precedente y clasificar las secuencias de observaciones, identificando el tipo de trayectoria seguida por el peatón. El primer objetivo se alcanza mediante el algoritmo de Viterbi y el segundo mediante el cálculo del modelo más probable. No obstante, el primer paso antes de poder realizar ambos cálculos consiste en el entrenamiento de los modelos.

5.8.2. Entrenamiento de los modelos

Se va a emplear dos HMMs para modelar las secuencias de etiquetas correspondientes a los dos tipos de trayectorias. Resulta complicado determinar qué parámetros representan de un modo eficaz esas trayectorias. La solución consiste en estimar esos parámetros, a partir de una base de datos conteniendo secuencias de observaciones, de un modo supervisado o no-supervisado. El tipo de entrenamiento realizado en esta tesis es supervisado.

Dado un par de conjuntos de entrenamiento, conteniendo respectivamente siluetas de peatones caminando frontalmente y lateralmente a la cámara, se quiere entrenar un HMM para cada una de las dos posibles direcciones tomadas al caminar. Debido a que es un aprendizaje supervisado, cada imagen de la base de datos está etiquetada, estableciéndose así los límites entre las siluetas observadas durante cada fase del ciclo de caminar.

5.8.2.1. Arquitectura de los modelos

Después de analizar secuencias de imágenes conteniendo a peatones caminando en las dos direcciones consideradas, a continuación se exponen los motivos que han llevado a implementar las siguientes estructuras para cada modelo.

Los HMMs construídos representan la apariencia de las siluetas, donde cada estado describe la silueta en diferentes fases del ciclo de caminar. Por tanto, se asume que las observaciones asociadas a cada fase o estado siguen densidades de probabilidad específicas. Esto quiere decir que las observaciones son separables en clases.

- Modelo para el reconocimiento del movimiento lateral:

  Para el caso de peatones caminando lateralmente, se han considerado 4 clases o estados. Esto significa que las observaciones (imágenes de entrenamiento tomadas de secuencias FIR, que contienen siluetas de peatones laterales), se han agrupado en 4 clases en base a la posición de las piernas. Estas clases son las mismas que se han considerado en el módulo probabilístico: piernas abiertas, semicerradas, cerradas y semicerradas.

  Así, el movimiento lateral puede considerarse el resultado de observar siluetas obtenidas de la distribución de piernas abiertas, después transitar a la distribución de piernas semicerradas y observar alguna silueta, a continuación transitar al estado de piernas cerradas y observar, para finalmente alcanzar el estado de piernas semiabiertas y observar. Parece razonable modelar una secuencia de movimiento lateral mediante un HMM de izquierda-a-derecha o modelo de *Bakis*, compuesto de 4 estados (ver fig. 5.23).

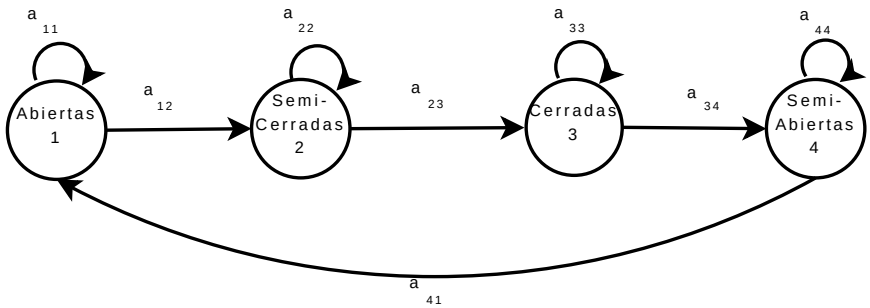


Figura 5.23: Arquitectura del modelo izquierda a derecha o modelo de *Bakis* para el reconocimiento del movimiento lateral .

- Modelo para el reconocimiento del movimiento frontal:

  De forma análoga, para el caso del movimiento frontal, las observaciones consisten en imágenes FIR tomadas del conjunto de entrenamiento que contienen siluetas de peatones de frente. Dichas siluetas se han agrupado en 4 clases o estados: piernas abiertas, semiabiertas, semicerradas y cerradas.

  El movimiento frontal está formado por observaciones que siguen un patrón menos estructurado que el movimiento lateral, observándose transiciones aleatorias entre los distintos estados, como refleja el grafo de la figura 5.24.

De manera que se ha modelado el caminar humano como un conjunto cíclico de transiciones entre 4 estados discretos.

5.8.2.2. Estimación de los parámetros

Un HMM puede describirse como un modelo generativo que modela un proceso (el caminar humano), como una sucesión de estados unidos por transiciones. Para la estimación de los parámetros que definen cada HMM (a saber, $\Theta = \{\pi, A, B\}$), se han empleado secuencias FIR

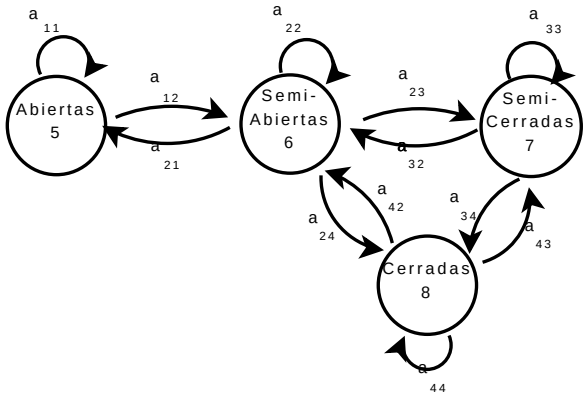


Figura 5.24: Arquitectura del modelo casi ergódico para el reconocimiento del movimiento frontal .

de peatones caminando en ambas direcciones. A partir de esas imágenes se han realizado los siguientes cálculos:

Figura 5.25: Ejemplos de imágenes de entrenamiento de peatones divididos en 4 clases; *Abiertas, semiabiertas, cerradas* y *semicerradas*, para cada uno de los movimientos (lateral y frontal). Además, cada clase se ha etiquetado con el estado al que pertenece (E1-E8).

- **Cálculo de las probabilidades a priori (Vector π):**
  Se considera que un peatón tiene la misma probabilidad de iniciar a caminar en cualquiera de los estados. Por tanto;

$$P(S_0 = i) = \frac{1}{N} \tag{5.13}$$

siendo $N = 4$ el número de estados considerados.

- **Cálculo de las probabilidades de emisión o del modelo de observación (Matriz B):**

  La secuencia de estados, que corresponden a los objetos de interés, pueden ser observados sólo a través de los procesos estocásticos definidos en cada estado; es necesario conocer las distribuciones de probabilidad de cada estado antes de poder asociar una secuencia de estados $Q = \{q_1...q_t\}$ a una secuencia de observaciones $O = \{o_1...o_t\}$.

  Las probabilidades de emisión, también conocidas como las probabilidades de observación, son las funciones de distribución de probabilidad que caracterizan cada estado.

  Para el cálculo del modelo de emisión se toma como punto de partida un conjunto de variables aleatorias de*Bernoulli* independientes. El proceso es el siguiente:

  1. Las observaciones de entrenamiento(siluetas binarias extraídas de secuencias FIR) se dividen en clases. Es un proceso manual y como resultado se obtiene 2 conjuntos de entrenamiento (uno conteniendo a peatones laterales y el otro, frontales) compuestos de 4 clases de siluetas cada uno de ellos (ver fig. 5.26).

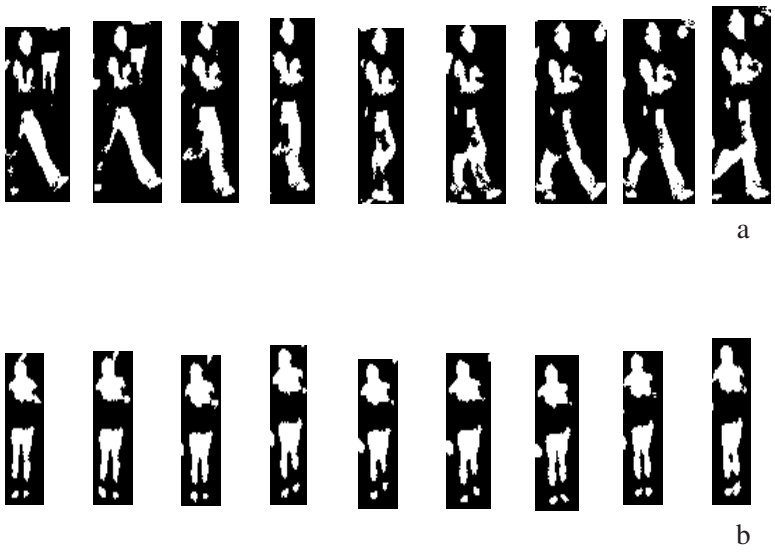


Figura 5.26: Ejemplos de observaciones extraídas de las imágenes FIR, antes de ser alineadas y redimensionadas al mismo tamaño; (*a*) Trayectoria lateral y (*b*) frontal de un peatón.

  2. Como se tiene un conjunto de observaciones para un estado dado, se puede estimar el parámetro de *Bernoulli* contando el número de píxeles activos para cada conjunto de observaciones. El resultado es un modelo probabilístico, donde cada píxel representa la media de las siluetas asignadas a un estado dado. Estos nuevos modelo probabilísticos describen una apariencia similar a la obtenida en el módulo previo. Sin embargo, el conjunto de imágenes utilizadas para su creación es distinto (ver tabla 5.5). Se han vuelto a calcular los modelos probabilísticos porque

es necesario seleccionar con cuidado el conjunto de siluetas agrupados en cada clase. De lo contrario, el entrenamiento no es válido. Por eso, si se compara el conjunto de entrenamiento empleado ahora (formado por 120 imágenes), con el empleado para el reconocimiento (formado por 240 imágenes), se observa que se han eliminado las siluetas que podían asignarse a varias clases.

Tipo de imágenes	
Movimiento Lateral o Frontal	Número de Imágenes
Piernas abiertas	30
Piernas semicerradas	30
Piernas cerradas	30
Todo tipo de peatones	30

Tabla 5.5: Conjunto de entrenamiento utilizado para el movimiento lateral o frontal. En ambos casos cada clase contiene 30 imágenes.

En total se obtiene 8 modelos probabilístico (ver fig. 5.27).

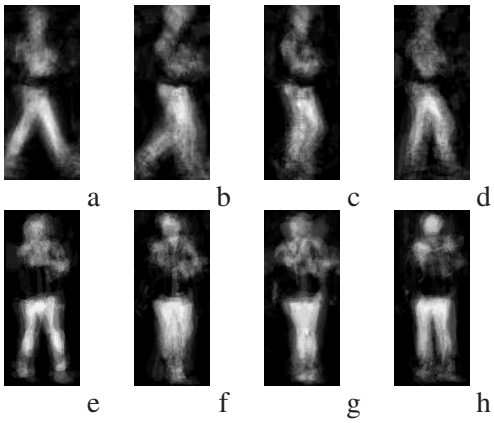


Figura 5.27: Detalle de los 8 modelos probabilísticos generados mediante el enfoque HMM Para el movimiento lateral y frontal, respectivamente; (*a*) y (*e*) abiertas, (*b*) y (*f*) semicerradas, (*c*) y (*g*) cerradas, (*d*) y (*h*) semiabiertas.

3. Una vez que se tiene la distribución de probabilidad que identifica cada estado, se ha utilizado esos modelos probabilísticos para etiquetar cada observación de entrenamiento. Si el canal de información no fuese ruidoso, cada estado sólo emitiría observaciones de su clase. Sin embargo, debido a que las siluetas son propensas a fallos, es habitual estar en un estado y observar siluetas pertenecientes a otros estados. Sin embargo, no resulta sencillo determinar la probabilidad de estar en un estado y observar una silueta defectuosa $P(O|Q)$, que de hecho es lo que refleja la probabilidad de emisión, cuyo cálculo se persigue.

   Para obtener esa probabilidad se va a asignar etiquetas O al conjunto de siluetas observadas en cada estado. Sobre cada silueta se calcula la correlación con cada

modelo probabilístico, asignando la etiqueta de máxima probabilidad. Es un procedimiento idéntico al empleado en el módulo probabilístico previo.

Así, se va a calcular la máxima verosimilitud de cada etiqueta, contando el número de veces que en un estado Q ={ *piernas abiertas*, *piernas semiabiertas*, *piernas cerradas*, *piernas semicerradas* } se observa cada una de las etiquetas *O*.

$$P(o_i|q_i) = \frac{C(q_i, o_i)}{q_i} \tag{5.14}$$

Aplicando la ecuación 5.14 se obtienen los modelos de emisión correspondientes a cada tipo de dirección. Las matrices de emisión 5.15 y 5.16 reflejan las distribuciones discretas de las etiquetas y sus valores son los siguientes:

$$b_{i_{\text{lateral}}}(O_j) = \begin{pmatrix} Estado/Observ & O1 & O2 & O3 & O4 & O5 & O6 & O7 & O8 \\ E1 & 0{,}5263 & 0{,}0526 & 0{,}0 & 0{,}42 & 0{,}0 & 0{,}0 & 0{,}0 & 0{,}0 \\ E2 & 0{,}0 & 0{,}747 & 0{,}1 & 0{,}1 & 0{,}0 & 0{,}0 & 0{,}0526 & 0{,}0 \\ E3 & 0{,}0 & 0{,}105 & 0{,}4785 & 0{,}1 & 0{,}0 & 0{,}105 & 0{,}1578 & 0{,}052 \\ E4 & 0{,}1052 & 0{,}0525 & 0{,}0 & 0{,}684 & 0{,}0526 & 0{,}105 & 0{,}0 & 0{,}0 \end{pmatrix} \tag{5.15}$$

$$b_{i_{\text{frontal}}}(O_j) = \begin{pmatrix} Estado/Observ & O1 & O2 & O3 & O4 & O5 & O6 & O7 & O8 \\ E5 & 0 & 0 & 0 & 0{,}105 & 0{,}789 & 0{,}105 & 0 & 0 \\ E6 & 0 & 0 & 0 & 0{,}0526 & 0{,}105 & 0{,}7368 & 0{,}1052 & 0 \\ E7 & 0 & 0 & 0{,}1 & 0 & 0 & 0{,}105 & 0{,}6315 & 0{,}1631 \\ E8 & 0 & 0 & 0 & 0 & 0 & 0{,}1578 & 0{,}105 & 0{,}7368 \end{pmatrix} \tag{5.16}$$

- **Cálculo de las probabilidades de transición o del modelo de evolución (Matriz A):** Para calcular las probabilidades de transición, bajo un enfoque supervisado, se podría usar la fórmula C.26. Sabiendo las etiquetas asignadas en cada estado, se puede contar el número de veces que estando en un estado se transita a otro. La matriz de transición quedaría determinada dividiendo ese valor por el número total de transiciones efectuados desde un estado dado. Esta es la base del algoritmo *Baum-Welch*. Esta puede ser una buena opción, si sólo se dispone de las observaciones para llevar a cabo en aprendizaje del resto de parámetros. Sin embargo, a efectos prácticos, se han obtenido mejores resultados especificando los valores de las probabilidades en lugar de dejar que el algoritmo estime esos valores a partir de las observaciones. Además, los valores aprendidos del conjunto de entrenamiento dependen excesivamente de la información contenida en dicho conjunto. Se ha podido comprobar cómo pequeños cambios en las secuencias de etiquetas afectaban a la matriz resultante.

  Las transiciones entre los estados en un HMM contienen información sobre la cantidad relativa de tiempo que un peatón permanece en un estado, y por tanto condicionan el

ciclo del periodo de caminar. Aunque esta restricción no es estricta y permite cambios para adaptarse a distintas velocidades al caminar.

Para calcular el tiempo medio que un peatón permanece en un estado, se han analizado secuencias de peatones en ambas direcciones, realizando una estimación del tiempo (el número de imágenes) que transcurre en cada ciclo (considerando que ha transcurrido un ciclo cuando las piernas del peatón, a partir de una posición inicial vuelven a dicha posición).

Considerando $f = 7$ como la media del número de imágenes en cada ciclo y N el número de estados, las probabilidades de transición para cada modelo se han obtenido aplicando:

$$P_{\text{lateral}}(S_{t+1} = i|S_t = j) = \begin{cases} 1 - \frac{1}{\frac{f}{N}} & \text{si i = j} \\ \frac{f}{N} & \text{si i = j+1 mod N} \\ 0 & \text{en el resto de los casos} \end{cases} \tag{5.17}$$

correspondiente a un HMM de izquierda a derecha. Siendo $f > N$, $\frac{f}{N}$ es el número medio de imágenes por fase o estado. Por tanto, $\frac{1}{\frac{f}{N}}$ es la probabilidad de transitar fuera de un estado después de haber observado una imagen.

$$P_{\text{frontal}}(S_{t+1} = i|S_t = j) = \begin{cases} 1 - \frac{1}{\frac{f}{N}} & \text{si i = j = 1} \\ \frac{f}{N} & \text{si i = j+1 = 1} \\ \frac{1}{N} & \text{en el resto de los casos} \end{cases} \tag{5.18}$$

que se corresponde con un HMM casi ergódico.

Y se han obtenido los siguientes valores para las matrices de transición lateral 5.19 y frontal 5.20 respectivamente:

$$a_{ij_{\text{lateral}}}(O) = \begin{pmatrix} Estado/Estado & E1 & E2 & E3 & E4 \\ E1 & 0{,}4286 & 0{,}5714 & 0{,}0 & 0{,}0 \\ E2 & 0{,}0 & 0{,}4286 & 0{,}5714 & 0{,}0 \\ E3 & 0{,}0 & 0{,}0 & 0{,}4286 & 0{,}5714 \\ E4 & 0{,}5714 & 0{,}0 & 0{,}0 & 0{,}4286 \end{pmatrix} \tag{5.19}$$

De estos valores se puede interpretar que la probabilidad de transitar a un estado distinto del actual es mayor que el de permanecer en dicho estado. A esta conclusión se ha llegado después de analizar las secuencias de observaciones de entrenamiento.

$$a_{ij_{\text{frontal}}}(O) = \begin{pmatrix} Estado/Estado & E5 & E6 & E7 & E8 \\ E5 & 0{,}4286 & 0{,}5714 & 0{,}0 & 0{,}0 \\ E6 & 0{,}25 & 0{,}25 & 0{,}25 & 0{,}25 \\ E7 & 0{,}25 & 0{,}25 & 0{,}25 & 0{,}25 \\ E8 & 0{,}25 & 0{,}25 & 0{,}25 & 0{,}25 \end{pmatrix} \tag{5.20}$$

En este caso todos los estados tienen la misma probabilidad de ser alcanzados y emitir observaciones.

Una vez que se tienen las estimaciones iniciales de $\Theta = \{\pi, A, B\}$, se puede entrenar cada HMM para refinar esas probabilidades.

5.8.3. Evaluación del entrenamiento: El algoritmo de Viterbi

El reconocimiento del caminar humano se plantea como una tarea de predicción de estados (posiciones de las piernas), a partir de una secuencia de siluetas extraídas de las imágenes FIR (conjunto de observaciones). Para resolver este problema, un primer paso consiste en saber qué siluetas pertenecen a qué estados. Así, se establecen los límites entre las distintas posiciones de las piernas en el tiempo. Es un proceso conocido como "alineamiento de las secuencias de observaciones".

En la solución planteada en esta tesis, las observaciones se han asignado de manera supervisada a los estados que se han considerado más probables. Después se han estimado los parámetros que definen cada modelo. Sin embargo, puede haberse cometido errores a la hora de realizar la asignación de estados. El algoritmo de *Viterbi* permite distribuir etiquetas dada una secuencia de observaciones. Es por tanto un aprendizaje no-supervisado, que se ha utilizado para comprobar la bondad de las estimaciones de las probabilidades realizadas, pudiendo refinarlas en caso de ser necesario.

El procedimiento es el siguiente:

1. Asignación inicial de estados: Se realiza una asignación inicial de la observaciones a los estados que se estime más probables. La figura 5.26 muestra el resultado de este proceso.

2. Estimación inicial de los modelos de emisión: Se calcula la matriz de emisión a partir de la estimación de estados del paso previo. El resultado obtenido corresponde a la matriz 5.15 y 5.16.

3. Reestimar la asignación de estados: Empleando las estimaciones de $\{\pi, A, B\}$ para cada modelo, se utiliza el algoritmo de *Viterbi* para reasignar etiquetas (estados) a las secuencias de observaciones de entrenamiento. Si la asignación inicial no difiere mucho de la nueva asignación de etiquetas, se ha terminado el entrenamiento.

4. Reestimar los modelos de emisión: Si la nueva asignación de etiquetas difiere mucho de la previa, se vuelve a estimar las probabilidades de emisión. Las observaciones se reagrupan en los nuevos estados asignados y se calcula la probabilidad de la ecuación 5.14. Una vez reestimados los modelos de emisión, se vuelve al paso 3, a menos que la evolución de la probabilidad del conjunto de entrenamiento sea asintótico a un límite superior y por tanto, la variación no sea significativa.

Esta propuesta, mediante el uso de *Viterbi* es una versión *hard* de asignación de etiquetas. Existe una versión *soft*, que consiste en emplear el algoritmo *Baum-Welch* o *forward-backward*, que es un algoritmo de máxima expectación (EM) específicamente adaptado para

el entrenamiento de HMMs. Es uno de los algoritmos de entrenamiento más utilizado para el reconocimiento del habla.

El diseño del conjunto de entrenamiento requiere especial atención, ya que las probabilidades en un modelo estadístico, como en el caso de un HMM, se obtienen de las imágenes sobre las que se entrena. Si el conjunto es demasiado específico a un tipo de secuencias, las probabilidades pueden no generalizar correctamente ante secuencias distintas. Pero, si las imágenes de entrenamiento son demasiado genéricas, las probabilidades pueden no representar las secuencias a reconocer.

Para evaluar los modelos bigramas, se ha dividido el conjunto de imágenes en un conjunto de entrenamiento y en un conjunto de test. Para evaluar el entrenamiento de cada HMM, se han empleado 26 secuencias de longitud 7 correspondientes al movimiento lateral (182 imágenes) y 30 secuencias de longitud 7 para el movimiento frontal (210 imágenes). En la tabla 5.6 se muestran algunas de las cadenas utilizadas durante el entrenamiento.

Sec	**t1**	**t2**	**t3**	**t4**	**t5**	**t6**	**t7**
1	1	1	2	2	3	4	1
2	1	2	2	3	4	1	1
3	2	2	3	4	1	1	2
4	2	3	4	1	1	2	3
5	3	4	1	1	2	3	4
6	4	1	1	2	3	4	1
7	1	1	2	3	4	1	1
8	1	2	3	4	1	1	2
9	2	3	4	1	1	2	3
10	3	4	1	1	2	3	4
11	4	1	1	2	3	4	1
12	1	1	2	3	4	1	1
13	1	2	3	4	1	1	1
14	2	3	4	1	1	1	2
15	3	4	1	1	1	2	3

(a)

Sec	**t1**	**t2**	**t3**	**t4**	**t5**	**t6**	**t7**
1	6	6	6	7	7	7	6
2	6	6	7	7	7	6	6
3	6	7	7	7	6	6	8
4	7	7	7	6	6	8	8
5	7	7	6	6	8	8	8
6	7	6	6	8	8	8	7
7	6	6	8	8	8	7	8
8	6	8	8	8	7	8	6
9	8	8	8	7	8	6	6
10	8	8	7	8	6	6	6
11	8	7	8	6	6	6	6
12	7	8	6	6	6	6	7
13	8	6	6	6	6	7	8
14	6	6	6	6	7	8	8
15	6	6	6	7	8	8	8

(b)

Tabla 5.6: Secuencia de observaciones etiquetadas de modo supervisado; (*a*) Movimiento lateral (*b*) Movimiento frontal. Etiquetas: 1 y 5, abiertas; 2 y 7, semicerradas; 3 y 8, cerradas; 4 y 6, semiabiertas.

Cada una de las secuencias de observaciones de entrenamiento son analizados por el algoritmo de *Viterbi*. Si las nuevas etiquetas (estados) calculados por el algoritmo, difieren mucho de las etiquetas asignadas por el operario a las observaciones, significa que se ha cometido un error de etiquetado y habrá que asignar la nueva etiqueta a la observación. Una vez reasignadas todas las etiquetas que sean necesarias, habrá que volver a calcular los modelos de emisión. El proceso de entrenamiento termina cuando no haya que reasignar estados o la probabilidad del entrenamiento no varíe.

En este caso, no ha sido necesario reasignar etiquetas por dos motivos; bien porque los

estados asignados por el operario coincidían con los asignados aplicando *Viterbi*, o bien porque las observaciones ”mal” clasificadas podían ser asignadas a varios estados debido a que la silueta era confusa. Se ha decidido no reasignar esos casos, ya que su etiquetado es no-determinista. Por todo esto, se ha considerado el entrenamiento satisfactorio si las predicciones de los estados se aproximan o coinciden con las etiquetas observadas. No ha sido necesario reestimar los modelos de emisión.

5.8.4. Corrección de Errores

El algoritmo de *Viterbi*, además de permitir reestimar las etiquetas, en esta tesis se ha utilizado para corregir los errores de etiquetado. El conjunto de test empleado está formado por 52 secuencias de longitud 7 (364 imágenes) correspondientes al movimiento lateral (considerando el movimiento de dos peatones distintos) y 30 secuencias de longitud 7 (210 imágenes) para el movimiento frontal (correspondientes a un peatón).

Comparando la predicción de las etiquetas obtenidas para ambos movimientos, con el etiquetado esperado (supervisado) se permite evaluar la bondad de las correcciones. Además, los resultados obtenidos empleando HMM se contrastan con los obtenidos mediante el *matching* probabilístico. Los resultados experimentales han demostrado cómo el algoritmo es capaz de hacer frente a las incorrecciones heredadas del módulo probabilístico. Por un lado, los gráficos de la figura 5.28 permiten comprobar cómo la curva que representa una trayectoria horizontal, tiene una representación que se adapta más a la realidad si se aplica HMM que mediante el *matching* probabilístico. Se puede verificar esto, si se considera que durante el ciclo del caminar humano, y en concreto, en el caso de el movimiento horizontal, el conjunto de posiciones de las piernas se repiten a lo largo del ciclo. Es por ello que una representación del mismo se aproximaría a la curva obtenida por el HMM, que pasa por cada uno de los estados (posiciones de las piernas) de manera secuencial.

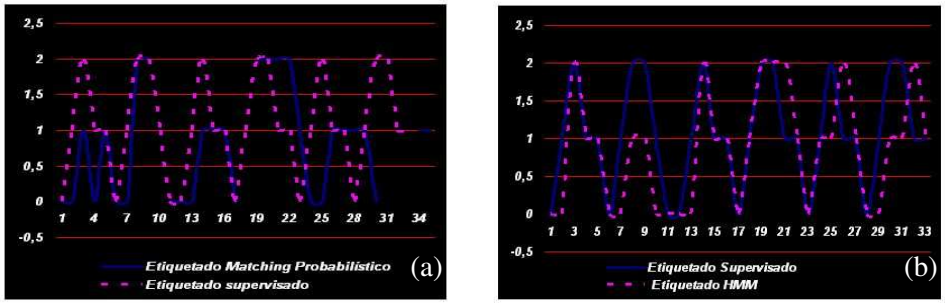


Figura 5.28: Confrontación de los estados etiquetados contra los obtenidos mediante; (*a*) HMM lateral y (*b*) matching probabilístico, aplicados a la misma secuencia de imágenes.

Por otro lado, se han creado matrices de confusión, que permiten verificar cómo el HMM clasifica mejor las observaciones pertenecientes a una trayectoria determinada, que la técnica basada en el *matching*. En cambio, el *matching* confundía etiquetas del movimiento frontal en el horizontal y viceversa. Se han considerado un total de 114 regiones de interés conteniendo a peatones, 49 de las cuales pertenecían al movimiento frontal y las 65 restantes, al horizontal.

A pesar de que el HMM consigue eliminar la mezcla de etiquetas de ambas trayectorias, sigue cometiendo errores en la clasificación. Sin embargo, el objetivo perseguido era la identificación del tipo de movimiento, alcanzado con esta técnica. Los errores de etiquetado podrían resolverse mediante un modelo del movimiento que se ajuste mejor a la velocidad del modo de caminar de cada persona. Las tablas 5.7 y 5.9 representan la clasificación obtenida por el *matching* probabilístico para el movimiento horizontal y frontal, respectivamente. Las tablas 5.8 y 5.10 son los correspondientes resultados obtenidos mediante los HMM.

MOVIMIENTO HORIZONTAL				MOVIMIENTO FRONTAL				
Abiertas	Semicerradas	Cerradas	Semiabiertas	Abiertas	Semicerradas	Cerradas	Semiabiertas	Tipo de las Observaciones
17	0	0	4	2	0	0	0	Horizontal Abiertas
3	10	2	2	0	1	0	0	Horizontal Semicerradas
0	0	16	0	0	0	0	0	Horizontal Cerradas
0	1	4	1	0	2	0	0	Horizontal Semiabiertas

Tabla 5.7: Matriz de confusión obtenida para el movimiento horizontal clasificado mediante el matching probabilístico

MOVIMIENTO HORIZONTAL				MOVIMIENTO FRONTAL				
Abiertas	Semicerradas	Cerradas	Semiabiertas	Abiertas	Semicerradas	Cerradas	Semiabiertas	Tipo de las Observaciones
17	0	0	0	0	0	0	0	Horizontal Abiertas
6	11	1	0	0	0	0	0	Horizontal Semicerradas
0	2	14	0	0	0	0	0	Horizontal Cerradas
0	1	5	3	0	0	0	0	Horizontal Semiabiertas

Tabla 5.8: Matriz de confusión obtenida para el movimiento horizontal clasificado mediante HMM

MOVIMIENTO HORIZONTAL				MOVIMIENTO FRONTAL				
Abiertas	Semicerradas	Cerradas	Semiabiertas	Abiertas	Semicerradas	Cerradas	Semiabiertas	Tipo de las Observaciones
0	0	0	0	4	0	0	0	Frontal Abiertas
0	0	1	0	0	11	1	1	Frontal Semicerradas
0	0	1	0	0	1	6	4	Frontal Cerradas
0	0	0	0	1	0	0	18	Frontal Semiabiertas

Tabla 5.9: Matriz de confusión obtenida para el movimiento frontal clasificado mediante el matching probabilístico

MOVIMIENTO HORIZONTAL				MOVIMIENTO FRONTAL				
Abiertas	Semicerradas	Cerradas	Semiabiertas	Abiertas	Semicerradas	Cerradas	Semiabiertas	Tipo de las Observaciones
0	0	0	0	4	0	0	0	Frontal Abiertas
0	0	0	0	0	12	1	1	Frontal Semicerradas
0	0	0	0	0	1	7	4	Frontal Cerradas
0	0	0	0	0	1	0	18	Frontal Semiabiertas

Tabla 5.10: Matriz de confusión obtenida para el movimiento frontal clasificado mediante HMM

5.8.5. Clasificación de las secuencias de movimiento

Una vez que se ha entrenado y evaluado cada modelo, la fase final consiste en clasificar cada secuencia de observaciones en las dos posibles trayectorias. Se asume que ambos modelos son equiprobables, es decir, las trayectorias a reconocer tienen la misma probabilidad de ocurrir en la realidad. Con lo que la tarea de reconocimiento consiste en una clasificación de máxima verosimilitud de las secuencias estocásticas de etiquetas (ver C.15).

La máxima verosimilitud de una secuencia dado un HMM ya se ha formulado en las ecuaciones C.9 y C.11:

$$\hat{\Theta}_{MAP} = \underset{\Theta_i}{\operatorname{argmax}}\ P(O|\Theta) = \underset{\Theta_i}{\operatorname{argmax}}\, P(O|Q,\Theta)P(Q|\Theta) \tag{5.21}$$

Estos modelos cumplen la función de "patrones estocásticos" contra las que comparar las observaciones. Así, a partir de una secuencia de observaciones se puede inferir el modelo dinámico más probable θ.

El gráfico 5.30-a muestra las probabilidades de que los movimientos de dos peatones hayan sido generados por el HMM lateral. Han sido calculadas aplicando el algoritmo *forward* a la secuencia de observaciones (con errores de etiquetado).

Por el contrario, el gráfico 5.30-b representa la máxima verosimilitud de las predicciones obtenidas después de aplicar *Viterbi* a las observaciones.

La secuencia de imágenes contenidas en la imagen 5.31 muestran el etiquetado realizado por el módulo probabilístico, que constituyen las observaciones (propensas a errores) de entrada a los HMM. El color de las cajas superpuestas identifica el tipo de movimiento; rojo para el movimiento lateral y azul para el frontal.

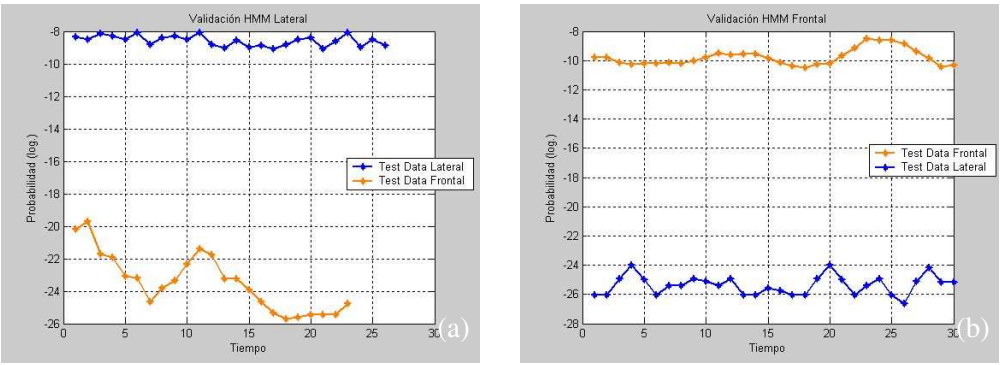


Figura 5.29: Para evaluar cada HMM una vez han sido entrenados, se les da como entrada secuencias correspondientes a ambos movimientos. Cada HMM debe dar mejores probabilidades ante secuencias para las que ha sido entrenado. (*a*) Probabilidades logarítmicas que el HMM lateral y (*b*) el HMM frontal dan a cada secuencia de test.

5.9. Limitaciones de los modelos ocultos de Markov

El enfoque bayesiano es un marco de trabajo bien definido para tratar la incertidumbre. Para aplicaciones en las que hay tratar con dinámicas temporales del entorno, los HMM permiten

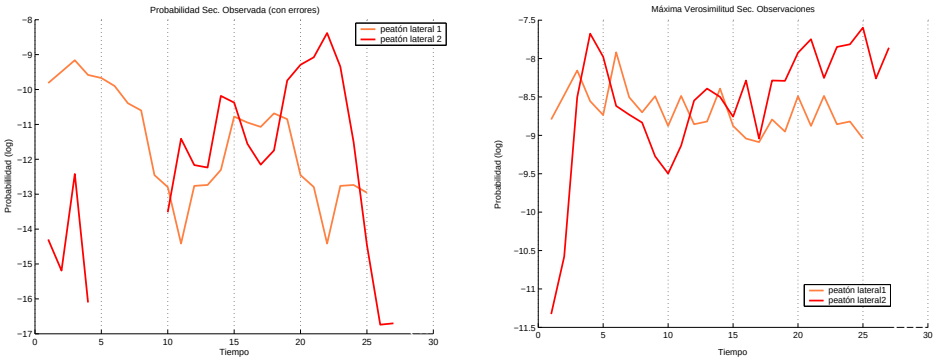


Figura 5.30: Secuencias de dos peatones moviéndose lateralmente a la cámara. (*a*) Probabilidad de que cada secuencia haya sido generada por cada modelo y (*b*) Máxima verosimilitud de las mismas secuencias.

representar la evolución de las variables del dominio en el tiempo. En este tipo de modelos, cada instante de tiempo refleja el estado del entorno en ese momento; sólo existe un estado y una observación. Por tanto, son adecuados cuando el dominio está localmente restringido. De lo contrario, resulta complicado modelar entornos complicados, que sean amplios o que requieran de largos intervalos de tiempo. Añadido a lo anterior, la suposición de que las distribuciones de probabilidad sean estacionarias hace que los HMM no puedan manejar entornos complejos.

Esto se debe, en gran medida, a la propiedad de Markov. Como consecuencia, cualquier relación entre dos estados separados (q_1 y q_4), deben ser comunicados a través de todos los estados intermedios. Un modelo de Markov de primer orden, donde $P(q_t)$ depende sólo de q_{t-1}, no puede en general capturar relaciones complejas.

Para el caso del caminar humano, resulta complicado modelar correctamente la velocidad de una persona. Las duraciones de cada estado se modelan de forma inexacta (ver ecs. 5.20 y 5.19) y hacen que las predicciones obtenidas mediante Viterbi no se ajusten de una forma precisa a los estados observados en la realidad. No obstante, esta limitación no impide la correcta identificación del movimiento de cada peatón, así como la asignación de las etiquetas coherentes con ese movimiento.

Otra causa de limitación, se debe a que un HMM genera cada observación o_i sólo a partir del correspondiente estado q_i. En teoría se podría sustituir $P(o_i|q_i)$ por una distribución más complicada $P(o_i|q_{t-1}, q_t, q_{t+1})$ que permitiría que una observación o_i influyese a tres estados distintos. Sin embargo, no está claro como representar una distribución tan complicada de manera compacta.

Esta suposición de independencia indica que no existe correlación entre imágenes consecutivas. Para el caso en estudio, la silueta extraída en una imagen pueden considerarse independiente de la siguiente. La variación de la forma del peatón se describe mediante las probabilidades de emisión asociadas a cada estado. Como se ha explicado anteriormente, esas probabilidades se calculan como procesos independientes de *Bernoulli*. Por tanto resulta adecuado considerar que no existe dependencia entre las siluetas.

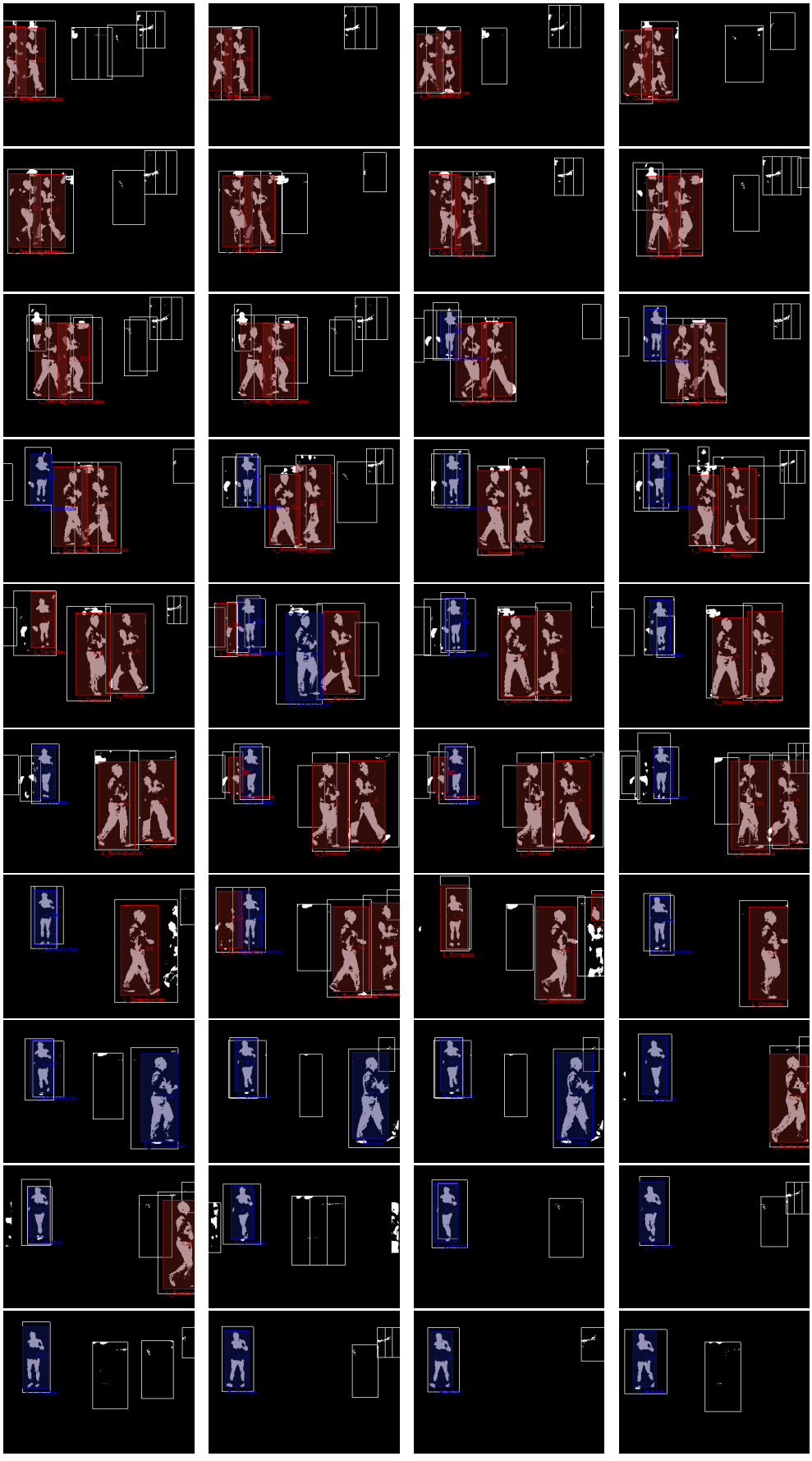


Figura 5.31: Secuencia FIR empleada durante la fase de test de los HMM. Las imágenes muestran las etiquetas asignadas por el matching probabilístico, que son la entrada a los modelos ocultos de Markov.

Sin embargo, la suposición de que la observación actual sólo dependa del instante (estado) actual, y considerar que un instante de tiempo concreto es independiente de lo que haya sucedido en el pasado o en el futuro se adapta a las necesidades del dominio en estudio.

5.9.1. Análisis de errores

Existe un problema importante con el cálculo de C.14, a saber, con la estimación de la secuencia más probable, y es la escasez de datos. Puede ocurrir que una secuencia concreta de estados $\{q_1 q_2 ... q_T\}$ que ocurra en el conjunto de pruebas, nunca se haya dado en el conjunto de entrenamiento. Esto implica que no se puede calcular la probabilidad bigrama $P(q_i|q_{i-1})$ como la máxima verosimilitud estimada mediante las ecuaciones C.26 y 5.14. Existen alternativas basadas en interpolación lineal para estimar estas probabilidades [Bra00].

Sin embargo, en el algoritmo desarrollado, el aprendizaje del HMM se realiza sobre datos etiquetados a mano. Como ya se ha comentado, se podría dejar que el algoritmo EM aprendiese las probabilidades de transición y de emisión. Pero se ha verificado que se obtienen mejores resultados con un aprendizaje supervisado. Otros autores [Mer94] llegaron a la misma conclusión, constatando que siempre que se tenga un conjunto de entrenamiento, aunque sea pequeño, es preferible realizar un aprendizaje basado en datos etiquetados a mano, que en el algoritmo EM.

Para el cálculo de las probabilidades de transición, las estimaciones obtenidas con la ecuación C.26 han dado resultados adecuados. Al fijar la arquitectura de los dos modelos a mano, se está forzando de alguna manera las secuencias de observaciones que se van a admitir. Las arquitecturas de izquierda-a-derecha 5.23 y semi-ergódico 5.24 se aproximan a las secuencias de movimiento que esperamos observar y permite alcanzar los objetivos perseguidos, no siendo vital un ajuste más preciso.

En lo referente a las probabilidades de emisión, resulta muy complicado considerar el amplio abanico de posibles observaciones que pueden ser observados en un estado. Un análisis de las matrices de emisión obtenidas (5.16 y 5.15) reflejan este hecho. Se puede comprobar cómo existen estados en los que ciertas observaciones tienen una probabilidad nula de ocurrir. O lo que es lo mismo, en el conjunto de entrenamiento nunca ocurrieron. Esto puede ser un problema, pudiendo incluso impedir la correcta clasificación de la secuencia.

Los resultados experimentales han demostrado la validez de este método, al rechazar todas las cadenas que sean asignadas al modelo equivocado. Así mismo, el método es capaz de corregir errores de etiquetado debidos a falsos negativos (p.ej. cuando un movimiento lateral admite como válidas etiquetas características del movimiento frontal).

Capítulo 6

Conclusiones y Aportaciones

El objetivo de esta tesis ha sido la realización de un sistema avanzado de asistencia a la conducción (ADAS) para la detección de peatones en entornos urbanos. La visión es una opción natural como sensor aplicado a la detección de peatones porque se basa en cómo las personas perciben a los humanos, es decir, en función de rasgos visuales. El uso de tecnologías de visión para vehículos "inteligentes" se ha estado utilizando durante los últimos 20 años. Sin embargo, ha evolucionado rápidamente en los últimos años y hoy en día es considerada como una de las tecnologías de percepción más prometedoras [BBB+07b]. Pero, para desarrollar aplicaciones de visión en tiempo real, que puedan trabajar en entornos altamente dinámicos, como son los entornos urbanos, es necesario combinar una gran cantidad información.

El sistema de detección de peatones propuesto, combina el uso de dos sistemas de visión distintos; uno, proporciona información del dominio visible y el otro, del infrarrojo lejano. La idea es tratar de aprovechar los puntos fuertes de cada dominio, ya que ninguno de los dos, por sí sólo, es capaz de resolver el problema.

En este capítulo, se plantean los puntos fuertes y débiles de los métodos utilizados y se extraen una serie de conclusiones. Después, se resumen las aportaciones de la misma al ámbito de los ADAS. Este apunte se realiza teniendo presentes los hitos marcados para el desarrollo de esta tesis. Finalmente, se realiza una propuesta de trabajos futuros, algunos enfocados a la mejora de los resultados, y otros, como consecuencia de nuevas líneas de investigación abiertas durante el desarrollo de esta tesis.

6.1. Sistema de Detección de Peatones desarrollado

En muchos escenarios, las cámaras de infrarrojo lejano (FIR), resultan más adecuadas que las del dominio visible para detectar peatones. Sobre todo, cuando el fondo emite menos calor que las personas en la imagen. Además, dependen poco de la iluminación, lo cual favorece su uso en entornos exteriores. Otra de las ventajas que ofrece, es un menor nivel de ruido, facilitando las etapas iniciales del proceso de detección. Pero, en contrapartida, carecen de texturas y colores y se pierden los detalles de los objetos. En este sentido, las cámaras del dominio

visible pueden ser útiles, ya que su resolución es mayor. Esto hace que la detección sea más difícil debido a la presencia de detalles, sombras y cambios en la iluminación o precisión de las imágenes. De cualquier modo, los detalles pequeños son cruciales para la clasificación de las detecciones.

Se han implementado distintos métodos para la detección de los peatones en cada uno de los dominios de la imagen. En el dominio visible; la detección de bordes verticales, el cálculo de mapas de disparidades (denso y no-denso) y el filtrado basado en simetrías generan una lista de cajas (*bounding boxes*) que contienen peatones potenciales. Después, se emplean dos técnicas de validación que evalúan las presencia humana en el interior de las *bounding boxes*. Ambos buscan rasgos de la forma humana, mediante la detección de los bordes y el uso de un método basado en contornos activos.

En el dominio infrarrojo, se parte de la lista de (*bounding boxes*) proporcionada por el sistema GOLD desarrollado en la Universidad de Parma. Al igual que en el dominio visible, se evalúa la presencia de la forma humana en el interior de cada caja, considerando la forma global mediante el uso de modelos probabilísticos. Este módulo ha sido integrado y probado en el sistema final GOLD. Por último, los modelos ocultos de Markov (HMM) permiten corregir errores de clasificación y reconocer el tipo de trayectoria seguida por el peatón.

6.2. Dominio visible

Se han usado distintos métodos para detectar peatones potenciales y después, validar las hipótesis.

6.2.1. Detección de regiones de interés

El sistema implementado en el dominio visible, realiza la detección de obstáculos en base a la información contenida en el mapa de disparidad y el uso de rasgos característicos de los humanos. Se ha desarrollado un algoritmo estéreo basado en rasgos, que sólo considera un subconjunto de píxeles en la imagen (los bordes verticales) para realizar la correspondencia, permitiendo así alcanzar el requisito de procesamiento en tiempo real. El mapa de disparidad obtenido es no-denso, que tiene el inconveniente de complicar la posterior fase de segmentación. En el caso estudiado, los bordes verticales de los objetos (p.ej. peatones) suelen aparecer separados, no como líneas continuas y en el caso de objetos que estén próximos, resulta difícil saber qué bordes pertenecen a cada uno de ellos.

En esta tesis, para hacer frente a los errores del mapa de disparidad, se ha utilizado esa información como pista para localizar las regiones de interés en la imagen, en lugar de confiar en una segmentación basada en la disparidad. Las siguientes etapas de la detección filtran los resultados obtenidos por el algoritmo estéreo, utilizando para ello:

- **Medidas de distancias:** El mapa de disparidad no-denso se filtra en base a medidas de distancia. Como resultado se obtiene un conjunto imágenes de mapas de disparidad, cada uno conteniendo aquellos píxeles que se encuentren a una cierta distancia.

- **Simetrías:** Después, se realiza la búsqueda de simetrías verticales en cada mapa de disparidad filtrado, considerando sólo puntos pertenecientes a los bordes verticales, siendo por tanto, rápido. El tamaño de la ventana de búsqueda queda establecido en función del rango de disparidades contenido en cada imagen. De este modo, se obliga a que la búsqueda de simetrías cumpla unas restricciones en cuanto al tamaño estimado del peatón.

- **Restricciones geométricas:** Para determinar las regiones de interés (ROI), sólo se consideran aquellas ventanas de búsqueda que cumplen las condiciones geométricas antes citadas. Pero en entornos urbanos, es habitual la presencia de otros objetos con una simetría vertical fuerte (como postes, farolas o árboles). Para filtrar las regiones candidatas, se han impuesto restricciones en cuanto al ancho mínimo que deben tener los bordes considerados para obtener el eje de simetría. De este modo se reduce el número de falsos positivos. Además, se exige que el número de votos de cada ventana esté por encima de un umbral, que varía en función del tamaño de la ventana en evaluación. Así, para la selección del mejor candidato se considera la relación existente entre, el número de puntos de borde simétricos y el tamaño de la ROI que los contiene.

Como resultado, el módulo de detección de obstáculos proporciona una lista de *bounding boxes*. El conocimiento de los parámetros de calibración permite resolver la relación existente entre las coordenadas de la imagen y el mundo. Para ello, se ha empleado la suposición de ”mundo plano”. Este supuesto puede dar lugar a errores debido a que la carretera no sea plano o al hecho de que la cámara está instalada en un vehículo en movimiento, y su posición varía continuamente (el *yaw, pitch,roll*). Sin embargo, el movimiento de la cámara afecta poco a este sistema. Esto se debe a que, para cada imagen, se calcula el mapa de disparidad no-denso correspondiente. Después, se calcula el punto de contacto entre cada objeto contenido en una ROI y la carretera en base a la información estéreo. La posición de las ROIs en la imagen (columna) se obtiene mediante la simetría de los bordes verticales.

El módulo desarrollado ha sido probado en diferentes situaciones, utilizando el vehículo experimental IvvI equipado con el sistema de visión del dominio visible. El rango de detección se limita a distancias entre 3 y 15 metros. El correspondiente rango de disparidades va de 4 a 20 píxeles. Este rango es suficiente como para cubrir las necesidades de un ADAS orientado a evitar colisiones.

Cabe decir que el algoritmo da un elevado número de falsos positivos, pero pocos falsos negativos. El objetivo perseguido con este módulo era detectar el mayor número de peatones posibles, ya que las falsas alarmas pueden ser eliminadas en las siguientes etapas.

6.2.2. Detección de la forma humana

Después de la fase de análisis de hipótesis basado en simetrías, la lista de ROIs conteniendo a peatones potenciales sufre aún más comprobaciones para validar la presencia humana. En primer lugar, se lleva a cabo la clasificación de los peatones mediante el Análisis del Componente Principal (PCA). La decisión final se toma considerando la probabilidad de que una ROI contengan una forma humana y desechando aquellas con una probabilidad demasiado

baja. Después, y únicamente en aquellas regiones que han sido etiquetadas como peatones, se extrae la silueta mediante contornos activos (*snakes*). En trabajos futuros se va a analizar la posición de las piernas para determinar la actividad que está realizando el peatón (p.ej. correr o andar).

6.2.2.1. Reconocimiento basado en PCA

La clasificación final se realiza aplicando una técnica basada en el aprendizaje supervisado. el análisis del componente principal (PCA). Sin embargo, son técnicas muy sensibles a los cambios de iluminación y por tanto, no aplicable a entornos exteriores, como es el caso en estudio. Se ha optado por emplear imágenes de bordes y distancias para representar la forma humana, en lugar de las intensidades, ya que son más robustas ante dichos cambios. Después de la fase de aprendizaje, el algoritmo es capaz de clasificar el contenido de las ROIs como peatón o no peatón. De este modo se han reducido el número de falsos positivos.

El algoritmo propuesto ha sido ejecutado en tiempo real en varias secuencias tomadas con el sistema estéreo del dominio visible, siendo capaz de clasificar correctamente hasta un 80 % de las ROIs, sin imponer ninguna restricción en cuanto a la iluminación, posición o tamaño de los peatones (ver cap.4).

6.2.2.2. Reconocimiento basado en *Snakes*:

Finalmente, los contornos activos o *snakes* se emplean para extraer la silueta de los peatones, a partir de las ROIs proporcionadas por el módulo del PCA. Dos de los mayores inconvenientes de este método son: la inicialización y la propensión a fallos ante otros rasgos fuertes en la imagen. Las medidas adoptadas para reducir estos problemas son:

- **Algoritmo de inicialización:** Se ha resuelto al realizar la siembra dentro de las ROIs proporcionadas después del análisis del PCA.

- **Mapa de disparidad denso:** Este análisis de la forma se efectúa en un rango de disparidad concreto, establecido en función de la distancia a la que esté cada ROI o peatón potencial. Así se eliminan los elementos del fondo, que podrían dar lugar a errores. Se ha desarrollado un algoritmo estéreo basado en regiones, construyendo un mapa de disparidad denso. Se ha tomado esta decisión porque los mapas densos contienen unos bordes mejor definidos que los no-densos, y los contornos activos extraen mejor la siluetas de los objetos cuanto más precisos sean los límites de los mismos.

- **Simetrías y bordes:** La selección de las energías externas es crucial, ya que determinan el modo en el que el *snake* va a ser atraído hacia los rasgos de interés. Se vuelven a emplear los bordes (tanto verticales como horizontales) y las simetrías verticales, pero limitadas un rango de disparidad concreto. Para ello se filtran las imágenes empleando como máscara aquel rango de disparidad que sea coherente con la distancia estimada de cada ROI.

El objetivo es obtener la forma del peatón contenido en cada ROI, para su posterior evaluación. Esta última fase de reconocimiento está sin implementar en el dominio visible, pero su finalidad sería reconocer la posición de las piernas. Una posible propuesta para realizar esta tarea ha sido implementada en el dominio infrarrojo.

6.3. Dominio infrarrojo

El módulo probabilístico desarrollado forma parte del proceso de validación utilizado por el sistema GOLD [BBF+07b, BBF+07a]. El objetivo es reducir el número de falsos positivos, al igual que ocurre con los clasificadores empleados en el dominio visible. Finalmente, la integración temporal de los resultados permite corregir errores de etiquetado y clasificar el tipo de trayectoria, mediante el empleo de modelos ocultos de Markov (HMM).

6.3.0.3. Reconocimiento basado en modelos probabilísticos

Con el fin de clasificar el contenido de las *bounding boxes*, se realiza una correlación basada en modelos de la forma humana en el dominio infrarrojo. Como resultado, se descartan aquellas ROIs con una probabilidad baja de contener a un peatón. Sin embargo, la detección de la forma humana en imágenes FIR no es sencilla. Se han tomado varias medidas:

- **Siluetas:** Se ha decidido describir la apariencia en base a la silueta humana, ya que contiene la información esencial de la forma humana, el dominio FIR permite obtener una silueta más homogénea que el dominio visible y la naturaleza cíclica de la acción de caminar, asegura que la silueta de cada individuo será repetida a intervalos regulares.

- **Modelos probabilísticos:** A partir de las siluetas se crean modelos probabilísticos. La propuesta original de Nanda *et al.* [ND02] crea un único modelo, que aplicado al caso en estudio daba una tasa alta de errores. Además es incapaz de diferenciar las posiciones de las piernas. Se ha comprobado que al utilizar un mayor número de modelos (*piernas abiertas, semiabiertas, cerradas y semicerradas*) y considerar distintos tipos de trayectorias (lateral y frontal), la clasificación de los peatones es más robusta y permite realizar un reconocimiento de la posición de las piernas. Es un método basado en correlación, pero no es un simple *pattern matching*, ya que cada píxel del modelo contiene información sobre su probabilidad de ocurrencia.

- **Reestimación de las ROIs:** La lista de *bounding boxes* obtenida de las fases anteriores, pueden no ser correctas. Por ello, se redefinen considerando el *aspect-ratio* y se aumenta su tamaño, para permitir que la correlación con los modelos pueda efectuarse en varias posiciones próximas a la posición de la ROI original. Este método ha demostrado ser eficaz para detectar partes del peatón que estaban fuera de la ROI original.

El algoritmo propuesto ha sido ejecutado en tiempo real en varias secuencias FIR tomadas con el sistema Tetravision, siendo capaz de detectar con éxito más del 85 % de los peatones en la escena, con un número muy bajo de falsas detecciones por imagen (ver cap.5). La mejora de

la robustez del algoritmo se debe a las modificaciones realizadas a la implementación original de Nanda *et al.*. El rango de detección del sistema GOLD es de 7 a 43,5 metros.

Una de las limitaciones encontradas consiste en la necesidad de un gran número de imágenes para poder separar el ciclo de caminar en distintas clases. Se ha comprobado que la bondad de los modelos probabilísticos mejora al incrementar el conjunto de entrenamiento. Además, es un proceso laborioso, al tener que seleccionar y etiquetar a los peatones del conjunto de entrenamiento "a mano".

6.3.0.4. Reconocimiento de la trayectoria basado en HMM

Finalmente, las detecciones que superan un nivel de confianza son evaluadas mediante modelos ocultos de Markov (HMM). Se marcaron dos objetivos: corregir los errores de etiquetado cometidos por el módulo probabilístico y reconocer el tipo de trayectoria del peatón (frontal o lateral). Se ha conseguido mediante:

- **Siluetas:** La apariencia humana se vuelve a representar en base a la silueta. Esto se debe a que este módulo debe asignar la silueta correcta a cada silueta defectuosa observada en la imagen. De este modo se corrigen los errores previos.

- **Integración temporal:** Por motivos de sencillez, se ha implementado un HMM de primer orden, donde la probabilidad de un estado sólo depende del estado anterior. Si se analiza el movimiento de las piernas de un humano, esta suposición es adecuada. Por otro lado, el hecho de añadir más estados complica los cálculos de las probabilidades y no aporta, en este caso, una mayor robustez.

- **Movimiento:** Para cada tipo de trayectoria se ha creado un HMM distinto, que representan la apariencia de las siluetas durante las fases del ciclo de caminar. Se han diferenciado 4 fases distintas:*piernas abiertas, semiabiertas, cerradas y semicerradas*. Se ha comprobado que los resultados son más precisos al incluir más fases.

Para el caso del caminar humano, resulta complicado modelar correctamente la velocidad de una persona usando HMM. Las duraciones de cada estado se modelan de forma inexacta y hacen que las predicciones no se ajusten a la realidad. No obstante, esta limitación no impide la correcta identificación del movimiento de cada peatón, así como la asignación de las etiquetas coherentes con ese movimiento.

Esto se debe a que los HMM implementados estiman la trayectoria en función de la secuencia de observaciones, por lo que considera la evolución en el tiempo. Si se pierde una detección o si se comete algún fallo a la hora de etiquetar una de las observaciones, ello no impide que el algoritmo determine el tipo de trayectoria.

6.4. Aportaciones

A continuación se detallan las aportaciones realizadas en cada uno de los dominios.

6.4.1. Dominio Visible

- **Uso exhaustivo de las restricciones geométricas ofrecidas por la visión estéreo:**

  El uso del sistema de visión estéreo, aporta un conjunto de cualidades que pueden resultar de gran ayuda aplicadas a los sistemas de detección. En concreto en esta tesis, se ha sacado el máximo beneficio a las restricciones geométricas impuestas por las cámaras del dominio visible.

  Por un lado, el sistema estéreo permite obtener la profundidad a la que están los objetos del entorno. Tomando como punto de apoyo esta idea, se ha desarrollado la segmentación del mapa de disparidad basada en distancias. Otros autores han utilizado anteriormente los mapas de disparidad como método para la segmentación de los objetos del entorno, pero es la primera vez que se filtra dicho mapa de disparidad en varias imágenes, cada una de ellas conteniendo a los objetos que se encuentran en un rango de distancias determinado.

  Se han utilizado tanto técnicas estéreo denso como no-denso, sacando el mayor partido a cada una de ellas en función de las necesidades; el mapa denso, sirve de guía para la extracción de los contornos basados en contornos activos. Resulta necesario la obtención de un contorno aproximado de los objetos en la imagen, por lo que resulta inviable aplicar técnicas estéreo no-densas. En cambio, el mapa no-denso, resulta robusto y veloz para la localización de las regiones de interés (ROI) durante la fase de detección de obstáculos. La correspondencia se realiza considerando los bordes que presentan un grado de inclinación importante (más de 60 grados), pero se considera las intensidades en la vecindad de cada píxel de borde. Por tanto, se trata de un algoritmo que combina las ventajas del estéreo basado en rasgos y del estéreo basado en áreas.

  Además, la visión estéreo permite determinar el tamaño de las regiones de interés en la imagen. De este modo, y junto con el uso del mapa de profundidad filtrado por distancias, para cada imagen de profundidad filtrada sólo se usa un tamaño de ROI, evitando búsquedas a distintas resoluciones. Este modo de operar ha demostrado ser eficaz.

- **Búsqueda de rasgos basados en la morfología del peatón:**

  Un peatón posee rasgos propios que pueden explotarse a la hora de desarrollar sistemas que los detecten. Son varios los aspectos empleados en este trabajo. La fuerte simetría vertical que presentan ha dirigido el foco de atención durante la fase de detección de obstáculos. Esta idea ha sido empleada por otros autores. Sin embargo, en este propuesta se ha calculado la simetría restringido al tamaño de cada ROI. Así, las restricciones geométricas proporcionadas por el sistema estéreo limitan el análisis de los rasgos de interés al interior de cada ROI, exigiendo a las regiones de interés que, además de ser simétricas, posean unas proporciones coherentes con el tamaño esperado de un peatón en la imagen.

  Sin embargo, la simetría no resulta un rasgo suficientemente robusto, ya que existen otros objetos en los entornos urbanos con una simetría vertical importante. Aunque para la fase de detección inicial de obstáculos resulta sencillo basarse en esta cualidad,

para las fases posteriores se requiere rasgos más robustos, capaces de filtrar las falsas detecciones. Es por ello que se utiliza la forma, que contiene una información global del peatón, tanto durante la fase de reconocimiento como durante la etapa de segmentación de la silueta.

- **Empleo del PCA basado en bordes aplicado a entornos urbanos:**

  El análisis del componente principal (PCA) posee una cualidad muy atractiva, permitiendo representar la clase de los peatones con un conjunto significativamente menor de características que la imagen original. Sin embargo son muy sensibles a los cambios de iluminación [FGG+98]. Al utilizar los bordes de la forma del peatón en lugar de las intensidades se obtienen resultados que permite aplicar este método en entornos exteriores.

- **Segmentación de la forma del peatón aplicando modelos deformables:**

  Los modelos deformables son potentes métodos para la extracción de objetos en una imagen. Dentro de estos modelos, los contornos activos o *snakes* poseen una serie de cualidades que se adaptan a las condiciones tratadas en esta tesis: por un lado, la sencillez de su formulación y por otro, la rapidez de ejecución hizo que se seleccionase esta técnica frente a otras. Por ello, esta técnica permite procesar tantos *snakes* como peatones potenciales en tiempo real, requisito fundamental de los ADAS.

  Muy recientemente, el VISLAB [BBGR08] ha utilizado esta técnica con el mismo fin. Al igual que en el algoritmo propuesto, la inicialización se limita a las ROI y se definen eneregías externas similares (simetrías y bordes verticales). Sin embargo, la idea de segmentar las ROIs en base a la información de los mapas densos evita que el *snakes* se vea atraído por objetos del fondo, dotándolo de una mayor robustez.

6.4.2. Dominio Infrarrojo

- **Reconocimiento de la posición de las piernas basado en un conjunto de modelos probabilísticos:**

  A partir de imágenes tomadas con el sistema estéreo infrarrojo, se ha realizado un exhaustivo análisis del ciclo del caminar humano. Como resultado, se ha desarrollado una técnica basada en el aprendizaje, que de manera supervisada clasifica a cada peatón en función de la posición que tengan sus piernas a lo largo del ciclo. Una vez finalizado el aprendizaje, se obtienen un conjunto de modelos probabilísticos, cada uno de ellos entrenado para reconocer una posición característica de las piernas. Los modelos permiten además, distinguir la trayectoria seguida por el peatón, clasificándola como frontal o horizontal a la dirección de la cámara.

- **Formulación nueva aplicada a los modelos probabilísticos:**

  La correlación de los modelos probabilísticos aplicado a las imágenes FIR no era demasiado robusto. El hecho de que los peatones no sean los únicos objetos que emiten calor hace que la probabilidad devuelta por los modelos en esas regiones con píxeles

brillantes, sea alta. Con el objetivo de reducir las falsas detecciones, se propone una nueva formulación que trata de diferenciar entre los píxeles pertenecientes a los objetos y los píxeles pertenecientes al fondo, haciendo que la contribución a la probabilidad del modelo se calcule de manera independiente y normalizada. Así se evita que zonas con muchos píxeles de fondo den una probabilidad alta, que es lo que ocurría al aplicar la formulación clásica [ND02].

- **Integración temporal y reconocimiento de la trayectoria aplicando HMM:**

  La técnica basada en la correspondencia con los modelos, permite etiquetar a los peatones, pero no considera la coherencia del etiquetado en el tiempo. El hecho de estar analizando una secuencia de imágenes hace que surja la necesidad de introducir cierta referencia temporal. En concreto, se ha considerado que la posición de las piernas en un instante de tiempo, depende de la posición de las mismas en el instante previo, existiendo entre ambas una relación.

  Los modelos de Markov ocultos (HMM) permiten representar de un modo sencillo esta relación, sin tener que recurrir a modelos de movimiento complicados ni aplicar técnicas basadas en la estimación de la posición, como el filtro de Kalman. Los HMM permiten eliminar los errores cometidos por la técnica de la correlación. Esta técnica es propensa a fallos, ya que al basarse en la información en la imagen, el ruido contenido en la imagen o los errores de segmentación de la silueta hacen que la clasificación sea incorrecta. En cambio, el enfoque basado en HMM, considera las siluetas extraídas durante el ciclo completo del proceso de caminar. Así, establece el tipo de trayectoria que sigue el peatón, corrigiendo las etiquetas más probables considerando varias imágenes en lugar de una sola.

6.5. Trabajos futuros y conclusiones

Esta tesis ha investigado las ventajas e inconvenientes del uso de dos sistemas de visión distintos aplicados a la detección de peatones. Como resultado se proponen los siguientes trabajos futuros:

- **Dominio Visible:**

  - **Reconocimiento de la forma basado en *Snakes*:** El uso una representación discreta para representar al *snake*, hace que el modelo sólo se evalúa en los puntos discretos y como resultado, el modelo puede degenerar en su forma o bien, puede quedar atrapado en otras zonas de la imagen que presenten unos rasgos parecidos. Como solución se ha propone el uso de *splines*, que son curvas que permiten ser evaluadas no sólo en sus puntos de control, sino también a lo largo de la curva.

    Por último, queda por implementar el clasificador, encargado de tomar la decisión final a partir de los contornos extraídos con los *snakes*. Se puede pensar en usar un algoritmo de correspondencia de formas para comparar la configuración final de

cada *snake* con algunos modelos de peatones basados en la forma contenidos en una base de datos. [BBF+07a] han probado este enfoque, obteniendo un número reducido de falsos negativos pero, al mismo tiempo, un número inaceptable de falsos positivos. Otro posible enfoque puede basarse en el uso de redes de neuronas, siendo los puntos del *snake* las entradas a la red, y la salida, la probabilidad de que pertenezca a una persona.

- **Integración en el vehículo IvvI:** La detección de obstáculos ha sido integrado en el vehículo IvvI y evaluado su funcionamiento. Sin embargo, el módulo del PCA y el de los *snakes* han sido evaluados en secuencias de imágenes tomadas por el sistema de visión IvvI, pero no han sido integrados en el vehículo.
- **Seguimiento de detecciones *Tracking*:** El *tracking* de las ROIs reduciría los falsos positivos y negativos. Pero no es una tarea fácil. Por un lado, los movimientos del coche hacen que la inclinación de la cámara varíe, afectando a la posición de las cajas. Por otro, el algoritmo debe poder hacer frente a los falsos negativos, ya que si en una imagen se pierde una detección, deberá estimarse su movimiento. El algoritmo de seguimiento deberá ser robusto para solucionar estos problemas así como resolver la asignación de cada caja al peatón correspondiente.
- **Integración temporal basada en HMM:** La idea implementada en el dominio infrarrojo, podría aplicarse al dominio visible para reconocer la trayectoria de los peatones. Los beneficios serían, una reducción de errores y el reconocimiento del movimiento.

▪ **Dominio Infrarrojo:**

- **Clasificación de la forma basado en PCA y *Snakes*:** De forma parecida al dominio visible, se puede realizar el reconocimiento de la forma en base al PCA. Por otro lado, puede resultar interesante aplicar *snakes* al dominio FIR, ya que el nivel de ruido y detalles es menor que en el dominio visible. Podría considerarse usar, además, los mapas de disparidad para filtrar los objetos del fondo. Por último, al implementar las mismas técnicas en dos dominios distintos se podría hacer un análisis de los resultados.

▪ **Combinación de ambos dominios:**

El hecho de que las cualidades de ambas cámaras se complementen favorece la integración de las mismas, en un intento por cubrir los puntos débiles de cada dominio con los puntos fuertes del otro. Los resultados obtenidos por otros autores han demostrado que el uso de sensores que trabajan en distintos dominios permiten incrementar la tasa de detecciones, comparado con sistemas que sólo confían en un dominio.

La complejidad de este tipo de sistemas surge cuando la información obtenida en un dominio se quiere convertir al otro dominio, como es el caso de la fusión de las regiones de interés. La arquitectura del sistema final debe poder soportar el flujo de información y procesamiento de datos.

- **Integración en el vehículo IvvI:** Además de las ventajas de combinar los puntos fuertes de cada dominio, al disponer de cámaras de infrarrojo se podría evaluar el módulo probabilístico y el módulo basado en HMM en el vehículo IvvI.

Bibliografía

[AN89] J. Aggarwal and N. Nandhakumar. On the computation of motion from sequences of images: A review. *Proc. IEEE*, 76(8):917–935, aug 1989. [cited at p. 36]

[ANF05] ANFAC. Asociación española de fabricantes de automóviles y camiones. Available on the Internet: http://www.anfac.com/, 2005. [cited at p. 9]

[BBB+07a] M. Bertozzi, L. Bombini, A. Broggi, P. Cerri, P. Grisleri, and P. Zani. GOLD: A Complete Framework for Developing Artificial Vision Applications for Intelligent Vehicles. *IEEE Intelligent Systems*, 23(1):69–71, November-December 2007. [cited at p. 82, 133]

[BBB+07b] M. Bertozzi, A. Broggi, L. Bombini, C. Caraffi, Stefano Cattani, P. Cerri, A. Fascioli, Mirko Felisa, R. I. Fedriga, Stefano Ghidoni, P. Grisleri, P. Medici, M. Paterlini, P. Paolo Porta, M. Posterli, and P. Zani. Vision Technologies for Intelligent Vehicles. In *11th Intl. Conf. on Knowledge-Based and Intelligent Information and Engineering Systems*, Vietri sul Mare, Italy, September 2007. in press. [cited at p. 16, 179]

[BBC+07] M. Bertozzi, A. Broggi, C. Caraffi, M. del Rose, M. Felisa, and G. Vezzoni. Pedestrian Detection by means of Far-infrared Stereo Vision, Computer Vision and Image Understanding. *Computer Vision and Image Understanding*, pages 194–204, June 2007. [cited at p. 27, 44]

[BBCF01] A. Broggi, M. Bertozzi, G. Conte, and A. Fascioli. ARGO Prototype Vehicle. In Ljubisa Vlacic, Fumio Harashima, and Michel Parent, editors, *Intelligent Vehicle Technologies*, chapter 14, pages 445–493. Butterworth–Heinemann, London, UK, June 2001. ISBN 0750650931. [cited at p. 133]

[BBDL05] M. Bertozzi, A. Broggi, M. Del Rose, and A. Lasagni. Infrared Stereo Vision-based Human Shape Detection. *In Procs. IEEE Intelligent Vehicles Symposium*, pages 23–28, Las Vegas, USA, June 2005. [cited at p. 27, 43]

[BBF99a] M. Bertozzi, A. Broggi, and A. Fascioli. Autonomous Vehicles. *Linux Journal*, 59:40–45, 1999. [cited at p. 11]

[BBF99b] A. Broggi, M. Bertozzi, and A. Fascioli. ARGO and the MilleMiglia in Automatico Tour. *IEEE Intelligent Systems*, 14(1):55–64, 1999. [cited at p. 133]

[BBF00] M. Bertozzi, A. Broggi, and A. Fascioli. Vision-based intelligent vehicles: State of the art and perspectives. *Robotics and Autonomous Systems*, 32:1–16, 2000. [cited at p. 16, 27]

[BBF+04] M. Bertozzi, A. Broggi, A. Fascioli, T. Graf, and M.-M. Meinecke. Pedestrian Detection for Driver Assistance Using Multiresolution Infrared Vision. *IEEE Transactions on Vehicular Technology*, 53(6), 2004. [cited at p. 28, 43, 44, 45, 54, 157, 233, 235]

[BBF+06] M. Bertozzi, A. Broggi, M. Felisa, G. Vezzoni, and M. del Rose. Low-level Pedestrian Detection of Visible and Far Infra-red Tetra-vision. *In Procs. IEEE intellugent Vehilces Symposium*, pages 231–236, Tokyo, Japan, June 2006. [cited at p. 27, 45, 51]

[BBF+07a] M. Bertozzi, A. Broggi, M. Felisa, S. Ghidoni, P. Grisleri, G. Vezzoni, C. Hilario, and M. Del Rose. *Multi Stereo-based Pedestrian Detection by means of Daylight and Far Infrared Cameras*. Springer, 2007. [cited at p. 27, 76, 77, 134, 135, 150, 183, 188, 234]

[BBF+07b] Alberto Broggi, Massimo Bertozzi, Rean Isabella Fedriga, Cristina Hilario Gomez, Guido Vezzoni, and Michael Del Rose. Pedestrian Detection in Far Infrared Images based on the use of Probabilistic Templates. In *Procs. IEEE Intelligent Vehicles Symposium 2007*, pages 327–332, Istanbul, Turkey, 2007. [cited at p. 44, 51, 77, 134, 183, 234]

[BBFC99] A. Broggi, M. Bertozzi, A. Fascioli, and G. Conte. *Automatic Vehicle Guidance: the Experience of the ARGO Vehicle*. World Scientific, Singapore, April 1999. ISBN 9810237200. [cited at p. 133]

[BBFS00] A. Broggi, M. Bertozzi, A. Fascioli, and M. Sechi. Shape-based Pedestrian Detection. *In Procs. IEEE Intelligent Vehicles Symposium*, pages 215–220, 2000. [cited at p. 27, 28, 43, 44, 45, 51, 54, 82, 133, 233]

[BBG+03] M. Bertozzi, A. Broggi, T. Graf, P. Grisleri, and M.-M. Meinecke. Pedestrian Detection in Infrared Images. *In Procs. IEEE Intelligent Vehicles Symposium 2003*, pages 662–667, 2003. [cited at p. 78, 234]

[BBGD07] A. Broggi, M. Bertozzi, S. Ghidoni, and M. Del Rose. Pedestrian Shape Extraction by means of Active Contours. In *Procs. Intl. Conf. on Field and Service Robotics*, Chamonix, France, July 2007. [cited at p. 44]

[BBGM07] A. Broggi, M. Bertozzi, S. Ghidoni, and M. M. Meinecke. A Night Vision Module for the Detection of Distant Pedestrians. In *Procs. IEEE Intelligent Vehicles Symposium 2007*, pages 25–30, Istanbul, Turkey, June 2007. [cited at p. 28, 50]

[BBGR08] M. Bertozzi, A. Broggi, S. Ghidoni, and M. D. Rose. Pedestrian Shape Extraction by means of Active Contours. *Procs. Intl. Conf. on Field and Service Robotics*, pages 265–274, June 2008. [cited at p. 186]

[BCFG05] A. Broggi, C. Caraffi, R. I. Fedriga, and P. Grisleri. Obstacle detection with stereo vision for on-road vehicle navigation. *In Procs. Intl. IEEE Wks on MAchine Vision for Intelligent Vehicles*, June 2005. [cited at p. 28, 43, 44, 45, 52, 135]

[BCZ93] A. Blake, R. Curven, and A. Zisserman. A framework for spatiotempoarl control in the tracking of visual contours. *In Procs. IEEE Intelligent Vehicles Symposium*, pages 127–145, 1993. [cited at p. 42]

[BD00] A. Broggi and E. D. Dickmanns. Applications of Computer Vision to Intelligent Vehicles. *Image and Vision Computing*, (18):365–366, 2000. [cited at p. 16]

[BH94a] A. Baumberg and D. Hogg. An eficient method for contour tracking using active shape models. *in Proc. of IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pages 194–199, 1994. [cited at p. 45, 53]

[BH94b] A.M. Baumberg and D.C. Hogg. Learning flexible models from image sequences. 1994. [cited at p. 38, 43]

[BH96] A. Baumberg and D. Hogg. Generating Spatiotemporal Models From Examples. *IVC*, pages 525–532, 1996. [cited at p. 43]

[BHR97] U. Kressel B. Heisele and W. Ritter. Tracking non-rigid, moving objects based on color cluster flow. *in Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 257–260, 1997. [cited at p. 38, 40, 41]

[BK99] D. Beymer and K. Konolige. Real-time Tracking of Multiple People using Continuous Detection. *In Procs. Intl. Conf. on Computer Vision*, Kerkyra 1999. [cited at p. 27]

[Bou00] J. Y. Bouquet. Camera calibration toolbox for matlab. Available on the Internet: http://www.vision.caltech.edu/bouguet/, 2000. [cited at p. 67, 68, 70, 72, 83, 234]

[BP66] L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics*, pages 1554–1563, 1966. [cited at p. 219]

[Bra97] M. Brand. Learning concise models of human activity from ambient video. Technical Report 97-25, Mitsubishi Electric Research Labs, 1997. [cited at p. 225]

[Bra00] T. Brants. Tnt- a statistical part-of-speech tagger. *In Procs. of the 6th ANLP*, 2000. [cited at p. 178]

[BRF+03] A. Broggi, M. Del Rose, A. Fascioli, I. Fedriga, and A. Tibaldi. Stereo-based Preprocessing for Human Shape Localization in Unstructured Environments. *In Procs. IEEE Intelligent Vehicles Symposium 2003*, pages 410–415, 2003. [cited at p. 27, 83, 133]

[Bro71] D.C. Brown. Close Range Camera Calibration. *Photo. Eng.*, 37(8):855–866, 1971. [cited at p. 67]

[BS99] A.F. Bobick and S.S.Intille. Large occlusion stereo. In *International Journal of Computer Vision*, volume 33, pages 181–200, 1999. [cited at p. 90]

[BvS03] C. Curio J. Edelbrunner C. Igel D. Kastrup I. Leefken G. Lorenz A. Steinhage Bücher, T. and W. von Seele. Image processing and behaviour planning for intelligent vehicles. *IEEE Transactions on Industrial Electronics*, 90(1):62–75, 2003. [cited at p. 53, 54, 61]

[CD98] R. Cutler and L. S. Davis. View-based detection and analysis of periodic motion. In *ICPR '98: Proceedings of the 14th International Conference on Pattern Recognition-Volume 1*, page 495, Washington, DC, USA, 1998. IEEE Computer Society. [cited at p. 38, 39]

[CD00] R. Cutler and L.S. Davis. Robust Real-time Periodic Motion Detection, Analysis and Applications. *IEEE Trans. Pattern Analysis Machine Intell.*, 22:781–796, August 2000. [cited at p. 25, 36, 38, 39, 40, 233]

[CEK+00] C. Curio, J. Edelbrunner, T. Kalinke, C. Tzomakas, and W. von Seelen. Walking Pedestrian Recognition. *IEEE Transactions on Intelligent Transportation systems*, 1(3):155–163, 2000. [cited at p. 29, 40, 41, 42, 51, 53, 233]

[Cel01] M. Cellario. Human-centered intelligent vehicles: Toward multimodal interface integration. *IEEE Intelligent Systems*, 16(4):78–81, 2001. [cited at p. 51]

[CLK99] R. Collins, A. Lipton, and T. Kanade. A System for Video Surveillance and Monitoring. In *American Nuclear Society 8th Internal Topical Meeting on robotics and Remote Systems*, 1999. [cited at p. 22]

[Com00] European Comission. Chamaleon. pre-crash application all around the vehicle. http://www.chameleon-eu.org, 2000. Accessed 18 June 2008. [cited at p. 55]

[Com01a] European Comission. European transport policy for 2010: Time to decide. http://ec.europa.eu/transport/, 2001. White paper. Accessed 19 August 2007. [cited at p. 8]

[Com01b] European Comission. White book. european transport policy for 2010: Time to decide. http://www.fomento.es/, 2001. Accessed 14 August 2007. [cited at p. 14]

[Com02] European Comission. Edel. enhanced driver perception in poor visibility. http://www.edel-eu.org, 2002. Accessed 18 June 2008. [cited at p. 55]

[Com03] European Comission. Directive 2003/102/ec of the european parliament and of the council. http://eur-lex.europa.eu/LexUriServ/, 2003. Accessed 19 August 2007. [cited at p. 8, 10]

[Com05] European Comission. In-vehicle emergency call system ecall(second esafety communication). http://europa.eu/scadplus/leg/en/lvb/l31103a.htm, 2005. Accessed 15 June 2008. [cited at p. 8]

[Com06] European Comission. Watch-over project. http://www.watchover-eu.org/index.html, 2006. Accessed 15 June 2008. [cited at p. 55]

[Com07] European Comission. The seventh framework programme. http://cordis.europa.eu/fp7//, 2007. Accessed 14 June 2008. [cited at p. 14]

[Con03] Consumer. Revista Consumer Eroski. Available on the Internet: http://revista.consumer.es/, 2003. [cited at p. 3]

[DCdB+02] V. Depoortere, J. Cant, B. Van den Bosch, J. De Prins, R. Fransens, and L. Van Gool. Efficient pedestrian detection: a test case for svm based categorization. *In Proceedings Cognitive Vision Workshop*, 2002. [cited at p. 31]

[DM01] H. Delingette and J. Montagnat. Shape and topology constraints on parametric active contours. *CVIU*, pages 140–171, 2001. [cited at p. 215]

[dSG03] Informe del Secretario General. Resolucion 57/309. Technical report, Asamblea General de Naciones Unidas, 2003. Available on the Internet:http://www.who.int/world-health-day/2004/infomaterials/. [cited at p. 2, 3]

[dSV00] Instituto MAPFRE de Seguridad Vial. Available on the Internet: http://www.mapfre.com/, 2000. [cited at p. 3]

[dT04] Dirección General de Tráfico. Anuario estadístico. http://www.dgt.es/, 2004. (accessed 9 August 2007). [cited at p. 3]

[DT05a] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *In CVPR*, pages 886–893, 2005. [cited at p. 31, 32, 33, 36, 48, 233]

[dT05b] Dirección General de Tráfico. Estudio multicéntrico sobre morbilidad derivada de los accidentes de tráfico en españa. http://www.educacionvial.dgt.es/, 2005. (accessed 13 August 2007). [cited at p. 6]

[dT06] Dirección General de Tráfico. Los coches introducen nuevos sistemas de protección a los peatones. www.dgt.es/revista/archivo/pdf/num176-2006-Peatones.pdf, 2006. (accessed 27 December 2006). [cited at p. 10]

[Edw98] C.J.; Cootes T.F. Edwards, G.J.; Taylor. Interpreting face images using active appearance models. *Proceedings. Third IEEE Automatic Face and Gesture Recognitione*, (14-16):300 – 305, 1998. [cited at p. 99]

[ELW03] H. Elzein, S. Lakshmanan, and P. Watta. A motion and shape-based pedestrian detection algorithm. *In Procs. IEEE Intelligent Vehicles Symposium*, 22:500– 504, June 2003. [cited at p. 24, 30, 46, 51]

[eSa03] eSafety. European new car assessment programme. http://www.esafetysupport.org/, 2003. Accessed 12 August 2007. [cited at p. 14]

[ETS03] ETSC. Transport safety performance in the EU - a statistical overview, 2003. [cited at p. 1, 2]

[Eur95] EuroNCAP. European new car assessment programme. http://www.euroncap.com, 1995. Accessed 7 June 2008. [cited at p. 7]

[FGaG+01] U. Franke, D. Gavrila, A. Gern andS. Gorzig , R. Janssen, F. Paetzold, and C. Wohler. From Door to Door: Principles and Applications of Computer Vision for Driver Assistant Systems. *Intelligent Vehicle Technologies*, pages 131–188, 2001. [cited at p. 82]

[FGG+98] U. Franke, D. Gavrila, S. Gorzig , F. Lindner, F. Paetzold, and C. Wohler. Autonomous driving goes downtown. *in Procs. IEEE Intelligent Vehicles Symposium*, pages 40–48, October 1998. [cited at p. 27, 29, 30, 40, 41, 48, 54, 82, 186, 233]

[FH02] U. Franke and S. Heinrich. Fast Obstacle Detection for Urban Traffic Situations. *IEEE Transactions on Intelligent Transportation Systems*, 3(3), 2002. [cited at p. 110]

[FhM+93] O. Faugeras, B. hotz, H. Mathieu, T. Vivielle, Z. Zhang, P. Fua, E. Theron, L. Moll, G. Berry, J. Vuillemin, P. Bertin, and C. Proy. Real -time correlation-based stereo: algorithm, implementation and application. Technical Report 201, INRIA technical report, 1993. [cited at p. 112]

[FI07] FITSA-IDIADA. Tecnologías vehiculares para la mejora de la protección de peatones y ciclistas. Technical report, Fundación Instituto Tecnológico para la seguridad del automóvil, 2007. [cited at p. 10]

[FL98] H. Fujiyoshi and A. Lipton. Real-time Human Motion Analysis by Image Skeletonization. *In Procs. IEEE WACV'98*, pages 15–21, 1998. [cited at p. 38]

[FR95] W. T. Freeman and M. Roth. Orientation histograms for han gesture recognition. *in Procs. of Workshop on Autom. Face and Gesture Recognition*, pages 296–301, October 1995. [cited at p. 32]

[FS06] Frost and Sullivan. Frost: Legislation creates dramatic uptake for pedestrian safety systems. http://auto.ihs.com/news/2006/frost-pedestrian-protection.htm, 2006. accessed 2 January 2007. [cited at p. 11]

[fT05] Department for Transport. Road accidents statistics: Pedestrian casualties in road accidents: Great britain 2005. http://www.dft.gov.uk/pgr/statistics, 2005. (accessed 16 August 2007). [cited at p. 9]

[FTV00] A. Fusiello, E. Trucco, and A. Verri. A compact algorithm for rectification of stereo pairs. *Machine Vision and Applications*, pages 16–22, 2000. [cited at p. 67, 69, 90, 119]

[Fua91] P. Fua. Combining stereo and monocular information to compute dense depth maps that preserve depth discontinuities. *in Procs. of the 12th International Joint Conference on Artificial Intelligence*, pages 1292–1298, 1991. [cited at p. 91]

[Fue05] K.C. Fuerstenberg. Preventive Safety and Accident Avoidance. *5th European Congress on ITS*, June 2005. [cited at p. 3]

[FYN+] Y. Fang, K. Yamada, Y. Ninomiya, B. Horn, and I. Masaki. Comparison between Infrared-image-based and Visible-image-based Approaches for Pedestrian Detection. *IEEE*. [cited at p. 22]

[Gav98] D. M. Gavrila. Multi-feature hierarchical template matching using distance transforms. *Proc. of IEEE International Conference on Pattern Recognition*, pages 439–444, 1998. [cited at p. 48]

[Gav99] D. M. Gavrila. The visual analysis of human movement: a survey. *Comput. Vis. Image Underst.*, 73(1):82–98, January 1999. [cited at p. 22]

[Gav00] D. M. Gavrila. Pedestrian Detection from a Moving Vehicle. *In Procs. Eur. Conf. Computer Vision*, pages 37–49, 2000. [cited at p. 48, 51, 54]

[GG01] D. M. Gavrila and J. Giebel. Virtual Sample Generation for Template-based Shape Matching. *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 676–681, 2001. [cited at p. 48]

[GG02] D. M. Gavrila and J. Geibel. Shape-based Pedestrian Detection and Tracking. *In Procs. IEEE Intelligent Vehicles Symposium*, Paris, France, June 2002. [cited at p. 48, 52, 61]

[GGM04] D.M. Gavrila, J. Giebel, and S. Munder. Vision-based pedestrian detection: the protector system. pages 13–18, 2004. [cited at p. 55]

[GM07] D. M. Gavrila and S. Munder. Multi-Cue Pedestrian Detection and Tracking from a Moving Vehicle. *International Journal of Computer Vision*, 73(1):41–59, 2007. [cited at p. 51, 82]

[GML+03] D. M. Gavrila, P. Marchal, L. Letellier, M.-M. Meinecke, R. Morris, and M. Töns. Save-u : An innovative sensor platform for vulnerable road user protection. Intelligent Transport Systems and Services, 2003. [cited at p. 55]

[GP99] D. M. Gavrila and V. Philomin. Real-time object detection for smart vehicles. *Proc. of IEEE International Conference on Computer Vision*, pages 87–93, 1999. [cited at p. 48, 54]

[GZNR04] G. Grubb, A. Zelinsky, L.Ñilsson, and Ribbe. Pedestrian detection for driver assistance systems: Single-frame classification and system level performance. *In Proc. of the IEEE Intelligent Vehicle Symposium*, pages 19–24, Parma, Italy 2004. [cited at p. 27, 46, 52, 83]

[HAD+94] M. Hansen, P. Anandan, K. Dana, G. van der Wal, and P. Burt. Real-time scene stabilization and mosaic construction. *DARPA Image Understanding Workshop*, 1994. [cited at p. 25]

[HBP05] HBPO. Llegan los coches capaces de evitar daños graves a los peatones en caso de atropello. http://www.belt.es/noticias/2005/, 2005. accessed 7 June 2008. [cited at p. 10]

[HF82] D. Hoffman and B. Flinchbaugh. The interpretation of biological motion. *Biol. Cybernet*, pages 195–204, 1982. [cited at p. 36]

[HG93] R. Hartley and R. Gupta. Computing matched epipolar projections. *Proceedings of IEEE CVPR*, pages 549–555, 1993. [cited at p. 67]

[HHD98] I. Haritaoglu, D. Harwood, and L. S. Davis. W4s: A real-time system detecting and tracking people in 2 1/2d. In *ECCV '98: Proceedings of the 5th European Conference on Computer Vision-Volume I*, pages 877–892, London, UK, 1998. Springer-Verlag. [cited at p. 51, 53]

[HIG02] H. Hirschmuller, P.R Innocent, and J.M. Garibaldi. Real-time correlation-based stereo vision with reduced border errors. *Systems, Man and Cybernetics, Part C, IEEE Transactions on*, 47(1/2/3):229–246, 2002. [cited at p. 90, 91, 119]

[Hog83] D. Hogg. Model based vision: A program to see a walking person. *Image Vision Comput*, 1:5–20, 1983. [cited at p. 36]

[HTWM04] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *Systems, Man and Cybernetics, Part C, IEEE Transactions on*, 34(3):334–352, 2004. [cited at p. 22]

[Hu06] M. Hu. Multiple probabilistic templates based pedestrian detection in night driving with a normal camera. *in Procs. IEEE Intl. Conf. on Innovative Computing, Information and Control 2006*, 2:574–577, 2006. [cited at p. 50]

[KER95] W. Kruger, W. Enkelmann, and S. Rossle. Real-time estimation and tracking of optical flow vectors forobstacle detection. *In Procs. of the Intelligent Vehicles '95 Symposium.*, pages 304–309, 1995. [cited at p. 24]

[KER06] W. Kruger, W. Enkelmann, and S. Rossle. Real-Time Dense Stereo for Intelligent Vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 7(1):38–50, 2006. [cited at p. 90, 110, 114, 234]

[KFS05] M. Kuehn, R. Froeming, and V. Schindler. Assessment of vehicle related pedestrian safety. Technical Report 05-0044, Institute for Automotive Engineering.Technical University of Berlin,Germany, 2005. [cited at p. 6, 7, 233]

[KO94] T. Kanade and M. Okutomi. A stereo matching algorithm with an adative window: theory and experiment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(4):920–932, 1994. [cited at p. 90]

[KTZ92] B. B. Kimia, A.R. Tannenbaum, and S. W. Zucker. On the evolution of curves via a function of curvature, I: The classical case. *JMMA*, pages 438–458, 1992. [cited at p. 214]

[KWT88] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active Contour Models. *Intl. Journal of Computer Vision*, 1(4):321–331, 1988. [cited at p. 109, 120, 121, 213]

[KYO+96] T. Kanade, A. Yoshida, K. Oda, H. Kano, and M. Tanaka. A stereo matching for video-rate dense depth mapping and its new applications. *Procs. of Conference on Computer Vision and Pattern Recognition*, (4):196–202, 1996. [cited at p. 113]

[LAT02] R. Labayrade, D. Aubert, and J.-P Tarel. Real time obstacle detection on non flat road geometry through v-disparity representation. *In Proc. of the IEEE Intelligent Vehicle Symposium*, Versailles, France 2002. [cited at p. 27, 110]

[LAT03] R. Labayrade, D. Aubert, and J.-P Tarel. A single framework for vehicle roll, pitch, yaw estimation and obstacles detection by stereovision. *In Procs. of the IEEE Intelligent Vehicle Symposium 2003*, pages 31–36, Columbus, USA 2003. [cited at p. 135]

[LDT03] L. Lee, G. Dalley, and K. Tieu. Learning Pedestrian Models for Silhouette Refinement. *In Procs. IEEE International Conf. on Computer Vision*, Nice, France 2003. [cited at p. 144]

[Low04] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Computer Vision*, pages 91–110, 2004. [cited at p. 32]

[LRY02] S. Lakshmanan, N. Ramarathnam, and Teck Beng Desmond Yeo. A Side Collision Awareness System . *IEEE Intelligent Vehicle Symposium*, 2:640–645, June 2002. [cited at p. 24]

[MA95] B. Moghaddam and Pentland A. Probabilistic Visual Learning for Object Detection. *International Conference on Computer Vision*, pages 786–793, June 1995. [cited at p. 99]

[MDM02] K. Mühlmann and and R. Männer D. Maier, and J. Hesser. Calculating Dense Disparity Maps from Color Stereo Images, an Efficient Implementation. *International Journal of Computer Vision*, 47(1-3):79–88, April - June 2002. [cited at p. 91]

[Mer94] B. Merialdo. Tagging english text with a probabilistic model. *Computational Linguistics*, 20:155–171, 1994. [cited at p. 178]

[MJD+97] S.J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler. Non-intrusive Person Authentication for Access Control by Visual Tracking and Face Recognition. *Computer Vision and Image Understanding*, 80:42–56, 1997. [cited at p. 26]

[ML96] C.J. L. Murray and A. D. Lopez. *Global Burden of Disease: A comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020*. Harvard University Press, 1996. [cited at p. 2]

[MN95] H. Murase and S.Ñayar. Visual Learning and Recognition of 3D Objects from Appearance. *Interantional Journal of Computer Vision*, 14:5–24, 1995. [cited at p. 99]

[MPP01] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. *IEEE Transactions on Pattern Analysis andMachine Intelligence*, (23), April 2001. [cited at p. 30, 31, 46, 47]

[MS96] R. Malladi and J. A. Sethian. Level set methods for curvature flow, image enhancement, and shape recovery in medical images. *Proc. of International Conference on Mathematics and Visualization*, pages 255–267, Berlin 1996. [cited at p. 213]

[ND02] H. Nanda and L. Davis. Probabilistic Template Based Pedestrian Detection in Infrared Videos. *In Procs. IEEE Inteligent Vehicles Symposium*, 1:15–20, Paris, Francia, June 2002. [cited at p. 28, 29, 49, 50, 51, 61, 140, 141, 142, 147, 151, 152, 154, 183, 187, 233, 235]

[NDS06] B. Triggs N. Dalal and C. Schmid. Human detection using oriented histograms of flow and appearance. *In European Conference on Computer Vision*, 2006. [cited at p. 32, 33, 50]

[NHT97] NHTSA. National highway traffic safety administration. crashes arent accidents. Available on the Internet: http://www.nhtsa.dot.gov/, 1997. Archive v3.11. [cited at p. 6]

[OPS+97] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio. A trainable system for people detection. In *IUW 97*, pages 207–214, 1997. [cited at p. 30, 45, 46, 47, 49, 233]

[OS88] S. Osher and J. A Sethian. Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations. *Journal of Computation Physics*, 79:12–49, 1988. [cited at p. 213]

[PA97] T.Grove T.Tan D.Hogg K.Baker P.Remagnino, A.Baumberg and A.Worrall. An integrated traffic and pedestrian model-based vision system. *in Proceedings of British Machine Vision Conference*, pages 380–389, 1997. [cited at p. 53]

[PD96] D. V. Papadimitriou and Tim J. Dennis. Epipolar line estimation and rectification for stereo image pairs. *IEEE Transactions on Image Processing*, pages 672–676, 1996. [cited at p. 67]

[PDD00] V. Philomin, R. Duraiswami, and L. Davis. Pedestrian Tracking from a Moving Vehicle. *In Procs. IEEE Intelligent Vehicles Symposium*, pages 350–355, Detroit, USA, October 2000. [cited at p. 45, 46, 51, 53, 233]

[PN97] R. Polana and R. C. Nelson. Detection and Recognition of Periodic, Nonrigid Motion. *Intl. J. Computer Vision*, 23(3):261–282, 1997. [cited at p. 38, 39]

[POP98] Constantine P. Papageorgiou, Michael Oren, and Tomaso Poggio. A general framework for object detection. In *ICCV '98: Proceedings of the Sixth International Conference on Computer Vision*, page 555, Washington, DC, USA, 1998. IEEE Computer Society. [cited at p. 30, 35, 36, 45, 46, 51, 54]

[Rab89] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989. [cited at p. 224]

[RAC04] RACC. Real automóvil club de cataluña. test de seguridad activa del vehículo. Available on the Internet: http://www.racc.es/pub/ficheros/actualidad/, 2004. [cited at p. 6]

[RBH95] L. Robert, M. Buffa, and M. Hebert. Weakly Calibrated Stereo Perception for Rover Navigation. *Proceedings of ICCV-95*, pages 46–51, 1995. [cited at p. 67]

[RD96] L. Robert and R. Deriche. Dense Depth Map Reconstruction: A Minimization and Regularization Approach which Preserves Discontinuities. *ECCV*, 1:439–451, 1996. [cited at p. 83]

[RN03] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, 2 edition, 2003. [cited at p. 162]

[RWZD07] Yang Ran, Isaac Weiss, Qinfen Zheng, and Larry Davis. Pedestrian Detection via Periodic Motion Analysis. *International Journal of Computer Vision*, 71(2):143–160, 2007. [cited at p. 29, 38, 39, 40, 50, 51, 233]

[SARB06] F. Suard, and A. Bensrhair A. Rakotomamonjy, and A. Broggi. Pedestrian detection using infrared images and histograms of oriented gradients. *In Procs. IEEE Intelligent Vehicles Symposium 2006*, pages 206–212, 2006. [cited at p. 32]

[SB95] S. M. Smith and J. M. Brady. Asset-2: real-time motion segmentation and shape tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17(8):814–820, 1995. [cited at p. 24]

[SBP02] and J. Malik S. Belongie and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE T. Pattern Analysis and Machine Intelligence*, 24:509–522, 2002. [cited at p. 32]

[SG00a] C. Stauffer and W.E. Grimson. Learning patterns of activity using real-time tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2000. [cited at p. 61]

[SG00b] C. Stauffer and W.E. Grimson. Similarity templates for detection and recognition. *IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 221–28, 2000. [cited at p. 50]

[SGH04] A. Shashua, Y. Gdalyahu, and G. Hayun. Pedestrian Detection for Driving Assistance Systems: Single-frame Classification and System level Performance. *In Procs. IEEE Intelligent Vehicles Symposium*, Parma, Italy, June 2004. [cited at p. 30, 31, 47, 48, 51, 52, 83, 233]

[SK87] L. Sirovich and Kirby. Low-dimensional procedure for the characterization of human faces. *J. Optical Soc.*, pages 519–524, 1987. [cited at p. 101]

[SM00] L. Di Stefano and S. Mattoccia. Fast Stereo Matching for the VIDET System using a General Purpose Processor with Multimedia Extensions. *Fifth IEEE International Workshop on Computer Architectures for Machine Perception*, pages 356–362, September 2000. [cited at p. 91, 112]

[SM02a] N. T Siebel and S. Maybank. Fusion of Multiple Tracking Algorithms for Robust People Tracking. *In Proceedings of the 7th European Conference on Computer Vision (ECCV 2002)*, pages 373–387, 2002. [cited at p. 53]

[SM02b] L. Di Stefano and S. Mattoccia. Real-Time Stereo within the VIDET Project. *Real-Time Imaging*, 8(5):439–453, October 2002. [cited at p. 110]

[SMMN03] L. Di Stefano, M. Marchionni, S. Mattoccia, and G.Ñeri. A Fast Area-Based Stereo Matching Algorithm. *15th IAPR/CIPRS International Conference on Vision Interfacem*, pages 27–29, May 2003. [cited at p. 91]

[TCG95] and D. Cooper T. Cootes, and C. Taylor and J. Graham. Active shape models. Their training and applications. *Comput. Vision Image Understanding*, 51:38–59, 1995. [cited at p. 43, 45]

[THWN02] T. Tsuji, H. Hattori, M. Watanabe, and N. Nagaoka. Development of Night-vision System. *IEEE Transactions on ITS*, 3(3), September 2002. [cited at p. 28]

[TP91] M. Turk and A. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991. [cited at p. 99, 101]

[TT04] The World Health Organization and The World Bank. Informe mundial sobre prevención de los traumatismos causados por el tránsito. Technical report, 2004. available from the World Health Organization. (accessed 7 June 2008). [cited at p. 2, 16]

[VJ01a] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *International Journal of Computer Vision*, 2001. [cited at p. 32]

[VJ01b] P. Viola and M. Jones. Robust real-time object detection. *2nd Int.Workshop on Statistical and Computational Theories of Vision-Modelling, Learning, Computing, and Sampling*, 2001. [cited at p. 31, 50]

[VJS05] P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *Int. J. Comput. Vision*, 63(2):153–161, 2005. [cited at p. 27, 32, 33, 50, 51]

[WAPF98] C. Wohler, J. K. Aulanf, T. Portner, and U. Franke. A Time Delay Neural Network Algorithm for Real-time Pedestrian Recognition. *Int. Conf. on Intelligent Vehicle*, 1998. [cited at p. 33, 34, 233]

[WKA00] C. Wohler, U. Kressler, and J. K. Anlauf. Pedestrian Recognition by Classification of Image Sequences. Global Approaches vs Local Spacio-temporal Processing. *In Procs. IEEE Intl. Conf. Pattern Recognition*, 2000. [cited at p. 40, 41, 233]

[Wor08] World Bank Group. Road Safety [Online]. Available on web: www.worldbank.org/html/fpd/transport/roads/safety.htm. accessed 7 June 2008. [cited at p. 1]

[WS92] D. J: Williams and M. Shah. A fast algorithm for active contours and curvature estimation. *CVGIP: Image Understanding*, 55(1):12–26, 1992. [cited at p. 121, 129, 130, 235]

[XF02] F. Xu and K. Fujimura. Pedestrian detection and tracking with night vision. *in Procs. IEEE Intelligent Vehicles Symposium 2002*, pages 63–71, 2002. [cited at p. 50]

[YM00] S. Yasutomi and H. Mori. A method fot discriminating of pedestrian based on rhythm. *in Proc. of IEEE Intl. Conference on Intelligent Robots and Systems*, pages 988–995, 2000. [cited at p. 53]

[Zha99] Z. Zhang. Flexible Camera Calibration by Viewing a Plane from Unknown Orientations. *International Conference on Computer Vision (ICCV'99)*, pages 666–673, 1999. [cited at p. 67]

[Zha00] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000. [cited at p. 67, 82]

[Zha01] L. Zhao. *Dressed Human Modeling, Detection and Parts Localization*. PhD thesis, Carnegie Mellon University, 2001. [cited at p. 49, 51, 54]

[ZT00a] L. Zhao and C. Thorpe. Recursive context reasoning for human detection and parts identification. *IEEE Workshop on Human Modelling, Analysis and Synthesis*, 2000. [cited at p. 26, 83]

[ZT00b] L. Zhao and C. Thorpe. Stereo and Neural Network-based Pedestrian Detection. *IEEE Transactions on Intelligent Transportation Systems*, 1(3):148–154, September 2000. [cited at p. 26, 233]

Apéndices

Apéndice A

Algoritmo de Correspondencia Basado en Áreas

El algoritmo estéreo implementado en esta tesis, se basa en el *Videt Stereo Algorithm* abreviado como VSA, desarrollado en la Universidad de Bolonia en el contexto del proyecto VIdeo DEcoder by Touch (VIDET). Además de la optimización que supone el uso del cálculo incremental de la implementación VSA, adicionalmente se ha empleado el conjunto de instrucciones SSE2 así como los *intrínsecos* proporcionados por la tecnología MMX. Se consiguen así acelerar los tiempos de ejecución aprovechando las capacidades SIMD de los procesadores de propósito general.

A.1. Estrategia de correspondencia: Correlación basada en el SAD

La idea de recorrer toda la imagen para el cálculo del SAD queda desechada en cuanto se cae en la cuenta de la gran cantidad de operaciones que supone el proceso. En una primera optimización del VSA, se pretende usar operaciones realizadas en un píxel para las necesarias en píxeles futuros, consiguiendo así reducir el tiempo de operación.

Considerando que $SAD(x, y, d)$ es el valor SAD entre una ventana de tamaño $(2n+1)(2n+1)$ centrada en las coordenadas (x, y) en la imagen izquierda y la correspondiente ventana centra en $(x + d, y)$ en la imagen derecha:

$$SAD\,(x, y, d) = \sum_{i,j=-n}^{n} |I_{izda}\,(x + j, y + i) - I_{dcha}\,(x + d + j, y + i)| \tag{A.1}$$

Observando la figura A.1, es fácil notar que $SAD(x, y + 1, d)$ puede ser obtenida a partir de $SAD(x, y, d)$:

$$SAD(x, y + 1, d) = SAD(x, y, d) + U(x, y + 1, d) \tag{A.2}$$

con $U(x, y+1, d)$ representando la diferencia entre los SADs asociados con las fila más bajas y más altas de la ventana de combinación (mostradas en gris suave en la figura A.1):

$$U(x, y+1, d) = \sum_{j=-n}^{n} |I_{izda}(x+j, y+n+1) - I_{dcha}(x+d+j, y+n+1)| - \sum_{j=-n}^{n} |I_{izda}(x+j, y-n) - I_{dcha}(x+d+j, y-n)| \tag{A.3}$$

Además, $U(x, y+1, d)$ puede ser operado de $U(x-1, y+1, d)$ sencillamente considerando que las contribuciones aportadas por los cuatro puntos mostrados en gris oscuro en la figura A.1:

$$\begin{aligned} &U(x, y+1, d) = U(x-1, y+1, d) \\ &+ |I_{izda}(x+n, y+n+1) - I_{dcha}(x+d+n, y+n+1)| \\ &- |I_{izda}(x-n-1, y+n+1) - I_{dcha}(x+d-n-1, y+n+1)| \\ &+ |I_{izda}(x-n-1, y-n) - I_{dcha}(x+d-n-1, y-n)| \end{aligned} \tag{A.4}$$

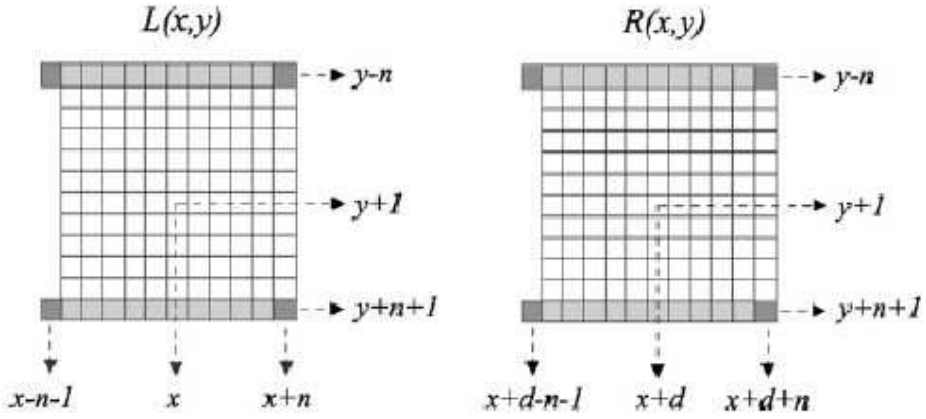


Figura A.1: Cálculo incremental de valores SAD.

Intentando aplicar las anteriores ecuaciones a la utilización de los intrínsecos se remarca que no es conveniente implementar la ecuación 4.15 ya que los intrínsecos trabajan con datos que se encuentran consecutivamente en memoria, por lo que al acceder a cuatro valores tan distanciados (en memoria) se perdería la posibilidad de trabajar con múltiples datos al mismo tiempo. Así, el valor de SAD se obtendría de el valor de SAD de una fila anterior sumándole la fila $y+n+1$ y restándole la fila $y-n$, tal y como se muestra en la figura A.2.

Por otra parte, utilizando la misma técnica recursiva con los movimiento verticales dentro de la imagen se observa que de igual modo que para calcular $SAD(x, y+1, d)$ partimos del valor de $SAD(x, y, d)$; para calcular $SAD(x+1, y, d)$ también se puede partir de $SAD(x, y, d)$. De tal forma que:

$$SAD(x+1, y, d) = SAD(x, y, d) + U(x+1, y, d) \tag{A.5}$$

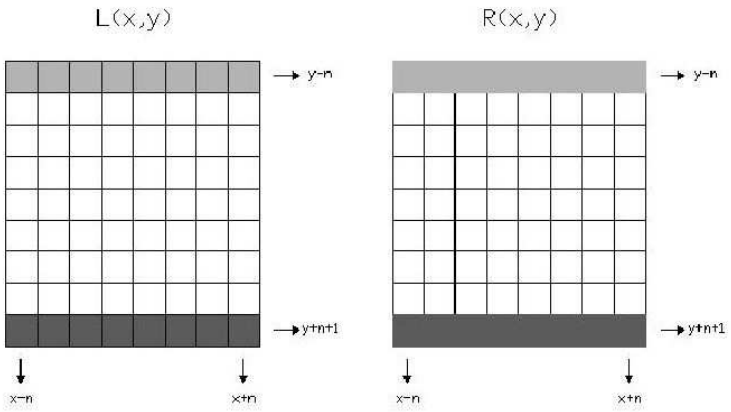


Figura A.2: Cálculo incremental de valores SAD por filas.

siendo $U(x+1, y, d)$:

$$U(x+1,y,d) = \sum_{j=-n}^{n} |I_{izda}(x+n+1, y+j) - I_{dcha}(x+d+n+1, y+j)| \\ - \sum_{j=-n}^{n} |I_{izda}(x-n, y+j) - I_{dcha}(x+d-n, y+j)| \tag{A.6}$$

Con estas consideraciones previas se consigue reducir considerablemente el número de operaciones a realizar optimizando de este modo el cálculo del mapa de disparidad. De esta forma se recorrerá la imagen del siguiente modo:

- Cálculo de SAD de las ventanas en $x = 0, y = 0, d \subset [0, d_{máximo}]$. En este caso es necesario el cálculo SAD completo ya que no se puede utilizar valores de datos anteriores.

- Cálculo de SAD recorriendo la matriz en horizontal. En este caso se puede obtener el valor a través de SAD anteriores tal y como explica la ecuación A.3. Moviéndose por columnas.

- Cálculo de SAD recorriendo la matriz en vertical. Una vez que se llega al valor de $x_{máximo}$ con $d_{máxima}$ será necesario pasar a la ventana de$x = 0, y = y + 1$ y $d = 0$. Aplicando la ecuación A.6 se obtendrá este valor usando datos calculados anteriormente. Moviéndose por filas.

La implementación del método del SAD podría realizarse recorriendo todos los píxeles de las imágenes. Sin embargo, resulta mucho más eficaz emplear la técnica SIMD al cálculo del mapa denso.

A.2. Implementación del Sistema Estéreo Mediante Instrucciones SIMD

La incorporación, en 1997, de la tecnología multimedia MMX a los procesadores Pentium ha sido la mejora más significativa introducida hasta entonces en sus procesadores. Aunque los fundamentos de la tecnología MMX se desarrollaron e implementaron anteriormente en otros procesadores, lo cierto es que Intel consiguió llevar dichos adelantos a la informática personal.

Aún así, el conocimiento del procesador y de su lenguaje ensamblador seguía siendo necesario. Posteriormente se ha ido incluyendo ampliaciones de flujo SIMD; En 1999, con la aparición del Pentium III se introduce el conjunto de instrucciones SSE (*Streaming SIMD Extensions*), que trabajaba sobre elementos de datos en coma flotante de 32-bits, procesando 4 de ellos en paralelo (4x32 =128 bit). Esta proposición está confeccionada con precisión para procesadores de juegos de 3D, pero resulta insuficiente para otras aplicaciones. Con el Pentium 4 o el AMD de Athlon, se soluciona este problema al añadir el conjunto de intrucciones SSE2, capaces de trabajar con mayor número de datos además de añadir estas últimas los *intrínsecos*. Los beneficios clave de SSE2 son que las instrucciones MMX pueden trabajar con bloques fijos de datos de 128-bits, y que las instrucciones SSE ahora soportan valores en coma flotante de 64-bits. Esto significa que las instrucciones SSE2 sólo pueden procesar dos datos de 64-bit en paralelo.

Los intrínsecos se incorporan con la función de facilitar el uso y la optimización del código permitiendo utilizar la tecnología MMX y sus ampliaciones sin recurrir a la programación en código ensamblador. Con esto se da paso a la posibilidad del uso de compiladores genéricos simplemente con añadir una serie de librerías. El conjunto de intrucciones SSE2, disponible a partir de los Pentium 4 de Intel o de los AMD de Athlon 64.

A pesar de los beneficios incluidos por la tecnología MMX y sus ampliaciones conviene remarcar que su uso conlleva ciertas restricciones. Entre ellas hay que destacar que todos los múltiples datos con los que se trabaja en una instrucción deben estar situados en memoria de forma consecutiva. Además la operación que se va a realizar sobre estos datos será la misma para todos ellos y para garantizar que las operaciones se realizen a la mayor velocidad posible, es necesario que los datos estén en la caché del procesador, no en la memoria principal. Por último, es necesario decir que no todas las operaciones están disponibles ni dentro de la tecnología MMX ni con los intrínsecos. Estas instrucciones realizan operaciones básicas e incluso los intrínsecos no comprenden todas ellas.

Para la obtención de los mapas densos, se trabaja con múltiples datos (los píxeles) que se encuentran agrupados, ya que pertenecen a una misma imagen, y se va a realizar sobre cada uno de ellos las mismas operaciones. Por tanto, el uso de funciones multimedia parece idóneo para este caso.

A.2.1. Conjunto de Instrucciones SSE2

Procesando los elementos de datos en paralelo, se mejora significativamente el rendimiento. La tecnología SIMD permite el desarrollo de aplicaciones multimedia optimizadas. El camino más directo para usar las instrucciones SIMD es introducirlas en lenguaje ensamblador

en el código fuente. Sin embargo, esto puede llevar mucho tiempo y ser tedioso, y el lenguaje ensamblador introducido en la programación no es soportado por todos los compiladores. En su lugar, Intel proporciona una implementación sencilla a través del uso de conjuntos de ampliaciones API llamadas *intrínsecos*. Los *intrínsecos* son ampliaciones especiales de código que permiten el uso de la sintaxis de C en vez de tener que acceder a registros *hardware*. Con esto se permite el uso de compiladores genéricos siendo suficiente con añadir una serie de librerías. Usando estas instrucciones se libera a los programadores de tener que programar en lenguaje ensamblador y de controlar los registros. Además, el compilador optimiza la programación de las instrucciones, por lo que los ejecutables funcionan más rápido.

El conjunto de intrucciones SSE2 utilizan datos de 128 bits. El tipo de dato $\_m128$ representa el contenido de un registro de la ampliación de flujos SIMD usado por los intrínsecos. En concreto, se va a utilizar el tipo $\_m128i$, que puede contener dieciséis valores enteros de 8-bit, ocho de 16-bit, cuatro de 32-bit o dos de 64-bit.

A.2.1.1. EL uso de intrínsecos

La primera mejora introducida al algoritmo VSA en C, consiste en la introducción de intrínsecos utilizando una ventana de 8x8 píxeles para el cálculo de la disparidad. Una vez cargada la imagen en memoria es necesario cargar la ventana de 8x8 en los tipos de datos utilizados por los intrínsecos,$\_m128i$, utilizando el intrínseco $\_mm\_loadu\_si128$. Este intrínseco almacena 8 valores consecutivos que se encuentren en memoria. Ante este hecho se obtendrá los 8 datos de una fila perteneciente a la imagen, por lo que usará 8 veces esta función para conseguir la ventana de $8x8$. El problema se presenta cuando se hace necesario obtener los 8 datos de una columna, ya que los datos no están consecutivos en memoria y como ya se ha dicho, la función de carga sólo puede seleccionar datos consecutivos. Esto es resuelto transmutando las matrices imágenes, obteniendo así dos imágenes, la normal y la transmutada. Con esto se permite el libre recorrido de las imágenes por filas o por columnas.

Aclarado el recorrido por las imágenes sólo queda implementar el cálculo de SAD con el uso de intrínsecos. Se explicará únicamente el cálculo de una de las filas de la ventana ya que el resto de los cálculos se llevan a cabo de forma similar. En cuanto al hecho de que fuera necesario trabajar con columnas, simplemente se obtendrían las filas de la imagen transmutada, pudiendo proceder a trabajas con los datos deseados.

```
    // Carga de los 8 valores de la imagen de la derecha
*lnd1=_mm_loadu_si128((__m128i *)(derecha));

// Carga de los 8 valores de la imagen de la izquierda
*ln1=_mm_loadu_si128((__m128i *)(izquierda+YMAX*(d)));

// a-b y b-a
*AB=_mm_sub_epi16(*ln1,*lnd1);
*BA=_mm_sub_epi16(*lnd1,*ln1);
```

```
// Valor máximo de (a-b, b-a) que nos da el valor absoluto de la diferencia
*ValorAbs=_mm_max_epi16(*AB,*BA);


// Suma de los valores absolutos
*Suma=_mm_add_epi16(*ValorAbs,*Suma);
```

Posteriormente es necesario sumar los valores absolutos de todas las filas de la ventana para obtener el SAD de la misma. Una vez obtenidos todos los valores de SAD necesarios para cada uno de los píxeles de las imágenes es necesario determinar el mapa de disparidad. La disparidad para cada píxel vendrá determinada por el valor de d con el que se consigue un valor de SAD mínimo.

A.2.1.2. Intrínsecos de alto nivel

A pesar de la introducción del uso de los intrínsecos y de las reducciones de cálculos se obtienen unos resultados no muy alentadores. Aunque el mapa de disparidad es aceptable, los tiempos de ejecución son prácticamente similares a los obtenidos con un cálculo de disparidad implementado en C. Ante esta hecho, se opta por cambiar el método de la obtención del valor de SAD, que se realizaba con intrínsecos sencillos como sumas, restas y comparaciones, por el uso de intrínsecos de más alto nivel.

Entre los intrínsecos se logra encontrar uno de ellos, ${}_m m_s ad_e pu8$. Éste toma dos grupos, de 16 valores cada uno, hace el valor absoluto de las diferencias de las parejas de cada grupo, suma 8 y 8 de estos valores y los devuelve en dos datos distintos de 64 bits cada uno, agrupados en una variable de tipo ${}_{-m}128i$, tal y como se indica en la figura A.3 .

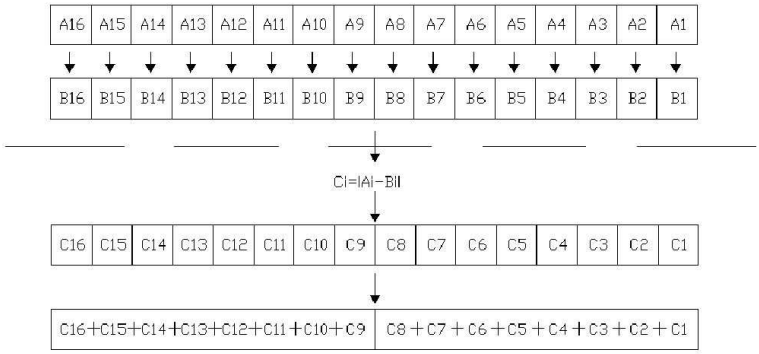


Figura A.3: Descripción de la función intrínseco

De este modo con un solo intrínseco se realiza la mayor parte de los cálculos del valor de SAD. Para poder usarlo, los datos deben ser de 8 bits. Obligando a que los valores de las

imágenes tomen valores positivos, se evita el uso de datos de 16 bits. Es suficiente con sumar el valor 127 a todos los píxeles de la imagen, tal y como refleja la ecuación 4.18.

Con el uso de este intrínseco se logra reducir notablemente el tiempo de procesado de las imágenes en el cálculo del mapa de disparidad, a pesar de ser de más alto nivel y por ello suponer un mayor coste computacional (para consultar tiempos ver tabla cap. 4).

Apéndice B

Modelos deformables o flexibles: Introducción a los *Snakes*

A la hora de representar un modelo de contorno activo se pueden seguir dos enfoques totalmente opuestos: El enfoque clásico propuesto por Kass *et al.* [KWT88], representa los contornos de manera explícita o paramétrica, siguiendo la formulación Lagrangiana. Por el contrario, el segundo enfoque representa los contornos de manera implícita, siguiendo la formulación Euleriana. La representación de contornos activos implícita fue planteado por Malladi et al. hacia 1996 [MS96], tras el trabajo sobre curvas de nivel (*Level Set*) de Osher y Sethian ([OS88]). Más tarde Caselles et al. [**?**] crean el concepto de *Geodesic snakes* o contornos activos geodésicos.

El enfrentamiento entre estas dos representaciones corresponde a la oposición existente entre adoptar una solución de Lagrange o de Euler. Bajo un enfoque Lagrangiano, la deformación de un cuerpo se describe considerando al cuerpo en reposo. En el enfoque Euleriano esa deformación se describe en función de la posición del cuerpo en cada instante.

En general, las representaciones implícitas son menos eficientes que las paramétricas. Esto es así porque la actualización de un contorno implícito requiere de la actualización, al menos, de una estrecha banda alrededor de cada contorno. Además, los contornos paramétricos no exigen que el muestreo de los vértices sea uniforme, mientras que en los contornos implícitos la resolución depende de la resolución de una malla regular. Si la implementación de un contorno paramétrico suele ser inmediato, en el caso de la implementación de algoritmos level-sets es más costoso.

La mayor ventaja de una representación implícita es, sin ninguna duda, su habilidad para cambiar automáticamente la topología del contorno durante su deformación. Esta cualidad hace que sean adecuados para reconstruir contornos de geometría compleja como por ejemplo, tipo árbol. Además permite la intersección de varios contornos. En contraposición, el mayor inconveniente de un contorno paramétrico, es que no soporta cambios en su topología. Sin embargo se han propuesto distintas soluciones a esta limitación.

Antes de pasar a detallar el modelo implementado, es importante explicar ciertos conceptos

relativos a un contorno paramétrico, que es el tipo de representación empleado.

B.1. Contornos activos paramétricos

B.1.1. Discretización del contorno

Para un contorno activo continuo la parametrización del contorno se caracteriza por la siguiente métrica:

$$f(s) = \left|\frac{\partial v}{\partial s}\right| \tag{B.1}$$

Si $f(s) = 1$, el parámetro de $v(s)$ coincide con la longitud de arco del contorno.

Para un contorno discreto, la parametrización se corresponde con el espaciado relativo entre puntos y se caracteriza por:

$$f_i = |v_i - v_{i-1}| \tag{B.2}$$

Si todos los puntos están separados a una misma distancia, entonces la parametrización discreta es proporcional a la longitud de arco.

Para una representación continua, la parametrización es independiente de la forma del contorno. Sin embargo, para un contorno discretizado por diferencias finitas, la forma y la parametrización no son completamente independientes, como puede verse en la figura B.1.

Figura B.1: Se muestran dos círculos discretos con diferente parametrización. La figura de la derecha ilustra cómo la parametrización afecta a la forma.

Por tanto, controlar la parametrización de los contornos discretos es crucial. Para modificar dicha parametrización, sólo debería considerarse la componente tangencial de la fuerza interna, como se explica a continuación.

B.1.2. Componente normal y componente tangencial de la fuerza interna

Kimia et al. [KTZ92] demostraron que únicamente la componente normal de la fuerza interna aplicada sobre una curva afecta a la forma. Este es el motivo por el que en el método de level-set sólo se considera la deformación en la dirección del gradiente (en la dirección normal). Sin embargo, cuando se trata de contornos paramétricos, es importante definir, además

de la componente normal, la componente tangencial de la fuerza interna para controlar el espaciado entre los puntos del contorno. Ya se ha visto en la figura B.1, el modo en el que la parametrización de los puntos afecta a la forma de un contorno paramétrico. La figura B.2 representa la componente tangente y normal del punto v_i.

Diversos autores han propuesto descomponer las fuerzas tanto internas como externas en sus componentes tangenciales y normales. Sin embargo, en una implementación clásica (paramétrica) de contorno activo, la forma y el muestreo de los vértices están interrelacionados debido a que los componentes tangencial y normal de la fuerza interna no se consideran por separado. Delingette et al. [DM01] fueron los primeros en considerarlos de manera independiente.

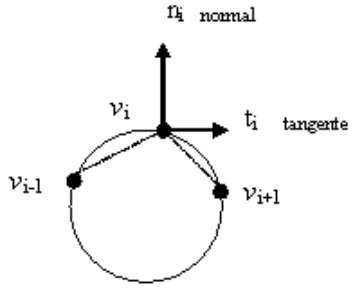


Figura B.2: Definición de la componente tangente y normal del punto v_i. La dirección tangente en un punto, es la dirección de la línea que une sus dos vecinos.

B.1.3. Deformación de un modelo paramétrico

El modelo convencional de *snakes*, representa la curva deformable , como un contorno 2D cerrado, donde hace referencia al parámetro del contorno, que en general no coincide con la longitud de arco del contorno.

La deformación del contorno sigue la ley de movimiento de Newton:

$$\underbrace{\left(\frac{\partial(\omega_1 v)}{\partial s} - \frac{\partial^2(\omega_2 v)}{\partial s^2}\right)}_{\substack{Fuerzas\\Internas}} + \underbrace{\nabla F}_{\substack{Fuerzas\\Externas}} = 0 \Rightarrow \frac{\partial^2 v}{\partial t^2} = -\gamma\frac{\partial v}{\partial t} + f_{internas} \tag{B.3}$$

donde la expresión de la izquierda es la ecuación de equilibrio del sistema y la de la derecha representa la ley de movimiento de Newton. Los subíndices s y t hacen referencia a la diferenciación con respecto al espacio y tiempo, y ∇F es el gradiente de F.

Las ecuaciones de equilibrio para $v(s)$ se establecen de tal forma que la curva tiende a valores altos de las fuerzas externas. Esta tendencia a maximizar F es compensada por las fuerzas internas. Las fuerzas externas son debidas a las energías presentes en la imagen. Esos campos de potenciales atraen al snake hacia ellos. Las fuerzas internas actúan en su contra, imponiendo restricciones de suavidad a la curva. Bajo un enfoque Lagrangiano, la curva se

deforma hasta que alcanza una situación de equilibrio. Tal estado se alcanza cuando la curva presenta velocidad y aceleración cero.

Los cálculos prácticos sobre $v(s)$ deben ser discretos en el tiempo y espacio. La discretización tanto espacial como temporal de la curva se basan en diferencias finitas. El snake tradicional se representa por un conjunto discreto de puntos o muestras, distribuidas de manera uniforme a lo largo del contorno $v(s) = (x(s), y(s))$. La ecuación B.3 se discretiza, aplicando diferencias finitas para aproximar las derivadas espaciales:

$$\left|\frac{dv_i}{ds}\right|^2 \approx |v_i - v_{i-1}|^2 = (x_i - x_{i-1})^2 + (y_i - y_{i-1})^2 \tag{B.4}$$

$$\left|\frac{d^2v_i}{ds^2}\right|^2 \approx |v_{i-1} - 2v_i + v_{i+1}|^2 \tag{B.5}$$

Al emplear aproximaciones por diferencias finitas, se trabaja con muestras de la curva tomadas en ciertos puntos discretos, careciendo de información sobre la forma entre las muestras. Análisis numéricos más modernos apuestan por el método de los elementos finitos, donde cada muestra$v(s_i)$ son consideradas como variables nodales a partir de las cuales se puede reconstruir la curva continua por completo. La forma más simple de representar una curva mediante elementos finitos, es un polígono, siendo sus vértices los elementos nodales.

Se pueden obtener aproximaciones más suaves si se modela la curva como un spline, ya que estas curvas polinomiales pasan cerca pero no necesariamente a través de los puntos o nodos. Esta característica es particularmente eficaz, porque un spline, por definición, mantiene cierto grado de suavidad (su representación ya cumple ciertas condiciones de continuidad), tarea que de otra manera recae completamente en la fuerzas internas. De manera que, empleando B-splines, aquellos términos de la formulación cuyo fin es proporcionar suavidad a la curva, pueden omitirse. Además el número de nodos requeridos se reduce, obteniendo una mejora en el tiempo de cómputo considerable.

B.2. Robustez y estabilidad de los *snakes*

Los términos de regularización (presentes en las energías internas) en las ecuaciones dinámicas ayudan a estabilizar los *snakes*, pero son bastantes limitados en sus acciones. Imponen unas restricciones muy generales a la forma, favoreciendo que el snake sea corto y suave. Sin embargo, muy a menudo, esto no es suficiente y se necesita incluir más conocimiento a priori en el modelo para conseguir un comportamiento estable.

Si además de llevar a cabo la detección de objetos, se quiere realizar el seguimiento de los mismos como un proceso continuo, las condiciones de robustez son aún más exigentes. Es de vital importancia recuperarse de los errores producidos en tiempo-real. Los mecanismos generales para establecer la forma de los modelos son suficientes. Se necesitan mecanismos más precisos que representen conocimiento a priori específico sobre las clases de objetos y sus movimientos.

B.2.1. Regularización de la forma

Como se ha comentado en el apartado anterior, las fuerzas tangenciales no afectan demasiado a la evolución de la forma del contorno, ya que tienen que ver con la parametrización del mismo. Es la componente normal de las fuerzas internas la que influye en la forma. Dicha componente trata de regularizar el contorno, imponiéndole restricciones para que el snake modifique su forma según unos criterios.

Las fuerzas internas más utilizadas en contornos activos son:

$$
\begin{aligned}
&f_{interna} = k\vec{n} && \text{Minimiza la longitud del contorno.} \\
& && \text{Se emplea en el método de level-set } \textit{(Mean curvature motion)} \\
&f_{interna} = \frac{\partial^2 v}{\partial s^2} && \text{Minimiza la energía elástica. } \textit{(Weak string or Laplacian smoothing)} \\
&f_{interna} = -\frac{\partial^4 v}{\partial s^4} && \text{Minimiza la energía de flexión } \textit{(Thin rod smoothing)} \\
&f_{interna} = p\vec{n} && \text{Minimiza el área encerrado en el contorno } \textit{(Balloon force)} \\
&f_{interna} = \sum_j \left(l_j - l_j^0\right)\left(v_j - v_i\right) && \text{Fuerza de muelles } \textit{(Spring force)}
\end{aligned}
\tag{B.6}
$$

Sin embargo, al aplicar estas fuerzas no siempre se obtiene el efecto deseado. Su mayor inconveniente es que tienden a hacer que el contorno se encoja, favoreciendo la segmentación de estructuras que estén dentro del contorno y no fuera. Es más, esa tendencia del contorno a encogerse a veces impide que el snake entre en el interior de estructuras que presentan cierta concavidad. La técnica de GFV (Gradient vector flow) trata de hacer frente a esa limitación.

En general, cuanto mayor sea el grado de suavizado que se aplique al contorno, menos se acortará su longitud. Por ejemplo, si se implementan fuerzas internas basadas en la cuarta derivada a lo largo del contorno, la curva se encoge menos que si se aplican fuerzas internas basadas en regularizar la curvatura (como en el caso de mean curvature motion o laplacian smoothing).

B.2.2. Cambios de topología de los contornos

Los contornos clásicos no soportan cambios topológicos. En cambio, es una de las ventajas de los contornos implícitos. Sin embargo, bajo el enfoque de level-set, los cambios topológicos están influenciados por el tamaño de la malla, que también determina la resolución del contorno. Se han propuesto distintas soluciones para que los contornos paramétricos puedan soportar cambios automáticos en su topología. Los T-*snakes* o *Topology Adaptive Snakes* son una de las primeras soluciones planteada en 1995 por McInerney y Terzopoulos.

Apéndice C

Introducción a los Modelos Ocultos de Markov

Los modelos ocultos de Markov, más conocidos como HMM, se han venido aplicando a la tecnología del habla y al campo relacionado del reconocimiento de caracteres (OCR) manuscritos obteniendo resultados prometedores. El modelo oculto de Markov fue creado por Baum *et al.* a mediados de los 60 [BP66].

Este anexo es una introducción a los procesos de Markov y a los modelos ocultos de Markov. En él se explican las nociones básicas necesarias para la comprensión del modo en que se han usado los HMM en esta tesis.

C.1. Inferencia Bayesiana

La inferencia Bayesiana o la clasificación Bayesiana fue aplicada con éxito a los problemas del reconocimiento del habla y de caracteres a finales de 1950. Este paradigma nace como resultado del trabajo de Bayes (1763). Bajo este paradigma, el razonamiento probabilístico o inferencia tiene lugar mediante el cálculo de la probabilidad condicional de una variable, siendo conocidos los valores de otro conjunto de variables o evidencias. Para una tarea de clasificación, este problema se traduce en determinar la clase, de entre un conjunto de clases, a la que pertenece una observación dada. Si se considera que en lugar de una única observación se tiene una secuencia de observaciones, la clasificación deberá asignar una secuencia de clases.

Cabe preguntarse cuál es la mejor secuencia de clases que corresponde a una secuencia o conjunto de observaciones. La interpretación Bayesiana de esta tarea comienza considerando todas las posibles hipótesis. De entre todas las secuencias de clases, se quiere seleccionar la hipótesis más probable dado el conjunto de n observaciones $x = \{x_1 x_2 ... x_n\}$. Dicho de otro modo, se quiere encontrar, de entre todas las hipótesis $C = \{C_1 C_2 ... C_n\}$ de longitud n, la secuencia concreta C tal que la probabilidad $P(C|x)$ es la más alta. De esta forma, a la hipótesis más probable se le suele denominar hipótesis maximum a posteriori o MAP.

Esto se puede expresar como:

$$\hat{C}_{MAP} = \underset{C_i}{\text{argmax}}\ P(C|x) \tag{C.1}$$

donde la notación ˆsignifica que la secuencia MAP es una estimación.

A pesar de que la ecuación C.1 garantiza que proporciona una secuencia óptima, no se sabe cómo calcular $P(C|x)$. La regla de Bayes permite transformar esa ecuación en un conjunto de probabilidades que resultan más sencillas de calcular;

$$P(C|x) = \frac{P(x|C)P(C)}{P(x)} \tag{C.2}$$

Se puede simplificar C.2 eliminando del denominador $P(x)$, ya que la búsqueda de la secuencia de hipótesis más probable se realiza para la misma secuencia de observaciones.

Por tanto, el problema de la clasificación se puede resolver maximizando la siguiente fórmula:

$$\hat{C}_{MAP} = \underset{C_i}{\text{argmax}}\ P(x|C)P(C) \tag{C.3}$$

Por lo que la secuencia más probable de clases para un conjunto de observaciones puede calcularse a través del producto de dos probabilidades, seleccionando la secuencia de etiquetas que como resultado del producto proporcione la mayor probabilidad. Los dos términos en cuestión son, la probabilidad a priori de la secuencia C, $P(C)$ y la verosimilitud del conjunto de observaciones x, $P(x|C)$. $P(C)$ es el conocimiento inicial que se tiene sobre que la secuencia C sea la correcta. $P(x|C)$ denota la probabilidad de observar el conjunto de observaciones x dada la secuencia de etiquetas C.

Desafortunadamente, la ecuación C.3 sigue siendo difícil de calcular. Para simplificar la expresión se suelen tomar dos suposiciones. Sin embargo, para poder definir ambas correctamente es necesario realizar primero una introducción a los procesos de Markov.

C.2. Procesos estocásticos de tipo Markov

Las cadenas de Markov y los modelos ocultos de Markov , son extensiones de los autómatas finitos. Un autómata finito se define como un conjunto de estados y un conjunto de transiciones entre estados, obtenidos ambos a partir de las observaciones. Un autómata de estados finitos ponderado no es más que un autómata donde cada transición – representada mediante un arco – tiene asociada una probabilidad, indicando cómo de probable es que se tome ese camino. La probabilidad de todos los arcos que salen de un estado deben sumar 1. Se dice que un autómata finito probabilístico constituye una gramática N-grama, siendo N el número de estados de que se compone el modelo.

Una cadena o proceso discreto de Markov es una secuencia de eventos, también llamados estados, donde la probabilidad de cada uno de ellos sólo depende del evento inmediatamente precedente. Es un caso especial de un autómata ponderado donde únicamente la secuencia de entrada determina que estados visitará el autómata.

C.2.1. Formalización de las cadenas de Markov

Se puede considerar una cadena de Markov como un tipo de modelo gráfico probabilístico; una manera de representar hipótesis probabilísticas en un grafo. Una cadena o proceso de Markov se especifica mediante los siguientes componentes:

- Espacio de estados del sistema: Formado por un conjunto de N estados distintos $S = \{S_1 S_2 ... S_n\}$. Cada estado corresponde a un evento observable.

- Probabilidad condicional de transición: Define la probabilidad de transición desde un estado i al estado j: $a_{ij} = P(q_{i+1} = S_j | q_i = S_i)$.
  Los instantes de tiempo asociados a cada cambio de estado se definen como $t = \{1, 2...\}$, identificando al estado actual en el instante t como q_t. Normalmente, el tiempo es el parámetro del sistema que mide la evolución del sistema.

  El conjunto de estas probabilidades de transición, para todos los estados del modelo, definen una matriz de probabilidades de transición A.

  La salida del proceso es la secuencia de eventos producidos en cada estado visitado.

Se puede tratar de predecir cuál va a ser el evento observado en un estado, basándose en los eventos u observaciones pasadas. Esto se puede formalizar como la búsqueda de la siguiente probabilidad condicional:

$$P(q_i | q_{i-1} q_{i-2} .... q_1) \tag{C.4}$$

Pero a mayor longitud de la secuencia, mayor número de observaciones deberán considerarse, lo que supone guardar estadísticos de todas ellas. Por tanto se suele realizar una simplificación, denominada la condición de Markov.

- **Condición de Markov:**
  La condición de Markov establece que para una secuencia de estados $Q = \{q_1 q_2 ... q_n\}$,

  $$P(q_i | q_{i-1} q_{i-2} .... q_1) = P(q_i | q_{i-1}) \tag{C.5}$$

  Se denomina condición de Markov de primer orden, ya que la probabilidad de un estado sólo va a depender del estado anterior. Es una probabilidad bigrama, porque sólo considera $N = 2$ estados. Si la condición de Markov es de segundo orden, la probabilidad de un estado o evento dependerá de los dos estados anteriores, $P(q_i | q_{i-1} q_{i-2})$. Es una probabilidad trigrama, ya que considera $N = 3$ estados. En general cuando se habla de la suposición o condición de Markov se refiere a la condición de Markov de primer orden. Un sistema para el cual la ecuación C.12 es cierta es un modelo de Markov (de primer orden) y la secuencia de salida q_i de dicho sistema es una cadena de Markov (de primer orden).

De manera que se puede expresar la probabilidad conjunta de una secuencia de eventos $Q = \{q_1 q_2 ... q_n\}$ aplicando la suposición de Markov.

$$P(q_1 q_2 ... q_n) = \prod_{i=1}^{n} P(q_i | q_{i-1}) \tag{C.6}$$

C.3. Introducción a los modelos ocultos de Markov

Los procesos de Markov son procesos estocásticos. Un proceso estocástico es un modelo matemático que describe el comportamiento de un sistema dinámico sometido a un fenómeno de naturaleza aleatoria. En concreto, los procesos de tipo Markov tienen gran importancia en el análisis y estudio de sistemas dinámicos. Las dinámicas del entorno pueden ser codificadas en base a los vínculos y parámetros existentes desde un instante de tiempo al siguiente. Constituye el modelo de evolución. Así mismo, cada instante de tiempo puede contener algunas observaciones, cuyos valores pueden ser observados en ese instante. Constituye el modelo de observación.

Todo proceso estocástico cumple con las siguientes características:

- Describe el comportamiento de un sistema que varía con respecto de un determinado parámetro (el cual es el objeto del modelo).
- Existe un fenómeno aleatorio que evoluciona según un parámetro t, normalmente es el tiempo.
- El sistema presenta estados definidos y observables, a los cuales se les puede asociar una variable aleatoria $x(t)$ que represente una característica medible de los mismos.
- El sistema cambia probabilísticamente de estado.
- Todo estado del sistema tiene una probabilidad de estado asociada S_i, la cual indica la probabilidad de estar en el estado i en el instante t.

Una cadena de Markov es útil cuando necesitamos calcular la probabilidad de una secuencia de eventos que pueden ser observados. Debido a que no pueden representar problemas ambiguos, una cadena de Markov sólo resulta útil para asignar probabilidades a secuencias no ambiguas o deterministas. Es por esto que las cadenas de Markov también reciben el nombre de modelos de Markov observados (*observed Markov model*).

Sin embargo, en muchas ocasiones, los eventos que nos interesan no pueden ser directamente observables. Por ejemplo, en el caso del reconocimiento del habla, a partir de lo eventos acústicos (las observaciones) en el mundo, se tienen que inferir las palabras que son las fuentes causales que las han generado, pero esas palabras no se pueden observar en el mundo. Se dice que las palabras son ocultas porque no se observan.

Un modelo de Markov oculto (HMM) permite hablar tanto de eventos observados como de eventos ocultos, considerándose estos últimos como factores causales del modelo probabilístico representado. Un HMM permite representar secuencias estocásticas como cadenas de Markov, donde los estados no pueden observarse directamente, pero tienen asociada una función de distribución de probabilidades. Dicho de otro modo, la secuencia de estados sólo

puede ser observada a través de los procesos estocásticos definidos en cada estado. La verdadera secuencia de estados está, por tanto, oculta por una primera capa de procesos estocásticos.

Un modelo oculto de Markov es un proceso doblemente estocástico que encierra un proceso estocástico fundamental que no es observable (es oculto), pero se puede descubrir a través de otro conjunto de procesos estocásticos que producen la secuencia de símbolos observados.

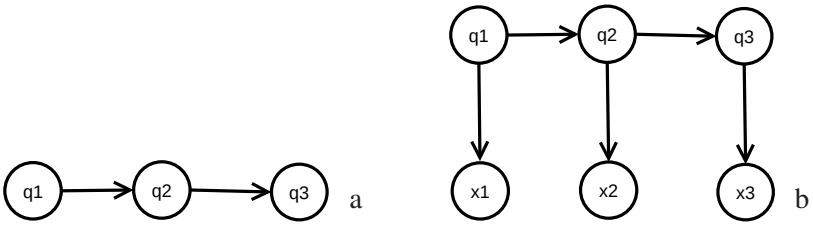


Figura C.1: Esquemas de (a) las cadenas de Markov, donde $q_1, q_2 y q_3$ son los parámetros directamente observables y (b) los modelos ocultos de Markov , donde $q_1, q_2 y q_3$ son los parámetros ocultos, mientras que $x_1, x_2 y x_3$ son los parámetros observables.

C.3.1. Formalización de los HMM

Conviene especificar la formulación de los HMM para ver en qué se diferencias de las cadenas de Markov.

Un HMM se especifica mediante:

- Espacio de estados del sistema: Consta de un conjunto de N estados distintos $S = \{S_1 S_2 ... S_N\}$. Cada estado corresponde a un evento no observable.
- Un conjunto de parámetros $\Theta = \{\pi, A, B\}$.
  - Probabilidad incondicional de estado: Son las probabilidades a priori $\pi_i = P(q_i = S_i)$, son las probabilidades de que S_i sea el primer estado de una secuencia de estados. Se almacenan en el vector de estado inicial π.
  - Probabilidad condicional de transición o modelo de evolución: Las probabilidades de transición representa la probabilidad de ir del estado i al estado j: $a_{ij} = P(q_{t+1} = S_j | q_t = S_i)$. Se almacenan en la matriz **A**.
  - Probabilidad condicional de emisión o modelo de observación: Las probabilidades de emisión caracterizan la verosimilitud de que se observe una cierta observación O, si el modelo está en el estado S_i. Constituyen un conjunto de verosimilitudes de las observaciones, que dependiendo del tipo de observaciones O pueden ser:
    - Si las observaciones son discretas, $O_t \epsilon \{v_1 ... v_k\}$, siendo v_i valores de un alfabeto de símbolos:
      $b_i(k) = P(O_t = v_k | q_t = S_i)$ representa la probabilidad de observar v_k si el estado actual es $q_t = S_i$. Los valores de $b_i(k)$ se pueden almacenar en una matriz **B**.

- Si las observaciones son continuas, $O_t \epsilon R^D$: Se emplea un conjunto de funciones $b_i(O_t) = P(O_t|q_t = S_i)$ que describen las funciones de densidad de probabilidad en el espacio de las observaciones, suponiendo que el sistema esté en el estado S_i. Se suelen almacenar en un vector $B(O)$ de funciones. Son habituales el empleo de Gaussianas o mezclas de Gaussianas para representar las probabilidades de densidad.

Resumiendo, un HMM es un modelo probabilístico temporal en donde cada estado del proceso se describe por una o más variables estocásticas. Por ser un conjunto de variables aleatorias, sus posibles valores dependen una probabilidad de distribución. Cada instante discreto de tiempo, el proceso asume estar en un estado y al recibir las observaciones el proceso cambia de estado según su matriz de probabilidades de transición.

En un HMM, se asume que todas las funciones de densidad de probabilidad que lo caracterizan son invariantes en el tiempo, lo que se conoce como un proceso estacionario.

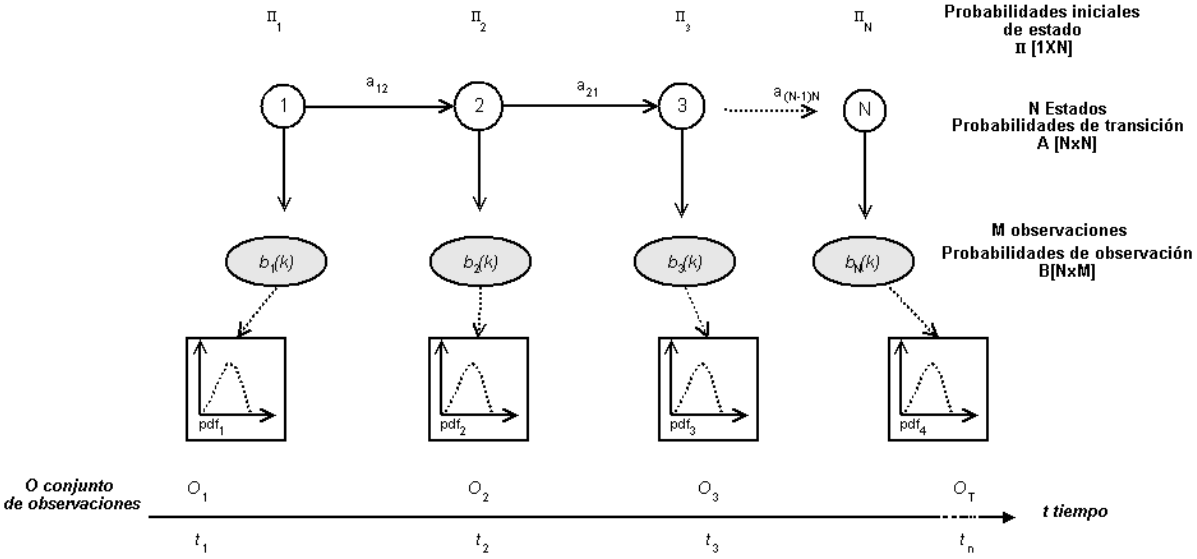


Figura C.2: Esquema general de los elementos que caracterizan completamente un HMM.

Antes de pasar a explicar cómo se ha realizado el reconocimiento de las secuencias de peatones, conviene hacer referencia a los tres problemas fundamentales de un HMM, ya que se ha tenido que resolver cada uno de ellos en el algoritmo desarrollado.

C.3.2. Los 3 problemas fundamentales de un HMM

Rabiner [Rab89] introdujo la idea de que los HMM deberían caracterizarse por tres problemas fundamentales:

Problema 1: Evaluación Dada una observación $O = \{O_1 O_2 ... O_T\}$ y un modelo $\Theta = \{\pi, A, B\}$, determinar la verosimilitud $P(O|\Theta)$ de la secuencia de observaciones dado el modelo. Es el Cálculo de la verosimilitud.

Problema 2: Descifrado Dada una observación $O = \{O_1 O_2 ... O_T\}$ y un modelo $\Theta = \{\pi, A, B\}$, descubrir la secuencia de estados correspondientes $Q = \{q_1 q_2 ... q_T\}$ que mejor explique las observaciones.

Problema 3: Aprendizaje Dada una observación $O = \{O_1 O_2 ... O_T\}$ y un conjunto $Q = \{q_1 q_2 ... q_T\}$ de estados, aprender los parámetros $\Theta = \{\pi, A, B\}$ del modelo que maximizan $P(O|\Theta)$.

C.4. Reconocimiento de patrones con modelos ocultos de Markov

El problema de seleccionar la secuencia de hipótesis más probable puede formalizarse como un modelo de Markov oculto (HMM). De hecho, un HMM puede considerarse un clasificador de secuencias probabilístico. Son modelos estadísticos basados en aprendizaje automático. Además de aplicarse en el campo del habla y del lenguaje, también se han aplicado al campo de la visión artificial para el reconocimiento de patrones. En concreto se han empleado para el reconocimiento y clasificación del comportamiento de personas [Bra97].

C.4.1. Razonamiento probabilístico. Inferencia Baseyiana en HMM

Los modelos ocultos de Markov constituyen un caso especial de inferencia Bayesiana. De manera que se puede formular el problema de la clasificación de una secuencia como la búsqueda del modelo de Markov que mejor clasifique esa secuencia. Reescribiendo la fórmula C.1 del maximum a posteriori se obtiene:

$$\hat{\Theta}_{MAP} = \underset{\Theta_i}{\operatorname{argmax}} \; P(\Theta|O) \tag{C.7}$$

representando Θ el conjunto de parámetros que definen el modelo dinámico.

El razonamiento probabilístico tiene lugar a través del cálculo de la probabilidad condicional del modelo dada una secuencia de observaciones O. Como ya se ha dicho antes, este cálculo no es sencillo, aplicándose Bayes para simplificar la tarea. Como resultado se obtiene la siguiente expresión:

$$\hat{\Theta}_{MAP} = \underset{\Theta_i}{\operatorname{argmax}} \; P(O|\Theta)P(\Theta) \tag{C.8}$$

El primer término corresponde a la verosimilitud de la secuencia de observaciones dado el modelo y el segundo, a la probabilidad a priori del modelo. Si la probabilidad a priori es uniforme, es decir, si todos los modelos que se van a evaluar tienen la misma probabilidad a priori, la ecuación C.8 se simplifica;

$$\hat{\Theta}_{MAP} = \underset{\Theta_i}{\operatorname{argmax}} \; P(O|\Theta) \tag{C.9}$$

El valor de Θ que maximiza $P(O|\Theta)$ se denomina valor de máxima verosimilitud (maximum likelihood). Por tanto se puede abordar la tarea de clasificación como el cálculo de la verosimilitud de una secuencia de observaciones para un modelo dado.

C.4.2. Verosimilitud de una secuencia dado un HMM

La verosimilitud de una secuencia de observaciones $O = \{O_1 O_2 ... O_T\}$ con respecto a un HMM con parámetros θ se expande del siguiente modo:

$$P(O|\Theta) = \sum_{Q_i} P(O, Q|\Theta) \tag{C.10}$$

Es la suma de la probabilidad condicional de la secuencia calculada sobre todas las posibles secuencias de estados generadas por el modelo. Y la probabilidad condicional dada una secuencia de observaciones O y una secuencia de estados o camino Q se calcula mediante;

$$P(O, Q|\Theta) = P(O|Q, \Theta)P(Q|\Theta) \tag{C.11}$$

Esta expresión no es más que la regla de Bayes (ver ecuación C.3), que como se ha comentado, se simplifica tomando dos suposiciones; la condición de Markov y la independencia de las observaciones.

- **Condición de Markov:**
  La condición de Markov establece que para una secuencia $Q = \{q_1 q_2 ... q_T\}$, la probabilidad de alcanzar un nuevo estado sólo depende del estado anterior. Se considera un HMM de primer orden.

$$P(q_i|q_{i-1}q_{i-2}....q_1) = P(q_i|q_{i-1}) \tag{C.12}$$

- **Independencia de las observaciones:** La segunda suposición considera que las observaciones son independientes con respecto al tiempo y que la probabilidad de una observación O_i sólo depende del estado q_i que la ha generado.

$$P(O_1 O_2 ... O_T|q_1 q_2 ... q_T) = \prod_{i=1}^{t} P(O_i|q_i) \tag{C.13}$$

Aplicando ambas hipótesis a la ecuación C.11 y recordando la probabilidad conjunta para una secuencia de estados C.6, se obtiene la siguiente expresión para la estimación de la secuencia más probable dado un modelo:

$$\hat{\Theta}_{MAP} = \underset{\Theta_i}{\operatorname{argmax}} \sum_{Q_i} P(O|\Theta) = \underset{\Theta_i}{\operatorname{argmax}} \prod_{i=1}^{t} P(O_i|q_i) \prod_{i=1}^{t} P(q_i|q_{i-1}) \tag{C.14}$$

Es una regla de decisión basada en la máxima verosimilitud, que es el criterio que se ha empleado en esta tesis para la clasificación de hipótesis. En este caso, se obtendría el mismo resultado aplicando una clasificación bayesiana, expresada en C.15. Esto es debido a que cada modelo Θ_i tiene la misma probabilidad de ocurrir, por lo que no es necesario calcular las

probabilidades a priori $P(\Theta_i|\Theta)$ de cada uno de ellos. En caso contrario, la regla de Bayes a aplicar sería:

$$P(\Theta_i|O,\Theta) = \frac{P(O|\Theta_i,\Theta)P(\Theta_i|\Theta)}{P(O|\Theta)} \propto P(O|\Theta_i,\Theta)P(\Theta_i|\Theta) \propto \prod_{i=1}^{t} P(O_i|q_i) \prod_{i=1}^{t} P(q_i|q_{i-1})P(\Theta_i|\Theta) \tag{C.15}$$

El cálculo de C.10 efectuado para todas las posibles secuencias de estados resulta inviable, más aún en el caso de grandes modelos o largas secuencias, por lo que tradicionalmente se ha venido usando el algoritmo *forward*.

C.4.3. Cálculo de la verosimilitud: El algoritmo *forward*

El algoritmo recursivo *forward* reduce la complejidad del problema. Este algoritmo define una variable *forward* o *"hacia delante"* $\alpha_t(i)$ que corresponde a;

$$\alpha_t(i) = P(O_1O_2...O_T, q_t = S_i|\Theta) \tag{C.16}$$

donde $\alpha_t(i)$ representa la probabilidad de haber observado la secuencia $O = \{O_1O_2...O_T\}$ y estar en el estado i en el instante t, dado el modelo Θ.

El algoritmo recursivo *forward*, calculado sobre un modelo de N estados, consta de los siguientes pasos:

1. Inicialización:
$$\alpha_1(i) = \pi_i.b_i(O_1), 1 \leq i \leq N \tag{C.17}$$
donde,π_i es la probabilidad a priori de estar en el estado S_i en el instante de tiempo $t = 1$ y $b_i(O_1)$, representa la probabilidad de generar el símbolo O_1 en dicho estado.

2. Recursión o Inducción:
$$\alpha_{1+1}(j) = [\sum_{i=1}^{N} \alpha_t(i).a_{ij}].b_j(O_{t+1}), 1 \leq t \leq T-11 \leq j \leq N \tag{C.18}$$

3. Finalización:
$$P(O|\Theta) = \sum_{i=1}^{N} \alpha_T(i) \tag{C.19}$$
En este último paso, es decir, cuando se ha alcanzado el final de la secuencia observada, se suman las probabilidades de todos los caminos que convergen en el estado final N.

Este procedimiento plantea una questión muy importante de implementación. De hecho, el vector α_t se obtiene a partir del cálculo de productos de un gran número de valores menores de 1. Por tanto, después de un número pequeño de observaciones $t \approx 10$, los valores de α_t se acercan exponencialmente a 0, y la precisión en coma flotante se excede (incluso en el caso de que se considere doble precisión). Existen dos posibles soluciones a este problema. La primera de ellas consiste en el escalado de los valores, que se deshace al final del procedimiento. La otra solución consiste en aplicar logaritmos al cálculo de las probabilidades y verosimilitudes, a saber, $\log P(O|\Theta)$ en lugar de $P(O|\Theta)$. Esta última ha sido la solución empleada en esta tesis.

Para más detalles del algoritmo *forward* se puede consultar Rabiner93.

C.5. La secuencia de estados óptima: El algoritmo de Viterbi

Para un modelo que contiene variables ocultas, como es el caso del HMM, la tarea de determinar qué secuencia de variables es la generadora de la secuencia de observaciones se denomina tarea de decodificación o inferencia.

En el ámbito del reconocimiento del habla y en algunas otras aplicaciones de reconocimiento de patrones, resulta útil asociar una secuencia de estados "óptima.[a] una secuencia de observaciones, siendo conocidos los parámetros que definen el modelo.

Un criterio razonable de optimalidad consiste en seleccionar la secuencia de estados o camino que proporciona una máxima verosimilitud con respecto a un modelo dado. Esta secuencia puede obtenerse de manera recursiva mediante el algoritmo de *Viterbi*. El término *Viterbi* es común en el ámbito del procesamiento del habla y del lenguaje, pero en realidad es una aplicación estándar de clásico algoritmo de programación dinámica.

Este algoritmo utiliza dos variables:

- La máxima verosimilitud $\delta_t(i)$ a lo largo de un único camino entre todos los posibles caminos que terminan en el estado i, en el instante de tiempo t.

$$\delta_t(i) = \max_{q_1 q_2 \ldots q_{t-1}} P(q_1 q_2 ... q_t = S_i, O_1 O_2 .. O_T | \Theta) \tag{C.20}$$

- Una variable $\psi_t(i)$ que permite mantener una memoria del *"mejor camino"* que termina en el estado i, en el instante de tiempo t.

$$\psi_t(i) = \operatorname*{argmax}_{q_1 q_2 \ldots q_{t-1}} P(q_1 q_2 ... q_t = S_i, O_1 O_2 .. O_T | \Theta) \tag{C.21}$$

  Ambas variables son vectores de N elementos, siendo N el número de estados generadores de observaciones.

Con la ayuda de estas dos variables, el algoritmo de Viterbi realiza los siguientes pasos:

1. Inicialización:

$$\delta_1(i) = \pi_i . b_i(O_1), 1 \leq i \leq N \psi_1(i) = 0 \tag{C.22}$$

   donde, al igual que antes, π_i es la probabilidad a priori de estar en el estado S_i en el instante de tiempo $t = 1$ y $b_i(O_1)$, representa la probabilidad de generar el símbolo O_1 en dicho estado.

2. Recursión o inducción:

$$\delta_t(j) = \max_{1 \leq i \leq N} [\delta_{t-1}(i) . a_{ij}] . b_j(O_n), 2 \leq t \leq T, 1 \leq j \leq N \psi_t(j)$$
$$= \operatorname*{argmax}_{1 \leq i \leq N} [\delta_{t-1}(i) . a_{ij}], 2 \leq t \leq T, 1 \leq j \leq N \tag{C.23}$$

   El camino óptimo se compone de subcaminos óptimos. Para ello se busca el camino de máxima verosimilitud, considerando el camino de mejor verosimilitud alcanzado

un paso antes y todas las transiciones a partir de él; el resultado se multiplica por la verosimilitud actual alcanzada en el estado actual. Por tanto el mejor camino se alcanza por inducción.

3. Finalización:

$$P^*(O|\Theta) = \max_{1\leq i\leq N}[\delta_T(i)]q_T^* = \max_{1\leq i\leq N}[\delta_T(i)] \tag{C.24}$$

La máxima verosimilitud se obtiene al alcanzar el final de la secuencia de observaciones.

4. Backtracking:

$$Q^* = \{q_1^*...q_T^*\}q_t^* = \psi_{t+1}(q_{t+1}^*)t = T-1, T-2..., 1. \tag{C.25}$$

La secuencia de estados se decodifica a partir de los vectores ψ_t.

Por tanto el algoritmo de Viterbi proporciona dos resultados de gran utilidad, dada una secuencia de observaciones $O = \{O_1...O_T\}$ y un modelo Θ:

- La selección del mejor camino $Q^* = \{q_1^*...q_T^*\}$ de entre todos los posibles caminos generados por el modelo, que se corresponden con la secuencia de estados que proporciona la máxima verosimilitud para la secuencia de observaviones O.
- La verosimilitud a los largo del mejor camino $P(O, Q^*|\Theta) = P^*(O|\Theta)$. Al contrario que el algoritmo *forward*, donde todos los posibles caminos son considerados, el algoritmo de Viterbi calcula la verosimilitud únicamente a lo largo del mejor camino.

Para más detalles del algoritmo de *Viterbi* se puede consultar Rabiner93.

C.6. El entrenamiento de los modelos ocultos de Markov : El algoritmo *Baum-Welch*

El tercero de los problemas de todo HMM tiene que ver con el aprendizaje de los parámetros que definen al modelo, $\Theta = \{\pi, A, B\}$. Formalmente, la solución consiste en estimar dichos parámetros, dada una secuencia de observaciones 0 y un conjunto de posibles estados. El aprendizaje puede ser supervisado o no-supervisado.

El algoritmo clásico para el entrenamiento es el algoritmo *Baum-Welch* o *forward-backward*, que es un caso particular del algoritmo de máxima expectación (Expectation-Maximization or EM algortihm). Este algoritmo permite entrenar tanto las probabilidades de transición A como las probabilidades de emisión B del modelo.

A partir de una secuencia de observaciones, se ve qué camino se sigue en el modelo en función de los estados que han generado cada uno de los símbolos observados. Se obtiene la estimación de máxima verosimilitud de la probabilidad a_{ij} para una transición concreta entre

los estados i y j, contando el número de veces que tuvo lugar la transición ($C(i \rightarrow j)$, y luego normalizando por el número total de veces que ocurre una transición desde el estado i:

$$a_{ij} = \frac{C(i \rightarrow j)}{\sum_{q \in Q} C(i \rightarrow q)} \tag{C.26}$$

En un HMM no se pueden realizar estos cálculos directamente a partir de las observaciones, ya que no se sabe qué camino (conjunto de estados) fue tomado por el modelo, dada una entrada. El algoritmo de *Baum-Welch* considera dos soluciones al problema. La primera consiste en realizar una estimación de C.26 iterativa. Se inicia con una estimación para las probabilidades de las transiciones y observaciones, y luego estas estimaciones se van refinando. La segunda idea obtiene esas estimaciones a través del cálculo de la probabilidad *forward* de una observación, y luego dividiendo el total de la probabilidad entre los diferentes caminos que han contribuido a la probabilidad *forward*.

Para comprender el algoritmo, es necesario definir la probabilidad *backward*.

La probabilidad *backward* β es la probabilidad de ver las observaciones desde el instante de tiempo $t+1$ hasta el final de la secuencia, suponiendo que se está en el estado j en el instante t y dado un modelo.

$$\beta_t(i) = P(O_{t+1}O_{t+2}...O_T|q_t = S_i, \Theta) \tag{C.27}$$

Se calcula mediante inducción de un modo similar al algoritmo *forward* .

1. Inicialización:

$$\beta_T(i) = 1, 1 \leq i \leq N \tag{C.28}$$

2. Recursión o Inducción:

$$\beta_t(i) = \sum_{i=1}^{N} a_{ij}.b_j(O_{t+1}).\beta_{t+1}(j), \&t = T-1, T-2..,11 \leq i \leq N \tag{C.29}$$

3. Finalización:

$$P(O|\Theta) = \alpha_t(N) = \beta_T(1) = \sum_{j=1}^{N} a_{1j}.b_j(O_1).\beta_1(j) \tag{C.30}$$

Ahora se puede comprender cómo las probabilidades *forward* y *"backward"* pueden ayudar al cálculo de las probabilidades de transición a_{ij} y las probabilidades de observación $B_i(O_t)$ a partir de una secuencia de observaciones, a pesar de que el camino tomado por la máquina es oculto.

Para reestimar a_{ij}, se estima $\hat{a}_{ij}$ como una variante de C.26:

$$\hat{a}_{ij} = \frac{\text{Número de transiciones esperadas desde el estado } i \text{ al estado } j}{\text{Número de transiciones esperadas desde el estado } i} = \frac{\sum_{t=1}^{T-1} \frac{\alpha_t(i).a_{ij}.b_j O_{t+1}.\beta_{t+1}(j)}{\alpha_T(N)}}{\sum_{t=1}^{T-1} \sum_{j=1}^{N} \frac{\alpha_t(i).a_{ij}.b_j O_{t+1}.\beta_{t+1}(j)}{\alpha_T(N)}} \tag{C.31}$$

De forma análoga, se recalcula la probabilidad de observación, que es la probabilidad de se observe un símbolo v_k de entre el vocabulario V, estando en el estado j:

$$\hat{b}_j(v_k) = \frac{\text{Número de veces esperadas de estar en el estado } j \text{ y observar el símbolo } v_k}{\text{Número de veces esperadas de estar en el estado } j}$$

$$= \frac{\sum_{t=1 s.t. O_t = v_k}^{T} \sum_{j=1}^{N} \frac{\alpha_t(i).a_{ij}.b_j O_{t+1}.\beta_{t+1}(j)}{\alpha_T(N)}}{\sum_{t=1}^{T} \sum_{j=1}^{N} \frac{\alpha_t(i).a_{ij}.b_j O_{t+1}.\beta_{t+1}(j)}{\alpha_T(N)}} \quad \text{(C.32)}$$

Suponiendo que se tenga una estimación previa de las probabilidades de transición A y de observación B, aplicando las fórmulas C.31 y C.32 se pueden reestimar dichas probabilidades a partir de una secuencia O de observaciones.

A pesar de que, en principio el algoritmo *"forward-backward"* puede realizar un aprendizaje no supervisado de los parámetros A, B y π, en la práctica las condiciones iniciales son muy importantes. Por este motivo, se suele fijar la estructura del HMM a mano y sólo se entrenan las probabilidades de emisión B y de transición A, a partir de la cadena de observaciones O.

Una vez expuestos los tres problemas fundamentales de un HMM así como el modo de hacerles frente, a continuación se expone el algoritmo implementado para el reconocimiento del modo de caminar humano.

Índice de figuras

Índice de tablas