UNIVERSITÀ DEGLI STUDI DI PARMA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Dottorato di Ricerca in Tecnologie dell'Informazione

XX Ciclo

Stefano Ghidoni

TECNICHE DI LOCALIZZAZIONE DI PEDONI E OSTACOLI IN AMBITO AUTOMOBILISTICO MEDIANTE VISIONE ARTIFICIALE

DISSERTAZIONE PRESENTATA PER IL CONSEGUIMENTO

DEL TITOLO DI DOTTORE DI RICERCA

GENNAIO 2008

Alla mia famiglia

Indice

Introduzione

L'ambiente in cui ogni essere umano vive è sorprendentemente ostile, tanto che, per poter sopravvivere, l'uomo ha bisogno di un apparato sensoriale estremamente completo e raffinato. Il senso maggiormente usato è la vista, perché è in grado di fornire una descrizione dettagliata del mondo circostante. Quando, in ambito robotico, è stato affrontato il problema di come far percepire l'ambiente alle macchine, la scelta di utilizzare sensori in grado di emulare la visione umana è stata perciò naturale.

La visione artificiale è la branca della robotica che si propone di ricavare informazioni a partire dalle immagini e dai flussi video. I dati provenienti da una telecamera sono i pixel che compongono l'immagine: si tratta, perciò, di un quantitativo davvero elevato di informazione di basso livello; in altre parole, i pixel sono pressoché inutilizzabili senza un'opportuna elaborazione, il cui scopo è quello di fornire in uscita pochi dati di alto livello. Questo compito può risultare davvero molto gravoso in termini di complessità computazionale.

La visione artificiale trova applicazione in numerosissimi campi, ed ha ormai raggiunto un certo livello di maturità, tanto da essere ampiamente utilizzata anche in ambito industriale, per il controllo qualità e la sorveglianza dei macchinari pericolosi, oltre alla videosorveglianza dei luoghi aperti al pubblico e la lettura targhe (sistemi *tutor* installati sulle autostrade italiane e varchi elettronici), per citare qualche esempio. Anche sul versante della ricerca, i sistemi di visione artificiale hanno raggiunto ormai un notevole livello di complessità; inoltre, contrariamente a quello che accadeva fino ad una decina di anni fa, oggi non è più necessario disporre di hardware

dedicato all'elaborazione di immagini, poiché la capacità di calcolo disponibile a bordo di un comune personal computer è sufficiente per l'esecuzione di algoritmi anche molto complessi in tempi ragionevoli.

Un ambito in cui la visione artificiale ha ricevuto un notevole impulso negli ultimi anni è quello automotive. Le industrie automobilistiche, infatti, sono particolarmente interessate a sistemi capaci di assistere il guidatore e di capire quando si sta verificando una situazione potenzialmente pericolosa, eventualmente intervenendo per mitigarne le conseguenze. Questo tipo di ricerca è attivamente sostenuto anche da importanti istituzioni civili, come la Comunità Europea, al fine di diminuire l'impressionante numero di decessi per incidenti stradali. La ricerca sta inoltre progredendo anche con lo sviluppo di sistemi di guida totalmente automatica, una prospettiva accarezzata dai ricercatori già negli anni '80, e ritornata attuale negli ultimi anni; a questo secondo ambito sono particolarmente interessate le organizzazioni militari, il cui obiettivo è quello di avere a disposizione flotte di veicoli autonomi da utilizzare negli scenari di guerra.

Sia che si tratti di assistenza al guidatore che di guida automatica, una delle funzioni più importanti che la visione artificiale può svolgere è il rilevamento di ostacoli; tra essi, quelli più vulnerabili sono senza dubbio i pedoni, e una particolare attenzione deve essere loro dedicata. Rilevare i pedoni è un compito particolarmente difficile da portare a termine, perché hanno una forma complessa che si modifica sia in funzione della postura che dei vestiti indossati; inoltre, alcune caratteristiche salienti, come la simmetria e lo sviluppo marcatamente verticale, sono comuni anche ad altri ostacoli, come i tronchi d'albero, i pali, le colonne, che costituiscono il maggior numero di falsi positivi dei sistemi di rilevamento pedoni.

In questa trattazione saranno presi in considerazione svariati sistemi di rilevamento pedoni (capitolo 2), e saranno analizzati e comparati i diversi sensori con cui tale compito viene portato a termine. Ci si concentrerà su alcune tecniche sviluppate per irrobustire il rilevamento, come l'analisi e la classificazione della sagoma, e il riconoscimento di caratteristiche peculiari come le gambe; saranno inoltre descritti alcuni moduli il cui scopo è quello di ampliare il range di funzionamento dei sistemi.

Nel capitolo 3 sarà preso in considerazione un algoritmo di calibrazione automatica, il cui scopo è quello di rendere affidabile un sistema di videosorveglianza per il rilevamento dei pedoni sulle strisce pedonali. Nel capitolo 4 si analizzerà il problema del rilevamento degli ostacoli in ambiente stradale, applicato ad un caso un po' particolare, ovvero al riconoscimento dei veicoli in fase di sorpasso.
Come appendice, è stata riportata un'esperienza interessante, strettamente collegata alla ricerca, ovvero l'equipaggiamento di un veicolo sperimentale, che si è rivelato molto utile per i test sugli algoritmi sviluppati.

Capitolo 1

Sensori per la visione artificiale

I sensori d'elezione per la visione artificiale sono, com'è ovvio, le telecamere. Attualmente, sul mercato se ne trovano di moltissime tipologie, e la scelta può essere a volte assai complicata, visto che bisogna tenere in considerazione molti fattori, che riguardano sia il sensore immagine vero e proprio, sia i circuiti elettronici che lo gestiscono. Inoltre, sempre più spesso i sistemi di visione artificiale fanno uso anche di altri sensori, per effettuare una fusione dei dati: ognuno, infatti, ha i propri punti di forza e di debolezza, e l'idea di creare sistemi che si basano su differenti tipologie di sensori si è dimostrata vincente.

Lo scopo di questo capitolo è di focalizzarsi sulle diverse tipologie di sensori, prendendo brevemente in esame i punti forti e deboli di ciascuno di essi, e analizzando, poi, quali sono le caratteristiche che un progettista deve considerare per effettuare una buona scelta. Questo capitolo non deve essere inteso come un'analisi delle attrezzature disponibili sul mercato, bensì come la sintesi dell'esperienza maturata sui sensori.

1.1 Telecamere

Una telecamera è un oggetto che contiene un elemento sensibile alla radiazione elettromagnetica, detto sensore, svariati circuiti elettronici che servono per controllarlo, e le interfacce verso l'esterno. Nella scelta di una telecamera, ciascuno di questi tre elementi deve essere accuratamente vagliato: per esempio, non è raro il caso in cui il sensore possiede delle caratteristiche particolari e interessanti, ma non è possibile utilizzarle, poiché non sono supportate dall'unità elettronica di controllo. In generale, comunque, la maggior attenzione deve essere posta nella scelta del sensore.

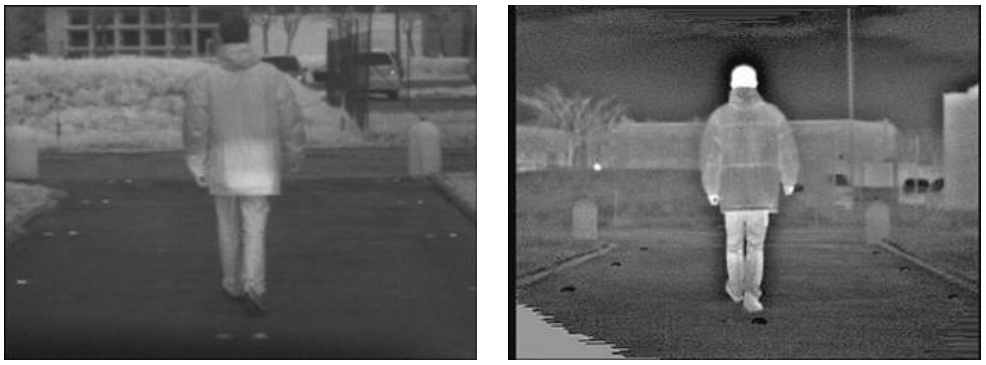
1.1.1 Range di frequenze

La prima scelta da fare riguarda la porzione dello spettro elettromagnetico che si desidera analizzare. Utilizzare immagini analoghe a quelle della visione umana, infatti, è solo una delle possibili scelte, che non necessariamente si rivela quella più efficiente. Oltre alle telecamere dette "nel visibile", ovvero sensibili alla radiazione elettromagnetica nelle stesse frequenze cui è sensibile l'occhio umano, ne esistono infatti altre che captano l'infrarosso e l'ultravioletto. Per quanto riguarda l'infrarosso, la banda che esso occupa è molto estesa, e va da 750 nm fino a 1000 µm, per cui le caratteristiche di questa radiazione variano molto a seconda che si abbia a che fare con frequenze vicine o lontane rispetto a quelle visibili. Per chiarezza, si distingue quindi tra il vicino infrarosso (NIR – Near InfraRed), nella banda 750 nm – 1400 nm, il medio infrarosso (MIR – Mid InfraRed) nel range $1,4\,\mu m - 15\,\mu m$, e il lontano infrarosso (FIR – Far InfraRed), che occupa lo spettro $15\,\mu m - 1000\,\mu m$. Si noti che questi range non sono definiti con precisione, e altre trattazioni potrebbero fornire una classificazione lievemente diversa.
Dal punto di vista della visione artificiale, è utile cogliere le differenze tra le varie immagini infrarosse: quelle NIR, infatti, sono abbastanza simili alle immagini a toni di grigio nel visibile; l'unica differenza apprezzabile è che i materiali, in quella banda, hanno un potere riflettente diverso, e quindi gli oggetti appaiono chiari o scuri diversamente da quanto accade nelle immagini nel visibile. Per esempio, l'erba

riflette moltissimo la radiazione NIR, e appare bianca (creando, a volte, la saturazione dell'immagine), mentre un vestito che all'occhio umano sembra colorato con un motivo può apparire in tinta unita. Le telecamere NIR, così come quelle nel visibile, captano la radiazione riflessa dagli oggetti, e quindi hanno bisogno di una sorgente che illumini la scena. Poiché il sole e quasi tutte le fonti luminose artificiali hanno emissione anche nell'infrarosso vicino, il controllo dell'illuminazione nel NIR è un problema molto simile a quello che si affronta quando si ha a che fare con sistemi che lavorano con la luce visibile: di giorno il sole illumina a sufficienza, mentre di notte è necessaria una fonte luminosa per poter vedere la scena inquadrata. Con l'infrarosso vicino, tuttavia, esiste un vantaggio: è infatti possibile realizzare illuminatori che emettono solo nella banda NIR, la cui radiazione, quindi, non è vista dall'occhio umano; in questo modo si riesce, per esempio, a costruire fari per auto molto luminosi ma che non abbagliano i guidatori dei veicoli che viaggiano nell'altro senso di marcia, oppure a sorvegliare una zona senza dare nell'occhio.

Le telecamere sensibili all'infrarosso lontano forniscono immagini completamente diverse. Si ricordi che ogni corpo emette una radiazione elettromagnetica la cui frequenza dipende dalla temperatura a cui si trova. Ebbene, i corpi alla temperatura ambiente emettono proprio nel range dell'infrarosso lontano: nelle immagini, quindi, ogni oggetto diventa sorgente luminosa, ed appare più chiaro o più scuro a seconda della temperatura a cui si trova, dalla quale dipende anche la quantità di energia che emette. Nelle immagini FIR, quindi, oggetti come i terminali di scarico delle auto e i radiatori, gli pneumatici, i fari, e anche le persone, appaiono più chiari dell'ambiente circostante, perché sono solitamente più caldi. Per questa ragione il FIR è anche detto infrarosso termico. In figura 1.1 sono mostrate due immagini della stessa scena, una NIR (a) e una FIR (b): le differenze tra i due domini sono evidenti.

Un altro range dello spettro elettromagnetico utilizzato in visione artificiale è quello dell'ultravioletto. Spesso accoppiato ad una specifica illuminazione, è sfruttato in ambito industriale e biomedicale; in figura 1.2 si può vedere un'immagine in cui alcune zone del viso, soggette a infezione, hanno caratteristiche di riflessione dei raggi ultravioletti diverse dalle zone sane. Le immagini ultraviolette, tuttavia, non sono



(a) (b)

Figura 1.1: Esempi di immagine nel dominio dell'infrarosso vicino (a) e lontano (b). Si noti che, nel secondo caso, gli oggetti caldi appaiono più chiari dell'ambiente circostante.

usate in ambito automobilistico, poiché non offrono nessun sostanziale vantaggio rispetto alle altre.

1.1.2 Caratteristiche del sensore immagine

Altre caratteristiche da tenere in considerazione per la scelta di un sensore sono:

dimensione e risoluzione: da essi dipendono la risoluzione delle immagini e la quantità di luce catturata;

guadagno e iris automatici: esprimono la capacità di adattarsi alternativamente a scene molto chiare o molto scure, sia modificando il guadagno elettronico del sensore, che aprendo o chiudendo il diaframma dell'ottica (iris);

dinamica ed eventuale HDR (High Dynamic Range): solitamente espressa in dB, misura la capacità di descrivere senza saturazioni sia zone molto illuminate che altre molto più scure nella stessa immagine;

shutter: in questo contesto non indica l'otturatore meccanico, come per le macchine fotografiche, bensì la lettura elettronica del sensore. Il tempo di shutter ha qui il

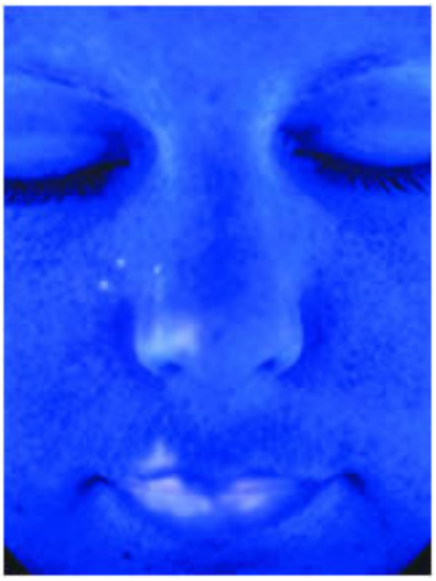


Figura 1.2: Esempio di immagine ultravioletta. Le zone più chiare indicano un'infezione in atto.

significato di tempo di integrazione della luce da parte degli elementi fotosensibili; il tipo di shutter indica l'ordine in cui essi sono letti: si parla di *rolling shutter* quando la lettura è sequenziale, riga per riga, mentre il *global shutter* fa sì che tutti gli elementi acquisiscano i dati nello stesso istante. Queste caratteristiche possono essere molto importanti se si desidera acquisire immagini di oggetti in movimento: dal tempo di shutter minimo, infatti, dipende la capacità della telecamera di "fermare" il movimento, ovvero di far apparire fermi gli oggetti che si stanno muovendo, evitando il cosiddetto effetto "mosso"; inoltre, il rolling shutter, più semplice a livello elettronico, in certe situazioni è causa di un effetto di distorsione dovuto al fatto che le righe dell'immagine non sono acquisite tutte nello stesso momento, come si vede in figura 1.3;

scansione: indica l'ordine di scansione del rolling shutter, e può essere progressiva o interallacciata: in quest'ultimo caso, l'immagine è divisa in due campi

Figura 1.3: Effetto generato dal rolling shutter: quando la telecamera è posta in movimento, l'acquisizione delle righe in istanti diversi provoca la tipica distorsione visibile nell'immagine.

(*field*), composti dalle linee pari e dispari, ed è scandita prima su un campo, poi sull'altro; questa tecnica fu inventata per migliorare la resa sugli schermi a tubo catodico, ma rende le immagini difficili da utilizzare nelle applicazioni di visione artificiale, che sono costrette a sfruttare uno solo dei due campi. La scansione progressiva, viceversa, non altera l'ordine delle righe;

framerate: è il numero massimo di immagini al secondo che il sensore è capace di acquisire e trasmettere;

bit per pixel: indica l'accuratezza della misurazione della luce di ogni elemento fotosensibile, e la conseguente necessità di un sufficiente numero di bit per esprimere tale precisione.

1.2 Sensori radar

I sensori radar (RAdio Detection And Ranging) sono elementi che emettono un'onda radio e che ricevono eventuali echi riflessi; misurando il tempo che intercorre tra

emissione e ricezione, sono in grado di valutare la distanza dell'oggetto che ha causato la riflessione. Sensori di questo tipo sono spesso usati per rafforzare l'affidabilità dei sistemi basati sulla visione. Ogni radar ha una sua circuiteria che provvede a trasformare i dati grezzi, cioè gli echi, nelle posizioni in cui si trovano gli ostacoli, cosicché l'output che è fornito al sistema è solitamente costituito da un insieme di punti che indicano dove si sono verificate le riflessioni più evidenti.
Il termine "radar" indica un concetto, più che un sensore specifico, e ci si rende conto di ciò osservando che, a seconda degli oggetti che si desidera rilevare, vi è una miriade di sensori radar, tutti basati sullo stesso principio, ma completamente diversi tra loro: alcuni sono grandi come palazzi, altri possono stare su un camion, e altri ancora sono talmente piccoli da poter essere tenuti in una mano. In questa sede si prendono in esame quelli di dimensioni minori, poiché sono più indicati per rilevare oggetti delle dimensioni di un'auto o un pedone, nonché gli unici che è possibile utilizzare come sensori per sistemi automotive.

1.2.1 Caratteristiche di un sensore radar

Le caratteristiche fondamentali di un sensore radar sono:

frequenza di funzionamento: indica la frequenza dell'onda radio utilizzata per il rilevamento; da essa dipende anche la finezza con cui è possibile rilevare gli ostacoli. Solitamente, i radar per applicazioni automotive hanno frequenze che vanno da 24 GHz a 80 GHz;

apertura: misura quanto è collimato il lobo di emissione dell'antenna alla frequenza di funzionamento; agendo su questa caratteristica è possibile rendere il radar più adatto al rilevamento di oggetti lontani che si trovano di fronte al sensore, con un'apertura piccola, oppure alla ricerca degli oggetti vicini che si possono trovare anche in posizioni laterali.

Nella realizzazione di un radar esistono, in realtà, anche molti altri dettagli di fondamentale importanza: un'eventuale modulazione; il numero di antenne, che possono

formare anche una piccola schiera; il tracking; il principio di emissione, che può essere continua (continuous wave – CW), oppure a impulsi. Tali dettagli sono tuttavia trasparenti a chi non si occupa di progettazione dei radar, ma si limita ad utilizzare il loro output.

1.2.2 Utilizzo del radar per il rilevamento dei pedoni

Nel corso di alcuni test con radar specificamente pensati per il rilevamento dei pedoni, ci si è accorti che il normale range di funzionamento dichiarato nelle specifiche è in realtà basato in buona misura sui risultati del tracking; questo fa sì che ci sia un certo ritardo nel primo rilevamento. Detto in altri termini, se un radar ha un range di funzionamento di 40 m, questo non significa che un ostacolo viene rilevato non appena è più vicino di tale distanza; così, se il radar è installato su un veicolo che si muove verso l'ostacolo, è possibile che la prima segnalazione si verifichi quando questo si trova a 15 o 20 m. Nel caso dei pedoni, le prestazioni sono peggiori rispetto a quanto accade per i veicoli e gli oggetti metallici, perché il corpo umano è un riflettore debole, cioè riflette una piccola porzione dell'onda radio che incide su di esso. Questo limita molto l'utilità di questo sensore, perché proprio nel caso più frequente e interessante si dimostra poco efficace, e l'alternativa di utilizzare radar con range di funzionamento maggiori è inapplicabile, perché, in termini pratici, significa ridurre l'apertura, rendendo quindi il radar ancora meno adatto a rilevare i pedoni ai bordi della strada.

I produttori di sensori radar stanno tuttora lavorando nel tentativo di migliorare le prestazioni con riflettori deboli, ma, in generale, si può concludere che questo tipo di sensore non è il più adatto al rilevamento dei pedoni, mentre risulta estremamente efficace per altre applicazioni, come la localizzazione dei veicoli. Ne è prova il fatto che alcune case automobilistiche hanno già commercializzato, per le loro vetture di punta, sistemi di rilevamento e di inseguimento del veicolo che precede basati sul radar.

1.3 Sensori laser

I sensori laser possono essere considerati un'evoluzione dei radar, perché sono basati sullo stesso principio dell'emissione e ascolto di echi, che avviene però utilizzando onde elettromagnetiche vicine al range della luce visibile, invece che onde radio: per questo motivo, questi sensori sono anche detti LIDAR (LIght Detection And Ranging). Il fatto di utilizzare frequenze molto più elevate e un raggio estremamente collimato offre il vantaggio di una precisione di rilevamento molto superiore rispetto al radar; in altre parole, è come avere un radar con un'apertura bassissima, larga quanto un raggio laser. Proprio questo fatto sembra costituire anche il limite dei lidar; tuttavia, l'idea di porre in rotazione il raggio ha segnato la svolta, generando i cosiddetti laserscanner: sensori che emettono un raggio laser rotante e ascoltano gli echi riflessi dagli oggetti che intersecano il piano interessato dalla misurazione (figura 1.4). Per ragioni pratiche, il laser è tenuto fisso, mentre il raggio è posto in movimento mediante uno specchio rotante che lo deflette.
Il laserscanner possiede, come il radar, un'elevata precisione nella valutazione delle distanze; inoltre offre una notevole accuratezza nella misurazione della direzione in cui si trova l'ostacolo, e permette di ricavarne la sagoma. Così, nel caso di rilevamento di un veicolo, se un radar fornisce solamente un punto quasi casuale della carrozzeria, il laserscanner fornirà tutti i punti in cui il piano di scansione intercetta il veicolo.

1.3.1 Caratteristiche di un sensore laserscanner

Le caratteristiche principali di un laserscanner sono:

frequenza di rotazione: misura la velocità di rotazione dello specchio che deflette il raggio, e, quindi, il numero di scansioni nell'unità di tempo;

risoluzione angolare: indica l'angolo posto tra due misurazioni consecutive durante la scansione;

(a) (b) (c)

Figura 1.4: Alcuni laserscanner: (a) il SICK LMS291, sensore da interno con un'apertura massima di 180°; (b) l'Alasca XT di Ibeo, a quattro piani di scansione, e (c) il Velodyne, avente ben 64 piani.

distanza massima di rilevamento: caratterizza la massima distanza a cui è possibile rilevare un ostacolo;

interallacciamento: è la possibilità di infittire le misurazioni facendo più scansioni con angoli iniziali diversi.

Sebbene il laserscanner offra notevoli vantaggi sul radar, esso rimane limitato per quanto riguarda l'apertura verticale, che è molto piccola: solo gli ostacoli che intersecano il piano di scansione possono essere rilevati. Per migliorare questo aspetto, alcuni produttori hanno ideato laserscanner aventi più piani di scansione: è il caso del sensore Alasca XT di Ibeo GmbH, visibile in figura 1.4.b, dotato di quattro piani, e del Velodyne (figura 1.4.c), che ne possiede ben 64.

1.3.2 Utilizzo del laserscanner per il rilevamento pedoni

Affrontando il problema del rilevamento dei pedoni, il laserscanner si dimostra decisamente più efficace del radar. Solitamente, questo sensore è montato a meno di un metro dal suolo, e il piano di scansione è quasi orizzontale: le parti del corpo che riflettono sono perciò le gambe; il pedone può quindi apparire come l'insieme di due ostacoli che si muovono alternativamente. Un particolare importante è la distanza

massima a cui un pedone può essere individuato, ben diversa dalla distanza massima di rilevamento: un laserscanner che segnala ostacoli fino a 80 m ben difficilmente localizzerà pedoni a tale distanza. Questo fenomeno è dovuto principalmente al fatto che il laserscanner non esegue una misurazione continua, mano a mano che lo specchio ruota, bensì campiona lo spazio a intervalli angolari costanti: poiché i raggi laser che effettuano le misurazioni divergono, la distanza tra di essi aumenta mano a mano che si allontanano dal sensore. Ora, si supponga che un laserscanner abbia una risoluzione di $0,25^\circ$; lo spazio d tra due misurazioni successive, entrambe alla distanza r dal sensore, e separate dall'angolo α, è data da:

$$d = 2r\sin\left(\frac{\alpha}{2}\right) \quad , \tag{1.1}$$

quindi, alla distanza di 1 m, due punti successivi distano tra loro 4,36 mm, a 10 m 4,36 cm, mentre a 50 m distano 21,82 cm. Con questi dati si può comprendere come sia difficile rilevare i pedoni a 50 m, anche se il sensore arriva a distanze anche maggiori. In questo discorso si è volutamente trascurato un effetto poco marcato, ma comunque presente, ovvero il lieve allargamento del raggio luminoso mano a mano che si allontana dal laser. Esistono anche altri fattori che ostacolano il rilevamento dei pedoni, come la bassa riflettività del corpo umano e degli abiti, ma la loro incidenza è minima; in generale, si è visto che utilizzando sensori con una risoluzione di $0,25^\circ$ è possibile rilevare pedoni con una certa sicurezza fino alla distanza di 30 m.

Capitolo 2

Tecniche di localizzazione e riconoscimento dei pedoni

La localizzazione dei pedoni è un argomento chiave della visione artificiale, sia per la sua complessità, che per l'elevato numero di applicazioni che esso ha. I pedoni sono infatti gli utenti della strada più deboli, e gli incidenti che li coinvolgono portano spesso a danni di notevole entità, proprio a causa della loro vulnerabilità; tali danni hanno dei costi sanitari, sociali e affettivi notevoli, ed è quindi interesse delle istituzioni cercare di ridurli il più possibile, sia intervenendo sulle infrastrutture stradali, sia investendo sui sistemi di sicurezza dei veicoli. In tal senso, la visione artificiale ha conseguito, negli ultimi anni, notevoli progressi, tanto che oggi si possono fare previsioni a medio-breve termine sull'adozione di sistemi di sicurezza basati sull'analisi di immagini; ne è prova l'interesse di tutte le case automobilistiche in queste tecnologie.

Lo stesso problema è affrontato anche al di fuori dell'ambito automobilistico: vi sono sistemi di videosorveglianza, per esempio, volti al riconoscimento di persone in luoghi in cui è vietato l'accesso, come le zone monumentali, in prossimità delle opere d'arte, o anche semplicemente per la sorveglianza degli edifici. In ambito industriale, si è pensato ad algoritmi di visione artificiale per riconoscere quando i lavoratori

si avvicinano troppo a macchinari pericolosi, per poter far scattare le procedure di sicurezza del caso.

Quello automobilistico rimane, comunque, uno degli scenari più complicati, perché l'ambiente è completamente destrutturato, e non è quindi possibile fare molte ipotesi su di esso: non si conosce la forma dell'ambiente, né il tipo di oggetti che sono presenti; inoltre, quando un flusso video proviene da una telecamera montata su un veicolo che si muove, ogni oggetto appare in movimento, e non è quindi immediato nemmeno riconoscere quelli fermi.

2.1 Il problema del riconoscimento dei pedoni

Poiché il riconoscimento dei pedoni è stato ed è tuttora così tanto studiato, esiste una fiorente letteratura al riguardo. I sistemi per la localizzazione dei pedoni trovano posto in quasi tutte le conferenze del settore, talvolta con specifiche sessioni; si veda [1] per una panoramica generale.

Gli algoritmi per il riconoscimento dei pedoni si fondano su molti approcci diversi, molti dei quali sono basati sull'analisi della forma o di alcune caratteristiche dell'immagine, oppure su classificatori AdaBoost, [2, 3, 4, 5]; altri, invece, osservano caratteristiche più specifiche, per esempio la camminata [6, 7, 8]. Il maggior elemento distintivo tra i vari sistemi riguarda la scelta del tipo di telecamere, in particolare il range di frequenze cui esse sono sensibili (si veda il paragrafo 1.1.1), e il loro numero; da questi fattori, infatti, dipende la scelta dell'approccio da utilizzare. I sistemi basati sulla visione stereoscopica, come [9, 10, 11, 12, 13, 14, 15], possono rilevare, per prima cosa, gli ostacoli e la loro distanza dalla telecamera, sfruttando la disparità, ovvero la differenza di posizione in cui si trova lo stesso oggetto nelle due immagini; gli approcci basati sulla visione monoculare, come [16, 17, 18, 19], viceversa, non possono sfruttare questo fenomeno, e fanno quindi ricorso a tecniche di pattern recognition più complicate, come il template matching [20] o la ricerca di caratteristiche [21, 22].

2.2 I sistemi preesistenti

Data la quantità di ricerca che è già stata svolta su questo argomento, una delle scelte che è possibile fare è quella di concentrarsi su un sistema preesistente, e cercare di migliorarne le prestazioni. Questo è possibile sia aggiungendo delle funzioni nuove, non previste dal sistema originale, come l'introduzione del tracking su un sistema che ne è privo, sia cercando di migliorare gli algoritmi esistenti, aumentandone le prestazioni.
Un notevole numero di sistemi per la localizzazione dei pedoni si basa su due passi fondamentali: l'individuazione di regioni di interesse (ROI – Region Of Interest), e la loro validazione, ovvero la verifica che tali regioni effettivamente contengano un pedone. La prima fase lavora sull'intera immagine, ed ha il compito di selezionare le aree in cui è possibile che sia presente una persona; in questo caso è importante che nessuna zona contenente un pedone sia trascurata, mentre non costituisce un problema se anche altre aree sono selezionate. La seconda fase, invece, lavora solo sulle regioni di interesse individuate, quasi sempre di forma rettangolare, chiamate anche *bounding box*; per ognuna di esse, in questo passo bisogna verificare se è presente un pedone o meno, ed è quindi necessaria un'accuratezza maggiore nell'eliminazione delle zone che non contengono un pedone. Solitamente, questa seconda fase è più complessa, sia dal punto di vista algoritmico che computazionale, e su di essa si concentra gran parte del lavoro qui presentato.

2.2.1 Sistema monoculare

Gran parte del lavoro sul riconoscimento dei pedoni ha avuto come obiettivo il miglioramento del sistema descritto in [23]. Si tratta di un sistema basato su una sola telecamera FIR a bassa risoluzione (320×240 pixel), integrata nella mascherina anteriore del veicolo sperimentale visibile in figura 2.1. Il nucleo fondamentale dell'algoritmo è costituito dall'analisi dei bordi verticali e della loro distribuzione e densità nell'immagine, fase durante la quale sono ricavate delle regioni di interesse rettangolari, che sono poi classificate dai passi successivi del sistema, per capire se si tratta di

Figura 2.1: Il veicolo sperimentale equipaggiato con una telecamera FIR integrata nella mascherina.

un pedone o meno. L'analisi dei bordi verticali è particolarmente efficiente nel caso di immagini nel dominio FIR, visto che i pedoni, essendo generalmente più caldi dell'ambiente circostante, lasciano un pattern molto evidente, e i bordi che ne derivano sono perciò molto marcati; l'idea di considerare solamente i bordi verticali è dovuta allo sviluppo principalmente verticale della forma dei pedoni.

Questo sistema adotta un approccio detto a multirisoluzione: poiché le prestazioni dell'algoritmo erano migliori per i pedoni lontani (fino a circa 40 m) che per quelli vicini (dai 6 ai 20 m), si è deciso di sottocampionare l'immagine per trovare i pedoni vicini, continuando, viceversa, a lavorare sull'immagine a piena risoluzione per la ricerca di quelli lontani; questo fatto ha anche portato dei vantaggi in termini di tempi di esecuzione, poiché parte dell'algoritmo lavora sull'immagine sottocampionata. Questo fenomeno è dovuto al fatto che i pedoni più vicini, paradossalmente, lasciano

una traccia troppo precisa sull'immagine, il che fa sì che l'algoritmo si concentri sui dettagli, e non sulla forma globale.

2.2.2 Sistema a doppio stereo

Un'altra parte del lavoro si è concentrata su un sistema diverso dal precedente, descritto in dettaglio in [24]. Si tratta di un sistema dotato di due coppie di telecamere, nel visibile e nell'infrarosso lontano, per il riconoscimento sia di ostacoli che di pedoni. Lo schema generale si trova in figura 2.2: si può vedere come l'elaborazione, in una prima fase, a causa delle differenze tra i domini visibile e FIR, sia separata per i due sottosistemi stereo; alla fine di questo primo passo, ognuno di essi individua delle regioni di interesse entro le quali si trovano gli ostacoli e, potenzialmente, i pedoni. In seguito, i risultati sono fusi assieme, e comincia l'elaborazione dedicata alla localizzazione dei pedoni, con la verifica della simmetria degli ostacoli trovati; successivamente agisce una serie di classificatori indipendenti, ciascuno dei quali fornisce un voto per ogni regione di interesse, nel range $[0; 1]$, che esprime la probabilità che essa contenga un pedone. Infine, un decisore effettua la classificazione finale, ovvero distingue tra pedoni e non pedoni, pesando opportunamente tutte le decisioni soft espresse dai classificatori. Per questo sistema è stato sviluppato uno dei classificatori; esso è basato sull'analisi di una sola delle due immagini FIR, quindi il modulo è intrinsecamente mono, anche se lavora su delle regioni di interesse provenienti da un'elaborazione stereo.

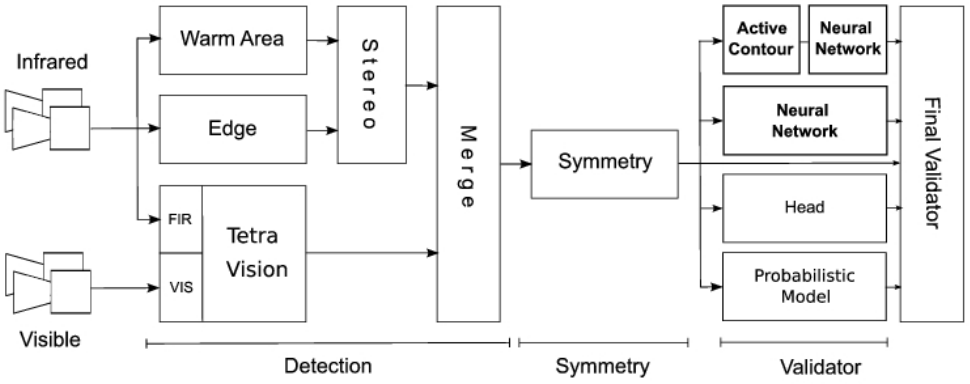


Figura 2.2: Schema del sistema di visione a doppio stereo.

2.3 Analisi e classificazione della sagoma

Gli ambiti su cui ci si è concentrati, nel tentativo di migliorare il riconoscimento dei pedoni, sono molteplici. Parte del lavoro ha riguardato l'analisi e la classificazione della sagoma intera o di alcune parti del corpo. Questa scelta è dovuta al fatto che la sagoma è una delle caratteristiche peculiari del corpo umano, ed analizzarla può quindi risultare utile per capire quali tra gli oggetti presenti nel flusso video sono pedoni, e quali no.
Per questo tipo di analisi sono state utilizzate due tecniche: il confronto tra i bordi presenti nell'immagine (trovati con l'algoritmo di Sobel) e un insieme di modelli, e l'estrazione della sagoma mediante contorni attivi, e successiva classificazione.

2.3.1 Validazione con modelli

Una delle idee più semplici per la validazione consiste nel confronto tra i bounding box trovati e un modello. Nel caso dei pedoni, questo modello risulta particolarmente complesso, a causa dell'elevato numero di forme e pose diverse in cui una persona può essere vista. Nonostante ciò, questa tecnica è ugualmente adottata da un buon numero di sistemi. Uno dei più famosi ed interessanti è quello descritto in [25, 26, 27]

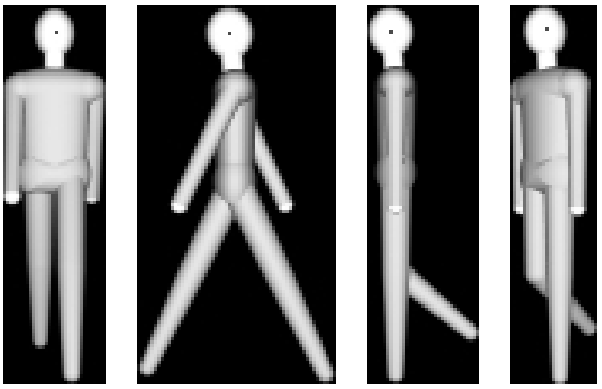
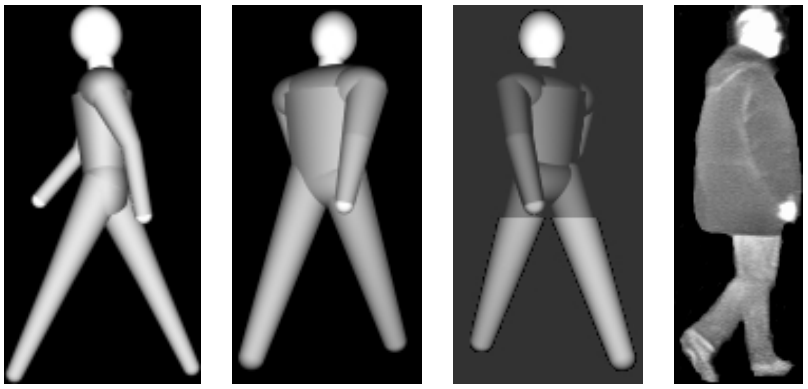


Figura 2.3: Alcuni esempi del primo insieme di modelli utilizzati per la validazione. Le forme risultano alquanto stilizzate, e si può notare l'assenza di dettagli come mani e piedi. Il punto all'interno della testa ne indica il centro.

e successive evoluzioni, basato su una gerarchia di modelli organizzata in modo tale da rendere più veloce la ricerca, raggruppando quelli simili.
Per questo motivo, in [23] è stato proposto un insieme di 72 modelli, ottenuto considerando otto diversi punti di vista, per ciascuno dei quali sono previste una posa da fermo e altre otto in fase di camminata, tutti con sfondo nero. I modelli, creati utilizzando il software PovRay 2.0, sono abbastanza stilizzati: come si vede in figura 2.3, mancano dei dettagli come le mani e i piedi, e le gambe hanno delle posizioni innaturali; inoltre, non tengono conto delle diverse emissioni di calore da parte del corpo: nelle immagini FIR, il busto è spesso più scuro perché maggiormente schermato dai vestiti. Nonostante ciò, si è ritenuto che le informazioni fondamentali fossero presenti, decidendo quindi di utilizzarli ugualmente.
Successivamente, per aumentare un po' le prestazioni, sono stati creati altri insiemi di modelli[1], rendendo maggiormente realistico il movimento, modificando le dimensioni del corpo e il tono di grigio delle varie parti del corpo, come si vede in figura 2.4.a-b-c; sono anche stati presi in considerazione 46 modelli ottenuti da immagini reali, con lo sfondo nero (d), come discusso in [28]. Si è preso in considerazione

[1] Questo studio è oggetto della tesi di laurea di Denis Nani.



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

Figura 2.4: Alcuni modelli della seconda generazione, aventi una camminata più naturale (a), un corpo più grande (b) e diversi toni di grigio per le varie parti del corpo (c); infine, modelli presi da immagini reali (d).

anche il cosiddetto *model tracking*, ovvero l'utilizzo di modelli ricavati dai bounding box validati nell'immagine precedente: l'idea alla base di questa tecnica è che in questo modo si utilizza, per ogni pedone, un modello che in realtà è lui stesso in una posizione lievemente diversa. Questa tecnica fornisce modelli in grado di fornire un match molto buono, ma ha subito mostrato anche il suo maggior limite: nel caso di un falso positivo, esso entra nell'insieme di modelli, e quindi provoca la validazione del bounding box errato anche nelle immagini successive, aumentando notevolmente il numero di falsi positivi.

2.3.1.1 Generazione dei nuovi modelli

Nel tentativo di ottenere risultati migliori, si è deciso di generare un ulteriore insieme di modelli, analizzando più nel dettaglio come appaiono i pedoni nelle immagini FIR, e tentando di riprodurre queste caratteristiche nelle immagini generate. In particolare, la testa e le mani sono i dettagli quasi sempre più caldi di un pedone, seguite dai piedi; il busto, viceversa, è maggiormente schermato, ed appare più scuro.
Sulla base di queste osservazioni si è creato il nuovo insieme di modelli, utilizzando

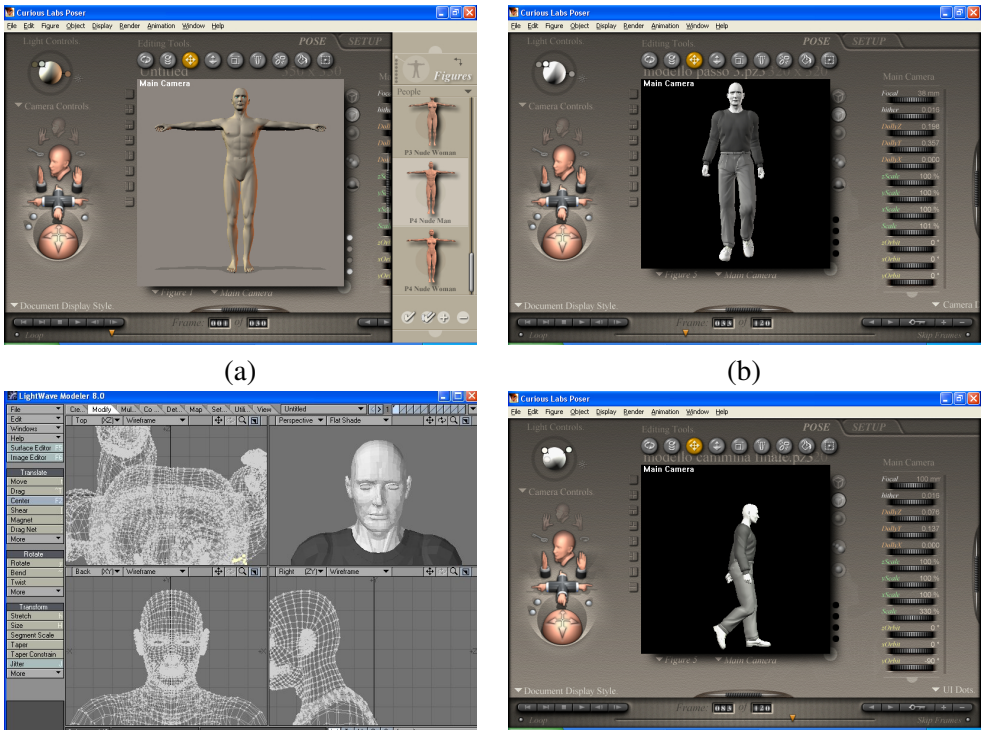


Figura 2.5: Le varie fasi della creazione dei nuovi modelli: scelta della corporatura (a), dei vestiti (b), rimozione dei dettagli del volto (c), generazione della camminata (d).

il software Poser 6.0 Pro. Le fasi della progettazione sono documentate in figura 2.5: dopo la scelta dell'altezza e della corporatura della figura umana (a) si è proceduto alla sua vestizione con "abiti FIR", ovvero optando per vestiti capaci di conferire all'uomo un aspetto simile a quello delle persone nelle immagini infrarosse (b). Il modello è quindi stato modificato con LightWave 3D 8.0 per togliere numerosi dettagli, specialmente sul volto, che non sono visibili nelle immagini FIR (c), e successivamente reimportato in Poser per generare la camminata (d). In questo caso si è scelto di mantenere gli otto punti di vista già utilizzati in precedenza, aumentando però il

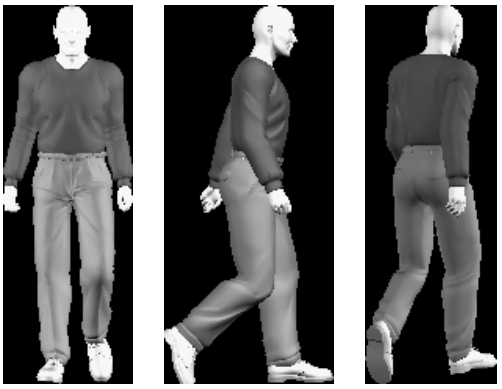


Figura 2.6: Alcuni esempi dei nuovi modelli generati.

numero di modelli in cui si discretizza la camminata, che diventano 12, più la posa statica; nel complesso, si tratta di 13 immagini per ogni punto di vista, per un totale di 104 modelli; in figura 2.6 si possono vedere alcuni esempi.
Per la generazione dei modelli è stato necessario anche impostare i parametri della telecamera virtuale di Poser: essa ha una lunghezza focale di 100 mm, e riprende il pedone da una distanza di 10 m; è stata anche scelta un'illuminazione in grado di rendere il modello il più simile possibile a come appaiono i pedoni nello spettro FIR.
Le immagini così ottenute sono poi state salvate in un formato a colori (il ppm), che contiene nel primo layer l'immagine a toni di grigio generata con Poser e nel secondo un'immagine binarizzata in cui ogni pixel è nero nei punti in cui si trova lo sfondo, e bianco in quelli occupati dal modello; il terzo layer è inutilizzato.
Visti i risultati incoraggianti con i nuovi modelli, sono stati creati altri tre set, modificando i vestiti e la capigliatura, come si vede in figura 2.7; l'ultimo dei tre, inoltre, è numericamente molto superiore agli altri, perché è stato creato utilizzando 16 angolazioni diverse, per ciascuna delle quali sono stati selezionati 40 fotogrammi per la camminata, più quello in cui il pedone è fermo, per un totale di 656 immagini: è stata fatta questa scelta per capire i margini di miglioramento delle prestazioni con un numero elevatissimo di modelli.
Le prestazioni della fase di validazione con i nuovi modelli saranno discusse in

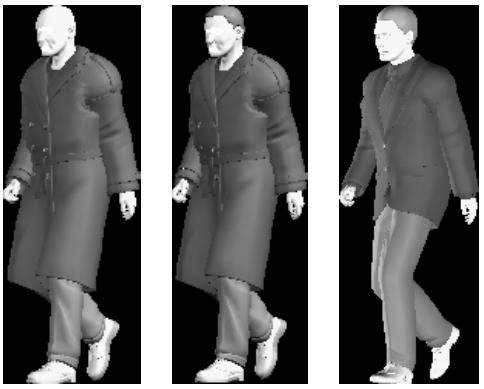


Figura 2.7: Ulteriori modelli generati modificando lievemente l'abbigliamento e la capigliatura.

un paragrafo a parte, perché si intrecciano con lo studio di una nuova funzione di correlazione, descritta di seguito.

2.3.1.2 Funzioni di correlazione

Dopo aver generato i nuovi modelli dei pedoni, è stata presa in considerazione la funzione utilizzata per effettuare i confronti. Quella utilizzata nel sistema preesistente, chiamata cross-correlazione a toni di grigio (GC – Grayscale Correlation), è calcolata come:

$$\mathrm{GC}(I,P) = \frac{\sum_{x=0}^{M}\sum_{y=0}^{N}(P(x,y)-m_{\mathrm{P}})(I(x,y)-m_{\mathrm{I}})}{\sqrt{\sum_{x=0}^{M}\sum_{y=0}^{N}(P(x,y)-m_{\mathrm{P}})^2\sum_{x=0}^{M}\sum_{y=0}^{N}(I(x,y)-m_{\mathrm{I}})^2}}\ , \tag{2.1}$$

in cui $I(x,y)$ e $P(x,y)$ rappresentano il tono di grigio del pixel di coordinate (x,y) dell'immagine e del pattern di confronto, mentre m_{I} e m_{P} sono i valori medi dei toni di grigio, sempre riferiti all'immagine e al pattern.

Osservando la formula si può notare che essa confronta le immagini pixel per pixel; tuttavia, è la forma dei pedoni, più che il loro contenuto, a dare il maggior contributo

informativo, e la cross-correlazione fa un'analisi sui contorni solo indirettamente. Nel tentativo di migliorare i risultati della validazione, sono state considerate alcune funzioni capaci di confrontare le forme dei pedoni, anziché le parti di immagine che li contengono. I contorni da utilizzare sono i bordi calcolati con l'algoritmo di Sobel, necessari per l'esecuzione dei passi iniziali dell'algoritmo di localizzazione dei pedoni, e perciò già disponibili, e i contorni dei modelli, ricavabili dal secondo layer di quelli nuovi.

Per il confronto delle sagome è stato preso in esame un approccio basato sulle *Edge Potential Function* (EP). Queste funzioni emulano quelle che descrivono un campo elettrostatico generato da un insieme di cariche; l'idea di base è di fare in modo che i pixel che compongono uno dei due contorni siano come le cariche che generano il campo, mentre l'altra sagoma è composta dalle cariche di prova, che sentono il campo esistente: tanto maggiore esso è, tanto superiore sarà il valore di match. Questo approccio è stato presentato in [29], in cui si definisce la *Binary Edge Potential Function* (BEPF) come:

$$\mathrm{BEPF}(x,y) = \frac{1}{4\pi\varepsilon_{\mathrm{eq}}} \sum_i \frac{1}{\sqrt{(x-x_i)^2+(y-y_i)^2}} \quad , \tag{2.2}$$

in cui $\varepsilon_{\mathrm{eq}}$, analogo della costante dielettrica, permette di scegliere quanto è estesa la funzione di potenziale. Tale funzione ha delle singolarità nel caso in cui alcuni punti del primo contorno coincidano con quelli del secondo, un fenomeno che in fisica non si verifica, ma che nel confronto tra sagome è possibile; per risolverlo, gli autori hanno introdotto una funzione che assume, in tali punti, un valore convenzionale, indicato con γ [30]. Si formalizza in questo modo la funzione $\mathrm{EP}(q,A)$, che esprime il valore del potenziale generato dall'insieme di punti $A = a_1,\ldots,a_m$ nel punto q:

$$\mathrm{EP}(q,A) = \begin{cases} \frac{1}{\varepsilon}\sum_{i=1}^{m} \frac{1}{\|q-a_i\|}, \ \forall a_i : a_i \neq q \\ \gamma, \ \exists a_k : a_k = q \end{cases} \quad . \tag{2.3}$$

L'equazione precedente permette di calcolare l'effetto che un insieme di punti, per esempio, quelli che compongono il contorno della sagoma da classificare, ha in una

precisa posizione, come può essere uno dei punti del contorno di un modello. Per confrontare due sagome è quindi necessario effettuare tale calcolo in tutti i punti del secondo contorno, e quindi, dati due insiemi di punti $A = a_1, \ldots, a_m$ e $B = b_1, \ldots, b_m$, si definisce la funzione EP dell'insieme B rispetto all'insieme A come:

$$\mathrm{EPF}(B,A) = \frac{1}{\|B\|} \sum_{b_i \in B} \mathrm{EP}(b_i,A) \ . \tag{2.4}$$

In questa formulazione rimane la dipendenza dai parametri ε e γ, per cui gli autori hanno proposto una nuova versione della funzione, chiamata WEP (*Weighted Edge Potential*), definita come:

$$\mathrm{WEP}(q,A) = w_q \mathrm{EP}(q,A) = w_q \left(\frac{1}{\varepsilon} \sum_{i=1}^{m} \frac{1}{\|q - a_i\|} \right) \ , \tag{2.5}$$

in cui:

$$w_q = \varepsilon \min_{a_i \in A} \|q - a_i\| \ ; \tag{2.6}$$

in questo modo, la funzione è indipendente da ε. Quindi, per due insiemi di punti A e B, la funzione pesata è definita come:

$$\mathrm{WEPF}(B,A) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{WEP}(b_i,A) \ . \tag{2.7}$$

Entrambe queste funzioni sono state provate, ed è risultato che, per il problema che si sta tentando di risolvere, la WEP non offre sostanziali miglioramenti rispetto alla EP, ma, al contrario, fa crescere il numero di falsi positivi, perché aumenta il valore di correlazione nei casi in cui ci sono pochi punti vicini tra le due sagome. Bisogna inoltre notare che la funzione EP è asimmetrica, ovvero non fornisce lo stesso risultato se si invertono i due insiemi di punti tra loro: per esempio, se B è un sottoinsieme di A, $\mathrm{EP}(B,A) = 1$, mentre $\mathrm{EP}(A,B) < 1$. Si è deciso di calcolare il match tra il contorno del modello, che ha un numero di punti limitato e predefinito, e l'immagine dei bordi, in cui il numero di punti è aleatorio.

Alcuni test sulle funzioni EP e WEP hanno fornito risultati non molto buoni, sia in termini di tempi di calcolo che di risultati ottenuti. È stata anche ideata una versione modificata, che pone a 0 la funzione (2.5) quando la distanza $\|q - a_i\|$ è superiore ad una certa soglia, però, ancora, non sono stati ottenuti risultati convincenti.
Si è dunque pensato ad una soluzione alternativa, sempre ispirata all'idea dei potenziali, ma in grado di garantire dei tempi di calcolo più rapidi rispetto alle funzioni EP e WEP; la nuova funzione è stata chiamata MMD, ovvero match a minima distanza. Molto semplicemente, se si confronta un punto q con un insieme di punti $A = a_1, \ldots, a_m$, ad esso è assegnato un valore che dipende dalla sua distanza dal punto di A più vicino, secondo la formula:

$$\mathrm{MMD}(q,A) = \begin{cases} \frac{1}{w_{min}}, \ \forall a_i : a_i \neq q \\ 1, \ \exists a_k : a_k = q \end{cases} , \tag{2.8}$$

in cui:

$$w_{min} = min_{a_i \in A} \|q - a_i\| \ . \tag{2.9}$$

Quindi, il confronto tra due insiemi di punti A e B si ottiene, come in precedenza, sommando tutti i contributi:

$$\mathrm{MMD}(B,A) = \frac{1}{\|B\|} \sum_{b_i \in B} MMD(b_i,A) \ . \tag{2.10}$$

Questa soluzione si è dimostrata migliore sia in termini di velocità di calcolo che di risultati forniti. Sono state allora apportate alcune modifiche che hanno migliorato ulteriormente il comportamento della funzione:

- sono stati esclusi dal calcolo i bordi orizzontali: così facendo, oltre a velocizzare la valutazione del match, si riesce ad eliminare una buona parte dei disturbi dovuti allo sfondo, anche se ciò porta a trascurare alcuni dettagli dei pedoni.

- È stato dato maggior peso ai punti di bordo sovrapposti rispetto a quanto previsto in (2.8); la formula di confronto tra un punto e un contorno diventa, in

questo caso:

$$\mathrm{MMD}(q,A) = \begin{cases} \frac{1}{w_{min}}, \ \forall a_i : a_i \neq q \\ \frac{1}{\gamma}, \ \exists a_k : a_k = q \end{cases} , \tag{2.11}$$

da cui deriva la funzione di confronto tra due contorni:

$$\mathrm{MMD}(B,A) = \gamma \frac{1}{\|B\|} \sum_{b_i \in B} MMD(b_i,A) \ . \tag{2.12}$$

- È stato modificato il calcolo, ponendo a 0 la funzione in (2.11) quando la minima distanza è superiore ad una certa soglia, in analogia con quanto fatto per la funzione WEP; inoltre, nella finestra di calcolo così individuata, è stato introdotto un limite minimo di punti di bordo che devono essere presenti: al di sotto di tale soglia, se i due pixel che si confrontano non sono coincidenti, il contributo è posto convenzionalmente a 0. Così facendo è possibile eliminare i contributi dovuti allo sfondo.

- Il calcolo della correlazione è stato suddiviso in tre, separando i contributi dovuti a testa, busto e gambe, e per ognuno di essi è stata fissata una soglia minima di correlazione; il valore globale è ottenuto come media pesata dei tre contributi, assegnando alla testa il peso maggiore, al busto quello minore, e alle gambe un peso intermedio.

2.3.1.3 Valutazione delle prestazioni

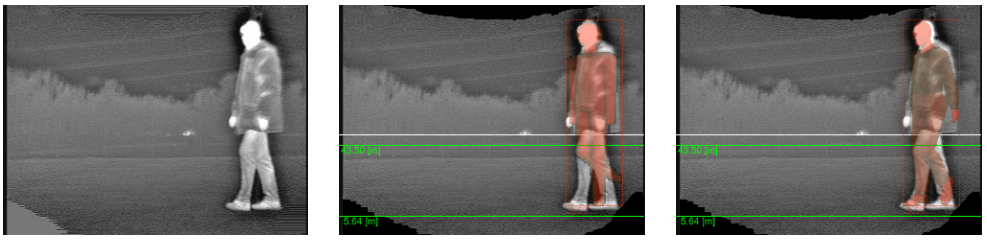
Le prestazioni ottenute grazie alla modifica dei modelli e della funzione di correlazione sono state valutate su una sequenza di test, contenente 5082 pedoni, a cui sono state tolte alcune scene contenenti gruppi di pedoni, che la fase di rilevamento dei bounding box non riesce a separare correttamente. In tale sequenza sono stati valutati due indicatori: *correct detection rate* (CDR) e falsi positivi; il primo è espresso come CD/(CD+FP), dove CD è il numero di pedoni presenti nelle immagini e correttamente classificati come tali, mentre FP è il numero di falsi positivi, cioè aree dell'immagine erroneamente classificate come pedoni. Per quanto riguarda il CDR,

	# modelli	CDR	FP	ms/frame
set originale	72	0,6664	0,1204	216
1° set creato	104	0,6386	0,0669	224
2° set creato	104	0,6255	0,0651	224
3° set creato	104	0,5948	0,0637	226
4° set creato	656	0,6929	0,1078	342

Tabella 2.1: Prestazioni per ciascuno degli insiemi di modelli creati.

esso ha un limite superiore a 0,7254, dovuto alla fase iniziale dell'algoritmo, che non individua tutti i bounding box contenenti i pedoni. I valori di CDR e falsi positivi per i vari insiemi di modelli creati sono riportati nella tabella 2.1, assieme al numero di modelli per ogni insieme, e ai tempi di calcolo, riferiti all'intero algoritmo di localizzazione dei pedoni, e non solo alla fase di validazione con i modelli. Come si vede, il primo insieme di modelli creati (figura 2.6) fa calare di circa il 3% il valore di CDR, ma fa contemporaneamente diminuire di metà il numero di falsi positivi; il tempo di calcolo aumenta, anche se in maniera contenuta, a causa del maggior numero di modelli da confrontare. Il secondo e terzo set di modelli (figura 2.7.b-c) hanno fornito prestazioni simili, e comunque lievemente peggiori, e sono perciò stati scartati. Infine, il massimo valore di CDR si ottiene con il quarto gruppo di modelli (figura 2.7.d), grazie all'elevatissimo numero di immagini presenti; tuttavia, questa soluzione fa anche crescere il numero di falsi positivi rispetto al primo insieme, e i tempi di calcolo aumentano in misura notevole. Con i nuovi modelli, inoltre, aumenta il numero di casi in cui quello scelto ha la postura corretta, come si vede in figura 2.8.

Il primo insieme di modelli è sembrato quello in grado di offrire il miglior compromesso tra CDR e FP, ed è stato perciò scelto per la sostituzione dei modelli preesistenti; questa decisione è stata presa per premiare la riduzione del numero di falsi positivi a discapito della diminuzione del CDR. Si tenga presente, infatti, che le statistiche discusse sono ottenute analizzando immagine per immagine i risultati dell'algoritmo; nel caso di un'applicazione reale, è importante che ogni pedone sia individuato almeno in un certo numero di fotogrammi, e il fatto che la sua localizzazione non avvenga



(a) (b) (c)

Figura 2.8: Esempio di match di un pedone reale (a): confronto tra il modello scelto dall'algoritmo di cross-correlazione tra l'insieme di quelli preesistenti (b) e quelli del nuovo set (c). Si noti come nell'ultimo caso la postura del modello sia corretta, cosa che non accadeva precedentemente. Le linee verdi indicano dove si trovano i bordi inferiori dei bounding box contenenti i pedoni alla minima e massima distanza di riconoscimento; la linea bianca indica l'orizzonte.

proprio in tutti è secondario, almeno finché il numero di casi in cui è individuato è comunque elevato: questo perché è ugualmente possibile valutare la pericolosità di una situazione che si sta verificando sfruttando le immagini correttamente elaborate. Al contrario, il numero di falsi positivi è un indicatore di quanti falsi allarmi il sistema dà, e individuare un pedone laddove non è presente fa sì che il sistema rilevi un pericolo inesistente. Preferire un algoritmo con CDR e FP minori significa quindi optare per un sistema che introduce un ritardo nella rilevazione dei pericoli, anche se si tratta comunque di tempi inferiori al secondo, ma che genera anche un numero di falsi allarmi inferiore.

Per l'analisi delle prestazioni ottenute con le funzioni di correlazione proposte, effettuata utilizzando già il nuovo insieme di modelli, si è partiti dalla EP. I risultati di CDR e FP ottenuti al variare di γ, ε e della soglia di correlazione oltre la quale i pedoni sono validati, sono riportati nella tabella 2.2, assieme ai valori ottenuti con la cross-correlazione, per confronto. I risultati ottenuti limitando il calcolo ad una finestra quadrata si trovano in tabella 2.3, in cui AR (Area di ricerca) è la percentuale della larghezza del bounding box utilizzata come lato per la finestra di calcolo. Cambiando funzione di correlazione, e adottando la MMD descritta nel paragrafo

Soglia	γ	ε	CDR	FP
Cross-correlazione				
0,4	–	–	0,6386	0,0669
EP				
0,4	1	200	0,7100	0,2667
0,599	1	200	0,5147	0,0864

Tabella 2.2: Statistiche ottenute utilizzando la funzione EP.

Soglia	γ	ε	AR	CDR	FP
0,4	1	100	20%	0,7094	0,2993
0,5	1	100	20%	0,6492	0,1705
0,5	1	100	33%	0,6664	0,1822
0,4	1,5	100	25%	0,7149	0,5790
0,4	1	150	25%	0,7102	0,3109
0,4	1	200	25%	0,7094	0,3076
0,4	1	200	20%	0,7082	0,2988
0,4	1	200	33%	0,7102	0,3155
0,5	1,5	200	20%	0,7147	0,4168
0,5	1,5	300	20%	0,7147	0,4159
0,5	1,5	600	20%	0,7147	0,4150
0,4	1	200	20%	0,6957	0,2203
0,5	1	200	20%	0,6054	0,1343

Tabella 2.3: Statistiche ottenute utilizzando la funzione EP con finestra di calcolo limitata.

Soglia	AR	CDR	FP
Confronto del corpo completo			
0,698	33%	0,6444	0,1738
0,698	20%	0,5424	0,1059
Confronto escludendo il busto			
0,698	20%	0,6907	0,1413
0,698	33%	0,7059	0,2193
0,753	33%	0,6904	0,1617
0,698	50%	0,7104	0,2505
0,753	50%	0,6965	0,1887
0,720	20%	0,6797	0,1245
0,698	25%	0,6945	0,1617

Tabella 2.4: Statistiche ottenute utilizzando la funzione MMD, sia su tutto il corpo, sia escludendo il busto dal calcolo del match.

precedente, è stato possibile ottenere dei risultati simili a quelli della funzione EP, migliorati poi con la scelta di escludere il busto dal confronto tra le sagome, come si desume dalla tabella 2.4. Si è quindi provveduto ad agire su molti parametri, per ottimizzare le prestazioni della funzione MMD. I valori ottimi trovati sono:

- soglia minima di correlazione globale pari a 0,544;
- valore di γ pari a 0,6;
- finestra quadrata in cui è limitato il calcolo della correlazione avente il 33% della larghezza del bounding box;
- numero minimo di pixel di bordo in ciascuna finestra di calcolo locale: 10;
- pesi da assegnare alle varie parti del corpo: testa 30%, busto 10%, gambe 60%;
- soglia minima di correlazione per le singole parti del corpo: busto 10%, gambe 40% (per la testa non è fissato il minimo).

Algoritmo	CDR	FP
Cross-correlazione, modelli vecchi	0,6664	0,1204
Cross-correlazione, modelli nuovi	0,6386	0,0669
MMD, modelli nuovi	0,7039	0,1330

Tabella 2.5: Sintesi dei risultati ottenuti.

Utilizzando questi valori, si ottiene che CDR=0,7039, FP=0,1330. Questo risultato non è migliore della funzione di cross-correlazione che fa uso dei modelli nuovi, perché CDR è aumentato, ma a scapito di un maggior numero di falsi positivi. Le prestazioni con i nuovi modelli e la funzione MMD sono comunque migliori della situazione iniziale, in cui CDR=0,6664 e FP=0,1204, perché il valore di CDR è decisamente migliorato con un lieve peggioramento di FP.
In conclusione, la situazione iniziale è stata migliorata in due modi diversi: utilizzando i modelli nuovi e la cross-correlazione si ottiene una decisa diminuzione dei falsi positivi, con un valore di CDR sostanzialmente invariato; adottando la funzione MMD con i modelli nuovi, invece, i falsi positivi si riportano ad un valore simile a quello iniziale, consentendo però un sensibile aumento del CDR. La tabella 2.5 sintetizza i risultati ottenuti.

2.3.2 Estrazione della sagoma mediante contorni attivi

Il secondo metodo sviluppato per l'estrazione della sagoma si basa sui contorni attivi, o *snake*; questa tecnica, presentata per la prima volta in [31], è stata oggetto di studio anche negli ultimi anni. Dal punto di vista analitico, uno snake è una curva parametrica, e può essere descritta dall'equazione:

$$\mathbf{v}(s) = (x(s), y(s)) \quad , \tag{2.13}$$

in cui s ha il significato di ascissa curvilinea normalizzata, e varia, perciò, nel range $[0;1]$. Nel dominio discreto, questa curva è campionata in un insieme di punti, detti *snaxel*, contrazione di snake e pixel, tra i quali sono presenti alcuni vincoli che li

tengono uniti e ordinati.

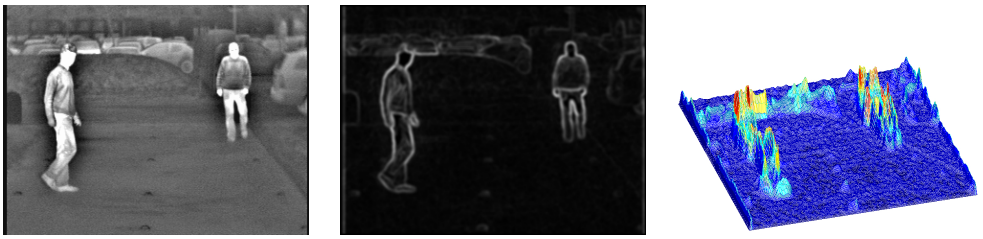
Gli snake sono utilizzati per l'estrazione delle sagome: essi devono essere posizionati attorno all'oggetto di cui si vuole trovare il contorno, e si adattano ad esso mediante una serie di iterazioni che muovono uno snaxel alla volta. Le forze che muovono i punti sono suddivise in due tipologie: quelle interne e quelle esterne. Le forze interne tengono assieme il contorno attivo, e gli danno una sorta di consistenza meccanica, facendo in modo che ogni snaxel sia attratto dai suoi vicini ed evitando che il contorno si pieghi troppo bruscamente; in altre parole, esse fanno sì che lo snake si comporti sia come un elastico, sia come una lastrina metallica. Le forze esterne, invece, sono quelle che attraggono gli snaxel verso le caratteristiche dell'immagine di interesse, e la loro formulazione dipende perciò dal problema in esame.

Dal punto di vista analitico, l'energia interna dello snake si esprime come la somma pesata di due contributi:

$$E_{\text{int}} = \alpha(s)|\mathbf{v}_s(s)|^2 + \beta(s)|\mathbf{v}_{ss}(s)|^2 \ , \tag{2.14}$$

in cui $\mathbf{v}_s(s)$ e $\mathbf{v}_{ss}(s)$ sono, rispettivamente, la derivata prima e seconda di $\mathbf{v}(s)$ rispetto a s. Il primo termine è detto tensione dello snake, ed è causa del comportamento elastico, mentre il secondo esprime la resistenza alle piegature; $\alpha(s)$ e $\beta(s)$ sono semplicemente dei pesi, variando i quali si possono scegliere le caratteristiche meccaniche del contorno. Si noti come la dipendenza da s, anche se analiticamente prevista, sia quasi sempre trascurata nei casi pratici, che utilizzano dei pesi costanti.

Come anticipato, lo snake si muove in maniera iterativa: ogni snaxel è preso in considerazione singolarmente, ed è spostato nel punto del suo vicinato in cui è minimo il bilancio energetico globale: proprio in questa decisione, quindi, entrano in gioco anche le forze esterne, che devono essere tali da attrarre lo snake verso le caratteristiche salienti dell'immagine. L'algoritmo di minimizzazione non è unico, ma quello più utilizzato, e che è stato adottato anche in questo lavoro, è quello presentato in [32], applicato ad un vicinato 5×5. Dopo un certo numero di iterazioni sull'intero snake,



(a) (b) (c)

Figura 2.9: Dati su cui si calcolano le forze esterne: immagine FIR originale (a); elaborazione di Sobel e successivo filtraggio gaussiano (b), e diagramma della funzione energia dei bordi (il segno è invertito per maggiore chiarezza grafica).

l'energia totale, data da:

$$E_{snake} = \int_0^1 \left(E_{\text{int}}\left(\mathbf{v}(s)\right) + E_{ext}\left(\mathbf{v}(s)\right)\right) \mathrm{d}s \ , \tag{2.15}$$

sarà minimizzata, e ci si aspetta che lo snake sia aderente al contorno dell'oggetto di cui si desidera estrarre la sagoma.

Per come è strutturata, la tecnica dei contorni attivi cerca un minimo locale: questo significa che la configurazione finale dipende dalla posizione iniziale dello snake; la scelta dell'inizializzazione deve pertanto essere fatta accuratamente. Nel caso in esame, ci si trova a lavorare su delle regioni di interesse già trovate dai passi precedenti del sistema (si veda lo schema di figura 2.2), e la scelta più naturale è quella di posizionare lo snake sul contorno dell'area di interesse.

Le forze esterne che entrano nel bilancio energetico dello snake sono basate su due immagini: quella FIR acquisita dalla telecamera, e la stessa dopo l'applicazione di un filtro di Sobel per l'estrazione dei bordi, e un operatore di smoothing gaussiano; quest'ultimo si è reso utile perché amplia i bordi, ed estende quindi la zona in cui gli snake risentono del loro effetto; si veda la figura 2.9. Poiché i contorni attivi devono posizionarsi sulla sagoma del pedone, le zone con i bordi più intensi devono avere un'energia molto bassa, in modo da attrarre gli snake: la funzione assume il valore del tono di grigio nel pixel corrispondente, cambiato di segno; un esempio di

funzione di questo tipo si trova in figura 2.9.c. Per quanto riguarda l'energia associata all'immagine FIR, essa è direttamente proporzionale al valore del tono di grigio di ogni pixel: in questo modo, le zone chiare respingono lo snake, effetto che aiuta a mantenerlo al di fuori del contorno del pedone durante la fase di contrazione. Infine, per migliorare il comportamento contrattivo attorno ai pedoni, è stata aggiunta una forza che attira gli snaxel verso il centro geometrico della regione di interesse. Alla fine, il bilancio energetico espresso dall'equazione (2.15), riferito al singolo snaxel, diventa:

$$E_{\text{snaxel}} = \alpha E_{\text{el}} + \beta E_{\text{ben}} + \gamma E_{\text{img}} + \delta E_{\text{edge}} + \eta E_{\text{hor}} + \theta E_{\text{vert}} \quad , \tag{2.16}$$

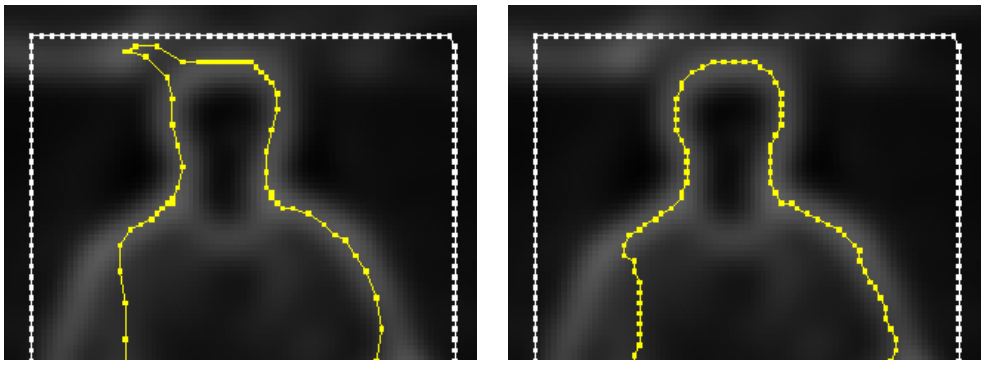
in cui le energie a secondo membro rappresentano, nell'ordine, quella elastica, di curvatura, dell'immagine, dei bordi, e quelle centripete orizzontale e verticale.

Mentre il contorno attivo si contrae, avvicinandosi alla sagoma del pedone, la sua lunghezza diminuisce, così come la distanza media tra gli snaxel, il che altera le proprietà meccaniche dello snake, che dipendono anche da tale distanza. Si possono inoltre creare degli accumuli di punti molto vicini, o dei tratti in cui essi sono troppo radi, situazioni che peggiorano la contrazione; un esempio si vede in figura 2.10.a, in cui il contorno bianco mostra la posizione iniziale, e quello giallo la posizione assunta alla fine del processo di contrazione. Per ovviare a questo effetto, il contorno attivo è periodicamente ricampionato a passo costante (figura 2.10.b).

In figura 2.11 è possibile vedere alcuni esempi di forme estratte utilizzando gli snake; si noti come la regione più difficile da analizzare sia quella delle gambe quando non sono unite: è difficile che lo snake riesca a contrarsi attorno alla concavità che si crea in questo caso. In figura 2.12 si vedono alcune sagome estratte in immagini acquisite d'estate, quando i pedoni non sono più caldi dell'ambiente circostante, e, quindi, non appaiono più chiari: le prestazioni sono comunque accettabili anche in questi casi.

2.3.2.1 Snake doppio

L'estrazione della sagoma mediante gli snake ha dato buoni risultati, tuttavia si è notato come ci possano essere dei problemi nella rilevazione corretta delle zone con-

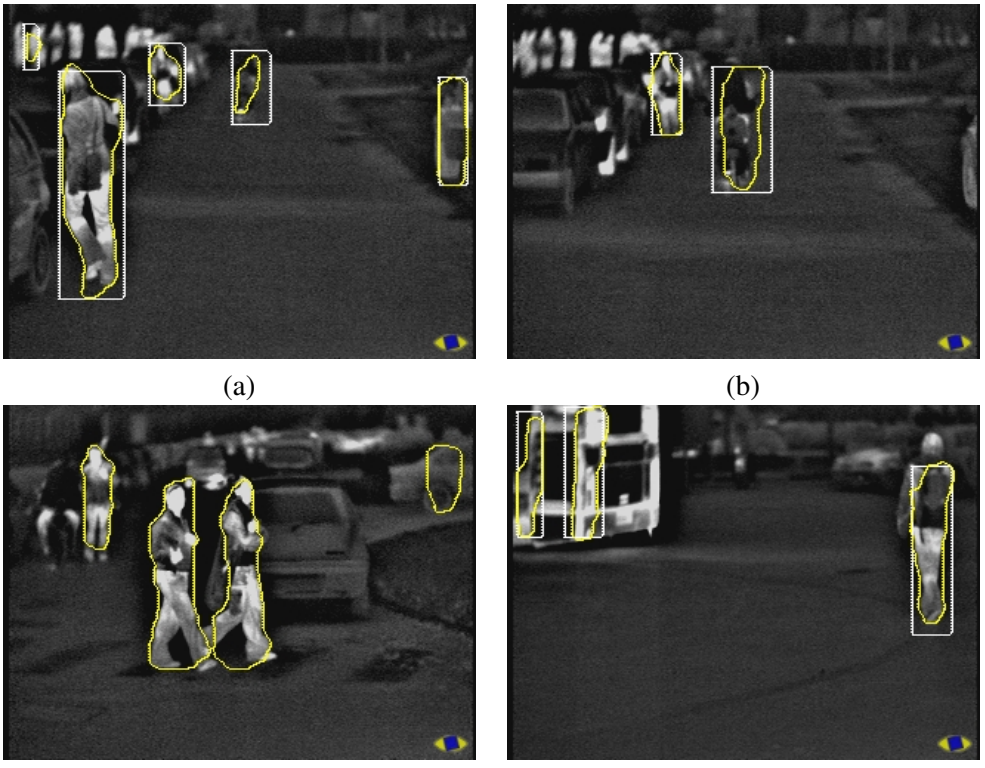


(a) (b)

Figura 2.10: Risultato finale della contrazione di uno snake attorno alla testa di un pedone; in (a) non è applicato nessun ricampionamento, e si osserva un accumulo di punti molto vicini, che è assente quando il ricampionamento è utilizzato, come si vede in (b).

cave, come quella tra le gambe. Per ovviare a questo problema, e, in generale, a tutti i casi in cui lo snake non è sufficientemente attratto dalla sagoma del pedone, è stato provato un approccio detto a "doppio snake", come proposto in [33]. Si tratta di una tecnica che prevede di utilizzare due snake, uno posto al di fuori della sagoma, dotato di una tendenza a contrarsi, e un altro piazzato all'interno, tendente ad espandersi. Le energie interne di ciascuno di essi sono le stesse dello snake singolo, mentre tra quelle esterne ci sono due differenze: innanzi tutti, lo snake interno ha il segno cambiato nella forza di contrazione, che diventa, in tal modo, forza di espansione; in secondo luogo, in entrambi si aggiunge una forza che fa in modo che ogni contorno sia attratto dall'altro. L'obiettivo è quello di far sì che i due snake entrino in contatto, individuando in tal modo il contorno.

Sebbene l'idea alla base della tecnica del doppio snake sia interessante, la realizzazione pratica di questo approccio ha presentato un certo grado di difficoltà. Innanzi tutto, la forza di attrazione tra gli snake è risultata molto complicata da regolare nel bilancio energetico: è stato difficile trovare sia una funzione di valutazione dell'energia che dei pesi capaci, nella maggioranza dei casi, di far attrarre i due snake, lasciando



(a) (b)

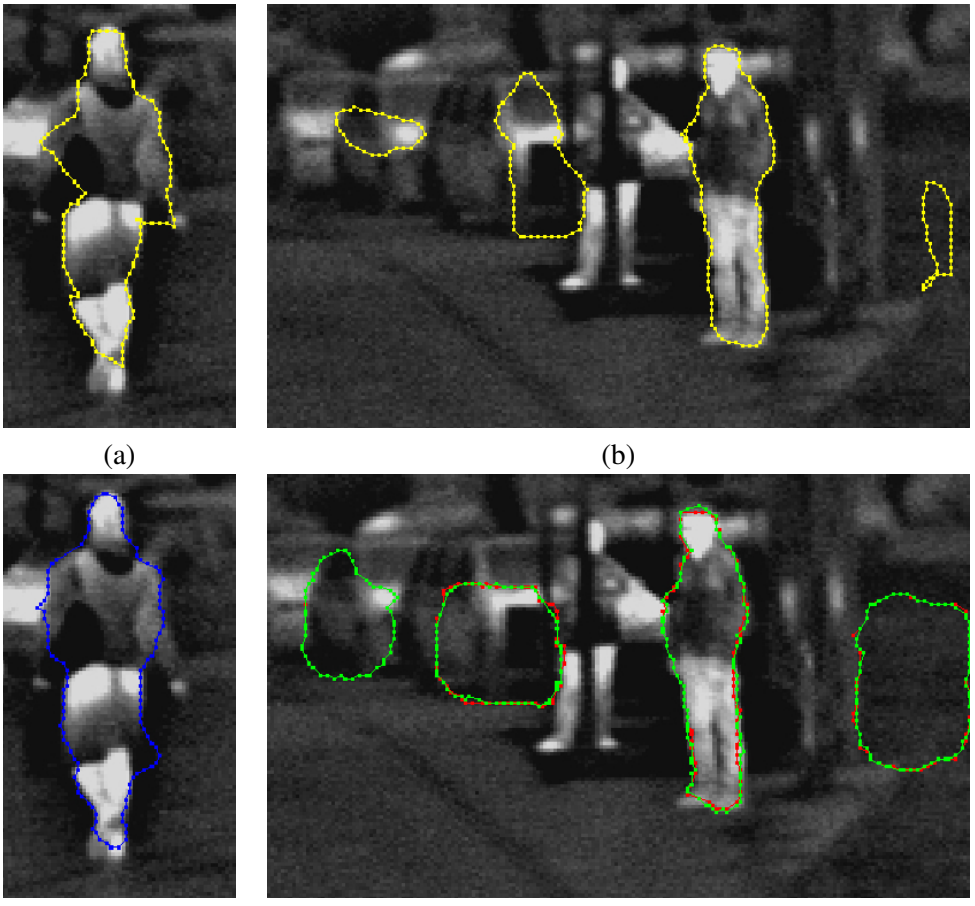
(c) (d)

Figura 2.11: Esempi di sagome estratte utilizzando i contorni attivi. In (a) e (b), un pedone vicino e uno lontano; lo snake bianco rappresenta la posizione iniziale del contorno, mentre quello giallo è la forma finale. In (c), esempio di analisi di un gruppo di pedoni; in (d), un tipico esempio di come la configurazione iniziale condizioni il risultato: poiché la testa è al di fuori dello snake nella sua posizione iniziale (a causa di un errore nella valutazione della regione di interesse), essa è tagliata anche dalla sagoma estratta.

(a) (b)

Figura 2.12: Esempi di sagome estratte quando gli snake lavorano su immagini acquisite d'estate, particolarmente problematiche a causa dell'elevato numero di oggetti caldi.

contemporaneamente spazio anche al contributo delle altre forze. Si è allora deciso di modificare il meccanismo di evoluzione dello snake, creando delle associazioni tra gli snaxel dei due contorni, chiamate molle. Ciascuna di esse ha alle estremità due snaxel, e obbliga uno di loro a muoversi verso l'altro, scegliendo di volta in volta uno o l'altro a seconda del movimento che minimizza l'energia totale. Così facendo, è stato ottenuto un risultato ragionevole: i due contorni vengono per forza a contatto, e l'estrazione della sagoma presenta risultati soddisfacenti, come si può vedere in figura 2.13; nel caso (d), poi, si può osservare che, quando non si ha a che fare con pedoni, la sagoma ricavata è più simile a un quadrato rispetto a quella ricavata con la tecnica dello snake singolo. Anche se lo snake doppio offre dei vantaggi su quello singolo, il miglioramento non è così netto da giustificare l'aumento della complessità computazionale, e si è deciso di abbandonare questo approccio, procedendo con la classificazione degli snake singoli.



(a) (b)

(c) (d)

Figura 2.13: Confronto tra le sagome estratte a partire dalla stessa immagine utilizzando la tecnica del singolo (a) e doppio snake (c). Si prende in analisi anche una scena più complessa; di nuovo, (b) mostra il risultato utilizzando un solo snake, mentre in (d) si fa uso di quello doppio. Il contorno rosso rappresenta lo snake interno, mentre quello esterno è colorato di verde.

2.3.3 Classificazione della sagoma mediante reti neurali

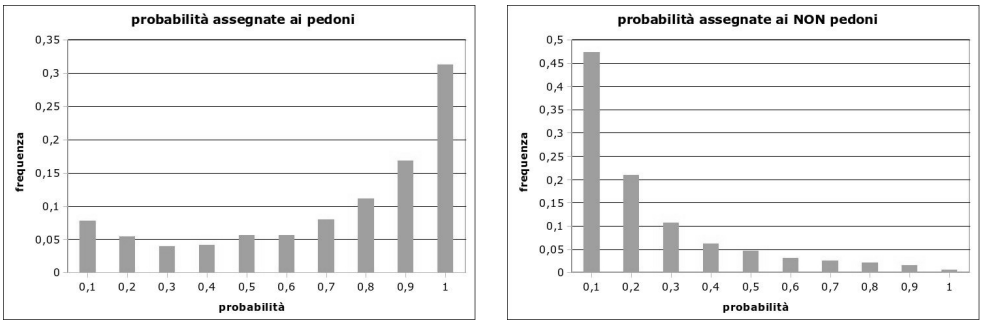
Una volta estratta la sagoma degli oggetti contenuti nelle regioni di interesse, si procede alla loro classificazione, ovvero a stabilire con che probabilità la sagoma appartiene ad un pedone. Nel caso degli snake, questo avviene mediante una rete neurale[2], creata utilizzando il software Lightweight Neural Network++.

L'insieme di punti di un contorno attivo non è immediatamente utilizzabile da una rete neurale, che richiede che il numero degli ingressi sia sempre lo stesso, e che ciascuno di essi vari in un range fissato. Al contrario, il numero degli snaxel dipende dalla lunghezza dello snake, a causa del ricampionamento, e le loro coordinate hanno l'immagine come riferimento.

Per poter utilizzare una rete neurale, è quindi necessario effettuare due adattamenti; il primo consiste in un ulteriore ricampionamento degli snake con un numero sempre uguale di punti, indipendentemente dalla lunghezza del contorno; con un po' di prove, si è visto che 30 punti sono sufficienti per descrivere una sagoma con un sufficiente livello di accuratezza, e la perdita di dettaglio delle sagome più grandi può essere visto come un vantaggio, perché evita alla rete neurale di doversi allenare anche su sagome molto dettagliate. Il secondo adattamento è una normalizzazione delle coordinate dei punti: si considerano l'ascissa e l'ordinata minime, e si assegna loro il valore 0; poi si cercano i valori massimi, e si assegna loro il valore 1, normalizzando poi entrambe le coordinate di tutti i punti. Questo procedimento deforma le sagome, perché, qualunque sia la proporzione del bounding box originale, questo diventa un quadrato di lato unitario; tuttavia, anche questo fenomeno può essere visto come un vantaggio, perché elimina la variabilità delle proporzioni dalla rete neurale, e non è quindi necessario utilizzare molte sagome di proporzioni diverse nel training set, facendo così in modo che la rete si concentri solo sulla forma.

L'addestramento della rete è stato effettuato utilizzando circa 1200 sagome di pedoni ed altrettante appartenenti ad altri oggetti, utilizzando brevi spezzoni di sequenze: in questo modo, per ogni pedone sono considerate diverse posizioni, e, al contempo, si

[2] Questo studio è oggetto della tesi di laurea di Alberto Agoletti.



(a) (b)

Figura 2.14: Valori dati in uscita dalla rete neurale. In ascissa si trova il range di valori del voto della rete neurale; in ordinata, viceversa, la frazione delle sagome che ha ricevuto un valore in ciascun range; questo andamento è stato calcolato sia per le sagome dei pedoni (a), che per quelle degli altri oggetti (b).

evita di utilizzare troppe sagome appartenenti alla stessa persona. Durante l'addestramento, il valore di output atteso è stato fissato a 0,95 per i pedoni, e 0,05 per gli altri oggetti. La rete neurale così creata, avente 60 neuroni nello strato nascosto, è stata testata su un migliaio di casi, ovviamente diversi da quelli utilizzati per l'addestramento; i valori associati ad ogni sagoma sono stati poi elaborati, e sono mostrati in figura 2.14: nel caso (a), istogramma dei valori che la rete ha associato ai pedoni; nel caso (b), i valori associati agli altri oggetti.

Questo classificatore fornisce in uscita un voto nel range $[0; 1]$, e infatti nel sistema di figura 2.2 non è prevista una soglia subito a valle della rete neurale: è la validazione finale che prende una decisione definitiva, sulla base dei voti di tutti i moduli di classificazione. Tuttavia, per capire che prestazioni ha questo classificatore, si è cercata la soglia ottima, che vale 0,4, utilizzando la quale il 79% dei pedoni, e l'85% dei non pedoni è correttamente classificato.

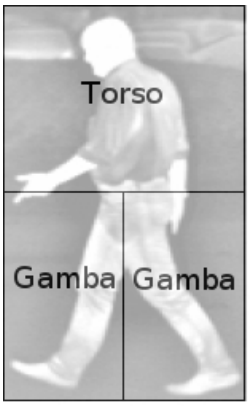


Figura 2.15: Suddivisione del bounding box: sono visibili le aree di ricerca delle gambe.

2.4 Ricerca degli arti inferiori

Gli arti inferiori sono una caratteristica peculiare dei pedoni, in grado di distinguerli dai pali e dagli alberi, che sono i casi più frequenti di falsi positivi; per questo motivo, si è deciso di dedicare uno studio specifico a questa caratteristica[3]. Anche in questo caso, si tratta di sviluppare un modulo capace di raffinare i risultati della ricerca dei pedoni in immagini FIR, e si ha quindi a che fare con delle regioni di interesse già individuate.

Dovendo cercare le gambe, ci si concentra nella parte inferiore della regione di interesse, più precisamente, nei $3/5$ inferiori del bounding box che racchiude il potenziale pedone (figura 2.15). In questa zona si procede con un'elaborazione di basso livello, che ha lo scopo di evidenziare le gambe, che sono zone dell'immagine più chiare rispetto allo sfondo. Per prima cosa, essa viene suddivisa in dieci fasce orizzontali di uguale dimensione: sulle prime nove si effettua una binarizzazione utilizzando come soglia la media aritmetica dei valori dei pixel che la compongono, mentre sull'ultima si utilizza una soglia di valore pari alla media delle altre. Si adotta questa procedura perché sulla fascia inferiore capita spesso che sia presente solo una delle due gambe,

[3] Questo studio è oggetto della tesi di laurea di Matteo Formaini.

(a) (b)

Figura 2.16: Esempio di binarizzazione della parte inferiore del bounding box quando la soglia è calcolata indipendentemente per tutte le fasce orizzontali (a), e quando alla fascia inferiore si applica una soglia pari alla media delle altre (b).

il che rende il numero di pixel chiari notevolmente inferiore rispetto a quello dei pixel scuri, come mostrato in figura 2.16.

A valle della sogliatura, la zona occupata dalle gambe sarà in gran parte bianca, mentre lo sfondo sarà di colore nero; si applica quindi un operatore morfologico per eliminare i punti bianchi sullo sfondo, mentre un'operazione di chiusura elimina i punti neri isolati nelle regioni chiare. Inizia quindi la ricerca delle gambe vera e propria: la parte inferiore dell'area di interesse viene idealmente suddivisa in tre fasce verticali di larghezza uguale; poi, nelle due esterne, si scorrono le varie colonne di pixel, partendo da quella più interna. Per ciascuna colonna si cerca di individuare in quale punto comincia la gamba, e si verifica che tali punti siano vicini; quindi, tutte le zone all'esterno di questi punti sono eliminate, portando il valore dei pixel a 0. Questo procedimento serve per eliminare oggetti chiari sullo sfondo, per esempio la gamba di un altro pedone, come mostrato in figura 2.17.

Giunti a questo punto, l'elaborazione di basso livello è conclusa, e si procede a determinare la forma delle gambe. Si cerca quindi il loro punto di giunzione, se esiste, e si procede con l'individuazione, riga per riga, delle zone chiare nelle vicinanze; in questo modo si riesce a capire se un pedone ha le gambe aperte o chiuse, e la loro

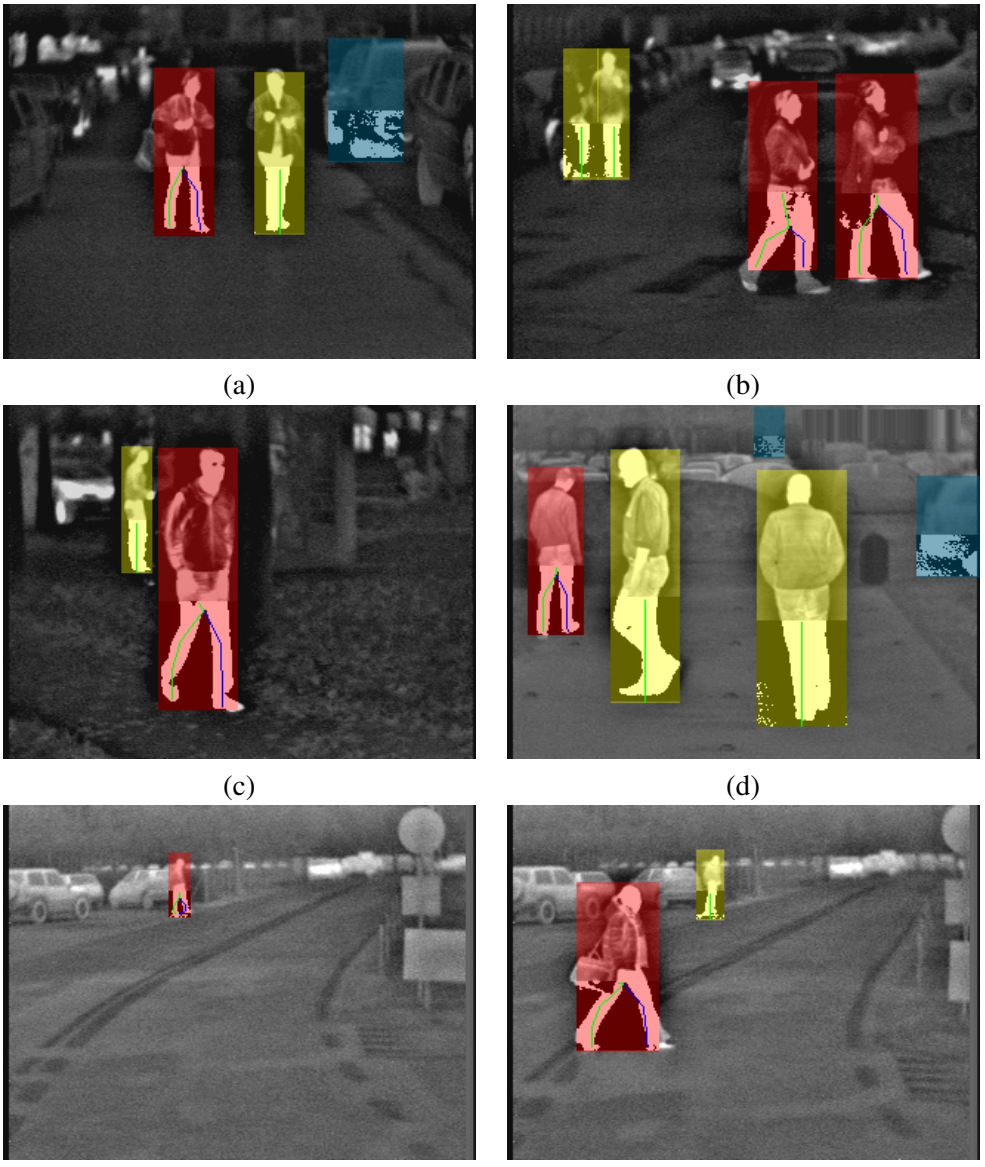
(a) (b) (c)

Figura 2.17: In (a), la zona dell'immagine interessata dall'elaborazione; in (b), il risultato della binarizzazione, mentre in (c) si osserva cosa accade dopo l'applicazione della ricerca degli arti inferiori: si noti come la gamba di un altro pedone, sulla destra, è stata eliminata.

postura.

Questa procedura di classificazione fornisce in uscita tre possibili risultati: "gambe non presenti", oppure "gambe unite", oppure "gambe separate". Nel primo caso, si tratta di un oggetto che non è un pedone, oppure di un'immagine di cattiva qualità; nel secondo si può avere a che fare con un uomo con le gambe unite, ma anche con un palo o un albero, quindi non si ha la certezza che si tratti di un pedone; quando ci si trova nell'ultimo caso, infine, si tratta quasi sicuramente di un pedone, perché ben difficilmente un oggetto ha una forma che assomiglia a quella delle gambe.

Alcuni risultati sono mostrati in figura 2.18; in blu, i bounding box in cui non sono state trovate le gambe, mentre gli altri colori indicano che sono state trovate le gambe chiuse (giallo) o aperte (rosso). Si può notare, in (e), che l'algoritmo funziona correttamente anche con i pedoni lontani; in (f), che la presenza di altri oggetti, come una borsetta, non compromette la corretta identificazione delle gambe. In generale, si può dire che questa tecnica è efficace per aumentare la confidenza con cui un oggetto è classificato come pedone, nel caso in cui abbia le gambe aperte. Può risultare particolarmente utile in un contesto in cui si fa il tracking dei pedoni, perché in tal caso



(a) (b)

(c) (d)

(e) (f)

Figura 2.18: Esempi di riconoscimento delle gambe a varie distanze.

il sistema è in grado di sapere che un certo ostacolo ha mostrato di avere le gambe nel passato, ed è quindi classificabile come pedone con una maggior sicurezza anche quando le gambe sono chiuse.

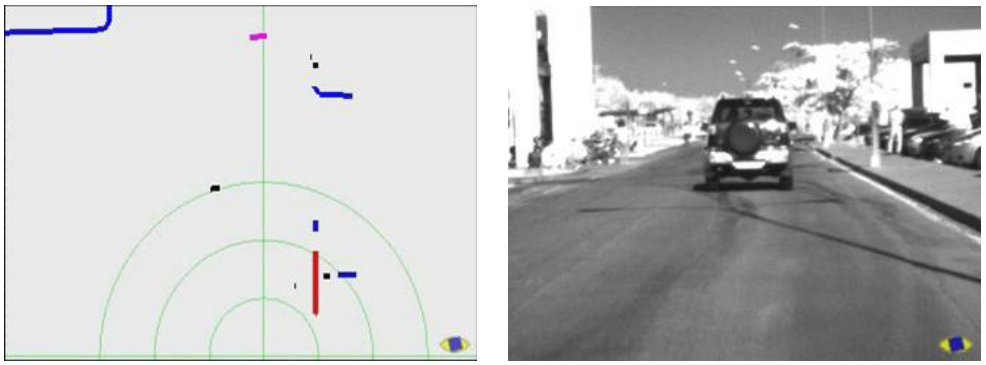
2.5 Riconoscimento dei pedoni usando i dati di un laserscanner

Il riconoscimento dei pedoni può essere migliorato fondendo i risultati di un algoritmo di visione artificiale con i dati provenienti da un laserscanner. Prima della fusione, può essere utile effettuare una classificazione preliminare degli ostacoli rilevati dal laser, in modo da trasformare i dati da un insieme contenente molti punti a uno con pochi ostacoli, di cui si conosce la dimensione.
Il laser con cui è stata sviluppata questa tecnica[4] è un SICK LMS211, simile a quello in figura 1.4.a, con la sola differenza dell'involucro, per renderlo di categoria IP67; in figura 2.19 si può vedere il sensore installato a bordo del veicolo sperimentale usato per i test. Questo sensore ha una risoluzione nativa di 1°, per aumentare la quale si fa ricorso all'interallacciamento: ogni scansione contiene dati acquisiti in quattro rotazioni successive dello specchio; questo fa sì che i profili degli oggetti che si muovono relativamente al sensore non siano netti, il che vale per quasi tutti gli oggetti quando è il sensore stesso ad essere in movimento. Il fenomeno descritto è distinguibile anche a velocità moderate: considerando che lo specchio del laserscanner ha un periodo di rotazione di 13,32 ms, quando il veicolo, per esempio, viaggia alla velocità di 36 km/h, ovvero 10 m/s, esso percorre, in un periodo di rotazione, 133 mm, ovvero uno spazio notevolmente superiore alla sensibilità del sensore, che è di 10 mm (in questa configurazione). Per ovviare a questo inconveniente è possibile applicare una correzione, visto che si conosce in che ordine e in che tempi il laserscanner effettua le misurazioni, ed è anche nota la velocità dell'automobile, grazie ai dati provenienti dalla centralina dell'auto.

[4] Questo studio è oggetto della tesi di laurea di Denis Simonazzi.

Figura 2.19: Veicolo sperimentale con il laserscanner montato nella sezione frontale, sotto la griglia del radiatore.

Dopo aver corretto i punti, si procede all'analisi degli stessi, e alla loro classificazione. In particolare, è possibile raggruppare i punti in modo da segmentare i vari ostacoli presenti nella scena. Per questo motivo, si analizzano dapprima i punti acquisiti durante un'unica rotazione dello specchio, si raggruppano quelli vicini e allineati, e, infine, si creano dei segmenti che ne approssimano l'andamento. Alla fine di questo passo rimangono pochi punti isolati, e una serie di segmenti che sostituiscono tutti gli altri punti. Come detto, ogni segmento rappresenta un insieme di punti allineati, quindi un profilo più complesso, che contiene dei cambi di direzione, sarà rappresentato da una spezzata. Poiché questo procedimento è effettuato separatamente sugli insiemi di punti acquisiti con una sola rotazione dello specchio, per ogni ostacolo saranno



(a) (b)

Figura 2.20: Esempio di dati laserscanner elaborati con la tecnica descritta (a): i segmenti colorati in nero rappresentano i possibili pedoni; quelli in blu gli ostacoli fissi di grandi dimensioni; quelli rosa gli ostacoli in movimento, mentre quelli rossi sono i possibili bordi della strada. In (b) è visibile la scena cui fanno riferimento le misurazioni del laserscanner.

presenti quattro spezzate dalla forma simile; esse si sovrappongono se l'oggetto è fermo rispetto al mondo (il movimento del veicolo è stato compensato), mentre, nel caso contrario, saranno disposte lungo la direzione del movimento. Per ottenere dei dati più semplici da gestire, per ogni gruppo di spezzate relative allo stesso oggetto si effettua una fusione, dopo la quale ogni ostacolo è segnalato da una sola spezzata.

Un esempio di output del sistema si trova in figura 2.20, in cui sono rappresentati i dati elaborati come descritto (a), e l'immagine acquisita dalla telecamera (b), riportata perché mostra la scena alla quale sono riferite le misurazioni del laserscanner. I dati così elaborati risultano molto più semplici da gestire, perché un insieme di poche spezzate ha sostituito 400 punti. A questo livello diventa possibile anche fare una prima classificazione degli ostacoli, basata sulla loro dimensione e sul movimento. In particolare, si distinguono gli ostacoli aventi dimensioni compatibili con quelle di un pedone, colorati in nero in figura 2.20.a.

2.6 Sistemi di tracking per i pedoni

I sistemi di tracking, molto usati per la visione artificiale e nel riconoscimento dei pedoni, permettono di introdurre informazioni su eventi che si sono verificati in passato. Il numero di sistemi sviluppati e testati è davvero elevato, anche solo considerando quelli per il rilevamento di pedoni [34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48]. Il vantaggio consiste nel poter rendere più stabile il riconoscimento, tentando di eliminare i falsi positivi e negativi, e di accumulare altri tipi di informazione mano a mano che l'osservazione va avanti. I sistemi di tracking fanno spesso uso di filtri predittivi, cioè in grado di predire il moto degli oggetti analizzando come questi si sono mossi nel passato; alcuni esempi sono i filtri α-β-γ [49] e quello di Kalman [50], con le successive modifiche e discussioni [51]. Per un tutorial sul filtro di Kalman si rimanda a [52], mentre in [53] si trova una trattazione estesa, semplice e rigorosa.

L'esistenza di un sistema di tracking può dare maggior valore ad alcuni algoritmi, come è stato già messo in luce parlando del rilevamento delle gambe dei pedoni: in quel contesto, la possibilità di sapere, in ogni momento, che di un certo ostacolo, nel passato, sono state individuate le gambe, permette di affermare con una probabilità alta che esso sia un pedone anche nei fotogrammi in cui le gambe non sono chiaramente distinguibili, per esempio perché sono unite. Inoltre, la disponibilità di informazioni su moto e posizione futura di un pedone permette di analizzarne la potenziale pericolosità, se confrontata con la dinamica del veicolo su cui il sistema è installato.

In un certo senso, è possibile vedere il tracking come un primo passo verso un sistema integrato, un vero e proprio framework, avente a disposizione numerosi algoritmi di medio-basso livello, strumenti tramite i quali esso acquisisce informazioni sulla scena inquadrata, sotto forma di valori di probabilità, che sono poi combinati assieme mediante algoritmi di alto livello, che non agiscono più sulle immagini, bensì su modelli di situazioni più complesse. Un sistema del genere è capace di gestire e combinare assieme visione artificiale e qualsiasi altro sensore in grado di indagare l'ambiente circostante. Questo tipo di sistema è oggetto di studio da parte di alcuni

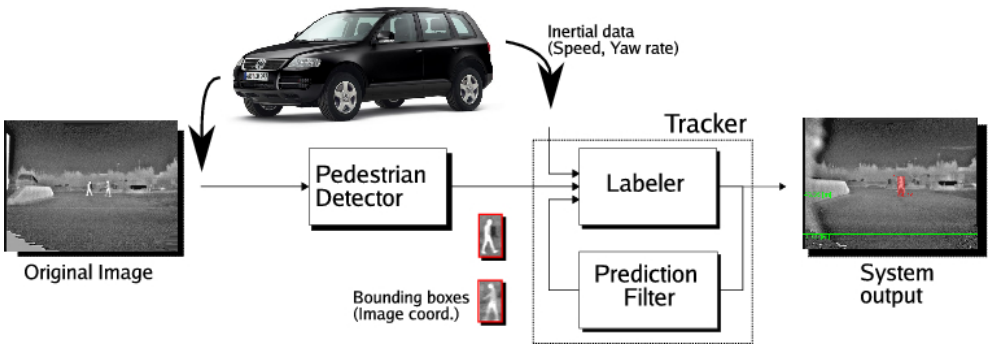
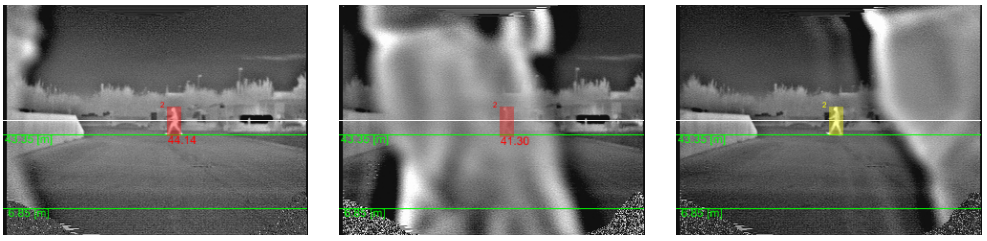


Figura 2.21: Schema del sistema modulare di tracking.

gruppi di ricerca: i primi risultati sono molto interessanti dal punto di vista teorico, anche se sotto il profilo pratico non sono ancora pronti per essere utilizzati, se non altro a causa dei tempi di calcolo molto elevati.

Il riconoscitore di pedoni discusso in [23] è stato utilizzato come base di partenza per lo sviluppo di un sistema modulare che integra funzionalità di tracking, presentato in [54], composto da tre blocchi: il localizzatore dei pedoni, l'etichettatore e il filtro predittivo, come di vede in figura 2.21.

Ogni volta che un pedone è rilevato, in fase di etichettatura si cerca di capire se esso era già presente nelle immagini precedenti: si considerano quindi i pedoni trovati nel passato, e si calcola un match con ciascuno di essi. Tale valore si ottiene considerando la sovrapposizione tra i bounding box che contengono i pedoni nei due fotogrammi, e tenendo presente la previsione del moto fatta nel passato, oltre che le informazioni sul moto del veicolo. Si confrontano anche le dimensioni, mentre le proporzioni non sono utilizzate perché la larghezza di un pedone varia molto a seconda che abbia le gambe unite o aperte, se è visto di profilo. Una volta stabilito se il pedone è associabile a uno di quelli riconosciuti nel passato, si provvede, in caso affermativo, ad aggiornare la descrizione del moto, assegnandogli la medesima etichetta di quello già riconosciuto, oppure a classificare quel pedone come nuovo, nel qual caso esso è dotato di un'etichetta nuova.



(a) (b) (c)

Figura 2.22: Esempio di riassociazione: il pedone trovato in (a) è perso in (b) a causa di un'occlusione, mentre in (c) esso è localizzato di nuovo, e l'etichettatore riesce a riassociarlo con con quello riconosciuto nel passato, e gli riassegna la stessa etichetta.

Mediante il sistema di tracking è possibile mantenere memoria dei riconoscimenti passati, e riassociare l'etichetta al pedone anche se esso è momentaneamente occluso, come si vede in figura 2.22: (a) mostra un fotogramma in cui il pedone è riconosciuto, mentre in (b) esso è temporaneamente occluso, ma l'etichettatore continua a segnalarne la presenza, muovendolo in accordo con il filtro predittore; in (c), infine, esso è riassociato ad un pedone trovato nell'immagine, e gli è assegnata la stessa etichetta di prima.

Misurando le prestazioni, si è visto che il tempo di calcolo non aumenta sensibilmente, mentre il correct detection rate cresce del 4%, a scapito, però, di un netto aumento dei falsi positivi; tuttavia, sarebbe opportuno ridefinire i parametri di misurazione delle prestazioni: nel caso di figura 2.22.b, per esempio, il pedone non è effettivamente presente nell'immagine, ma il sistema di tracking lo rileva ugualmente, perché ritiene che esso sia presente nella realtà, anche se non è visibile nell'immagine. Si tratta quindi di capire se è giusto segnalare pedoni che non sono visibili, o se sia più opportuno mantenere quest'informazione solo internamente alla struttura che si occupa del tracking; in seguito a queste decisioni, si può sviluppare un modo per misurare le prestazioni, ed utilizzarlo per confrontare i vari sistemi.

Questo lavoro si può considerare l'inizio della costruzione del framework cui si è fatto cenno. I possibili ampliamenti sono numerosissimi: per esempio, si potrebbe svi-

luppare un algoritmo che, quando un pedone scompare, effettua un'analisi specifica sull'immagine per verificare se si tratta di un errore; inoltre, un sistema di tracking, come si diceva, può effettuare svariate analisi su ogni pedone, e tenere traccia dei risultati in modo da irrobustire il riconoscimento.

2.7 Rilevamento di pedoni a grande distanza in immagini a bassa risoluzione

Un interessante sviluppo del sistema presentato in [23] consiste nell'aumento della distanza massima di riconoscimento dei pedoni, originariamente fissato in circa 43 m dalla telecamera, fino a 100 m, anche se limitatamente all'ambiente extraurbano. Questo ampliamento è giustificato dal fatto che la velocità dei veicoli aumenta sensibilmente al di fuori dei centri abitati, ed è quindi necessario avvistare gli ostacoli ad una maggiore distanza per mantenere lo stesso livello di sicurezza. Di pedoni fuori dai centri abitati non ce ne sono moltissimi, questo è vero, però sono comunque presenti: per esempio, nei pressi di alcuni locali pubblici lungo le strade extraurbane, o nelle zone appena al di fuori dagli ambienti cittadini, e la loro pericolosità è molto elevata, perché l'illuminazione pubblica è scarsa o addirittura assente.
Per estendere il range di rilevamento dei pedoni, è stato realizzato un modulo specifico, illustrato in [55], che può essere abilitato o meno; esso si concentra solo sui pedoni lontani, oltre i 40 m, e non deve comportare un aggravio eccessivo dei tempi di calcolo del sistema preesistente, che funziona intorno ai 5 Hz. Poiché questo modulo lavora su immagini all'infrarosso lontano, non ci sono problemi di illuminazione, visto che sono i pedoni stessi che emettono la radiazione captata dal sensore (come descritto nel paragrafo 1.1.1). Le difficoltà maggiori sono dovute, invece, alla bassa risoluzione dell'immagine, 320×240, che fa sì che le persone lontane dalla telecamera appaiano come un piccolo gruppo di pixel più chiari, per questo chiamati *hot spot*, come mostrato in figura 2.23 (dentro l'ovale).
Nel range di distanze che si sta considerando, l'altezza dei pedoni decresce velocemente, saturandosi dai 60 m in avanti, come descritto dalla tabella 2.6. Questo

Figura 2.23: Immagine contenente un pedone lontano (evidenziato dall'ovale).

fenomeno è inevitabile: non è possibile usare un'ottica con una focale più lunga, perché si ridurrebbe l'angolo di apertura, limitando il campo d'azione dell'algoritmo di localizzazione dei pedoni più vicini; utilizzare una telecamera con risoluzione più alta, invece, seppur possibile, avrebbe reso il sistema più costoso, e meno adatto ad essere installato a bordo delle automobili; in altre parole, si desiderava estendere le funzionalità del sistema installato a bordo del veicolo di figura 2.1 modificando solamente la parte software.

Distanza dalla telecamera (m)	Altezza pedone (pixel)
20	66
40	30
60	18
80	15
100	13

Tabella 2.6: Altezza di un pedone alto 1,80 m nelle immagini a bassa risoluzione.

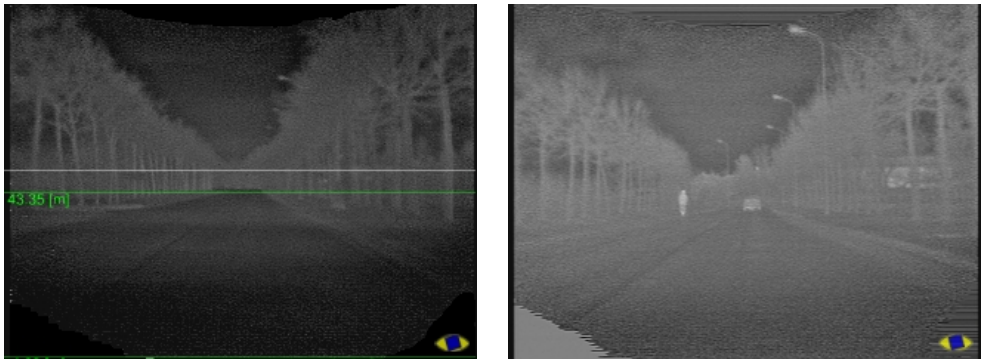


Figura 2.24: Esempio di immagini termiche acquisite in condizioni atmosferiche e di temperatura simili, ma con impostazioni diverse di luminosità.

2.7.1 Analisi dell'immagine e ricerca delle regioni di interesse

Il primo passo che si effettua consiste nell'analisi della luminosità e del contrasto, utile per scegliere alcune soglie; questo procedimento permette al sistema di svincolarsi dal modello e dalle impostazioni della telecamera: si è infatti osservato che sensori diversi hanno delle impostazioni di default di luminosità diverse tra loro, come si vede in figura 2.24, e su questo tipo di sensori, specie quelli non molto raffinati, la regolazione di tali parametri non è sempre possibile o agevole. L'analisi delle caratteristiche dell'immagine permette, entro certi limiti, di funzionare bene indipendentemente dalle condizioni di temperatura dell'ambiente.

La parte centrale dell'algoritmo sviluppato per questo modulo è diviso nelle due fasi

che accomunano molti sistemi di visione: selezione delle aree di interesse, e loro validazione; in questo caso, però, contrariamente a quanto avveniva per la classificazione delle sagome o delle gambe, le regioni di interesse non sono già state trovate. Per la loro ricerca, si prende in considerazione solo una fascia orizzontale dell'immagine, escludendo quindi le zone in cui non può apparire un pedone a grande distanza; questa regione è individuata conoscendo i dati di calibrazione della telecamera, già disponibili perché necessari anche per il rilevamento dei pedoni vicini, e ipotizzando di inquadrare una scena con terreno piatto, assunzione su cui è basata anche la restante parte del sistema.

In effetti, sia la necessità dei dati di calibrazione della telecamera che l'ipotesi di strada piatta sono caratteristiche abbastanza comuni nei sistemi di visione monoculare: ovviamente, si tratta di condizioni che limitano l'efficacia degli algoritmi. Si può tuttavia osservare che la calibrazione smette di essere un problema se si cura il montaggio della telecamera, e si ovvia alle oscillazioni del veicolo applicando un algoritmo di stabilizzazione del flusso video, un argomento ormai abbondantemente sondato dalla ricerca [56, 57, 58, 59, 60], ed utilizzato in molti dispositivi. Per quanto riguarda l'ipotesi di terreno piatto, è applicabile alla maggior parte degli scenari, e gli algoritmi che si basano su di essa non sono perciò troppo limitati, specie se sono progettati considerando una certa tolleranza.

Le fasce individuate nell'immagine sono rappresentate in figura 2.25: esse sono etichettate con le lettere A, B e C. Le ROI sono determinate cercando, in tutte e tre le fasce, i pixel più chiari, e quelli con un tono di grigio più scuro, ma connessi ad una regione chiara. Questa procedura serve a includere tutti quei pixel che contengono le zone meno calde dei pedoni, come il busto, il cui calore è spesso schermato dai vestiti; si tratta quindi di un'operazione simile a una sogliatura, ma un po' meno drastica, che permette di perdere pochi dettagli. Sulle regioni così trovate si calcola l'istogramma per colonne, e si scartano le zone chiare trovate che non hanno una componente verticale significativa; questo metodo è simile all'elaborazione di basso livello descritta in [61], con alcune differenze dovute alle piccole dimensioni dei pedoni cercati.

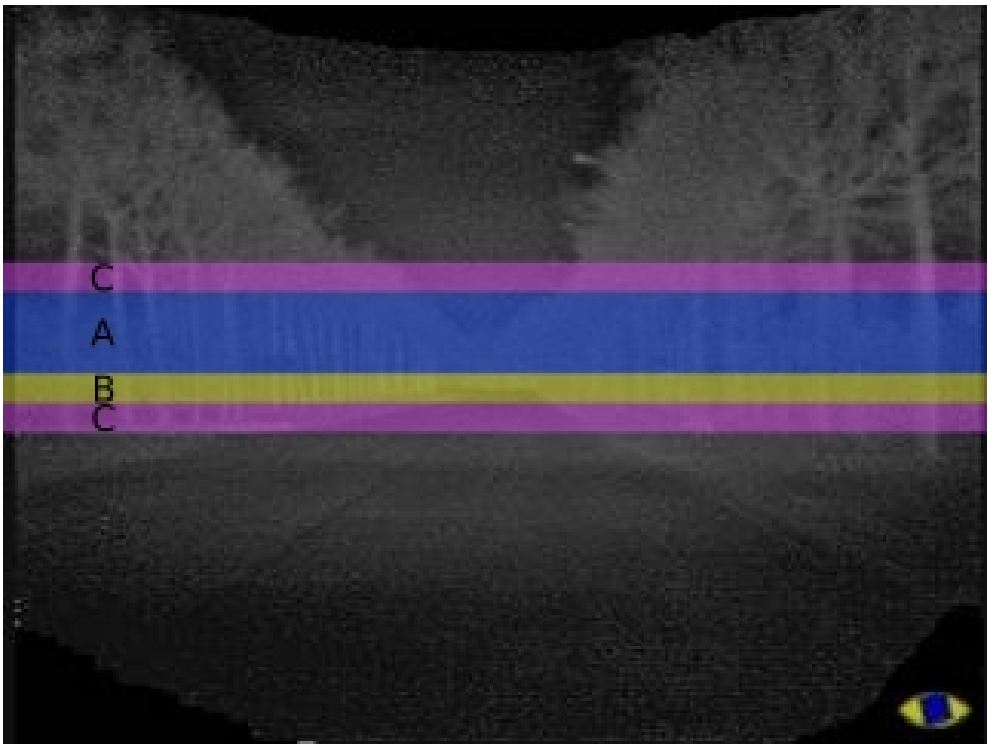


Figura 2.25: Regioni di ricerca dei pedoni lontani. Ogni pedone deve trovarsi nelle regioni A e B per essere localizzato.

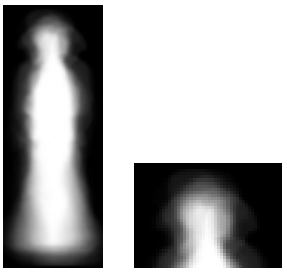
A questo punto avviene la fusione dei dati con l'algoritmo per la localizzazione dei pedoni vicini, che agisce prima di questo modulo: sono eliminate tutte le regioni calde che si sovrappongono ai pedoni vicini già trovati, il che permette di eliminare alcuni falsi positivi, come quelli presenti sulle mani e sui polsi, quando nell'immagine hanno dimensioni confrontabili a quelle di un pedone lontano, e si trovano nelle regioni A o B.

2.7.2 Validazione delle regioni di interesse

Il passo seguente è costituito da una serie di filtraggi molto semplici, in modo da eliminare gran parte delle regioni che non contengono un pedone utilizzando poche risorse di calcolo. Date le piccole dimensioni, non è significativo fare un'analisi della simmetria, un passo di solito presente nella validazione dei pedoni; si procede allora con una verifica sulla posizione. Per essere accettabili, gli hot spot devono trovarsi nelle zone A e B, mentre, se sconfinano nella regione C, sono scartati: in questo modo si eliminano molti falsi positivi, come alberi e pali, e, in generale, tutti gli oggetti di grandi dimensioni. La scelta di cercare gli hot spot non solo nella regione in cui si trovano i pedoni, ma in una lievemente più grande, risponde proprio all'esigenza di capire se un oggetto si espande anche al di fuori delle zone A e B. Sono inoltre scartati anche i bounding box la cui base non risiede nella regione B, perché contengono degli oggetti caldi che non appoggiano per terra, oppure che si appoggiano mediante delle strutture fredde; in entrambi i casi, non si tratta di un pedone. La zona B è scelta con un certo margine di tolleranza, in modo da non eliminare nessun pedone, anche se le gambe non sono molto evidenti nell'immagine, e anche in presenza di un certo beccheggio dell'automobile.

Altre verifiche sono condotte sulle dimensioni del bounding box, che devono essere in un certo intervallo; è tuttavia impossibile verificare che esse siano compatibili con la distanza del pedone, poiché quest'ultimo dato non è disponibile. Nei sistemi di visione monoculare con telecamera calibrata e assunzione di terreno piatto, la distanza degli oggetti può essere calcolata solo trovando il punto di contatto con il terreno, ma, nel caso in esame, esso varia di pochi pixel in tutto il range di funzionamento dell'algoritmo, ed è quindi impossibile pensare di valutare la distanza. Anche le proporzioni sono controllate, e si scartano gli hot spot che non sono sviluppati in verticale; questo rischia di eliminare i gruppi di molti pedoni vicini, che formano un'unica zona chiara sviluppata orizzontalmente, ma è molto raro che le persone si riuniscano in gruppi su strada extraurbana.

Il passo più importante per il rilevamento dei pedoni lontani è costituito da un confronto dei bounding box che hanno superato la fase di filtraggio con un modello.



(a) (b)

Figura 2.26: Modello probabilistico di un pedone (a) e la parte relativa alla testa (b), usata per la validazione.

Trattandosi di sagome molto piccole, e poco dettagliate, si è deciso di utilizzare un modello probabilistico, in analogia con quanto proposto in [62] per immagini infrarosse, e in [63] per lo spettro visibile. Sono state effettuate prove sia con il modello dell'intero corpo (figura 2.26.a), sia con una sua parte, quella contenente solo la testa e le spalle (figura 2.26.b), ed è poi stata scelta quest'ultima soluzione, poiché ha fornito dei risultati migliori.

Il confronto avviene utilizzando la semplice funzione di correlazione seguente:

$$\text{Match} = \frac{\sum_{i,j}\left(d_{\text{mod}}(i,j)\cdot d_{\text{img}}(i,j)\right)}{\sqrt{\sigma^2_{\text{mod}}\cdot\sigma^2_{\text{img}}}} \quad , \tag{2.17}$$

in cui $d_{\text{img}}(i,j)$ rappresenta la differenza tra il pixel in posizione (i,j) dell'immagine e il valor medio di tutti i pixel coinvolti nel calcolo della correlazione, ovvero μ_{image}. Il valore di $d_{\text{model}}(i,j)$ è l'analogo applicato al modello. I valori σ sono definiti come:

$$\sigma^2_{\text{img}} = \sum_{i,j}\left(\text{img}(i,j)-\mu_{\text{img}}\right)^2 \quad , \tag{2.18}$$

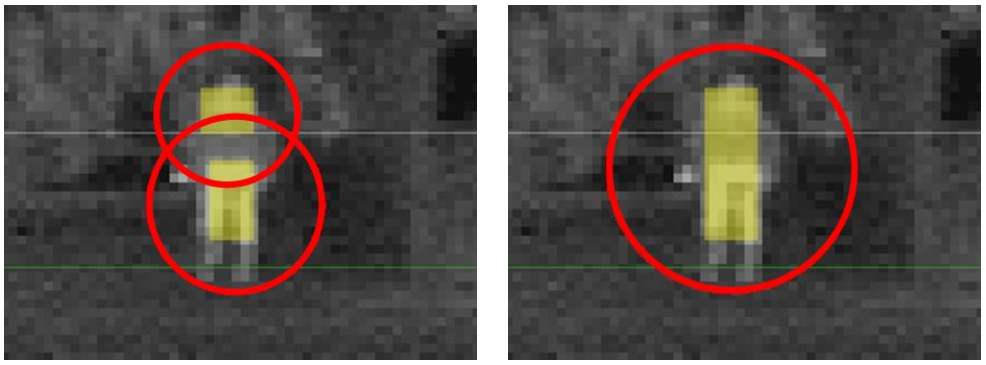
$$\sigma^2_{\text{mod}} = \sum_{i,j}\left(\text{mod}(i,j)-\mu_{\text{mod}}\right)^2 \quad . \tag{2.19}$$

Si noti come il valore del match, per come è definito nell'equazione (2.17), può anche

essere negativo: è il caso di un'immagine che è confrontata con il suo negativo.
Il confronto inizia scalando il modello della testa in modo che abbia la stessa larghezza del bounding box che si desidera analizzare, posizionandolo nella sua parte superiore, e calcolando il valore del match. In seguito, si modificano ancora le dimensioni del modello, e per ogni valore si prova a valutare la correlazione con l'hot spot, per capire qual è la dimensione che fornisce la maggior somiglianza: questa procedura è importante perché non c'è una proporzione fissa tra larghezza del bounding box e dimensione della testa.
Una volta trovata la dimensione con il miglior match, si procede traslando il modello di quella dimensione sia verticalmente che orizzontalmente; dopo ogni traslazione, si salva il valore di correlazione calcolato in quella posizione. Così facendo si crea un pattern di valori di correlazione, ed è stato osservato che quelli generati dai bounding box contenenti un pedone hanno alcune caratteristiche peculiari: i valori decrescono rapidamente a fronte di spostamenti verso l'alto, e meno se la traslazione avviene verso il basso; inoltre, la diminuzione è visibile ma non troppo marcata mano a mano che ci si allontana orizzontalmente dalla posizione centrale, ed è generalmente simmetrica. Gli stessi fenomeni generalmente non si riscontrano quando si ha a che fare con bounding box che non contengono pedoni. Questo modo di procedere richiede molte volte il calcolo della correlazione, però esso è applicato a delle porzioni di immagine davvero piccole, e i tempi di calcolo non sono dunque elevati.

2.7.3 Fusione delle regioni di interesse

Come illustrato precedentemente, l'algoritmo di ricerca delle zone chiare dell'immagine tenta di selezionare anche quelle vicine lievemente più scure, in modo da includere nel bounding box l'intero corpo. Tuttavia, in certe situazioni può capitare che un pedone sia trovato come due hot spot separati, per esempio, durante la stagione invernale, quando il busto è schermato dai vestiti più di quanto non accada per la testa e le gambe. Quando ciò si verifica, i due bounding box sono entrambi scartati nella fase di validazione, a causa delle loro caratteristiche geometriche.
Per ovviare a questo problema si considerano nuovamente gli hot spot scartati, veri-



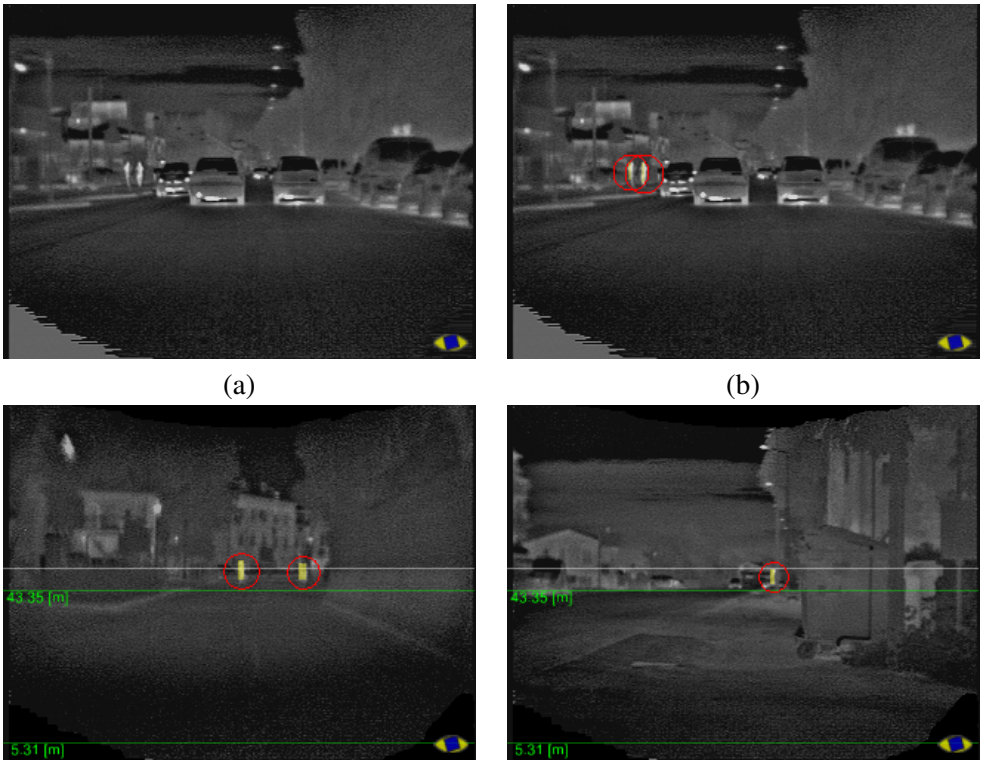
(a) (b)

Figura 2.27: Esempio di fusione delle regioni di interesse: i due hot spot (a), allineati verticalmente perché relativi allo stesso pedone, sono fusi assieme (b).

ficando se ci sono alcune coppie di elementi allineati verticalmente, e di dimensioni simili; ciascuna di esse è poi fusa in un unico bounding box, che è quindi sottoposto alla normale fase di validazione. In figura 2.27 si può vedere un esempio con due hot spot trovati per un unico pedone (a), e il risultato dopo la fusione (b), che è in grado di passare la fase di validazione. Con questa tecnica è stato possibile migliorare decisamente le prestazioni di questo modulo.

2.7.4 Esempi di funzionamento

Alcuni esempi di localizzazione dei pedoni lontani sono riportati in figura 2.28; ogni hot spot è disegnato come un rettangolo giallo con un cerchio rosso attorno, per facilitarne il riconoscimento ad occhio. In particolare, si può vedere l'immagine originale (a), in cui sono presenti due pedoni lontani, e il risultato dell'algoritmo (b), che dimostra un buon funzionamento; in (c), una scena in una zona urbana non troppo complessa, in (d), un'immagine acquisita al di fuori del centro abitato, in prossimità di una trattoria che si trova lungo una statale: il pedone è riconosciuto nonostante sia composto da soli 20 pixel. In figura 2.29 è possibile vedere un ciclista riconosciuto come pedone (a), un effetto dei rilevatori di pedoni solitamente considerato positivo,



(a) (b)

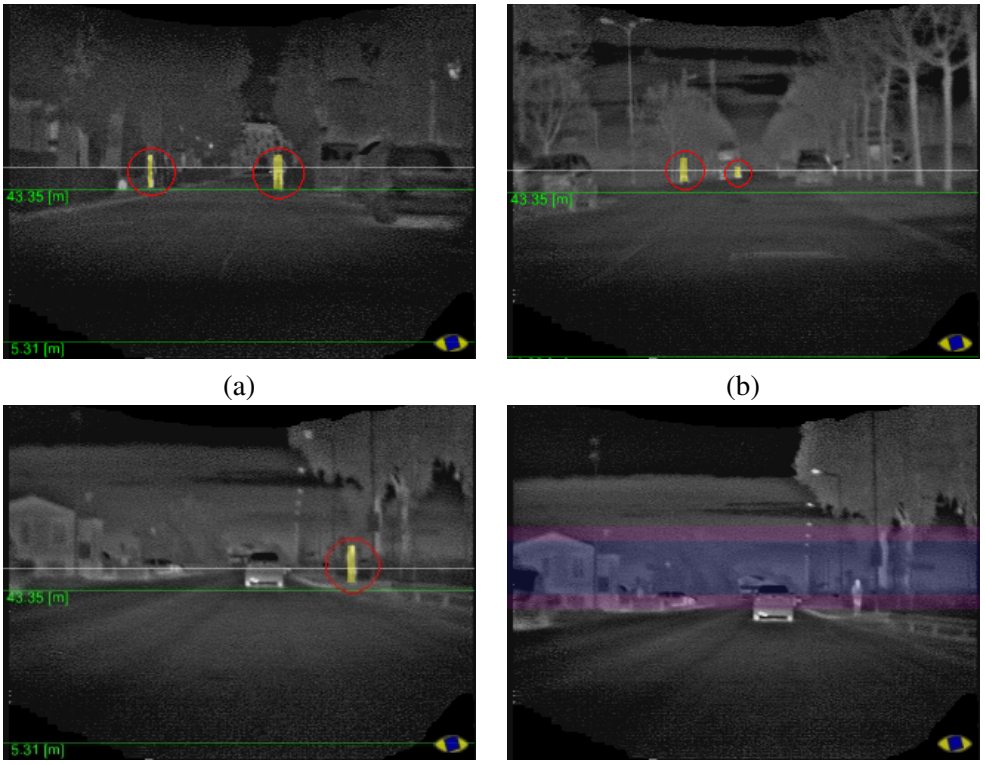
(c) (d)

Figura 2.28: Immagine non elaborata (a): in essa è presente un pedone lontano, riconosciuto dall'algoritmo (b); esempi di funzionamento in scene riprese in periferia (c) e in ambiente extraurbano (d).

o, comunque, non negativo, poiché i ciclisti hanno caratteristiche simili ai pedoni sia per quanto riguarda la loro forma nelle immagini che per quanto concerne la vulnerabilità; in (b) è presente un falso positivo trovato in corrispondenza del faro di un veicolo e del suo passaruota: si tratta del caso più comune, a motivo della forma circolare dei fari e della loro posizione nell'immagine, che è simile a quella di un pedone ad una certa distanza. In (c) si può vedere un pedone lontano localizzato correttamente, mentre in (d) è mostrato un fotogramma acquisito poco dopo, quando il veicolo si trovava su un dosso artificiale, e la calibrazione era momentaneamente errata: si può osservare come il pedone si trovi sovrapposto alla regione C (in viola), il che fa sì che l'hot spot sia scartato nella fase di validazione. Questo effetto, di per sé critico, è comunque risolvibile applicando un algoritmo di stabilizzazione, come già osservato.

Questo modulo è stato testato utilizzando il veicolo sperimentale di figura 2.1, sia in ambito urbano che extraurbano. Fare un calcolo delle prestazioni non è molto semplice, giacché anche una persona può non riconoscere i pedoni così lontani nelle immagini a bassa risoluzione, e diventa quindi difficile ottenere un termine di paragone ritenuto vero con cui confrontare i risultati dell'algoritmo. Ad ogni modo, la valutazione delle prestazioni è stata fatta, basata su 2938 immagini, acquisite sia in ambito extraurbano che in periferia; gli indicatori usati sono due: il primo è il correct detection rate, espresso come CD/(CD+FP), dove CD è il numero di pedoni presenti nelle immagini e correttamente classificati come tali, mentre FP è il numero di falsi positivi; tale valore si attesta a 69,2%. Il secondo indicatore è il numero di falsi positivi per ogni fotogramma, che risulta particolarmente basso: 3,6%; tali falsi, inoltre, non sono persistenti, e tutti i pedoni lontani sono riconosciuti in almeno un'immagine.

La fusione con altri sensori non è stata considerata, poiché il modulo preso in esame è in grado di svolgere un compito praticamente impossibile da portare a termine con altri sensori: un radar in versione automotive con un'apertura ragionevole non rileva i pedoni oltre i 40-50 m; un laserscanner non troppo sofisticato, invece, ha problemi di risoluzione, come messo in luce nel paragrafo 1.3.2.



(a) (b)

(c) (d)

Figura 2.29: Esempi di funzionamento: un ciclista è riconosciuto (a), falso positivo in corrispondenza del faro di un'auto (b), un pedone correttamente riconosciuto (c) e poi perso qualche immagine dopo a causa di un dosso (d).

L'ultima nota riguarda i tempi di calcolo, che sono risultati particolarmente bassi, come richiesto dal sistema di cui questo modulo fa parte: il tempo medio di analisi di un'immagine è di 3 ms su un personal computer dotato di processore Pentium 4 a 3 GHz.

Capitolo 3

Calibrazione automatica per sistemi di videosorveglianza

I sistemi di videosorveglianza intelligenti, cioè dotati di capacità di calcolo per l'esecuzione di algoritmi di visione artificiale, hanno riscosso un notevole successo, sia da parte dei centri di ricerca, che hanno destinato una buona parte dei loro sforzi per lo sviluppo di tali sistemi, sia da parte dell'industria e delle istituzioni pubbliche, che ritengono di poter sfruttare queste tecnologie per risolvere alcuni dei loro problemi. In particolare, la regione Emilia-Romagna ha dato impulso alla ricerca in questo ambito mediante un progetto, denominato LAICA – Laboratorio di Ambient Intelligence per una Città Amica – articolato in diversi sottoprogetti, che riflettono alcune delle esigenze di videosorveglianza di un ambiente cittadino. L'obiettivo di questo progetto, cui hanno preso parte tre università della regione (Parma, Modena e Reggio Emilia, Bologna) era quello di sviluppare:

- un sistema in grado di sorvegliare gli attraversamenti pedonali;
- un sistema per il controllo di un sottopassaggio;
- un sistema per la sorveglianza di un parco;
- un sistema capace di monitorare il traffico.

Il progetto andava anche al di là degli algoritmi di visione artificiale, e prevedeva, per esempio, lo sviluppo dell'infrastruttura di comunicazione e condivisione dei dati.

3.1 Sistemi di videosorveglianza di passaggi pedonali

Un passaggio pedonale è un luogo potenzialmente pericoloso, visto che uomini a piedi si trovano ad attraversare la sede stradale su cui normalmente viaggiano gli autoveicoli. Sebbene il codice della strada assegni la precedenza ai pedoni, i casi di investimento sulle strisce pedonali non sono, purtroppo, eventi molto rari. Un investimento può essere causato dalla distrazione del guidatore, ma anche dall'oggettiva difficoltà di vedere delle persone in condizioni di scarsa visibilità, come in caso di nebbia, di pioggia, o anche semplicemente di notte, visto che i pedoni sono gli unici utenti della strada privi di segnalatori luminosi. Questo ha fatto sì che in molte città siano stati adottati svariati metodi per limitare questo tipo di incidenti: passaggi pedonali illuminati di notte, segnalazioni luminose ai conducenti di veicoli, dossi artificiali per limitare la velocità in prossimità dei passaggi, semafori a chiamata.
Tra i vari componenti di una "città amica" si è quindi pensato di inserire un passaggio pedonale intelligente, cioè sorvegliato da un sistema in grado di comprendere quello che sta accadendo, e di comportarsi di conseguenza; alcuni esempi si trovano in [64, 65, 66, 67]. Nel caso di LAICA, il prototipo di attraversamento sorvegliato è composto da un sistema di visione artificiale per il rilevamento dei pedoni sulle strisce pedonali; il sistema è attivato dai pedoni stessi premendo il pulsante di chiamata, posto sul sostegno del semaforo: non appena scatta il verde per i pedoni, il sistema controlla il loro attraversamento, e ridà il verde alle automobili quando l'incrocio si libera. In questo modo, è possibile ottimizzare i tempi del semaforo, ridando il verde al traffico veicolare in anticipo se i pedoni sono veloci, e, soprattutto, ritardandolo se persone in difficoltà impiegano un tempo più lungo della media per attraversare la strada; si veda [68] per un'analisi approfondita sulle temporizzazioni dei semafori. Questa, comunque, è solo una delle possibili applicazioni: è ovvio che la parte più complessa e interessante del progetto è costituita dal sistema di videosorveglianza,

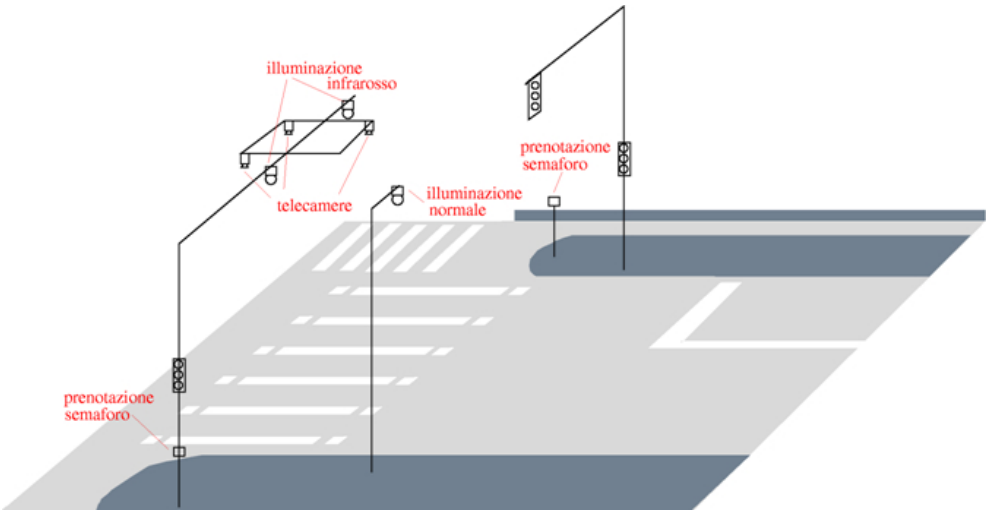


Figura 3.1: Struttura del passaggio pedonale sorvegliato. Le telecamere e gli illuminatori NIR sono posizionati sospesi sopra le strisce pedonali.

avendo a disposizione il quale è poi possibile scegliere quali misure adottare in caso di attraversamento.

Il sistema di visione utilizzato per il prototipo del passaggio pedonale sorvegliato, situato nel comune di Reggio Emilia, è composto da due telecamere sensibili nello spettro dell'infrarosso vicino (NIR) e da due illuminatori; tutta l'attrezzatura è montata sullo stesso supporto che regge il semaforo che sta sospeso sopra la strada. In figura 3.1 è illustrato lo schema dell'attraversamento, in cui sono presenti tre telecamere, inizialmente previste per valutare se fosse preferibile la coppia stereo parallela all'asse dell'incrocio, o quella perpendicolare ad esso. Questa scelta non è ottimale dal punto di vista della scena inquadrata, perché le telecamere non sono al centro dell'incrocio; tuttavia, in questo modo si limita il più possibile l'intervento sull'infrastruttura esistente, rendendo più semplice ed economica l'installazione del sistema.

Nella progettazione di questo sistema si è optato per una soluzione basata sulla visione stereoscopica in modo da eliminare alcuni problemi classici della videosorve-

glianza. Molti sistemi monoculari, infatti, sono basati sulla cosiddetta *background subtraction*, ovvero sul confronto tra la scena ripresa e lo sfondo: quest'ultimo, ovviamente, è ottenuto sempre osservando la scena, e individuando quali sono le zone dell'immagine che non cambiano. Questi algoritmi hanno raggiunto una notevole raffinatezza, ed esistono studi specifici per l'eliminazione di alcuni problemi connessi con questa tecnica, come le ombre e le piante lievemente mosse dal vento [69]; tuttavia, la visione stereo permette di approcciare il problema in maniera diversa, consentendo di rilevare gli ostacoli anche se questi rimangono nella stessa posizione per un tempo indefinito. Questo vantaggio, tuttavia, è conseguito al costo di dover controllare precisamente l'orientamento delle telecamere, un parametro da cui dipende il buon funzionamento degli algoritmi stereo.
I maggiori problemi realizzativi del sistema illustrato, alcuni dei quali dipendono dai vincoli sulle scelte di posizionamento delle telecamere, sono i seguenti:

- l'altezza da terra, vincolata dal palo di sostegno del semaforo, è di 6 m, una distanza che obbliga a scegliere ottiche dalla focale molto corta, quindi molto distorcenti, per poter inquadrare l'intero passaggio;
- l'aver posizionato i sensori non al centro del passaggio impone di ruotare le telecamere, in modo da inquadrare anche l'estremo dell'attraversamento più lontano, il che introduce un effetto prospettico, che non sarebbe stato presente se l'asse ottico delle telecamere fosse stato perpendicolare al suolo;
- il difficile accesso alle telecamere rende impossibile, nella pratica, ottenere un'orientazione precisa per entrambe le telecamere, un requisito fondamentale nei sistemi stereo;
- l'installazione all'aperto sottopone il sistema agli eventi atmosferici, alle variazioni di temperatura e alle vibrazioni dovute al traffico, fenomeni che, col tempo, modificherebbero comunque l'orientazione delle telecamere.

Nel seguito, ci si focalizzerà principalmente sulla tecnica sviluppata per rendere il sistema il più possibile immune dai problemi citati.

3.2 Sistema di calibrazione automatica

Per calibrazione si intende la conoscenza dei parametri di una telecamera [70], come le dimensioni del sensore, la lunghezza focale dell'ottica, la posizione e l'orientazione; tali parametri si dividono in due categorie: quelli estrinseci dipendono dalla posizione e dall'orientazione della telecamera, mentre quelli per cui non vale questa dipendenza sono detti intrinseci. Calibrare un sistema indica l'attività mediante la quale è possibile ottenere i dati di calibrazione.

Per quanto riguarda i parametri intrinseci, poiché dipendono solo dall'hardware impiegato, sono noti a priori e non devono essere misurati; in realtà, tuttavia, l'esperienza pratica mostra come ci siano delle discrepanze dai valori nominali, anche avendo a che fare con hardware di prim'ordine. Per fare un esempio, i valori di apertura orizzontale e verticale dovrebbero essere calcolabili con precisione, una volta note le dimensioni del sensore e la lunghezza focale dell'ottica, ma non è così: i valori misurati sperimentalmente sono diversi da quelli calcolati, e la differenza, seppur piccola, è apprezzabile; questo può essere causato, per esempio, da un accoppiamento dei filetti di ottica e telecamera non molto preciso. Ad ogni modo, rimane il fatto che queste discrepanze dipendono solo dalle caratteristiche costruttive della telecamera, e possono quindi essere misurate una volta per tutte. Un discorso diverso vale per i parametri estrinseci, che devono essere valutati ogni volta che si muove una telecamera. Per questo motivo, quando si calibra un sistema, spesso si effettuano misurazioni solo relative agli angoli di orientazione (si veda lo schema di figura 3.2) e, eventualmente, alla posizione.

Come detto precedentemente, gli algoritmi di visione stereoscopica sono molto sensibili agli errori di calibrazione, perché sono basati sul confronto tra le immagini provenienti da diverse telecamere, le cui posizioni e orientazioni reciproche devono essere note. Per questo, la difficoltà nel regolare precisamente i parametri di calibrazione, e la possibilità di una loro deriva col tempo, giustificano l'adozione di un sistema di calibrazione automatica, capace di calcolare autonomamente e tenere aggiornati i parametri di interesse, evitando l'intervento dell'uomo. In figura 3.3 si possono vedere

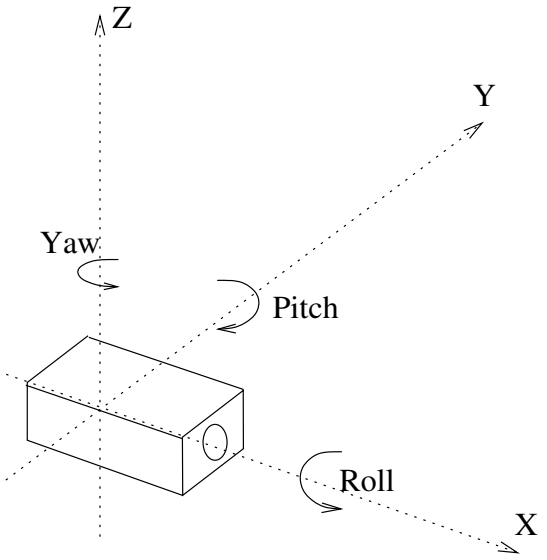


Figura 3.2: Schema che mette in evidenza i tre assi di orientazione di una telecamera.

le immagini acquisite, aventi risoluzione 768×288, poiché è stato tolto un field per eliminare l'interallacciamento; la differenza di orientazione è evidente.

La calibrazione dei sistemi di visione artificiale avviene, di solito, inquadrando un oggetto avente posizione e dimensioni note, e osservando come esso appare nelle immagini. Nel caso in esame, poiché il sistema di videosorveglianza inquadra immancabilmente un passaggio pedonale, ovvero un pattern di forma e dimensioni note, si è pensato di sfruttare questo elemento per realizzare la calibrazione automatica.

L'algoritmo di calibrazione sviluppato per questo sistema si basa su tre passi fondamentali. Il primo ha lo scopo di ricavare un certo numero di punti di calibrazione sulle strisce pedonali, riconoscendo anche in che posizione essi sono, e raggruppando i punti trovati in modo che ciascun insieme contenga solo punti che, nella realtà, sono allineati. Il secondo passo consiste nella rimozione della distorsione introdotta dall'ottica sfruttando i punti trovati. Il terzo, infine, agisce sulle immagini dedistorte, e ricava l'orientazione delle telecamere analizzando l'effetto prospettico presente in

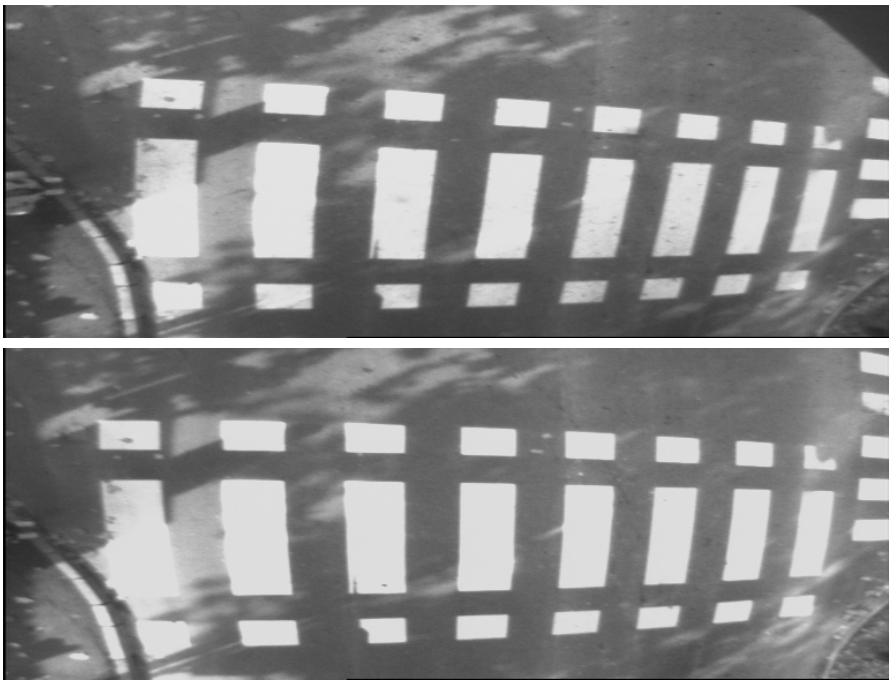


Figura 3.3: Immagini in ingresso al sistema. Sono evidenti la distorsione e la differenza di orientazione, a causa del difficile accesso alle telecamere.

esse; con la conoscenza di questi parametri, e applicando una trasformazione chiamata IPM (Inverse Perspective Mapping) [71], si ottengono delle immagini simili a quelle che sarebbero state ottenute se la telecamera avesse avuto l'asse ottico perpendicolare al suolo. Il sistema di calibrazione automatica, quindi, è in grado di fornire le immagini dedistorte e "dall'alto" indipendentemente dalla distorsione introdotta dall'ottica e dall'orientazione delle telecamere.

3.2.1 Localizzazione dei punti di calibrazione

Come anticipato, il sistema basa la calibrazione automatica su come appaiono le strisce pedonali nelle immagini acquisite; questa analisi è portata a termine osservando la disposizione di alcuni gruppi di punti che, nella realtà, sono allineati, ma che non

appaiono tali nelle immagini. La prima fase, quindi, consiste nel localizzare le strisce pedonali, valutare la bontà dei loro bordi, e selezionare i punti: questo è il compito del cosiddetto *zebra detector*.

Il problema della rilevazione dei passaggi pedonali non è nuovo: alcuni lavori [72, 73] lo affrontano dal più comune punto di vista dell'utente della strada: si tratta, cioè, di algoritmi che rilevano i passaggi pedonali come appaiono dalla prospettiva di una persona che cammina o di una telecamera montata a bordo di un veicolo. Il problema che si sta affrontando in questa sede, tuttavia, è sostanzialmente diverso. Per certi aspetti è più semplice, perché le strisce occupano una buona parte dell'immagine; tuttavia, il risultato che si desidera, in questo caso, non è solo il rilevamento del passaggio, bensì la localizzazione precisa di un buon numero di punti sui bordi delle strisce.

L'individuazione delle strisce inizia con la ricerca delle zone più chiare dell'immagine, tra le quali si selezionano quelle aventi un tono di grigio omogeneo, forma rettangolare e proporzioni appropriate. Successivamente, si esegue l'algoritmo di Sobel per l'estrazione dei bordi, tra i quali si selezionano quelli aventi una certa lunghezza e un piccolo spessore, e si controlla poi quali sono molto vicini ad uno dei rettangoli chiari individuati in precedenza. I primi punti cercati sono i quattro vertici di ciascun rettangolo: essi devono trovarsi vicino al vertice di una striscia bianca, e in una zona in cui un bordo verticale e uno orizzontale si incrociano. Partendo da questi, si procede con la ricerca degli altri punti lungo i lati delle strisce, che devono trovarsi sempre su bordi lunghi, e vicini ad una zona chiara. Quando un vertice non è trovato, la ricerca dei punti lungo i lati avviene ugualmente, espandendo le zone chiare sia in verticale che in orizzontale. Questa tecnica permette di selezionare solo i punti veramente appartenenti ai bordi di una striscia del passaggio pedonale: i vincoli imposti, anche come dettagli di basso livello, sono molti, tutti scelti tenendo conto che è di gran lunga preferibile perdere un punto piuttosto che trovarne uno sbagliato.

I punti localizzati dallo zebra detector sono divisi in gruppi, ciascuno dei quali, come detto, deve contenere punti che nella realtà sono allineati tra loro. Si crea quindi un insieme per ciascun lato lungo di ogni rettangolo, mentre i punti sui lati corti

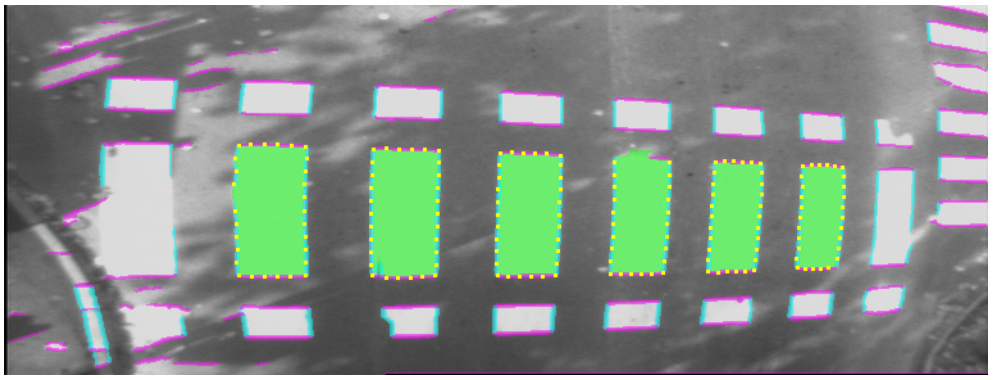


Figura 3.4: Risultato dello zebra detector. I punti di calibrazione (in giallo) sono localizzati sui bordi verticali (in azzurro) e orizzontali (in viola) vicini alle strisce dalla forma regolare (in verde).

dell'intera zebratura sono suddivisi in due insiemi; i punti sui vertici sono considerati due volte, poiché appartengono sia al lato minore che a quello maggiore.
I punti rilevati devono essere in numero sufficiente da poter descrivere sia l'effetto prospettico che la distorsione introdotta dalla lente; l'algoritmo di dedistorsione richiede che essi siano almeno tre per ogni retta, ma, ovviamente, per un buon funzionamento è preferibile un numero maggiore. In figura 3.4 si può vedere il risultato del rilevamento dei punti di calibrazione; la scena è in gran parte sotto l'ombra degli alberi, quindi il guadagno della telecamera è elevato, e causa alcuni fenomeni di saturazione nelle poche zone in cui la luce del sole filtra tra la vegetazione: per questo motivo, le strisce agli estremi, pur essendo rilevate, sono scartate, perché hanno una forma troppo irregolare. Tutte le altre strisce, viceversa, sono utilizzate per la calibrazione, il che è indicato dal colore verde al loro interno; l'azzurro indica le zone classificate come bordo verticale, il viola i bordi orizzontali; i punti gialli, infine, sono quelli che saranno utilizzati per la calibrazione. Partendo dalla striscia più a sinistra si può notare che:

- la seconda ha un punto di calibrazione in meno dovuto ad un'irregolarità locale del bordo sul lato sinistro;

- la terza ha un'imperfezione realmente presente sull'asfalto, e perciò vicino al vertice inferiore sinistro sono presenti due piccoli bordi, che, tuttavia, non compromettono la rilevazione dei punti nella posizione corretta;

- la quinta ha il bordo superiore in una zona in cui l'immagine è saturata: questo compromette il rilevamento dei punti solo sul bordo in questione.

I rettangoli più piccoli, esterni alla zebratura, indicano il passaggio ciclabile, e sarebbe stato conveniente sfruttarli, perché, essendo ancora più lontani dal centro dell'immagine, forniscono una descrizione della distorsione dell'ottica migliore rispetto alle strisce pedonali; si è tuttavia deciso di ignorarli per una questione di generalità: non è sempre presente un attraversamento ciclabile in corrispondenza di un passaggio pedonale.

3.2.2 Rimozione della distorsione

La rimozione della distorsione è stata abbondantemente studiata e formalizzata da tempo; in [74] si trova un ottimo riferimento per l'argomento, in cui sono analizzati e discussi svariati modelli. Per il problema in questione, si è deciso di ricorrere al modello denominato FOV (Field Of View), adatto alle lenti fish-eye utilizzate, che fa dipendere la distorsione cui ogni punto è sottoposto solo dalla sua distanza dal centro dell'immagine, chiamata r; per questo motivo, il FOV appartiene alla categoria dei modelli radiali. Le due funzioni che permettono di passare dalle coordinate nell'immagine distorta a quelle nell'immagine dedistorta, e viceversa, sono le seguenti:

$$r_{\mathrm{d}} = \frac{1}{\omega}\arctan\left(2r_{\mathrm{u}}\tan\frac{\omega}{2}\right) , \tag{3.1}$$

$$r_{\mathrm{u}} = \frac{\tan\left(r_{\mathrm{d}}\omega\right)}{2\tan\frac{\omega}{2}} , \tag{3.2}$$

in cui r_{d} è la distanza del pixel dal centro dell'immagine distorta, mentre r_{u} è la stessa distanza nell'immagine non distorta. Grazie a queste funzioni è possibile rimuovere

la distorsione, una volta determinato l'unico parametro che ne quantifica l'intensità, ω.

Dal punto di vista geometrico, si può dire che la distorsione è eliminata quando tutti i punti che si sa essere allineati nel mondo reale appaiono tali anche nell'immagine: si rende quindi necessario misurare l'allineamento dei punti, operazione effettuata applicando un algoritmo di approssimazione ai minimi quadrati, molto noto anche con il suo acronimo inglese, *least square fitting*. Questo procedimento trova contemporaneamente sia la retta che meglio approssima l'andamento dei punti, sia R^2, ovvero la somma dei quadrati degli scostamenti dei punti dalla retta, un parametro che indica la bontà dell'approssimazione. La misura della distorsione può quindi essere effettuata utilizzando una somma pesata dei valori di R^2 delle rette visibili nell'immagine; i pesi servono per dare più risalto alle rette formate da un maggior numero di punti.

Il parametro ω è trovato a partire da un valore di primo tentativo trovato empiricamente, ed effettuando, poi, una minimizzazione sulla somma pesata dei valori di R^2. Il valore ottenuto è quindi inserito in (3.1) e (3.2) per ottenere un modello di distorsione completamente specificato, sulla base del quale è infine creata una LUT (Look-Up Table) che sarà applicata a tutte le immagini per rimuovere la distorsione: l'effetto è mostrato in figura 3.5, in cui si nota che i punti di calibrazione (questa volta in rosso) sono allineati sulle rette best-fit verdi.

Riguardo all'approssimazione ai minimi quadrati, è opportuno fare una precisazione implementativa importante. Matematicamente, l'algoritmo considera gli scostamenti tra i punti e la retta best-fit, calcolati utilizzando la distanza tra punto e retta, che è definita come la lunghezza del segmento che parte dal punto e arriva perpendicolarmente sulla retta. Le implementazioni dell'algoritmo, tuttavia, per semplicità di calcolo, approssimano tale distanza con lo scostamento verticale, cioè la distanza tra il punto in questione e quello appartenente alla retta e avente la stessa ascissa. Questa approssimazione funziona generalmente bene, a meno che i punti non siano allineati quasi verticalmente, perché in questo caso gli scostamenti verticali hanno dei valori molto elevati, e il risultato della minimizzazione non ha senso. Nel caso delle rette quasi verticali bisogna utilizzare la versione duale, cioè quella che approssima la

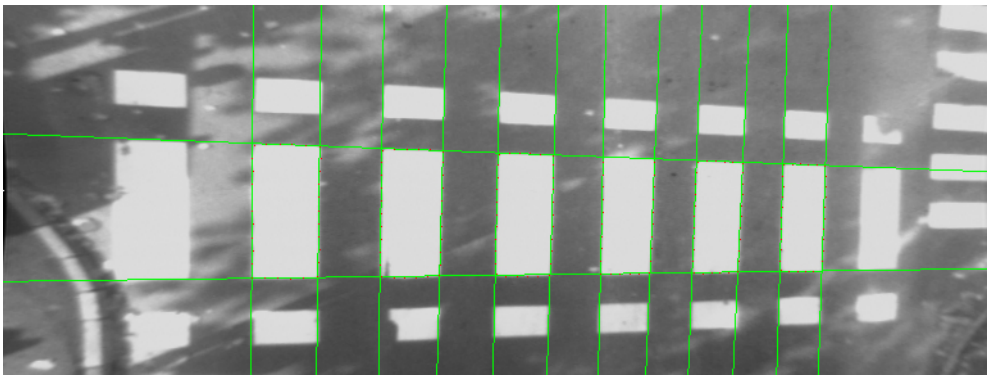


Figura 3.5: Risultato della dedistorsione: i punti di calibrazione (in rosso) sono ora allineati sulle rette best-fit (in verde).

distanza tra punto e retta con la distanza tra il punto e quello appartenente alla retta avente la stessa ordinata.

3.2.3 Rimozione dell'effetto prospettico

Nelle immagini dedistorte è ancora presente la prospettiva, dovuta all'orientazione delle telecamere, i cui assi ottici non sono perpendicolari al suolo. Come già anticipato, per rimuovere l'effetto prospettico si può applicare una trasformazione chiamata IPM, che necessita, però, dei dati di calibrazione, e di conoscere la geometria del suolo, che in questo caso si suppone piatto. Questo procedimento rientra nell'ambito della ricostruzione del mondo tridimensionale a partire dalle immagini, un campo di ricerca su cui esistono moltissimi lavori [70, 75].
Supponendo noti i parametri intrinseci, per l'applicazione dell'IPM bisogna ricavare gli angoli di orientazione della telecamera; ancora una volta, si fa ricorso a come le strisce appaiono nelle immagini per ottenere questi dati. Trascurando inizialmente la rotazione attorno all'asse ottico, si determinano gli altri due angoli, in maniera grossolana, analizzando i punti di intersezione delle rette best-fit trovate al passo precedente. Visto che le rette orizzontali sono due, non ci sono problemi, ma quelle verticali, molto più numerose, ovviamente, non si intersecano tutte in un punto, e si

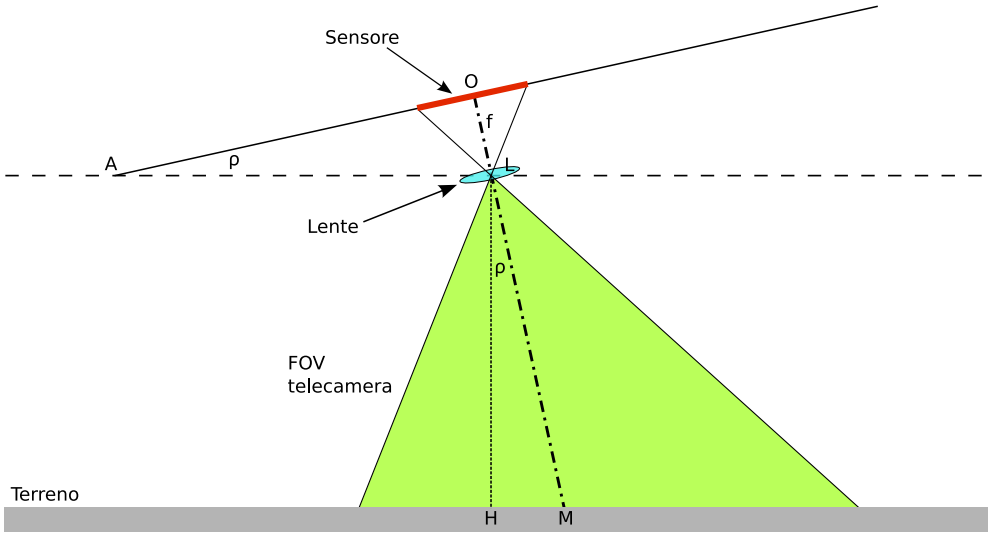


Figura 3.6: Schema che mostra come è possibile valutare gli angoli di orientazione della telecamera analizzando i punti di fuga.

utilizza quindi il baricentro dei punti di intersezione, escludendone alcuni. La necessità di non tenere conto di certi punti di intersezione si capisce considerando il rettangolo che contiene l'immagine, e i due semipiani generati dall'asse di simmetria che lo taglia perpendicolarmente ai due lati minori. Ebbene, può capitare che, per piccoli errori nella valutazione delle rette best-fit, due di esse si intersechino nel semipiano opposto rispetto a quello in cui si dovrebbero intersecare se non ci fossero delle approssimazioni: considerando anche quel punto di intersezione, il baricentro sarebbe totalmente sbagliato.

Conoscendo i punti di intersezione, è possibile calcolare gli angoli di orientazione della telecamera. Per capire come, si faccia riferimento allo schema bidimensionale in figura 3.6, che rappresenta il caso di un solo angolo da valutare; il segmento rosso indica il sensore (fuori scala per esigenze grafiche), OM è l'asse ottico, e ρ l'angolo da stimare. I triangoli LHM e AOL sono simili, perché entrambi rettangoli, e $\hat{OLA} = \hat{LMH}$, perché sono angoli corrispondenti formati da due rette parallele (il suolo e la retta parallela ad essa passante per L) tagliate dalla trasversale OM. L'an-

golo ρ può essere calcolato trovando il punto di fuga A, calcolandone la distanza dal centro dell'immagine O, e applicando la formula:

$$\rho = \arctan\left(\frac{\overline{OL}}{\overline{AO}}\right) \quad , \tag{3.3}$$

in cui $\overline{OL}$ è la distanza focale, che si ritiene nota.

L'accuratezza con cui si calcolano gli angoli dipende da quanto precisamente sono trovati i punti di intersezione; nel caso in esame, l'effetto prospettico è poco marcato, e i punti di fuga sono soggetti a forti variazioni anche per piccoli spostamenti dei punti di calibrazione, quindi gli angoli di rotazione non sono valutati molto precisamente. Per rendere più precisi gli angoli calcolati si ricorre, anche in questo caso, a un processo di minimizzazione, che usa come valori di partenza gli angoli calcolati con il procedimento illustrato. Il parametro su cui si effettua la minimizzazione deve misurare il parallelismo tra le rette: è stata quindi scelta la varianza degli angoli individuati dalle rette best-fit con il semiasse positivo delle ascisse dell'immagine. Le minimizzazioni sono state tenute separate per i due angoli presi in considerazione, perché la minimizzazione su due variabili è tendenzialmente più instabile, e deve essere usata solo quando è necessario agire sui due parametri contemporaneamente; in questo caso, invece, i due problemi sono indipendenti.

L'ultimo passo riguarda l'angolo che è stato trascurato fino ad ora, cioè quello di rotazione attorno all'asse ottico. Per calcolarlo, si sceglie un valore di primo tentativo pari a 0 (cioè nessuna rotazione), e si effettua una minimizzazione su un parametro che misura lo scostamento dall'allineamento delle strisce con l'asse verticale.

Dopo aver calcolato i tre angoli di orientazione della telecamera, si sfrutta la libreria che implementa l'IPM per generare una seconda LUT, applicando la quale si elimina l'effetto prospettico; essa può anche essere combinata con quella in grado di rimuovere la distorsione per generarne una terza, che elimina sia la distorsione della lente che l'effetto prospettico con un'unica trasformazione. In figura 3.7 si vede il risultato delle due trasformazioni applicate in cascata: l'immagine sembra acquisita dall'alto; i bordi irregolari dell'immagine sono causati dalle trasformazioni applicate, che,

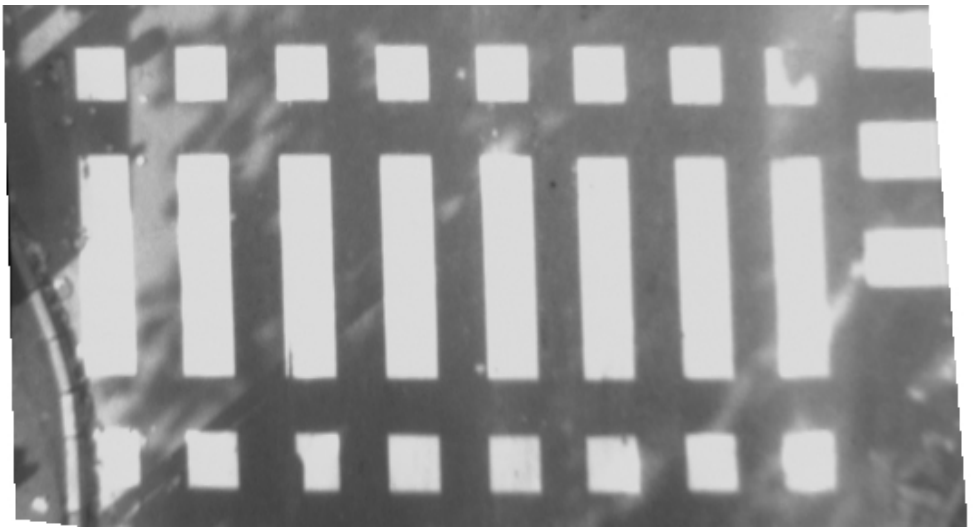
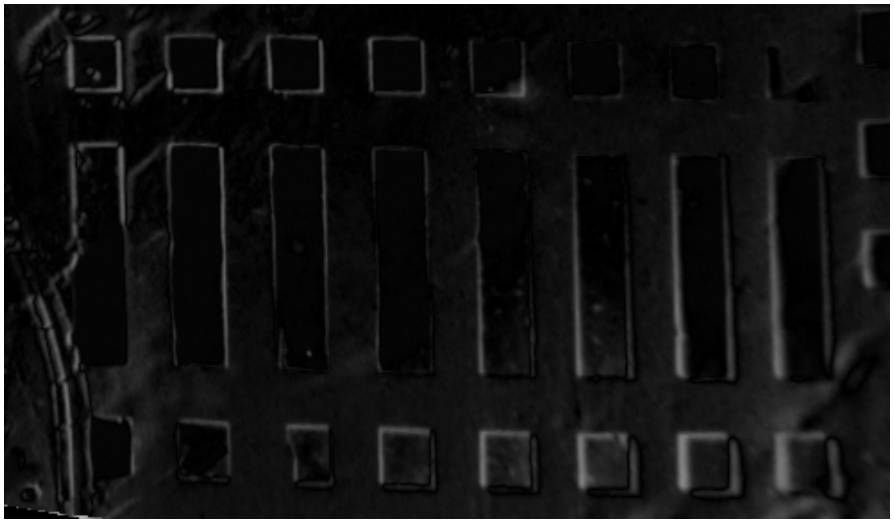


Figura 3.7: Immagine cui è stata applicata la rimozione della distorsione e dell'effetto prospettico: la scena appare come se fosse stata acquisita con una telecamera con ottica non distorcente e con l'asse ottico perpendicolare al suolo.

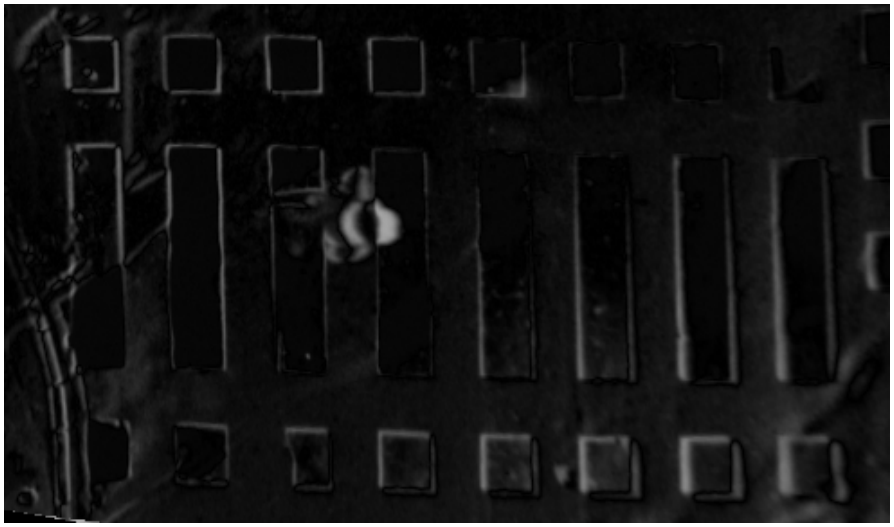
generando una deformazione, rendono obliqui i bordi orizzontali e verticali. Si noti, infine, che gli algoritmi descritti agiscono sulle singole telecamere: il match tra le immagini delle due è ottenuto solo alla fine, quando si effettua una traslazione per fare in modo che le strisce pedonali si sovrappongano.

3.2.4 Risultati

La valutazione delle prestazioni del sistema è stata fatta con due indicatori: nel caso della dedistorsione, è stata utilizzata la media degli scostamenti quadratici, che, nel corso degli esperimenti, ha raggiunto come valore minimo 3,3. Per valutare il risultato della rimozione dell'effetto prospettico, invece, è stata presa in considerazione la varianza degli angoli, cioè il valore oggetto della minimizzazione, che assume valori prossimi allo 0 alla fine del processo. Per una valutazione qualitativa dei risultati dell'intero sistema, compreso l'allineamento tra le due immagini, si veda la figura 3.8,



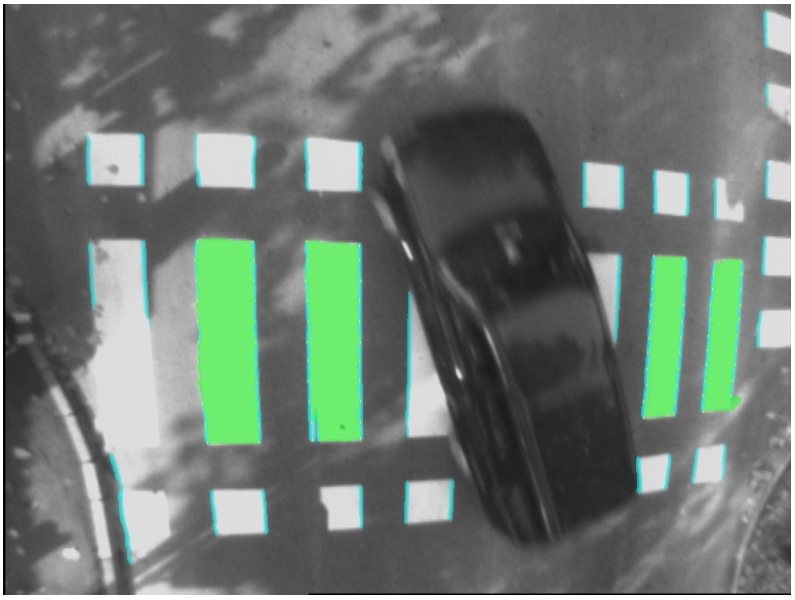
(a)



(b)

Figura 3.8: Differenza tra le due immagini in ingresso al sistema: si nota (a) come la cancellazione delle strisce, delle ombre e di tutti i pattern disegnati sul terreno sia effettuata a meno di pochissimi pixel, mentre gli oggetti che si elevano dal terreno, come un pedone (b) lasciano delle macchie chiare evidenti.

in cui è mostrata la differenza tra le due immagini allineate (a): si può vedere come la cancellazione delle strisce e delle ombre sia affetta da una imprecisione di pochissimi pixel, quindi soddisfacente, se si tiene conto che le immagini di partenza sono quelle di figura 3.3. In figura 3.8.b si vede l'immagine differenza quando è presente un pedone sul passaggio, che lascia la tipica macchia chiara dovuta alla diversa posizione che ogni oggetto che si eleva dal terreno ha nelle due immagini.
Il numero di punti di calibrazione è molto importante per poter ottenere un buon risultato. Nel caso quasi ottimale di figura 3.4, in cui solo le due strisce estreme sono state escluse dalla calibrazione, il numero di punti è elevato: 232 in un'immagine, e 229 nell'altra. Tuttavia, l'informazione sulla distorsione risiede principalmente nei punti più lontani dal centro dell'immagine, perché in essi tale fenomeno è più marcato. In figura 3.9.a si può osservare un caso particolare: un'automobile occlude la visuale della parte centrale del passaggio, per cui solo quattro segnali sono usati per la calibrazione, il che diminuisce sensibilmente il numero di punti trovati, che diventano 129. Ciò nonostante, la calibrazione fornisce comunque un buon risultato, come si vede qualitativamente dall'immagine differenza (figura 3.9.b), anche grazie al fatto che i pochi punti di calibrazione sono concentrati in zone lontane dal centro dell'immagine.
Nel sistema di calibrazione automatica è presente un modulo che valuta la qualità del risultato raggiunto ad ogni passo: esso controlla il numero di punti di calibrazione trovati, e misura i risultati delle fasi di dedistorsione e rimozione dell'effetto prospettico, in modo tale che, al termine del processo, si può decidere se la calibrazione è efficace, o se è opportuno rifarla per ottenere risultati migliori.
Poiché questo sistema deve compensare anche i movimenti cui le telecamere sono soggette con il passare del tempo, si è pensato di effettuare una calibrazione a intervalli regolari, come giorni o settimane.



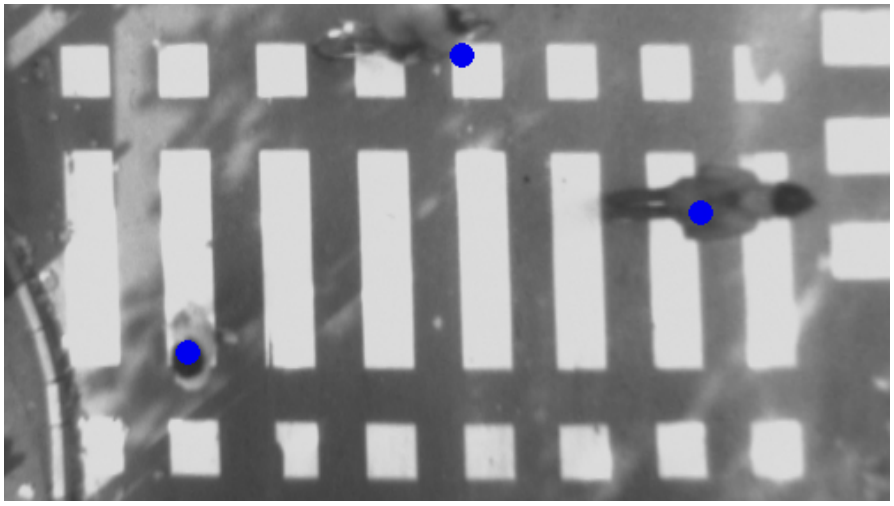
(a)

(b)

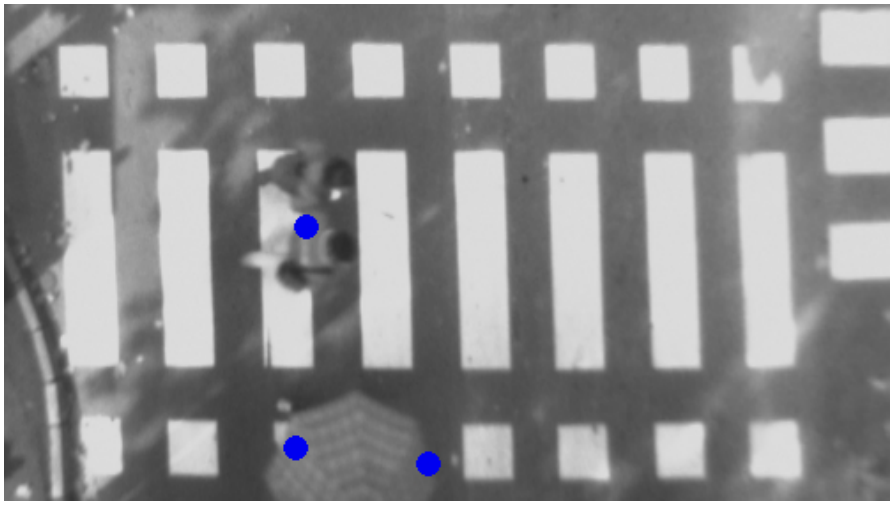
Figura 3.9: Nella scena ripresa, un'automobile occlude la visuale della zebratura, e il sistema di autocalibrazione seleziona 129 punti su solo quattro strisce (a). Il risultato è comunque accettabile (b), poiché i punti sono lontani dal centro dell'immagine, e la descrizione della distorsione è comunque efficace.

3.3 Sistema di rilevamento pedoni

Il sistema di rilevamento dei pedoni sui passaggi pedonali, nel suo complesso, offre delle buone prestazioni, e può contribuire all'aumento della sicurezza stradale. Sulle immagini ottenute dopo la calibrazione agisce un algoritmo di rilevamento pedoni che si basa sull'analisi dell'immagine differenza, come quella di figura 3.8.b, sul confronto con lo sfondo, e sull'osservazione della variazione del tono di grigio di alcune regioni dell'immagine, che si comportano come dei sensori. Questo algoritmo è basato su principi completamente diversi da quelli di cui si è parlato nel capitolo 2, perché, nelle immagini acquisite dall'alto, i pedoni appaiono con una forma differente: si vedono principalmente la testa e le spalle, mentre le gambe sono visibili solo quando sono ripresi di lato. La distinzione tra pedoni e altri oggetti è basata sull'analisi del loro moto, e non sulla morfologia, né sulle dimensioni; questo approccio si è rivelato più efficace e veloce di altri, più classici, basati su come i pedoni appaiono nelle immagini.
Un esempio del funzionamento dell'algoritmo si trova in figura 3.10, in cui si può osservare che sono rilevati anche i ciclisti (a), e le persone con l'ombrello (b): per quest'ultimo caso, la classificazione basata sul moto anziché sulla forma si è rivelata molto efficace, perché una persona con l'ombrello si muove nello stesso modo di un pedone normale, ma ha un aspetto completamente diverso, almeno se ripreso dall'alto. I tempi di calcolo sono molto rapidi, e consentono un'elaborazione alla frequenza massima di 23 Hz su un processore Pentium 4 a 2,80 GHz, pienamente compatibili con le esigenze di un sistema di visione.
Per ulteriori dettagli sull'algoritmo di rilevamento dei pedoni, che qui non si descrive in maniera dettagliata, si rimanda a [76].



(a)



(b)

Figura 3.10: Esempi di localizzazione dei pedoni (indicati dai punti) sul passaggio pedonale. Si può notare che sono localizzati anche ciclisti (a) e pedoni con l'ombrello (b).

Capitolo 4

Rilevamento dei veicoli in fase di sorpasso

In questo capitolo è descritto un sistema per il rilevamento di veicoli in fase di sorpasso, in grado fornire le stesse informazioni che ogni guidatore si procura osservando lo specchietto retrovisore. Questo sistema si inserisce in un contesto particolare, quello del progetto "TerraMax", che prevede l'allestimento di un veicolo totalmente automatico, in grado di destreggiarsi nel traffico urbano. TerraMax, il camion visibile in figura 4.1.a, è stato progettato per la partecipazione ad una gara, la "DARPA Urban Challenge", svoltasi il 3 novembre 2007 in una base militare americana parzialmente abbandonata, nei pressi di Victorville, CA; l'obiettivo era quello di portare a termine un percorso di quasi un centinaio di chilometri in ambito urbano, sapendosi muovere nel traffico, costituito sia dagli altri veicoli robotizzati, sia da un gran numero di automobili guidate da persone.
Le attività che è stato necessario automatizzare sono molte; di seguito si riporta una prima distinzione fatta per aree funzionali.

Sensoristica: è la parte, sia hardware che algoritmica, che si occupa di percepire il mondo; comprende i sistemi di visione artificiale, i laserscanner, il GPS, il sensore inerziale (INS).

(a)

(b)

Figura 4.1: Il veicolo autonomo TerraMax (a): le telecamere per il rilevamento dei veicoli in fase di sorpasso si trovano a fianco della parte posteriore della cabina di guida, sotto al ricevitore GPS. Una buona parte dei sistemi è alloggiata sotto al sedile dei passeggeri (b).

Navigazione: è composta dai sistemi che analizzano il percorso da compiere e calcolano la traiettoria, la velocità, il comportamento da tenere agli incroci, decidono se superare o incolonnarsi, e tutto quanto concerne il comportamento su strada. Questa funzione si basa sulla conoscenza della mappa, sulla capacità di localizzarsi in essa, e tenendo conto dei dati sull'ambiente circostante fornite dai sensori.

Attuazione: è l'insieme di tutti gli attuatori e gli apparati che li controllano, che permettono ai sistemi di gestire il camion mediante impulsi elettrici, chiamato anche X-by-wire.

La visione artificiale ha la responsabilità, assieme ai tre laserscanner, di dare una descrizione dell'ambiente circostante con un livello di dettaglio sufficiente a muoversi in sicurezza in una città. Questo è reso possibile grazie a quattro sistemi, specificamente sviluppati per TerraMax:

Trinocular: analizzando le immagini provenienti da tre telecamere poste dentro alla cabina, rileva gli ostacoli frontali e le linee dipinte sulla strada;

Stereo: si focalizza sull'area più vicina al camion, sia frontalmente che posteriormente, per segnalare gli ostacoli;

Lateral: è stato pensato per cercare i veicoli agli incroci, anche quelli al di fuori della portata del laserscanner;

RearView: lavora su immagini che inquadrano posteriormente le corsie alla destra e alla sinistra del veicolo; deve rilevare i veicoli in fase di sorpasso.

Questi sistemi fanno uso di 11 telecamere alloggiate in punti strategici del veicolo, e collegate a quattro computer ospitati sotto al sedile dei due passeggeri, come si vede in figura 4.1.b.
Il sistema che sarà descritto in questo capitolo è il RearView[1], le cui due telecamere sono montate dietro alla cabina di guida, ruotate di 90°, in modo da far comparire

[1] Questo studio è oggetto della tesi di laurea di Giuliano Maccherozzi.

nelle immagini anche gli oggetti molto vicini, e puntate verso il retro del camion; le inquadrature hanno una piccola zona di sovrapposizione, vicina all'orizzonte, ma si ha a che fare con due sistemi monoculari. Le immagini acquisite hanno risoluzione 384×512 e sono nello spettro visibile, a colori.

4.1 Rilevamento di ostacoli in movimento

L'obiettivo del sistema RearView è quello di rilevare i veicoli in fase di sorpasso, un problema studiato anche da altri gruppi [77], anche se su immagini riprese da una prospettiva differente. L'obiettivo dell'algoritmo è, in linea di principio, qualunque oggetto che si muove con una certa velcità sulla strada, fino a una distanza di 40 m, trascurando quelli che viaggiano nel senso di marcia opposto, e tutti gli altri ostacoli fermi. La caratteristica saliente diventa quindi il moto dell'oggetto, che deve essere estratto analizzando il flusso ottico degli oggetti, e conoscendo la velocità del camion. Di ogni veicolo trovato deve anche essere fornita la distanza, in modo che il sistema di navigazione possa decidere se è comunque possibile sorpassare, oppure no.
Per estrarre il flusso ottico, cioè le informazioni relative al movimento degli oggetti, si trasforma inizialmente l'immagine in ingresso usando un'omografia, e, sull'immagine così modificata, si procede con una suddivisione in regioni di colore uniforme, sulle quali agisce un algoritmo di tracking che estrae il flusso ottico.

4.1.1 Calibrazione e omografia

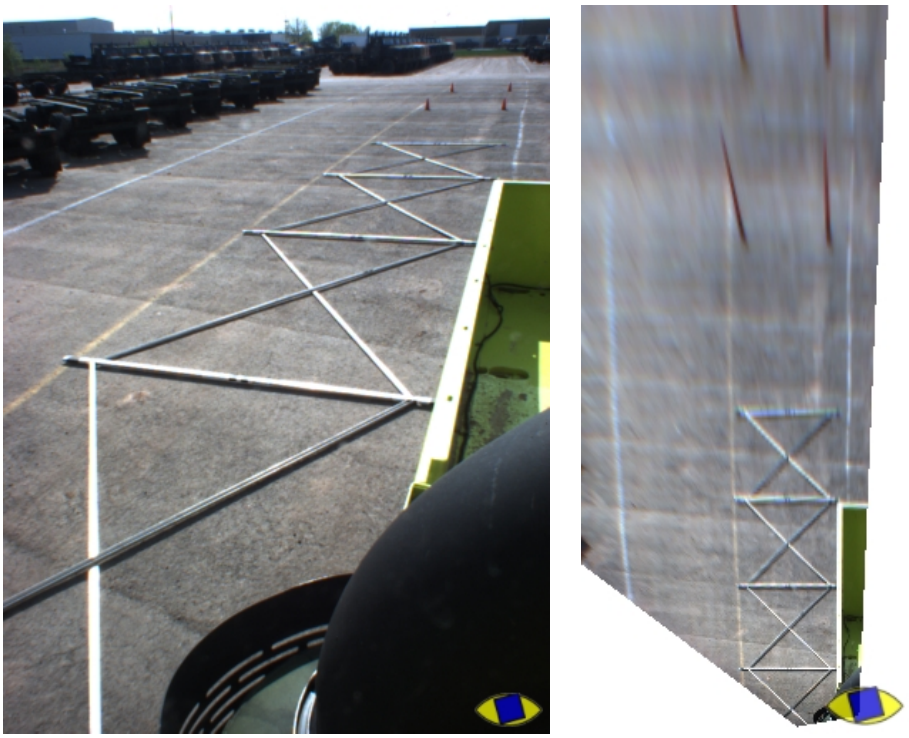
L'omografia è un'operazione che modifica la prospettiva da cui un piano è visto. Essa permette quindi di rimappare un certo numero di punti di un'immagine in altre posizioni, scelte a piacere, deformando di conseguenza tutto il resto dell'immagine. Nel caso specifico, si è fatto ricorso all'omografia per mappare la zona inquadrata dalle telecamere, fino alla distanza di 40 m, in modo che essa appaia come se fosse stata acquisita dall'alto. Si tratta di un procedimento un po' diverso da quanto descritto nel paragrafo 3.2.3, nel quale la rimozione della distorsione è basata sull'IPM, una trasformazione che fa uso del modello della telecamera; in questo caso, infatti,

l'omografia non fa uso di modelli, ma solo di tecniche di interpolazione applicate ai pixel dell'immagine.

Anche se basato su un principio diverso, il risultato finale dell'omografia è abbastanza simile a quello ottenibile con l'IPM; inoltre, il calcolo di questa trasformazione è strettamente connesso con la cosiddetta "calibrazione" del sistema. In questo caso non si tratta di una calibrazione vera e propria, perché non porta a conoscere i parametri di orientazione e posizione della telecamera, ma di una procedura che permette di ottenere la posizione nel mondo di un oggetto localizzato nell'immagine. Questo è possibile avendo a disposizione un fotogramma di una griglia che si trova in una posizione nota, scegliendo dei punti su di essa, e creando un'omografia che mappa quei punti in posizioni opportune, scelte sulla finestra di output. Per realizzare la trasformazione omografica sono necessari almeno quattro punti, ma un numero superiore permette di ottenere una trasformazione più accurata. Si riesce così, con una certa tolleranza, a risalire alla posizione nel mondo di ogni punto dell'immagine omografica.

La scelta della griglia di calibrazione dev'essere fatta in modo che vi sia una sufficiente concentrazione di punti che copre tutta la zona di interesse; ragioni pratiche consigliano di utilizzare almeno una decina di punti. Per il RearView, la griglia è quella che si vede in figura 4.2.a: i punti in questione si trovano agli estremi delle quattro barre di ferro parallele tra loro, e alla base dei coni rossi. La griglia e i birilli sono posizionati in modo che sei punti siano disposti lungo la retta parallela all'asse del veicolo che passa per la proiezione del sensore della telecamera sul terreno, e distanti dal punto di proiezione 5, 10, 15, 20, 30 e 40 m; gli altri sei sono su una retta parallela alla precedente, e distante da essa 3 m. Questa disposizione si è rivelata sufficiente per ottenere una trasformazione di buona qualità, come si vede in figura 4.2, caso (b), che mostra l'immagine (a) trasformata.

Poiché tutto l'algoritmo è basato sull'immagine omografica, i risultati rimangono gli stessi anche se le telecamere sono mosse, a patto di effettuare nuovamente una calibrazione.



(a) (b)

Figura 4.2: La griglia di calibrazione (a) e la stessa dopo l'applicazione della trasformazione omografica (b).

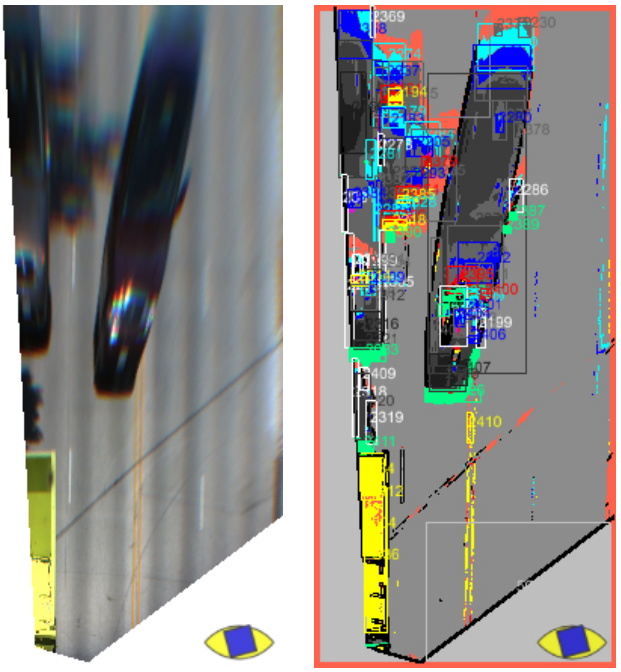
4.1.2 Clusterizzazione sul colore

Per riconoscere gli oggetti presenti sulla strada, si ricorre solitamente ad una loro classificazione, ovvero si definiscono dei criteri che devono essere soddisfatti per poter stabilire che un oggetto appartiene ad una certa categoria. Nel caso del sistema RearView, quest'esigenza in gran parte viene meno: poiché il moto degli oggetti presenti sulla scena deve comunque essere valutato, è lecito ritenere che, se qualcosa è in movimento sulla strada, questo sia un veicolo, e ulteriori classificazioni che vanno al di là di un semplice controllo delle dimensioni diventano superflue, visto

che, di qualunque veicolo si tratti, l'unico dato importante per decidere se è possibile cambiare corsia è la distanza a cui esso si trova. Quest'osservazione permette di sviluppare l'algoritmo cercando solo "oggetti" sulla strada.
La ricerca degli oggetti nella scena comincia dividendo l'immagine in zone di colore simile: i 16581375 colori possibili sono ridotti a 17, e l'immagine appare come in figura 4.3.b, ovvero come un insieme di cluster dalle forme più disparate. Per ciascuno di essi si considerano alcune caratteristiche geometriche:

- il colore;
- le proporzioni;
- le dimensioni, con particolare attenzione a quella verticale;
- il numero di pixel da cui è composto, e il suo rapporto con l'area del rettangolo che lo contiene;
- il rapporto tra area e perimetro;
- la posizione del baricentro dei pixel rispetto al centro geometrico del rettangolo circoscritto.

Sulla base di tali proprietà è possibile selezionare solo i cluster che possono rappresentare un veicolo o una parte di esso, escludendo gli altri: questo filtraggio si rende necessario perché tenere in considerazione tutti i gruppi di pixel, che sono alcune centinaia in ogni fotogramma, sarebbe inutile dal punto di vista algoritmico, e dispendioso da quello computazionale. In figura 4.3.b si vedono i cluster selezionati dalla fase di filtraggio sovrapposti all'immagine a zone di colore uniforme.
Per migliorare la stabilità del riconoscimento degli oggetti, si inserisce a questo punto l'analisi delle ombre, molto utile, e utilizzata in numerosi sistemi nella fase di elaborazione di basso livello [78, 79]. Le ombre sono infatti presenti, più o meno marcate, alla base di ogni oggetto sulla strada, e riconoscerle è relativamente semplice, perché si presentano come gradiente, dallo scuro al chiaro, di dimensione variabile.



(a) (b)

Figura 4.3: Un fotogramma trasformato con l'omografia (a) e la stessa immagine dopo la clusterizzazione sul colore (b); i rettangoli sono quelli circoscritti ai cluster dopo la fase di filtraggio.

Ovviamente, non tutti i gradienti sono ombre, però questa distinzione è lasciata all'analisi sul movimento. Ciò che accade in fase di clusterizzazione è semplicemente un'aggiunta di alcuni cluster nelle zone in cui si osserva un gradiente di una certa lunghezza.

4.1.3 Calcolo del flusso ottico

Il flusso ottico è un concetto noto da molto tempo, e vi sono numerosi metodi per calcolarlo [80, 81, 82, 83, 84]. Il risultato che tutte le tecniche sviluppate desiderano ottenere è un insieme di vettori che descrivono come si sono mossi gli oggetti nelle

varie zone dell'immagine. Nella scelta del metodo con cui calcolare il flusso ottico, bisogna tenere presente che alcuni algoritmi non consentono un'elaborazione rapida, e non sono perciò utilizzabili in questo contesto.

Nel sistema RearView, gli elementi che permettono di valutare il movimento degli oggetti nell'immagine sono i cluster rimasti dopo la fase di filtraggio. Quando si analizza un nuovo fotogramma, i cluster trovati in esso sono confrontati con quelli dell'immagine precedente, e si creano delle associazioni tra i due gruppi sulla base della posizione e della somiglianza tra i valori dei parametri utilizzati per il filtraggio. Ad ogni associazione creata corrisponde un vettore di moto, relativo all'oggetto contenuto nel cluster cui quel vettore si riferisce; inoltre, si tiene traccia di alcuni valori: un numero identificativo (*tracking ID*) che distingue quel cluster da tutti gli altri, gli spostamenti avvenuti nel passato, e il numero di fotogrammi per cui lo stesso ID è stato assegnato ad un cluster nell'immagine.

La fase di associazione dev'essere estremamente curata, perché ogni errore crea dei falsi movimenti, e può portare alla localizzazione di ostacoli inesistenti. Dal punto di vista implementativo, la funzione che calcola il valore del match tra due cluster è valutata per ogni coppia di elementi, e i valori ottenuti, ordinati, sono salvati in un'apposita lista, assieme agli identificatori degli elementi cui il match si riferisce. Quando tutte le coppie sono state analizzate, si scorre la lista, creando le associazioni fra i cluster quando nessuno dei due è ancora stato associato, purché il valore di match sia superiore ad una certa soglia; in questo modo si evita che i legami dipendano all'ordine in cui si considerano le coppie, cosa possibile quando vi sono dei cluster associabili con più di uno nell'altro gruppo. Si noti che tutti i parametri che entrano nel calcolo del match dipendono dalla frequenza di esecuzione dell'algoritmo: ogni blob, con il mutare della scena, cambia forma e posizione, e la differenza tra due immagini consecutive dipende dal tempo trascorso tra le acquisizioni.

Alla fine della fase di associazione sono disponibili i vettori movimento relativi alle zone dell'immagine di colore omogeneo: è quindi possibile individuare gli oggetti che si trovano nelle corsie adiacenti a quella su cui marcia il camion, che viaggiano nello stesso senso di marcia, e che hanno una velocità sostenuta, e quindi possono

costituire un pericolo nel caso si decida di effettuare un cambio di corsia. I veicoli sono solitamente scomposti in più cluster vicini, che sono quindi uniti in modo da segnalare un unico ostacolo al sistema di navigazione.
Il tracking insito nell'analisi del flusso ottico è sfruttato anche per limitare i falsi positivi sporadici: prima di segnalare la presenza di un ostacolo, si osserva il movimento per un certo numero di fotogrammi, in modo da irrobustire la localizzazione; questo, ovviamente, introduce un certo ritardo nel rilevamento, limitato però ad alcuni decimi di secondo.

4.1.4 Fusione con i dati laserscanner

Il TerraMax è dotato di tre laserscanner, due installati nella parte frontale del veicolo, e il terzo in quella posteriore; i dati provenienti da quest'ultimo sono stati utilizzati per effettuare una fusione con i risultati dell'algoritmo descritto. In questa fase, per ogni ostacolo rilevato dal sistema di visione, si verifica se ci sono dei dati provenienti dal laserscanner nelle vicinanze: in caso affermativo, si provvede a spostare l'ostacolo rilevato tramite la visione alla distanza misurata dal laserscanner; in caso contrario, non accade nulla, e gli ostacoli rilevati sono inviati al sistema di navigazione.
Questa tecnica è stata pensata in modo tale che eventuali falsi del laserscanner non si riflettano negativamente sul sistema di visione. Sebbene la misurazione laser sia molto affidabile, infatti, in alcune situazioni si creano del falsi positivi, come gli ostacoli rilevati quando il camion percorre una strada sterrata: la polvere alzata dalle ruote, se è molto densa, è rilevata come un ostacolo con una forma in continuo cambiamento, che insegue il sensore. Anche quando il laserscanner non trova alcun ostacolo, le rilevazioni del sistema RearView non sono rimosse: altrimenti, un malfunzionamento del laser comprometterebbe anche il sistema di visione.

4.2 Risultati ottenuti

Il sistema RearView è stato concepito per rilevare solo alcuni degli ostacoli presenti nella scena, sulla base del loro movimento, tralasciando tutti gli altri. Per questo

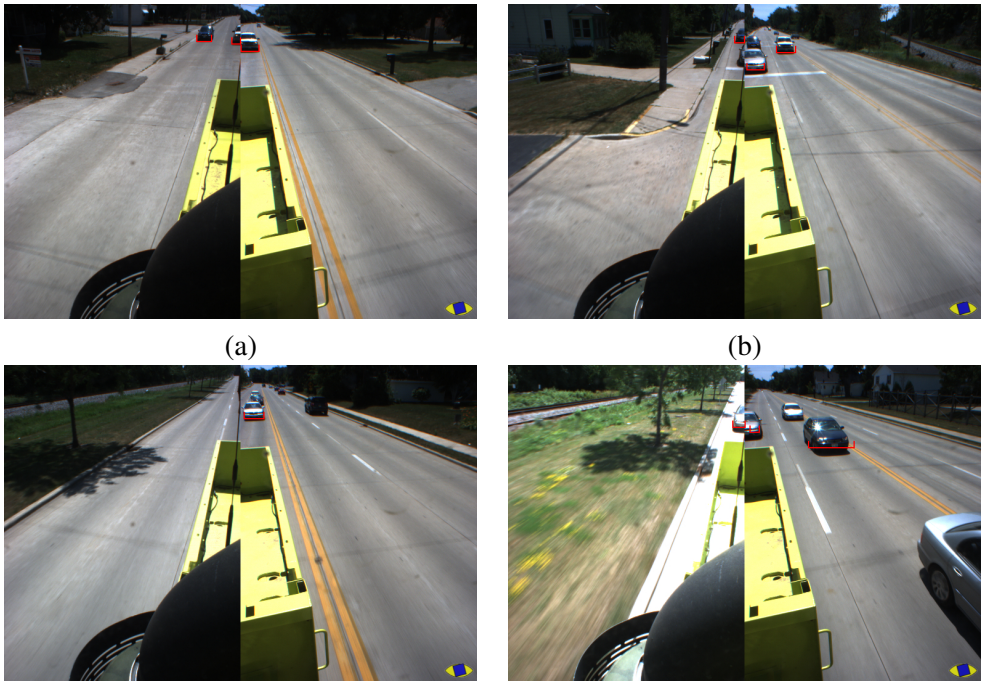
motivo non è stato possibile valutarne numericamente le prestazioni, operazione che avrebbe richiesto molto lavoro per creare il *ground truth*, ovvero l'insieme degli eventi ritenuti veri, generato manualmente da una persona, con il quale si confrontano i risultati forniti dall'algoritmo.

Da un punto di vista qualitativo, comunque, i risultati sono buoni. Sono presenti alcuni falsi positivi, nonostante il tracking, dovuti a cluster poco definiti, che si trovano quasi sempre fuori dalla sede stradale, e che hanno una componente di moto verso il camion per qualche immagine. Analizzando le sequenze di test, si è osservato che tutti i veicoli sono stati rilevati nella maggior parte dei fotogrammi; i falsi negativi sono sporadici, e concentrati alla massima distanza di localizzazione, quando il veicolo non appare ancora ben definito nell'immagine omografica. Nel caso di ostacoli molto vicini al camion, viceversa, cioè nella regione di spazio in cui escono dalla visuale della telecamera, la localizzazione avviene, ma basata su cluster trovati su parti dell'auto come la fiancata o il tetto: poiché tale parte del veicolo viene considerata come se fosse il frontale, la posizione dell'ostacolo risulta errata, ma tale da rendere comunque impossibile il sorpasso. Il beccheggio del camion non costituisce un particolare problema, perché causa solo una lieve alterazione nella valutazione delle distanze degli ostacoli.

L'algoritmo è risultato molto veloce da eseguire, grazie alla leggerezza della fase di tracking. I tempi di calcolo complessivi sono dell'ordine dei 50 ms sul computer installato sul TerraMax, un Core 2 Duo T7200 che lavora alla frequenza di 2 GHz; tale tempo è lo stesso anche se uno solo dei due sistemi mono (destro e sinistro) che costituiscono il RearView è fatto funzionare: questo perché essi sono indipendenti, e sul veicolo erano posti in esecuzione su due thread diversi su un computer biprocessore. L'algoritmo è quindi in grado di funzionare fino a 20 Hz, ma in realtà la frequenza effettiva era di 12,5 Hz, ovvero quella del segnale di sincronizzazione distribuito a telecamere e laserscanner.

In figura 4.4 sono illustrati alcuni esempi di output dell'algoritmo: (a) mostra come in alcuni casi lo stesso veicolo, se si trova nella stessa corsia del camion, è localizzato in entrambe le immagini del sistema; in (b), un veicolo è rilevato anche se sta passan-

do sopra a una riga bianca dipinta sull'asfalto, un punto in cui nelle prime versioni dell'algoritmo si verificava un falso negativo; in (c) si può osservare come il sistema segnali come ostacolo solamente i veicoli che viaggiano nello stesso senso di marcia; in (d), infine, è presentata una scena in cui, nella corsia di fianco a quella in cui si trova il TerraMax, è rilevato un veicolo relativamente vicino, mentre uno che lo segue è ignorato perché erroneamente fuso con quello che lo precede; un'altra auto, infine, non è rilevata perché ormai totalmente fuori dalla scena.



(a) (b)

(c) (d)

Figura 4.4: Esempi di localizzazione dei veicoli in fase di sorpasso; ogni figura è l'insieme delle due immagini su cui lavora il sistema. Si può notare in (a) che uno stesso veicolo è localizzato in entrambi i fotogrammi; in (b), la localizzazione non risente delle strisce bianche dipinte sull'asfalto; (c) mostra una serie di veicoli non rilevati, poiché viaggiano nel senso di marcia opposto; in (d), infine, un veicolo non è rilevato perché è erroneamente fuso con quello che lo precede.

Conclusioni

In questa tesi è stato preso in considerazione il problema del rilevamento dei pedoni nelle immagini, molto studiato nell'ambiente della ricerca scientifica, e sono stati illustrati alcuni sistemi sviluppati per affrontarlo. Per migliorare le prestazioni ottenute, sono stati ideati dei nuovi algoritmi, capaci di effettuare analisi più raffinate, nel tentativo di migliorare il riconoscimento e rendere i sistemi più affidabili. Un'altra parte del lavoro si è concentrata sull'ampliamento delle situazioni in cui è possibile localizzare i pedoni, per esempio aumentando la distanza massima di rilevamento. Per ogni singolo studio è stata effettuata un'analisi specifica sui miglioramenti ottenuti rispetto alla situazione preesistente.

Il confronto con lo stato dell'arte è stato evidenziato soprattutto nella fase di descrizione della struttura dei sistemi, mentre diventa, nei fatti, molto difficile confrontare i risultati numerici che descrivono le prestazioni dei sistemi. Questo problema è noto, tanto che alcuni gruppi hanno proposto alla comunità scientifica dei database di immagini da utilizzare per misurare le prestazioni; tuttavia, questa proposta non ha riscosso un grande successo, a causa della difficoltà di utilizzare immagini e sequenze comuni, vista la grande varietà dei sistemi di acquisizione, che differiscono per numero di telecamere, tipologia delle stesse, e fusione con altri sensori.

Oltre al problema delle immagini, manca un vero e proprio standard per il metodo con cui le prestazioni sono misurate: alcuni gruppi considerano le immagini singolarmente, e verificano se gli algoritmi si comportano correttamente su ciascuna di esse; altri, invece, considerano gli "eventi", e, quindi, sono portati a dire che, per esempio,

un pedone è riconosciuto se è localizzato nella maggioranza dei fotogrammi, senza focalizzarsi sulle singole immagini. Per queste ragioni, nella parte riguardante i pedoni, sono stati tenuti in considerazione principalmente i miglioramenti rispetto alla situazione preesistente.

Il lavoro svolto ha un certo carattere di eterogeneità, nel senso che sono stati affrontati alcuni problemi, connessi tra loro ma comunque diversi, il cui filo conduttore è lo sviluppo di sistemi e tecniche per rendere più sicuro l'ambiente stradale per tutti gli utenti, siano essi pedoni o automobilisti. Progettare un sistema complesso come quelli illustrati richiede un certo numero di competenze diverse tra loro, e guadagnare esperienza in più di un ambito può essere utile per avere una migliore visione d'insieme.

Gli studi che sono stati portati a termine e gli algoritmi sviluppati hanno permesso un miglioramento delle prestazioni dei sistemi preesistenti, oppure l'aggiunta di nuove caratteristiche, portando il loro piccolo contributo per il miglioramento dello stato dell'arte dei sistemi di visione artificiale applicati all'ambiente stradale.

Appendice A

Equipaggiare un veicolo sperimentale

Avere a disposizione un veicolo dotato di sensori e computer installati a bordo è estremamente utile: innanzi tutto, si possono fare delle prove sul posizionamento delle telecamere e sulla loro orientazione, per capire, per ciascun sistema che si sviluppa, quale sia l'inquadratura migliore; in secondo luogo, lo si può utilizzare per registrare delle sequenze e per provare estensivamente gli algoritmi e il sistema di acquisizione in casi reali, che pongono quasi sempre di fronte a problemi di cui non si era tenuto conto, o che erano stati trascurati.

Attrezzare un veicolo sperimentale è un'operazione che diventa facilmente dispendiosa sia in termini economici che di tempo, a meno che gli interventi non siano davvero limitati. In questa appendice sarà descritta l'attività svolta per equipaggiare un veicolo dimostrativo di proprietà di Mando Corp., un'azienda sudcoreana con cui è in atto una collaborazione.

L'automobile in questione, una Hyundai Grandeur, doveva essere equipaggiata per ospitare un sistema di rilevamento dei pedoni basato su visione monoculare nello spettro NIR e laserscanner. Visto ad alto livello, il sistema da installare è composto da telecamera, laserscanner, illuminatori infrarossi ed elettronica di controllo, ma si

vedrà che il numero di scelte di dettaglio che è necessario affrontare, anche su un sistema semplice, è consistente.

A.1 Scelta dei componenti

Le prime scelte riguardano i sensori da installare. È stata fatta una ricerca di mercato sulle telecamere industriali, ma di fascia economica, con sensibilità nello spettro NIR, aventi shutter globale, interfaccia firewire e High Dynamic Range (si veda il paragrafo 1.1). Alla fine, è stata scelta la AVT Guppy F-036B, visibile in figura A.1.a, sulla quale è stata montata un'ottica dalla focale molto lunga, 12 mm, in modo da vedere in dettaglio gli oggetti lontani, ovvero quelli che il laserscanner trova con maggior difficoltà.

Come sensore laser si è optato per il SICK LMS-211, figura A.1.b, un modello largamente venduto e presente sul mercato da molto tempo, avente una risoluzione angolare di $0,25\,^{\circ}$, quindi sufficientemente elevata da permettere il rilevamento di ostacoli sottili come i pedoni; il modello scelto, inoltre, è stato studiato per il montaggio in esterni: è conforme allo standard IP67, quindi resistente ad acqua e polvere, ed è dotato di una piastra riscaldatrice che si attiva quando la temperatura scende al di sotto del limite di funzionamento; l'unica manutenzione di cui necessita è la pulizia della finestrella da cui esce il raggio laser, e il cambio del disseccante.

Il laserscanner scelto mette a disposizione due interfacce di comunicazione con l'host: la comune porta seriale, RS-232, e una versione industriale, chiamata RS-422; si è optato per la seconda, perché è basata su un segnale differenziale, decisamente più robusto ai disturbi elettromagnetici, e quindi più adatto alla propagazione nella zona vicino al motore, sede di un certo rumore; inoltre, la lunghezza massima dei cavi può essere di 300 m, contro i 10 m della seriale. Per contro, non è prevista una piedinatura standard, ed è stato necessario costruire un convertitore per poter usare due schede RS-422 diverse. Il firmware del laserscanner supporta quattro velocità di comunicazione: 9600, 19200, 38400 e 500000 bps, ma solo usando l'ultima (disponibile solo su interfaccia RS-422) è possibile ricevere tutte le scansioni; negli altri casi i dati

sono prodotti ad una velocità superiore a quella di trasmissione, ed alcune scansioni sono perse.
Gli illuminatori NIR sono stati difficili da reperire, poiché quelli che si trovano facilmente in commercio sono pensati per la videosorveglianza di ambienti piccoli, e hanno una portata massima molto ridotta, mentre il sistema che si desidera sviluppare deve localizzare i pedoni a svariate decine di metri. Gli illuminatori più usati sono quelli a LED e quelli a lampada alogena; si è deciso di provarli entrambi. Sono così stati acquistati un illuminatore a LED Serinn LX9-IR, figura A.1.c, e una coppia di fari alogeni come quello in figura A.1.d. Questi sono stati creati utilizzando un comune faro di profondità della Hella, a cui è stato sostituito il bulbo alogeno con uno speciale (Philips HIR1), che ha un'alta emissione nello spettro infrarosso; infine, un filtro che assorbe la luce visibile rende quest'oggetto un faro NIR.
Il computer installato a bordo del veicolo, prodotto dalla SmallPC, è basato su una scheda madre di piccole dimensioni, e monta un buon numero di componenti per laptop (compreso il processore, un Core 2 Duo T7200), che consentono di avere ridotte dimensioni e pochi problemi di raffreddamento; il disco fisso è di tipo standard da 2,5", perché soluzioni come memorie flash o dischi *automotive grade* sono ancora molto costose. Al computer sono collegati un touch screen, considerato più pratico da gestire di una trackball mentre l'auto è in movimento, una tastiera wireless, e un piccolo schermo analogico, installato al posto dell'autoradio.

(a)

(b)

(c)

(d)

Figura A.1: Sensori e illuminatori installati sul veicolo: la telecamera (a), con il filtro per tagliare la radiazione visibile e le ottiche provate, da 8 e 12 mm; il laserscanner SICK LMS-211 (b); l'illuminatore infrarosso a LED (c), e quello alogeno (d).

A.2 Connessioni dati ed elettriche

Dal punto di vista elettrico, è stato necessario ricorrere a numerosi adattatori, a causa dell'eterogeneità delle alimentazioni richieste: il computer accetta una tensione in ingresso variabile tra 6 e 18 V DC, quindi la tensione di 12 V comunemente disponibile sulle auto è sufficiente; il touchscreen è alimentato a 12 V DC, ma con una tolleranza molto bassa, per cui, per evitare di danneggiarlo, si è fatto ricorso ad un alimentatore stabilizzato. Il laserscanner e la relativa piastra di riscaldamento accettano un'alimentazione a 24 V DC, ed è quindi stato acquistato un convertitore DC/DC 12-24 V, capace di erogare fino a 10 A in uscita; gli illuminatori alogeni sono alimentati a 12 V DC, quindi non hanno necessità di alcun adattatore. Un discorso diverso vale invece per l'illuminatore a LED, che dev'essere collegato ad una fonte a 15,3 V DC molto stabile, tanto che è venduto con un alimentatore specifico, che, però, accetta una tensione di 220 V AC; poiché il costo per sostituirlo è abbastanza elevato, si è deciso di acquistare un inverter e di collegarlo con l'alimentatore di serie.
Dato il gran numero di apparecchiature installate, è stata aggiunta una seconda batteria in parallelo con quella dell'auto, e un caricabatterie. Tutti gli apparati di potenza, gli interruttori per controllare le singole alimentazioni, e i fusibili sono stati raggruppati dentro un contenitore ventilato, fissato all'interno del bagagliaio dell'auto, all'esterno del quale è attaccato il computer, come si vede in figura A.2.
Per poter ricevere i dati odometrici dell'autoveicolo, sul computer è stata installata una scheda CAN, che si è provveduto a collegare con una presa situata vicino ai pedali di guida; una conoscenza parziale dello standard dei pacchetti che viaggiano sul bus CAN delle auto Hyundai ha permesso di ricevere i dati relativi alla velocità angolare delle ruote, e all'angolo di rotazione del volante.
I cavi di alimentazione per gli illuminatori e il laserscanner, così come il cavo dati di quest'ultimo, corrono sotto alla scocca dell'auto, mentre altri passano nell'intercapedine tra il tettuccio e il suo rivestimento interno: si tratta del cavo firewire che collega il computer alla telecamera, e di quelli che portano il segnale ad alcuni plug firewire, ethernet e USB installati sotto al bracciolo anteriore.

(a)

(b)

Figura A.2: Il bagagliaio dell'auto (a), in cui si trovano il computer, la batteria supplementare, e il contenitore ventilato che ospita i numerosi alimentatori (b), gli interruttori di alimentazione, e i fusibili.

Figura A.3: Intervento meccanico per l'installazione del laserscanner nella parte frontale del veicolo.

A.3 Interventi meccanici

L'installazione del sistema a bordo del veicolo ha richiesto alcuni interventi sulla struttura meccanica dell'auto, effettuati presso una carrozzeria specializzata. Il più importante riguarda l'ancoraggio del laserscanner nella parte frontale: date le dimensioni del sensore, è stato necessario sostituire alcune barre metalliche, che garantiscono la robustezza strutturale del veicolo, con altre, appositamente sagomate, che lasciassero uno spazio maggiore, come si può vedere in figura A.3. Interventi minori hanno riguardato la stesura dei cavi e l'installazione delle prese sotto al bracciolo anteriore.

A.4 Versione finale del veicolo

Gli interventi sul veicolo sono stati fatti cercando di ottenere una buona integrazione dei sensori, non solo per ragioni estetiche, ma anche per rendere evidente che installare il sistema finale a bordo di un'automobile è possibile anche senza sacrificare il design e l'aerodinamica. In effetti, a intervento ultimato, la sagoma dell'auto non è molto diversa da quella originale, come si vede in figura A.4, che mostra l'auto prima (a) e dopo (b) l'intervento.
Per quanto riguarda l'interno dell'auto, può essere migliorata l'integrazione della telecamera, attualmente montata su un supporto metallico in vista, mentre il touchscreen è uno strumento necessario solo per lo sviluppo, e non sarebbe quindi necessario in un veicolo di serie.
L'auto equipaggiata può essere usata anche per il test di altri sistemi: in assenza delle barre portatutto specifiche per questo modello, che non sono previste, sul tetto è possibile fissare delle piastre magnetiche per reggere una barra su cui si può installare un sistema di acquisizione qualsiasi.

(a) (b)

Figura A.4: Esempio di integrazione dei sensori: la differenza tra l'auto originale (a) e quella modificata (b), seppure visibile, è minima.

Ringraziamenti

È domenica mattina, sono le 9:12, e io mi trovo nella sala riunioni della palazzina 1, da solo. Aspetto che arrivino i miei compagni di scrittura di tesi. È il secondo fine settimana consecutivo che passo a lavorare, e da domani sarà tutto finito. Nella mia mente si affollano molti pensieri: rifiniture lasciate in sospeso, rilettura, rispetto delle convenzioni tipografiche, che a me stanno così tanto a cuore... ma quello più insistente è che, da domani, tutto questo mi mancherà. Davvero. Mi mancherà. Non farò un solo nome, voglio evitare di scrivere un elenco stucchevole. Ma chi legge queste righe sa bene se il suo nome, anziché su carta, è scritto dentro di me.

In questi anni, passati in un ambiente caldo e accogliente, ho soprattutto imparato. Il contatto con alcune persone dotate di impressionanti doti umane e tecniche ha lasciato su di me un segno indelebile, di cui sono gelosissimo.

Da parte mia, spero di aver contribuito a creare un buon clima, quello che si dovrebbe percepire all'università, fatto di rispetto reciproco, di infiniti dubbi, che, al contrario delle certezze, permettono di crescere e di imparare; fatto di confronto e di dialogo, e anche di scontro, ma mai di chiusura.

All'università si dovrebbe respirare un'aria carica di affetto per il sapere, e per la volontà di trasmetterlo alle nuove generazioni. Gli studenti sono un impiccio, perché mettono alla prova la conoscenza e la capacità di comunicazione. E sono una risorsa: in tre anni di esperienza, non ho fatto mai una lezione senza che mi ascoltasse un buon numero di ragazzi desiderosi di affrontare seriamente i problemi. Anche solo per loro, bisognerebbe provare la volontà di migliorare sempre, di riflettere di più, di

imparare di nuovo ciò che si pensava di conoscere.
Con coloro che affrontano la tesi di laurea, il contatto si fa più stretto e invadente, e bisogna armarsi di pazienza, ripetendo ancora ciò che l'altro ascolta per la prima, dubbiosa volta. Avere a che fare con tante persone diverse, motivarle, pretendendo e ottenendo il meglio e il miglioramento da loro, è davvero difficile, e richiede molto tempo e dedizione, e può non piacere, ma è il lavoro che bisogna fare (anche) all'università.
Sala riunioni della palazzina 1 – 9:46.

Bibliografia

[1] Tarak Gandhi and Mohan M. Trivedi. Pedestrian Collision Avoidance Systems: a Survey of Computer Vision based Recent Studies. In *Procs. IEEE Intelligent Transportation Systems 2006*, pages 976–981, September 2006.

[2] Yotam Abramson and Bruno Steux. Hardware-friendly pedestrian detection and impact prediction. In *Proc. IEEE Intelligent Vehicles Symposium 2004*, pages 590–595, Parma, Italy, June 2004.

[3] Urban Meis, Matthias Oberläender, and Werner Ritter. Reinforcing the Reliability of Pedestrian Detection in Far-infrared Sensing. In *Proc. IEEE Intelligent Vehicles Symposium 2004*, pages 779–783, Parma, Italy, June 2004.

[4] Yajun Fang, Keiichi Yamada, Yoshiki Ninomiya, Berthold K. P. Horn, and Ichiro Masaki. A Shape-independent Method for Pedestrian Detection with Far-infrared Images. *IEEE Trans. on Vehicular Technology*, 53(6):1679–1697, November 2004. ISSN 0018-9545.

[5] Takayuki Tsuji, Hiroshi Hattori, Nobuharu Nagaoka, and Masahito Watabane. Delopment of Night Vision System. In *Procs. IEEE Intelligent Vehicles Symposium 2001*, pages 133–140, Tokyo, Japan, May 2001.

[6] Basel Fardi, Ingmar Seifert, Gerd Wanielik, and Jens Gayko. Motion-based pedestrian recognition from a moving vehicle. In *Proc. IEEE Intelligent Vehicles Symposium 2006*, pages 219–224, Tokyo, Japan, June 2006.

[7] Paul Viola, Michael J. Jones, and Daniel Snow. Detecting Pedestrians using Patterns of Motion and Appearance. In *Procs. IEEE Intl. Conf. on Computer Vision*, pages 734–741, Nice, France, September 2003.

[8] S. Yasutomi and H. Mori. A Method for Discriminating of Pedestrian Based on Rythm. In *Procs. IEEE Intl. Conference on Intelligent Robots and Systems*, pages 988–995, 1994.

[9] Constantine Papageorgiou, Theodoros Evgeniou, and Tomaso Poggio. A Trainable Pedestrian Detection System. In *Procs. IEEE Intelligent Vehicles Symposium'98*, pages 241–246, Stuttgart, Germany, October 1998.

[10] H. Shimizu and T. Poggie. Direction Estimation of Pedestrian from Multiple Still Images. In *Procs. IEEE Intelligent Vehicles Symposium 2004*, Parma, Italy, June 2004.

[11] S. Tate and Y. Takefuji. Video-based Human Shape Detection Deformable Templates and Neural Network. In *Procs. of Knowledge Engineering System Conf.*, Crema, Italy, 2002.

[12] Ross Cutler and Larry S. Davis. Robust Real-time Periodic Motion Detection, Analysis and Applications. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(8):781–796, August 2000.

[13] Takayuki Tsuji, Hiroshi Hattori, Masahito Watanabe, and Nobuharu Nagaoka. Development of Night-vision System . *IEEE Trans. on Intelligent Transportation Systems*, 3(3):203–209, September 2002.

[14] Xia Liu and Kikuo Fujimura. Pedestrian Detection using Stereo Night Vision. *IEEE Trans. on Vehicular Technology*, 53(6):1657–1665, November 2004. ISSN 0018-9545.

[15] Xia Liu and Kikuo Fujimura. Pedestrian Detection using Stereo Night Vision. In *Procs. IEEE Intl. Conf. on Intelligent Transportation Systems 2003*, pages 334–339, Shangai, China, October 2003.

[16] 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 2004*, Parma, Italy, June 2004.

[17] G. P. Stein, O. Mano, and A. Shashua. Vision based ACC with a Single Camera: Bounds on Range and Range Rate Accuracy. In *Procs. IEEE Intelligent Vehicles Symposium 2003*, Columbus, USA, June 2003.

[18] L. Zhao. *Dressed Human Modeling, Detection, and Parts Localization*. Ph.D. dissertation, Carnegie Mellon University, 2001.

[19] Fengliang Xu and Kikuo Fujimura. Pedestrian Detection and Tracking with Night Vision. In *Procs. IEEE Intelligent Vehicles Symposium 2002*, Paris, France, June 2002.

[20] Ram Rajagopal. Pattern Matching Based on a Generalized Transform. Technical report, National Instruments, 2000.

[21] Paul Viola and Michael Jones. Rapid Object Detection using a Boosted Cascade of Simple Features. In *Intl. Conf. on Computer Vision & Pattern Recognition*, volume 1, pages 511–518, December 2001.

[22] Yajun Fang, Keiichi Yamada, Yoshiki Ninomiya, Berthold Horn, and Ichiro Masaki. Comparison between Infrared-image-based and Visible-image-based Approaches for Pedestrian Detection. In *Procs. IEEE Intelligent Vehicles Symposium 2003*, pages 505–510, Columbus, USA, June 2003.

[23] Alberto Broggi, Alessandra Fascioli, Marcello Carletti, Thorsten Graf, and Marc-Michael Meinecke. A Multi-resolution Approach for Infrared Vision-based Pedestrian Detection. In *Procs. IEEE Intelligent Vehicles Symposium 2004*, pages 7–12, Parma, Italy, June 2004.

[24] Massimo Bertozzi, Alberto Broggi, Stefano Ghidoni, and Mike Del Rose. Pedestrian Shape Extraction by means of Active Contours. In Christian Laugier

and Roland Siegwart, editors, *Procs. Intl. Conf. on Field and Service Robotics*. Springer-Verlag, Chamonix, France, March 2008.

[25] Dariu M. Gavrila. Pedestrian Detection from a Moving Vehicle. In *Procs. of European Conference on Computer Vision*, volume 2, pages 37–49, June–July 2000.

[26] D. M. Gavrila and J. Giebel. Virtual sample generation for template-based shape matching. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition 2001*, volume 1, pages 676–681, Kauai, USA, 2001.

[27] J. Giebel, D. M. Gavrila, and S. Munder. Vision-based pedestrian detection: the PROTECTOR System. In *Proc. IEEE Intelligent Vehicles Symposium 2004*, Parma, Italy, 2004.

[28] Alberto Broggi, Alessandra Fascioli, Paolo Grisleri, Thorsten Graf, and Michael-Marc Meinecke. Model-based Validation Approaches and Matching Techniques for Automotive Vision based Pedestrian Detection. In *Procs. Intl. IEEE Wks. on Object Tracking and Classification in and Beyond the Visible Spectrum*, San Diego, USA, June 2005.

[29] Minh-Son Dao, Francesco G. B. De Natale, and Andrea Massa. Edge potential functions and genetic algorithms for shape-based image retrieval. In *Procs. IEEE Intl. Conf. on Image Processing (ICIP'03)*, volume 2, pages 729–732, Barcelona, Spain, September 2003.

[30] Minh-Son Dao, Francesco G. B. De Natale, and Andrea Massa. Efficient Shape Matching Using Weighted Edge Potential Function. In *Procs. 13th Intl. Conf. on Image Analysis and Processing (ICIAP'05)*, Cagliari, Italy, September 2005.

[31] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active Contour Models. *Intl. Journal of Computer Vision*, 1(4):321–331, 1988.

[32] Donna J. Williams and Mubarak Shah. A Fast Algorithm for Active Contours and Curvature Estimation. *CVGIP: Image Understanding*, 55(1):14–26, 1992.

[33] S. R. Gunn and M. S. Nixon. A Robust Snake Implementation: a Dual Active Contour. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(1):63–68, January 1997.

[34] Congxia Dai, Yunfei Zheng, and Xin Li. Pedestrian Detection and Tracking in Infrared Imagery Using Shape and Appearance. *Computer Vision and Image Understanding*, 106(3):288–299, June 2007.

[35] Fengliang Xu, Xia Liu, and Kikuo Fujimura. Pedestrian Detection and Tracking With Night Vision. *IEEE Trans. on Intelligent Transportation Systems*, 6(1):63–71, March 2005.

[36] V. Philomin, R. Duraiswami, and L. Davis. Pedestrian Tracking from a Moving Vehicle. In *Procs. IEEE Intelligent Vehicles Symposium 2000*, pages 350–355, Detroit, USA, October 2000.

[37] David Lefée, Stéphane Mousset, Abdelaziz Bensrhair, and Massimo Bertozzi. Cooperation of Passive Vision Systems in Detection and Tracking of Pedestrians. In *Procs. IEEE Intelligent Vehicles Symposium 2004*, pages 768–773, Parma, Italy, June 2004.

[38] Osama Masoud and Nikolaos P. Papanikolopoulos. Robust pedestrian tracking using a model-based approach. In *Procs. IEEE Intl. Conf. on Intelligent Transportation Systems'97*, pages 338–343, November 1997.

[39] Shunsuke Kamijo and Masao Sakauchi. Simultaneous Tracking of Pedestrians and Vehicles by the Spatio-Temporal Markov Random Field Model. In *Procs. IEEE Intl. Conf. on Systems, Man, and Cybernetics*, volume 4, pages 3732–3737, October 2003.

[40] L. Davis, V. Philomin, and R. Duraiswami. Tracking Humans from a Moving Platform. In *Procs. 15th Intl. Conf. on Pattern Recognition*, volume 4, pages 171–178, Barcelona, Spain, September 2000.

[41] Zhijun Qiu, Danya Yao, Yi Zhang, Daosong Ma, and Xinyu Liu. Detecting Pedestrian and Bicycle using Image Processing. In *Procs. IEEE Intl. Conf. on Intelligent Transportation Systems 2003*, pages 340–345, Shangai, China, October 2003.

[42] Hee-Deok Yang and Seong-Whan Lee. Multiple Pedestrian Detection and Tracking based on Weighted Temporal Texture Features. In *Procs. IEEE Intl. Conf. on Pattern Recognition*, volume 4, pages 248–251, Cambridge, UK, August 2004.

[43] Julien Burlet, Olivier Aycard, Anne Spalanzani, and Christian Laugier. Pedestrian Tracking in Car Parks: An Adaptive Interacting Multiple Models Based Filtering Method. In *Procs. IEEE Intl. Conf. on Intelligent Transportation Systems 2006*, pages 462–467, Jhongli, Taiwan, June 2006.

[44] Dariu M. Gavrila and J. Geibel. Shape-Based Pedestrian Detection and Tracking. In *Procs. IEEE Intelligent Vehicles Symposium 2002*, Paris, France, June 2002.

[45] Magrit Betke, Esin Haritaoglu, and Larry Davis. Multiple Vehicle Detection and Tracking in Hard Real-time. In *Procs. IEEE Intelligent Vehicles Symposium'96*, pages 351–356, Tokyo, Japan, September 1996.

[46] Christopher E. Smith, Charles A. Richards, Scott A. Brandt, and Nikolaos P. Papanikolopoulos. Visual Tracking for Intelligent Vehicle-Highway Systems. *IEEE Trans. on Vehicular Technology*, 45(4):744–759, November 1996.

[47] Alberto Broggi, Massimo Bertozzi, Roland Chapuis, Frédéric Chausse Alessandra Fascioli, and Amos Tibaldi. Pedestrian Localization and Tracking System with Kalman Filtering. In *Procs. IEEE Intelligent Vehicles Symposium 2004*, pages 584–589, Parma, Italy, June 2004.

[48] Ulrich Scheunert, Heiko Cramer, Basel Fardi, and Gerd Wanielik. Multi Sensor based Tracking of Pedestrians: a Survey of Suitable Movement Models. In

Procs. IEEE Intelligent Vehicles Symposium 2004, pages 774–778, Parma, Italy, June 2004.

[49] Dirk Tenne and Tarunraj Singh. Analysis of alpha-beta-gamma Filters. In *IEEE International Conference on Control Applications*, Kohala Coast-Island of Hawai'i, Hawai'i, USA, August 1999.

[50] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Trans. ASME Journal of Basic Engineering*, 82(1):35–45, March 1960.

[51] Michel Verhaegen and Paul Van Dooren. Numerical Aspects of Different Kalman Filter Implementations. *IEEE Trans. on Automatic Control*, 21(1):68–72, December 1982.

[52] Greg Welch and Gary Bishop. An Introduction to the Kalman Filter. Technical Report NC 27599-3175, Department of Computer Science, University of North Carolina at Chape Hill, April 2004.

[53] Eli Brookner. *Tracking and Kalman Filtering Made Easy*. John Wiley Interscience, 1998.

[54] Emanuele Binelli, Alberto Broggi, Alessandra Fascioli, Stefano Ghidoni, Paolo Grisleri, Thorsten Graf, and Marc-Michael Meinecke. A Modular Tracking System for Far Infrared Pedestrian Recognition. In *Procs. IEEE Intelligent Vehicles Symposium 2005*, pages 758–763, Las Vegas, USA, June 2005.

[55] Alberto Broggi, Massimo Bertozzi, Stefano Ghidoni, and Marc Michael 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.

[56] Y. S. Yao and R. Chellappa. Selective Stabilization of Images Acquired by Unmanned Ground Vehicles. *IEEE Trans. on Robotics and Automation*, 13(5):693–708, October 1997.

[57] Z. Duric and A. Rosenfeld. Image Sequence Stabilization in Real-Time. *Real Time Imaging*, 2(5):271–284, October 1996.

[58] Alberto Broggi and Paolo Grisleri. A Software Video Stabilization System for Automotive oriented Applications. In *Procs. IEEE Vehicular Technology Conference*, Stockholm, Sweden, June 2005.

[59] A. Censi, A. Fusiello, and V. Roberto. Image stabilization by features tracking. In *Procs. Intl. Conf. on Image Analysis and Processing*, pages 665–667, Venice, Italy, September 1999.

[60] Luca Bombini, Pietro Cerri, Paolo Grisleri, Simone Scaffardi, and Paolo Zani. An Evaluation of Monocular Image Stabilization Algorithms for Automotive Applications. In *Procs. IEEE Intl. Conf. on Intelligent Transportation Systems 2006*, pages 1562–1567, Toronto, Canada, September 2006.

[61] Massimo Bertozzi, Alberto Broggi, Michael Del Rose, and Andrea Lasagni. Infrared Stereo Vision-based Human Shape Detection. In *Procs. IEEE Intelligent Vehicles Symposium 2005*, pages 23–28, Las Vegas, USA, June 2005.

[62] Harsh Nanda and Larry Davis. Probabilistic Template Based Pedestrian Detection in Infrared Videos. In *Procs. IEEE Intelligent Vehicles Symposium 2002*, Paris, France, June 2002.

[63] 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*, volume 2, pages 574–577, Beijing, China, August 2006.

[64] H. Vceraraghavan, O. Masoud, and N. Papanikolopoulos. Vision-based monitoring of intersections. In *Procs. 5th IEEE Intl. Conf. on Intelligent Transportation Systems*, pages 7–12, 2002.

[65] V. Bhuvaneshwar and P. B. Mirchandani. Real-time detection of crossing pedestrians for traffic-adaptive signal control. In *Procs. 7th IEEE Intl. Conf. on Intelligent Transportation Systems*, pages 309–313, October 2004.

[66] C. Conde, Á. Serrano, L. J. Rodríguez-Aragón, J. Pérez, and E. Cabello. An Experimental Approach to a Real-Time Controlled Traffic Light Multi-Agent Application. In *Procs. 3rd Intl. Conf. on Autonomous Agents and Multi Agent Systems*, July 2004.

[67] P. Vannoorenberghe, C. Motamed, J. M. Blosseville, and J. G. Postaire. Monitoring pedestrians in a uncontrolled urban environment by matching low-level features. In *Procs. IEEE Intl. Conf. on Systems, Man, and Cybernetics*, volume 3, pages 2259–2264, October 1996.

[68] F. N. Mc Leod, N. B. Hounsell, and B. Rajbhandari. Improving Traffic Signal Control for Pedestrians. In *Procs. 12th IEE Intl. Conf. on Road Transport Information and Control*, pages 268–277, April 2004.

[69] A. Prati, I. Mikic, M. M. Trivedi, and R. Cucchiara. Detecting moving shadows: Algorithms and evaluation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(7):918–923, July 2003.

[70] Olivier Faugeras. *Three-Dimensional Computer Vision: A Geometric Viewpoint*. The MIT Press, Cambridge, 1993.

[71] Massimo Bertozzi, Alberto Broggi, and Alessandra Fascioli. Stereo Inverse Perspective Mapping: Theory and Applications. *Image and Vision Computing Journal*, 8(16):585–590, 1998.

[72] Stephen Se. Zebra-crossing Detection for the Partially Sighted. In *Procs. IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 211–217, Hilton Head Island, SC, USA, June 2000.

[73] Mohammad Shorif Uddin and Tadayoshi Shioyama. Robust Zebra-Crossing Detection using Bipolarity and Projective Invariant. In *Procs. 8th Intl. Symp. on Signal Processing and Its Applications*, pages 571–574, Sidney, Australia, August 2005.

[74] Frederic Devernay and Olivier D. Faugeras. Straight Lines have to be Straight. *Machine Vision Application*, 13(1):14–24, 2001.

[75] William J. Wolfe, Donald Mathis, Cheryl Weber Sklair, and Michael Magee. The Perspective View of Three Points. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13(1):66–73, January 1991.

[76] Alessandra Fascioli, Rean Isabella Fedriga, and Stefano Ghidoni. Vision-based Monitoring of Pedestrian Crossings. In *Procs. 14th Intl. Conf. on Image Analysis and Processing*, Modena, Italy, September 2007.

[77] Wei Liu, XueZhi Wen, Bobo Duan, Huai Yuan, and Nan Wang. Rear Vehicle Detection and Tracking for Lane Change Assist. In *Proc. IEEE Intelligent Vehicles Symposium 2007*, pages 252–257, Istanbul, Turkey, June 2007.

[78] Christos Tzomakas and Werner von Seelen. Vehicle Detection in Traffic Scenes Using Shadows. Technical report, Institut für Neuroinformatik, Ruhr-Universität Bochum, June 1998.

[79] Marinus B. Van Leeuwn and Frans C.A. Goren. Vehicle Detection with a Mobile Camera. *IEEE Robotics & Automation Magazine*, 12(1):37–43, March 2005.

[80] Zehang Sun, George Bebis, and Roland Miller. On-Road Vehicle Detection: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:694–711, May 2006.

[81] Yan Zhang, Stephen J. Kiselewich, William A. Bauson, and Riad Hammoud. Robust Moving Object Detection at Distance in the Visible Spectrum and

Beyond Using A Moving Camera. In *Procs. Conference on Computer Vision and Pattern Recognition 2006*, June 2006.

[82] Chiou-Shann Fuh and Petros Maragos. Region-based Optical Flow Estimation. In *Procs. Conference on Computer Vision and Pattern Recognition 1989*, June 1989.

[83] B. Heisele and W. Ritter. Obstacle Detection Based on Color Blob Flow. In *Proc. IEEE Intelligent Vehicles Symposium 1995*, pages 282–286, Detroit, USA, September 1995.

[84] B. Heisele, U. Kreßel, and W. Ritter. Tracking Non-Rigid, Moving Objects Based on Color Cluster Flow. In *Procs. IEEE Conf. on Computer Vision and Pattern Recognition 1997*, pages 257–260, San Juan, Puerto Rico, June 1997.