UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

FACOLTÀ DI SCIENZE MATEMATICHE FISICHE E NATURALI

Corso di Laurea Magistrale in Informatica

Sistema di visione per la guida autonoma di veicoli

Relatore: Prof. Domenico G. Sorrenti
Correlatore: Dr. Axel Furlan
Correlatore: Dr. Daniele Marzorati

Tesi di Laurea Specialistica di:
Roberto FERMI
Matricola n°048091

Anno Accademico 2009 – 2010

Ai miei genitori

e ai miei nonni

Indice

Introduzione

L'obiettivo di questa tesi è di realizzare un sistema di visione da utilizzare nei veicoli a guida autonoma per determinare la posizione del veicolo nel mondo e riconoscere la presenza di ostacoli nel suo campo visivo.

Per guidare autonomamente sulle strade, o in qualsiasi altro luogo dove la presenza di ostacoli non è predeterminabile, un veicolo deve essere in grado di "vedere" quello che succede nel mondo in modo da riconoscere, ad esempio, una persona che attraversa la strada oppure una macchina parcheggiata in doppia fila che blocca la via.

Se per noi il concetto di vista è scontato, non lo è altrettanto per le macchine. Sapere riconoscere gli oggetti, identificarne il colore, la forma e altre proprietà fisiche, saperli collocare nello spazio, sono azioni che svolgiamo inconsciamente ogni giorno e che apparentemente sembrano semplici e scontate; per un calcolatore, riuscire a compiere queste azioni efficacemente ed efficientemente è ancora una sfida che molte volte porta a dei compromessi tra velocità di elaborazione e qualità del risultato.

Un calcolatore può disporre di numerosi sistemi di visione come, ad esempio, sonar, oppure scanner laser, ma oggigiorno l'approccio più promettente è quello basato sulla visione artificiale.

Con l'uso di sensori come gli scanner laser, oggi lo stato dell'arte in robotica, è infatti possibile sapere a quale distanza si trova un oggetto rispetto al sensore. Tuttavia, questi sensori non forniscono informazioni dettagliate sulla struttura a cui appartiene effettivamente il punto misurato: se questo fa parte di una configurazione più grande, di una persona, di un'autovettura o di qualsiasi altro oggetto; il sensore indica solamente che il punto si trova ad una data distanza. Un'altra grande limitazione di questi sensori nasce dal fatto che le misure vengono fatte su piani di scansione (generalmente uno o, per sensori molto costosi, in un numero molto limitato) e, quindi, non è possibile vedere degli oggetti

nella loro interezza, ma solo parti di essi nei particolari piani di scansione considerati.

L'utilizzo delle camere per la visione artificiale permette di affrontare alcuni di questi problemi, ma ne introduce, ovviamente, altri. Le informazioni generate dalle immagini delle camere consentono, infatti, di avere una visione globale degli oggetti e di distinguere, a seguito di specifiche elaborazioni, le diverse tipologie di ostacoli presenti, come persone, autovetture, alberi, etc. Di contro, dato che la descrizione del mondo reale tridimensionale viene proiettata su di un supporto bidimensionale, si ha la perdita di una dimensione ed, in particolare, della profondità. Quando catturiamo un'immagine con la camera, infatti, non siamo più in grado di capire a che distanza siano i diversi punti in essa rappresentati. Se conoscessimo le dimensioni degli oggetti quest'aspetto non costituirebbe un problema, ma ciò non è ipotizzabile.

Per ovviare a questo problema si tenta di simulare quello che normalmente avviene nella vista umana, ovvero si utilizzano due immagini della stessa scena, prese da due punti di vista, risultato, nella visione artificiale, dell'utilizzo di due camere. Le due immagini del mondo catturate dai sensori devono riferirsi allo stesso istante di tempo e quindi il comando di acquisizione delle immagini delle camere deve essere opportunamente sincronizzato, per non rischiare di osservare due scene che hanno poco in comune. I due dispositivi, inoltre, devono essere calibrati: è necessario, in altre parole, stabilire come sono disposti l'uno rispetto all'altro in modo da riuscire a determinare, per i vari punti dell'immagine della prima camera, i punti corrispondenti nella seconda.

Una volta in possesso dei dati di calibrazione delle due camere e di due punti immagine dello stesso punto scena, è possibile ricostruire la posizione tridimensionale dei punti nel mondo.

L'utilizzo delle camere al posto di altri sensori comporta anche altre problematiche, prima tra tutte la gestione della qualità dell'immagine. Per fare analisi delle immagini è infatti necessario che, in ogni condizione di utilizzo del sistema, gli oggetti siano ben definiti

e ben contrastati, ovvero ben distinguibili gli uni dagli altri e discriminabili dallo sfondo. Se in condizioni "controllate", come ad esempio in un laboratorio o in una stanza ben illuminata, questo può essere facilmente raggiunto, in condizioni normali è molto difficile da ottenere. Basti pensare, ad esempio, alla condizione di guida di una vettura in strada durante una giornata soleggiata mentre ci si avvicina all'ingresso di una galleria. In queste situazioni alcune volte è arduo anche per l'occhio umano adattarsi per vedere correttamente i dettagli esposti al sole fuori dal tunnel e contemporaneamente quelli in ombra dentro la galleria. Questa è una situazione di alto range dinamico, dove la dinamica è quella del segnale luminanza immagine. Per una camera questa è una circostanza molto difficile e con camere stato-dell'arte non si riuscirà ad avere un'immagine utilizzabile. Se non si dispone di un'immagine sfruttabile, non sarà possibile determinare gli ostacoli che ci precedono e capire cosa accade intorno al veicolo rendendo impraticabile la guida autonoma. Per questo motivo è di essenziale importanza avere sempre un output valido dalle camere.

Prima di svolgere qualsiasi elaborazione di alto livello con le immagini catturate dalle camere (come, ad esempio, il riconoscimento e la classificazione di oggetti), è necessario assicurarsi che la qualità sia sufficiente. La diversa risposta all'intensità luminosa in differenti condizioni di acquisizione, il rumore elettrico, l'ampiezza del range dinamico, il contrasto, sono solo alcuni degli aspetti fondamentali che caratterizzano i sensori e le immagini e che bisogna opportunamente valutare, correggere ed adattare per ottenere dei buoni risultati.

In questa tesi mostrerò le caratteristiche che distinguono le differenti tipologie di sensori presenti sul mercato, quali sono i principali vantaggi e svantaggi di ognuno di essi e su quali basi si è proceduto alla scelta di un particolare sensore ***[22] [23] [24] [25].***

Un particolare punto d'interesse nelle applicazioni di visione artificiale è la scelta di usare parametri delle camere fissi (come, ad esempio, tempo di esposizione, apertura del diaframma, guadagno,

etc.), oppure di farli variare nel tempo in relazione alle specifiche condizioni dell'ambiente in cui ci stiamo muovendo.

L'utilizzo di parametri fissi per il controllo delle camere e delle immagini ha il vantaggio di non dovere fare elaborazioni in background per analizzare la scena, ma ha, generalmente, dei limiti in termini di utilizzo. In questa tesi mostrerò come l'uso di particolari tecnologie dei sensori digitali in unione a tecniche di elaborazione delle immagini ha reso possibile questo scenario, permettendo di usare dei parametri fissi delle camere e, allo stesso tempo, di avere un risultato eccellente in ogni condizione di esposizione, dal buio della notte, al controluce intenso in una giornata soleggiata ***[15] [5] [10]***.

Verrà, inoltre, mostrato com'è possibile variare un solo parametro delle camere (il tempo di esposizione) con delle analisi in background che tengano conto, oltre che delle particolari condizioni d'illuminazione della scena, anche di specifici indicatori che consentono di definire quanto "*motion blur*" è presente nelle immagini ***[20] [21]***, al fine di rilevare quando il tempo di esposizione è alto. In questi casi, infatti, è importante non ottenere mai un'immagine mossa della scena, sopperendo all'eventuale tempo di esposizione troppo basso (con tempi di esposizioni troppo bassi, si ottengono immagini mediamente più scure) con tecniche di post elaborazione al fine di aumentare la luminosità dell'immagine.

Vi sono diverse tecniche di post-elaborazione e la più appropriata, nella nostra proposta, è il pixel-binning, che consiste nel raggruppare (o più precisamente sommare) la quantità di energia immagazzinata in elementi fotosensibili contigui del sensore digitale delle camere, in modo da ottenere una maggiore intensità, per scene prive di luce o con scarsa illuminazione, a discapito della risoluzione. Un'altra tecnica molto utile e molto usata anche in altri ambiti dell'elaborazione delle immagini, è l'equalizzazione dell'istogramma che permette di ridistribuire equamente i vari livelli di colore nell'intervallo di possibili valori, in modo da migliorare il contrasto dell'immagine stessa ***[10] [5] [11] [12]***.

Nella tesi si descrive anche come è possibile usare le informazioni provenienti dalle due camere per collocare nel mondo tridimensionale gli oggetti presenti nell'immagine (ricostruzione dei punti del mondo) e come vengono determinati ostacoli e spazio libero usando U-V disparity ***[1] [2] [3] [4] [13] [19] [18]***.

Per quest'ultimo aspetto va segnalato che il mio contributo si è limitato all'integrazione di software sviluppato dal D. R. Marzorati

Gli argomenti trattati nella tesi, considerati nel loro insieme, forniscono la base per lo sviluppo di un futuro sistema finalizzato alla determinazione della posizione del veicolo stesso nel mondo e nella intrapresa di azioni di aggiramento e/o arresto in presenza di ostacoli sulla propria traiettoria. Usando le immagini elaborate seguendo le tecniche presentate nella tesi, sarà anche possibile riuscire ad identificare e categorizzare gli oggetti nella scena per individuare quelli in movimento ed eventualmente tracciarne la traiettoria.

Nel primo capitolo della tesi vedremo lo stato dell'arte delle tecnologie e delle tecniche, analizzando anche il funzionamento dei sensori digitali per capirne al meglio le peculiarità e le differenze. Analizzeremo il modello pin-hole delle camere e come funziona una testa-stereo per la visione artificiale.

Nel secondo capitolo, verranno presentate le tecnologie scelte per creare questo sistema di visione da implementare sui veicoli a guida autonoma e le tecniche utilizzate per rendere usabili le immagini in ogni condizione della scena.

Nel terzo capitolo, si mostreranno i risultati ottenuti e le conclusioni del progetto, soffermandosi sui risultati sperimentali.

Infine, verranno illustrati i possibili sviluppi futuri di questo progetto e le applicazioni che potrebbero trarre beneficio o spunto da queste analisi e da questo lavoro.

Capitolo 1 – Stato dell'arte

In questo capitolo vedremo lo stato dell'arte delle tecnologie e delle tecniche oggetto di questa tesi, analizzando il funzionamento dei sensori digitali per capirne al meglio le peculiarità e le differenze. Verranno presentati il modello pin-hole delle camere, e come sia possibile ottenere la stereo visione partendo da due camere. Saranno quindi illustrate le tecniche di post-processing più comuni utilizzate nell'ambito della visione artificiale

Le camere

I sensori digitali

I sensori digitali sono la vera interfaccia verso il mondo delle camere poiché sono i dispositivi che si occupano di ricevere l'energia luminosa proveniente dalla scena e di tradurla in un segnale elettrico comprensibile da un calcolatore. Vi sono diverse tipologie di sensori digitali che differiscono tra loro per le proprietà costruttive, la sensibilità alla luce, la tecnologia di fabbricazione e numerose altre caratteristiche. Vediamo ora il principio di funzionamento di un sensore generico e delle principali tipologie di sensori che esistono attualmente per coglierne le differenze, i vantaggi e gli svantaggi.

Come vedremo, la scelta del sensore digitale con cui si svilupperà il progetto ha un enorme impatto sulla buona riuscita dello stesso. E' quindi importante sia comprendere il principio di funzionamento di questi dispositivi, sia capire quali sono le differenze tra le varie tipologie in modo da effettuare una valutazione ponderata sulle reali necessità del progetto.

Una delle peculiarità dei sensori digitali è la sensibilità alla radiazione elettromagnetica infrarossa che permette di catturare delle immagini anche in presenza di luce molto scarsa.

Generalmente, la temperatura del sensore è direttamente collegata al fenomeno della "*dark-current*[1]": più la temperatura è bassa, minore sarà questo fenomeno e viceversa, più alta è la temperatura, più accentuato sarà questo problema.

I sensori CCD

I primi sensori digitali comparsi sul mercato, i CCD, sono stati inventati da Willard S. Boyle e George E- Smith di Bell Laboratories nel 1969 e sono stati utilizzati per la prima volta nel 1975.

I sensori CCD (dall'inglese Charge Coupled Devices, ovvero dispositivi a carica accoppiata) sono dispositivi formati da una griglia di elementi semiconduttori (pixel) in grado di accumulare una carica elettrica proporzionale all'intensità della radiazione elettromagnetica che li colpisce. Questi elementi sono accoppiati in modo tale che ognuno di essi, quando riceve un opportuno comando elettrico sia in grado di trasferire la propria carica a un elemento adiacente. Il principio base dei sensori CCD "tradizionali" è quello di un registro a scorrimento: una volta che la scena è stata esposta al sensore, un apposito circuito elettrico, invia una serie temporizzata di comandi alle righe (o colonne, a seconda dell'implementazione della circuiteria del dispositivo), in modo che i dati memorizzati in ogni pixel vengano trasferiti in un apposito buffer. Il contenuto del buffer, a questo punto, è opportunamente elaborato e restituito all'utilizzatore per la visualizzazione/analisi dei dati. Non appena le righe del sensore CCD vengono "svuotate" ed il contenuto trasferito nel buffer il dispositivo è nuovamente pronto ad acquisire una nuova immagine.

Uno dei principali problemi dei sensori CCD nasce proprio dalla particolare lettura dei dati. Il trasferimento dei dati tra gli elementi

[1] Per "dark current" si intende quel fenomeno fisico ed elettronico che si verifica nei dispositivi fotosensibili. Una piccola quantità di corrente elettrica fluisce attraverso il dispositivo stesso anche quando non è esposto alla radiazione elettromagnetica. In Elaborazione delle immagini, questo difetto provoca un sensibile rumore a basse intensità luminose che può compromettere i buoni risultati.

fotosensibili alle celle del buffer necessita di un certo lasso di tempo e, mentre le cariche vengono spostate, altra energia elettromagnetica viene immagazzinata negli elementi fotosensibili non ancora copiati, provocando quello che viene definito problema di "*shuttering*", ovvero un effetto di trascinamento dell'immagine.

Ci sono tre tipologie di implementazione per i sensori CCD atte ad evitare o diminuire il problema dello "*shuttering*": full frame, frame-transfer ed interline transfer. In un sensore "*Full frame*", tutta l'area del sensore è sensibile alla luce e di conseguenza, il problema dello "*shuttering*" è molto accentuato; per ridurlo vengono utilizzati otturatori meccanici. Nella versione "frame-transfer" metà del sensore è coperto da una maschera opaca, generalmente in alluminio. L'immagine catturata dalla parte del sensore esposta alla luce, può essere trasferita in modo molto veloce dalla sua zona "fotosensibile" alla parte oscurata. A questo punto l'immagine può essere letta più lentamente attraverso lo shifting delle celle mascherate, mentre quelle libere sono pronte nuovamente ad immagazzinare le informazioni elettromagnetiche provenienti dal mondo. Lo svantaggio principale di questa architettura deriva dal fatto che, nonostante il sensore occupi il medesimo spazio di un sensore "*full frame*", la sua superficie utile effettivamente esposta alla luce, o meglio la sua risoluzione spaziale, è solamente la metà.

Nei sensori CCD "*interline"* si ha un'ulteriore evoluzione della precedente versione: ogni colonna (o riga) che viene esposta alla radiazione elettromagnetica, è affiancata da una colonna (o riga) opaca che viene usata come colonna di memorizzazione. In questa implementazione, per trasferire le informazioni dagli elementi fotosensibili agli elementi di memorizzazione, è necessario un solo trasferimento. Anche in questo caso, tuttavia, il sensore presenterà delle parti oscurate e si avrà perciò, una diminuzione di risoluzione rispetto alla versione full-frame. Per ovviare a questo inconveniente, nei sensori professionali vengono utilizzate delle micro-lenti per convogliare la maggior parte della radiazione elettromagnetica che colpisce il sensore, sulla parte non coperta dello stesso.

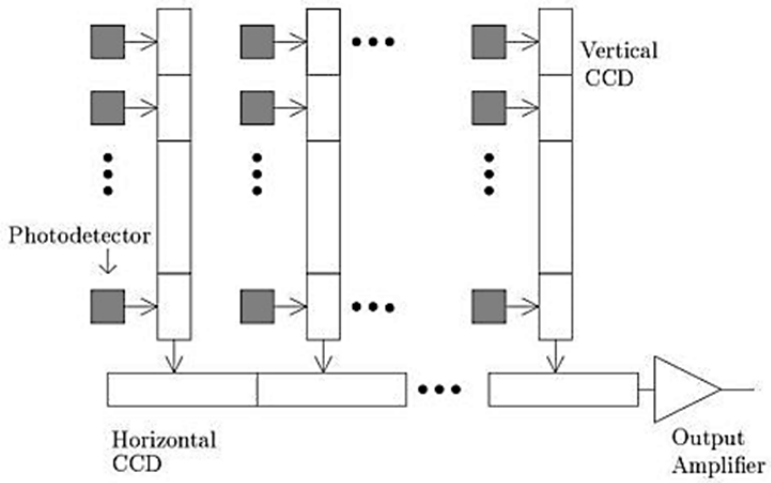


Figura 1 – La figura mostra la struttura concettuale di un sensore CCD interline-transfer. Ogni colonna viene mappata in una colonna cieca direttamente adiacente che viene a sua volta svuotata con degli appositi comandi elettrici. L'uscita viene quindi amplificata.

In un sensore CCD la tipica risposta alla luce diretta incidente è di circa il 70%. Nelle applicazioni di visione artificiale è opportuno optare per sensori "*frame transfer*" in modo da ovviare al problema dello "*shuttering*" visto in precedenza. Lo svantaggio principale di questo tipo di sensori è la complessità architetturale notevolmente più alta in quanto ogni cella deve avere la sua "controparte" cieca per la fase di copia dei dati al termine del periodo di esposizione.

Uno dei principali difetti dei sensori CCD è quello che viene definito "*smear*": quando un elemento fotosensibile è sottoposto ad una grande quantità di radiazione elettromagnetica, l'energia in esso accumulata tende a sconfinare nei pixel appartenenti alla stessa riga/colonna, saturandola per intero.

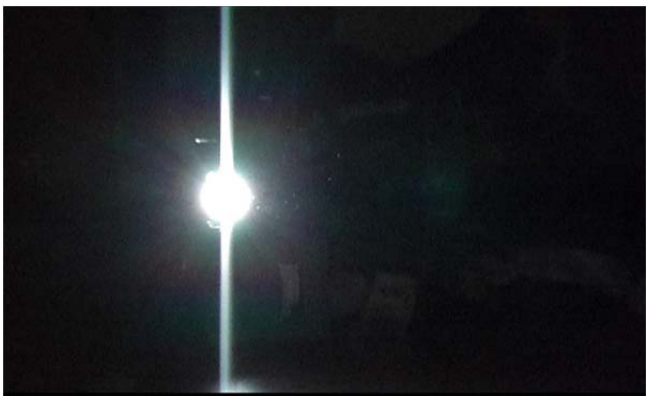


Figura 2 – La figura mostra il tipico effetto di "smearing" dei sensori CCD esposti ad una forte fonte luminosa.

I sensori CMOS

I sensori CMOS (Complementary metal-oxide semiconductor) fanno parte dei sensori APS (Active Pixel Sensor) caratterizzati dalla presenza in ogni pixel del sensore, di un circuito integrato contenente l'elemento fotosensibile ed un amplificatore attivo del segnale.

Nascono come alternativa ai sensori CCD e sono stati creati per risolverne i principali problemi di velocità e scalabilità. I sensori CMOS, inoltre, consumano meno energia per funzionare e richiedono tempi di lettura delle informazioni più brevi, che consentono la riduzione del ritardo nella cattura dell'immagine.

Uno dei principali vantaggi dei sensori CMOS è quello di potere indirizzare, mediante la selezione di riga e colonna, un singolo pixel della matrice e ciò determina, quindi, la possibilità di effettuare la lettura di ciascun elemento singolarmente. Dato che la velocità di lettura dei sensori CMOS è notevolmente superiore rispetto a quella dei sensori CCD, il problema dello "*shuttering*" è così nettamente ridotto da diventare trascurabile.

Esistono due principali tipologie di sensori CMOS: quelli definiti "*global shutter*" e quelli detti "*rolling shutter*". Nei primi, un apposito comando elettronico viene fornito a tutte le celle fotosensibili quando ha inizio l'esposizione; al termine della stessa, un altro comando simultaneo a tutte le celle interrompe la cattura della luce

da parte degli elementi fotosensibili e la quantità di energia immagazzinata viene convertita in tensione elettrica. Nei sensori CMOS "*rolling shutter*", invece, il comando di inizio e fine acquisizione viene fatto scorrere sulle varie celle (da qui il termine *rolling* ovvero rotolare) in tempi differenti. Il risultato è che diverse parti del sensore vengono esposte alla luce in tempi differenti. Com'è ovvio supporre, i sensori CMOS "*global shutter*" sono migliori rispetto a quelli "*rolling*", ma la loro maggiore complessità circuitale li rende mediamente più costosi.

Anche se apparentemente i sensori CMOS di tipo "*rolling shutter*" possono introdurre il fenomeno dello "*shuttering*", in realtà l'entità di quest'ultimo è relativamente limitata dalla maggiore velocità di lettura della circuiteria rispetto ai sensori CCD e dipende, in grande misura, dal tempo di esposizione scelto per catturare la scena. Nel caso in cui il tempo di acquisizione sia particolarmente elevato (ad esempio in condizioni di scarsa illuminazione), anche il fenomeno di "*shuttering*" presente nell'immagine sarà accentuato.

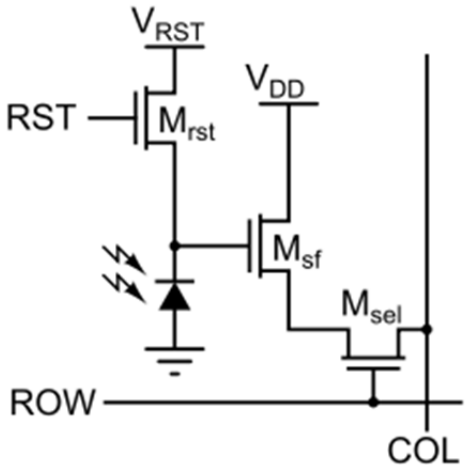


Figura 3 – L'immagine mostra la struttura tipo di un pixel di un sensore CMOS. ROW e COL sono i segnali di selezione di riga e colonna, Vdd è la tensione di alimentazione, Msel è il transistor di selezione del pixel, Mrst è il transistor di reset che viene comandato dal segnale RST. Msf è un transistor che funge da buffer per le informazioni catturate dall'elemento fotosensibile

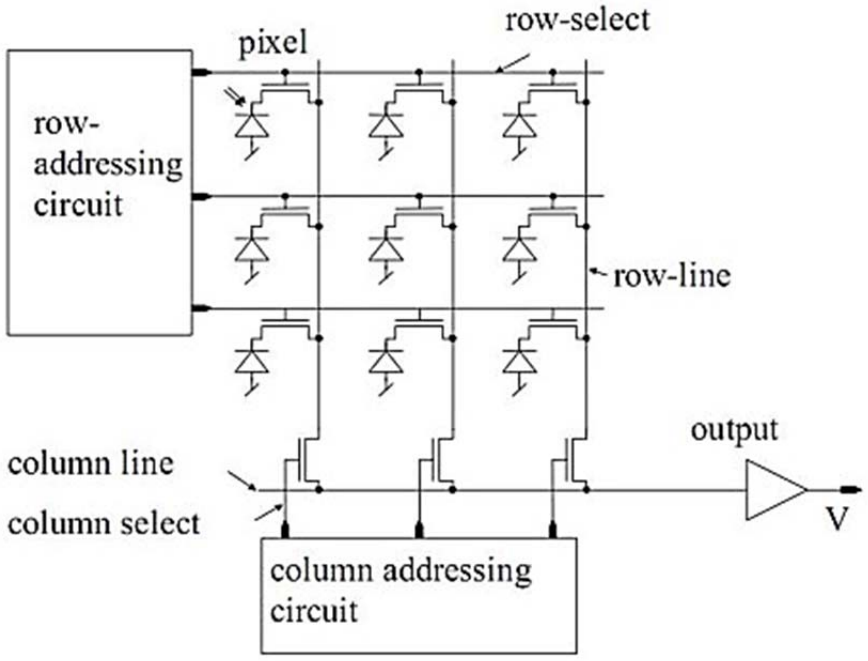


Figura 4 – L'immagine mostra la struttura tipica di un sensore CMOS. Ogni pixel del sensore può essere letto singolarmente indirizzando la rispettiva riga e colonna. Ogni elemento fotosensibile dispone di un buffer interno per la lettura e di un amplificatore di segnale.

I sensori consumer vs. sensori professionali

L'obiettivo di qualsiasi progetto per la guida autonoma di veicoli è quello di utilizzare dispositivi poco dispendiosi in termini economici. Questo per due motivi: da un lato i dispositivi economici sono più diffusi e quindi più facilmente reperibili di quelli professionali, e dall'altro la scelta di prodotti consumer permette di contenere il costo del prodotto finito. Tuttavia, in alcune circostanze è necessario arrivare a dei compromessi per ottenere dei buoni risultati, proprio come in questo progetto.

Nella nostra applicazione, infatti, è necessario che il sensore disponga di caratteristiche particolari:

- un ampio range dinamico per poter gestire efficacemente le scene con un elevato contrasto

- un convertitore ADC con buona definizione ed una risoluzione superiore ai consueti 8bit, in modo da riuscire a codificare con più precisione le varie intensità luminose
- un basso livello di "*black current*" e, quindi, una minima quantità di rumore alle basse intensità luminose
- tempi di risposta rapidi nel fornire le immagini, tali da minimizzare il problema dello "*shuttering*".

Molte di queste caratteristiche, purtroppo, non sono ancora disponibili nei sensori consumer perché sono troppo costose o tecnicamente troppo onerose da implementare.

Il compromesso che si può raggiungere è quello di utilizzare sensori CMOS che sono mediamente più economici dei sensori CCD, nonostante la loro maggiore complessità circuitale. Poiché la circuiteria dei sensori CMOS è più complicata e dato che integrano un amplificatore di segnale direttamente in ogni pixel della matrice, i sensori CMOS sono leggermente più sensibili al rumore provocato della "*dark current*". Tuttavia, l'isolamento intrinseco tra i pixel della matrice li rende molto meno sensibili al problema dello "*smearing*" rispetto ai sensori CCD, dove questo difetto è spesso un problema importante e fastidioso.

Inoltre, i sensori CMOS sono in grado di raggiungere velocità di lettura considerevoli. Infatti, dato che la circuiteria di ogni singolo pixel è più complessa, molte operazioni (come l'amplificazione e la conversione ADC) vengono eseguite direttamente nel pixel e non successivamente alla lettura del sensore. E' possibile, quindi, mantenere dei tempi di esposizione relativamente bassi pur avendo una frequenza d'immagini alta.

Lenti ed ottiche

Nell'utilizzo di sistemi di acquisizione immagini le lenti sono di fondamentale importanza. La lente, infatti, è l'elemento ottico, generalmente in vetro, che permette di concentrare, o meglio convogliare, i raggi della luce verso un punto predeterminato che, nel nostro caso, è rappresentato dal sensore digitale.

Per riuscire a convogliare quanta più luce possibile e tanto più precisamente possibile verso il sensore, una sola lente non è mai sufficiente: per questo motivo, vengono create le ottiche, cioè un insieme di lenti di differenti tipologie e forme, opportunamente accoppiate.

Vi sono varie forme di lenti ognuna con particolari proprietà di concentrazione della radiazione elettromagnetica che le attraversa. Alcune tra le forme più utilizzate sono illustrate nella seguente immagine:

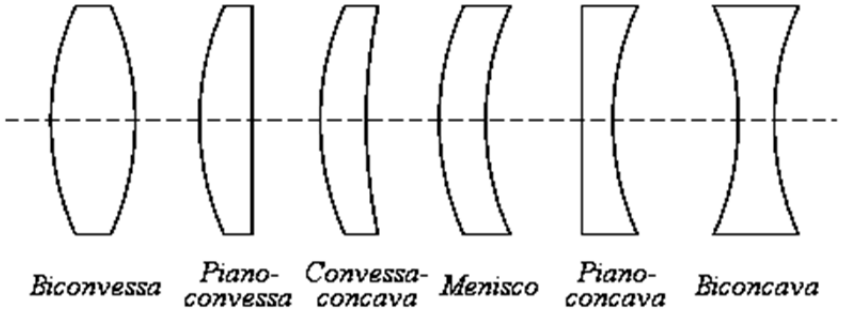


Figura 5 – L'immagine mostra alcune tra le più diffuse forme di lenti

Per calcolare la distanza focale di una lente (in aria), si ricorre alla seguente formula:

$$\frac{1}{f} = (n-1)\left[\frac{1}{R_1} - \frac{1}{R_2} + \frac{(n-1)d}{nR_1R_2}\right]$$

dove *f* è la lunghezza focale, *n* è l'indice di rifrazione del materiale utilizzato per costruire la lente, R_1 è il raggio di curvatura della lente sulla superficie più vicina alla sorgente luminosa, R_2 è il raggio di curvatura della lente sulla superficie più lontana dalla sorgente luminosa e *d* è lo spessore della lente stessa (la distanza viene calcolata sull'asse della lente nei punti di intersezione dello stesso con la superfice più esterna).

Per differenziare le lenti concave da quelle convesse è sufficiente inserire correttamente il segno del raggio di curvatura. Non vi è uno

standard predefinito per i segni dei raggi di curvatura, ma generalmente si tende ad indicare una lente convessa con un segno positivo di R_1 e una concava con un segno negativo; per la superficie più distante dalla sorgente luminosa i segni si invertono e si usa un segno positivo di R_2 per indicare una superficie concava e un segno positivo per indicare una superficie convessa.

Se lo spessore *d* della lente è piccolo rispetto ai raggi di curvatura, è possibile utilizzare la seguente formula semplificata per il calcolo della distanza focale:

$$\frac{1}{f} \approx (n-1)\left[\frac{1}{R_1} - \frac{1}{R_2}\right]$$

Distorsioni e aberrazioni delle lenti

Le lenti non formano mai immagini perfette e creano sempre un certo grado di distorsione o aberrazione causato dal processo produttivo della lente stessa, dalla non perfetta levigatura della superficie, dalle impurità contenute all'interno del materiale, etc. Più alta è la qualità di una lente, maggiore sarà il suo costo, ma minori saranno le distorsioni introdotte.

Esistono differenti tipologie di alterazioni all'immagine causate dalle lenti che introducono difetti più o meno rilevanti. L'aberrazione sferica nasce dal fatto che le superfici delle lenti non sono sfere perfette, ma la forma di lente più semplice da rettificare e lucidare per il vetro. I raggi paralleli all'asse della lente, ma che si trovano a diversa distanza da esso, vengono messi a fuoco in punti leggermente differenti, causando appunto, il fenomeno dell'aberrazione sferica e l'effetto di sfocatura sull'immagine risultante. Per ovviare a questo problema vengono create lenti con forme particolari capaci di ridurre questo difetto, tra cui quelle complesse chiamate *asferiche*.

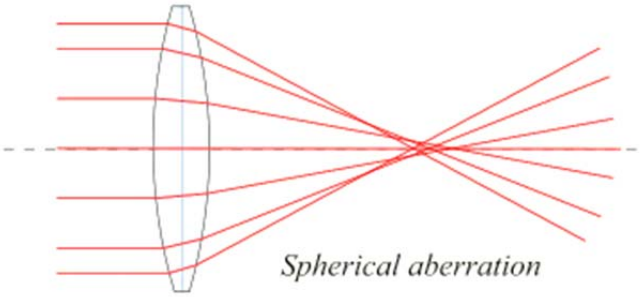


Figura 6 – L'immagine mostra il tipico effetto dell'aberrazione sferica sulla focalizzazioni di fasci di luce paralleli all'asse della lente, ma distanti da esso. L'effetto visibile è la sfocatura dell'immagine.

Un altro importante fenomeno legato alle lenti è quello dell'aberrazione cromatica, causata dalla dispersione del materiale con cui è realizzata la lente stessa. Più precisamente, l'aberrazione cromatica è causata dalla variazione dell'indice di rifrazione "*n*" della lente al variare della lunghezza d'onda della luce. Siccome *f*, come abbiamo visto nelle precedenti formule, dipende anche da *n*, ne deriva che differenti lunghezze d'onda saranno messe a fuoco in punti differenti, causando frange di diversi colori intorno agli oggetti. Questo problema può essere risolto accoppiando due materiali con differenti indici di dispersione.

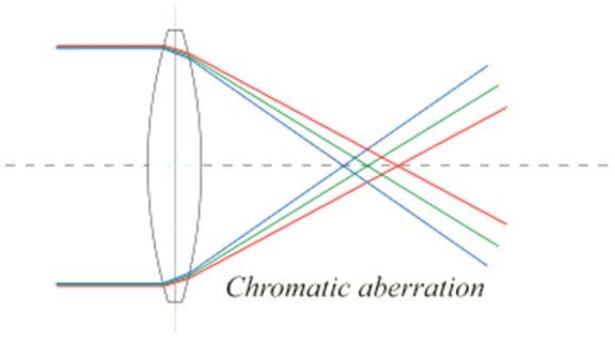


Figura 7 – L'immagine mostra il fenomeno dell'aberrazione cromatica che si verifica nelle lenti quando fasci luminosi di differente lunghezza d'onda vengono focalizzati in un punto.

Altri due importanti difetti causati dalle lenti sono le cosiddette distorsioni a barile e a cuscino. La prima è causata da un effetto d'ingrandimento dell'immagine nella parte centrale che diminuisce man mano che ci si allontana dall'asse ottico. L'effetto ottico è di un'immagine che è stata "avvolta" intorno ad una sfera (o barile). Questa distorsione è generalmente provocata da lenti con un ampio campo visivo, come grandangoli o fisheye.

La distorsione a cuscino è esattamente l'opposto della distorsione a barile: l'ingrandimento dell'immagine cresce all'aumentare della distanza dall'asse ottico. L'effetto ottico causa la tendenza delle linee non passanti per il centro ottico ad essere inclinate verso lo stesso, creando una forma simile, appunto, ad un cuscino.

Per eliminare le distorsioni e le aberrazioni esaminate in precedenza, che rappresentano non un elenco esaustivo, ma solo i casi principali, si possono combinare lenti con caratteristiche tali da determinare alterazioni dell'immagine che si compensino a vicenda.

Blur circle

A questo punto sorge spontaneo chiedersi perché sia necessario utilizzare le ottiche se queste, generalmente, introducono distorsioni ed aberrazioni. Come vedremo in seguito, il modello ideale di proiezione dell'immagine è rappresentato da un foro di dimensioni infinitesimali, attraverso cui viene fatta passare la luce che colpisce il piano di proiezione. Nella realtà, tuttavia, non è possibile creare un foro così piccolo e, anche nel caso in cui ciò fosse realizzabile, la quantità di luce che riuscirebbe a passare sarebbe talmente limitata da non poter essere rilevata dai sensori digitali.

Se si aumenta la dimensione di questo foro, si avranno delle interazioni con le onde elettromagnetiche (luce) che lo attraversano ed, inoltre, un singolo punto della scena verrà proiettato in un "cerchio" sul piano immagine. Proprio la formazione di questo "cerchio" di proiezione causa un effetto ottico di sfocatura dell'immagine, da cui il nome "*blur circle*".

Le ottiche vengono utilizzate per cercare di ridurre questo problema, creando un modello del foro di dimensione infinitesimale anche quando non è possibile realizzarlo nella realtà.

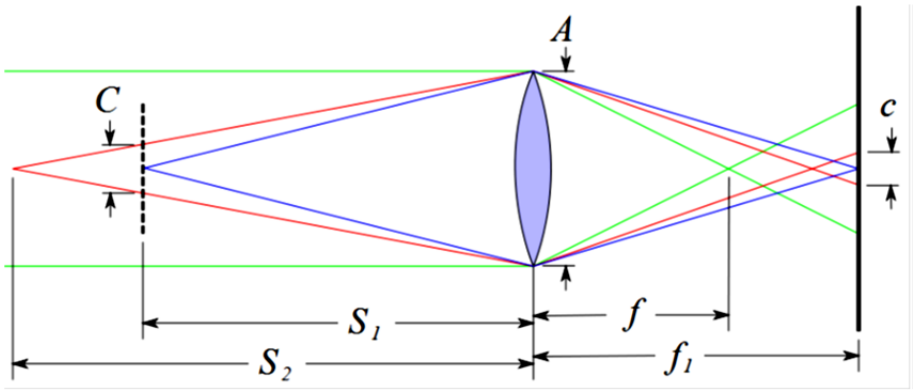


Figura 8 – L'immagine mostra la formazione del cerchio di sfocatura per un oggetto non a fuoco

La dimensione del blur-circle può essere calcolata utilizzando la formula:

$$C = A \cdot \frac{|S_2 - S_1|}{S_2}$$

dove A è il diametro della lente utilizzata, S_1 è la distanza del piano messo a fuoco dal centro ottico e S_2 è la distanza dell'oggetto, non a fuoco, che si vuole proiettare dal centro ottico.

Con dei semplici passaggi matematici, si ottiene che

$$c = A \cdot \frac{|S_2 - S_1|}{S_2} \cdot \frac{f}{S_1 - f}$$

da cui si evince che la dimensione dell'area di sfocatura è direttamente collegata al diametro della lente e alla distanza focale.

Le ottiche in visone artificiale

Dopo aver analizzato il funzionamento delle lenti e i principali problemi da esse derivanti esaminiamo quali sono i parametri da tenere in considerazione per l'applicazione della visione artificiale. E'

di fondamentale importanza che le distorsioni introdotte dalle ottiche siano limitate, soprattutto per quanto riguarda l'aberrazione cromatica e l'aberrazione sferica. Le distorsioni a barile e a cuscino possono essere efficacemente corrette in fase di post-processing utilizzando specifiche tecniche di *undistort*, dopo avere opportunamente calibrato la camera. Come vedremo in seguito, queste tecniche di post-processing sono in grado di eliminare quasi totalmente questi difetti; tuttavia, quando queste distorsioni sono eccessive, si rischia di ridurre troppo l'area utilizzabile delle immagini oppure di non riuscire completamente ad eliminarle.

Un altro requisito fondamentale per le ottiche usate in visione artificiale è la necessità che l'attenuazione della luce quanto passa attraverso le numerose lenti di cui è composta un'ottica, sia minima. In questo modo, una maggiore intensità di radiazione elettromagnetica colpirà il sensore e verrà trasformata in informazione. In visione artificiale, infatti, può capitare di trovarsi in ambienti con scarsa illuminazione ed è, quindi, di fondamentale importanza assicurarsi di perdere la minore quantità di luce possibile nei diversi passaggi che essa subisce.

Il range dinamico

Il range dinamico è un parametro fondamentale delle camere e, in particolare, dei sensori digitali. E' definito come il rapporto tra il massimo ed il minimo valore di luminosità presente nell'immagine.

Avere un alto range dinamico è un requisito fondamentale nelle applicazioni di guida autonoma mediante visione artificiale: tanto maggiore sarà tale range, tanto più sarà possibile rappresentare aree molto illuminate mantenendo, nel contempo, una buona sensibilità in aree scure della scena e viceversa.

L'occhio umano dispone di un range dinamico molto esteso che ci permette di osservare chiaramente sia il cielo stellato, sia una giornata molto soleggiata. In elaborazione delle immagini, o più precisamente in fotografia, il contrasto dinamico è anche indicato in

EV (*Exposure value*, o valore di esposizione), detti anche *stops*. Ad ogni incremento di EV (ovvero ad ogni incremento di stops) corrisponde un raddoppio della quantità di luce. Una camera professionale di altissimo livello può avere un range dinamico fino a 11 EV, mentre per l'occhio umano questo valore è compreso tra 10 e 14 EV, valore nettamente superiore anche ad una camera top di gamma.

Ci sono diversi modi per ottenere un maggiore contrasto dinamico nelle camere, ma il più conosciuto ed utilizzato è quello di scattare differenti immagini della stessa scena con diversi tempi di esposizione; in seguito, con particolari tecniche di post-processing si procede ad "unire" le differenti immagini a diverse esposizione al fine di ottenere un'unica immagine con un contrasto dinamico maggiore (i bassi tempi di esposizione cattureranno i dettagli dei toni chiari della scena, mentre quelli alti metteranno in evidenza le zone scarsamente illuminate).

In visione artificiale questa tecnica non è applicabile dato che la cattura di differenti immagini della stessa scena comporta notevoli difficoltà. Se la scena cambia velocemente, come normalmente avviene, le immagini catturate saranno diverse e non sarà possibile unirle correttamente. Per questo motivo viene privilegiato l'utilizzo di sensori in grado di gestire in maniera ottimale la quantità di luce che incide sugli elementi fotosensibili.

Le camere in visione artificiale

Dopo avere analizzato in cosa consistono i sensori digitali e le diverse tipologie a disposizione, esaminandone i vantaggi e gli svantaggi, cosa sono le ottiche e quali problemi introducono, approfondiamo ora le caratteristiche dell'elemento principale generato dall'unione di questi dispositivi: la camera.

Proiezione e modello pin-hole

Il passaggio dalla scena del mondo tridimensionale a un'immagine vista dalla camera avviene attraverso una trasformazione che è molto frequentemente ben approssimata da una trasformazione prospettica. Tale operazione proietta i punti dello spazio (ovvero del mondo tridimensionale) su di un piano (che è l'immagine della scena catturata dalla camera). Come abbiamo visto in precedenza, la luce, prima di colpire il sensore digitale, attraversa una serie di lenti che servono per convogliare opportunamente i raggi della radiazione elettromagnetica. La trasformazione prospettica da mondo tridimensionale a bidimensionale può essere, tuttavia, modellizzata in modo semplice ed efficace attraverso l'approssimazione teorica del pin-hole che non utilizza lenti ed ottiche.

L'approssimazione teorica del pin-hole (teorica, perché nella realtà non è riproducibile), consiste in una camera senza lenti né ottiche in cui la luce passa attraverso un foro di dimensione infinitesima (questo ultimo punto rende l'approssimazione prettamente teorica, dato che nella realtà una tale circostanza non è realizzabile). La luce che passa attraverso il pin-hole, proietta un'immagine invertita della scena su di una superficie posta ad una certa distanza del foro infinitesimale, ovviamente dalla parte opposta rispetto a quella da dove proviene la luce.

Il piano su cui la luce viene proiettata dal pin-hole viene detto piano immagine e la distanza tra il piano e il pin-hole stesso viene chiamata distanza focale ed è indicata, generalmente, con "f" o λ.

Il punto "C" rappresentato nella Figura 9 coincide con il pin-hole attraverso cui passa la luce della scena e viene chiamato "centro di proiezione". "π" è il piano su cui i punti della scena tridimensionale sono proiettati e, nel nostro caso, coincide con il sensore digitale della camera. C_π è l'intersezione tra l'asse ottico, passante per il foro di dimensione infinitesimale e perpendicolare al piano immagine, ed il piano immagine stesso. E' un punto molto importante e prende il nome di "punto principale" o "centro immagine".

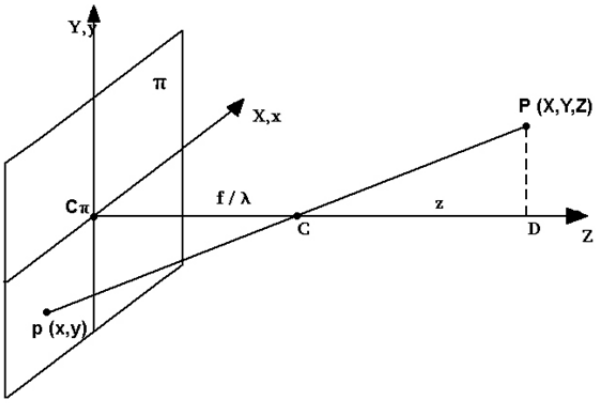


Figura 9 – L'immagine mostra il modello geometrico della proiezione pin-hole.

La retta che congiunge il punto "p", appartenente al piano immagine, ed il punto "P", appartenente al mondo, prende due nomi differenti a seconda che sia realizzata partendo da P e C, oppure da "p" e C. Nel primo caso, è chiamata "retta di proiezione" poiché permette di proiettare il punto del mondo sul piano immagine. Nel secondo caso è chiamata "retta di interpretazione" perché, partendo dal punto bidimensionale appartenente al piano immagine, permette di risalire alla posizione del punto nel sistema di riferimento mondo.

Uno dei problemi più ardui da affrontare in visione artificiale è proprio quello di riuscire a determinare la posizione originaria del punto P, avendo a disposizione solamente il punto *p* e la sua retta di interpretazione.

La proiezione del punto in coordinate mondo (3D) nel punto in coordinate camera (2D) è legato dalle seguenti equazioni:

$$\begin{cases} x = -f\,\dfrac{X}{Z} \\ y = -f\,\dfrac{Y}{Z} \end{cases}$$

Il valore f/Z può essere visto come un fattore di scala che dipende dalla profondità (Z) del punto scena. Se poniamo $m = f/Z$, otteniamo quella che viene definita come "*weak prospective scaled ortographic*"

$$\begin{cases} x = -mX \\ y = -mY \end{cases}$$

Calibrazione

Analizziamo ora il problema del sistema di riferimento assoluto utilizzato. In precedenza, abbiamo considerato un sistema di riferimento coincidente con quello della camera. In realtà, esso è normalmente posto in posizione fissa rispetto alla scena tridimensionale, mentre il sistema di riferimento della camera si muove con essa. Per porre in relazione questi due sistemi di riferimento è necessario eseguire una rototraslazione, che può essere a sua volta composta da una concatenazione di rototraslazioni.

Il tipo di trasformazioni necessarie per passare da un sistema di riferimento all'altro dipende dalla tipologia di camera utilizzata, dalla sua posizione e dalle sue caratteristiche intrinseche. E' quindi facile dedurre che la matrice di trasformazione dipende da una molteplicità di parametri (intrinseci ed estrinseci) e che tali parametri sono, il più delle volte, difficili da stimare con precisione e soggetti a variazioni nel tempo. Per questo motivo, la stima delle matrici di rototraslazione non deriva solamente da calcoli geometrici e matematici, ma viene determinata sperimentalmente in un delicato processo chiamato "calibrazione della camera".

La calibrazione è un processo fondamentale nelle applicazioni di visione artificiale ed è essenziale che venga svolta con particolare attenzione e cura: maggiore sarà l'accuratezza della calibrazione, tanto più sarà ottimale la stima dei parametri intrinseci ed estrinseci

della camera. In questo processo, nessuna delle informazioni eventualmente disponibili a priori riguardanti la posizione relativa dei due sistemi di riferimento viene utilizzata.

La matrice di rototraslazione risultante sarà una matrice 4x4 di parametri incogniti che sono ricavati facendo uso della relazione, vista in precedenza, esistente tra il punto scena e quello proiettato sul piano immagine.

Il passaggio dal punto tridimensionale 3D del mondo al punto bidimensionale 2D dell'immagine è legato dalla seguente relazione:

$$P_{2D} = M \cdot P_{3D}$$

dove le coordinate dei punti sono indicate in coordinate omogenee

$$P_{3D} = \begin{vmatrix} X \\ Y \\ Z \\ K \end{vmatrix} \text{ e } P_{2D} = \begin{vmatrix} w_u \\ w_v \\ w \end{vmatrix}$$

Da cui si ricava che

$$\begin{cases} wu = m_{1,1}X + m_{1,2}Y + m_{1,3}Z + m_{1,4} \\ wv = m_{2,1}X + m_{2,2}Y + m_{2,3}Z + m_{2,4} \\ w = m_{3,1}X + m_{3,2}Y + m_{3,3}Z + m_{3,4} \end{cases}$$

Il processo di calibrazione, ha come obbiettivo quello di stimare i parametri $m_{i,j}$ della matrice.

Esistono diverse tecniche di calibrazione, ma quelle più utilizzate sono quelle chiamate "*offline*", che richiedono una fase iniziale di raccolta dei dati, solitamente una serie di immagini campione opportunamente predisposte, che vengono analizzati per stimare i parametri $m_{i,j}$ della matrice di rototraslazione. I valori così ottenuti sono quindi utilizzati nell'applicazione reale, ma affinchè ciò determini risultati ottimali, è importante che i parametri calcolati non cambino, cioè che il riferimento delle camere rimanga lo stesso, e che non vengano apportate variazioni alle ottiche ed ai parametri intrinseci della camera stessa.

Un'altra tecnica di calibrazione è quella definita "*online*", dove dei processi software in background analizzano continuativamente le immagini della scena al fine di adattare e correggere la calibrazione nel tempo. In questo caso sono ammesse piccole variazioni ai parametri intrinseci della camera. E' sempre bene, tuttavia, che questi parametri rimangano il più possibile invariati nel tempo.

I parametri della camera (intrinseci ed estrinseci)

Come abbiamo visto in precedenza, la fase di calibrazione permette di stimare la matrice che rappresenta la trasformazione proiettiva del mondo sul piano immagine. Ogni elemento di questa matrice si riferisce a determinati parametri riferiti alla camera. In particolare è possibile distinguere due tipologie principali di parametri: quelli intrinseci, che sono strettamente legati alle caratteristiche della camera utilizzata, e quelli estrinseci, che permettono di mettere in relazione il sistema di coordinate della camera stessa con il sistema di riferimento del mondo.

I parametri intrinseci di una camera sono i seguenti:

- f = lunghezza focale
- u_0, v_0 = posizione del centro immagine rispetto alla proiezione del centro ottico, ovvero traslazione del centro immagine.
- s_x, s_y = dimensione dei pixel (va osservato che possono non essere di forma quadrata)
- $skew(\theta)$ = angolo tra gli assi del sistema di riferimento camera

I parametri estrinseci della camera, invece, sono formati dai valori della rototraslazione tra il sistema coordinate camera ed il sistema coordinate mondo:

- t_x, t_y, t_z = traslazione del sistema di riferimento camera
- r_1, r_2, r_3 = rotazione del sistema di riferimento camera

La stereo-visione

Dopo avere esaminato qual è il modello di proiezione di una camera, come è possibile calibrarla e quali sono i parametri che la caratterizzano, di seguito vedremo come è possibile utilizzare due camere per ottenere un sistema di visione (generalmente chiamato *testa-stereo*) che si comporta come l'occhio umano, cioè che è in grado di determinare la profondità degli oggetti.

Fondamenti della visione binoculare

La visione binoculare è la caratteristica propria del sistema visivo dell'essere umano (e di altre specie animali) per cui un'immagine viene proiettata sul piano retinico di entrambi gli occhi. La scena viene "vista" da due prospettive diverse ed il cervello analizza le differenze che esistono tra le viste al fine di calcolare la profondità e la prospettiva degli oggetti. Queste differenze sono causate dalla naturale traslazione che è presente tra i due punti di visione e che si propaga anche nelle scene proiettate.

In visione artificiale si tenta di ricostruire questa realtà utilizzando due camere.

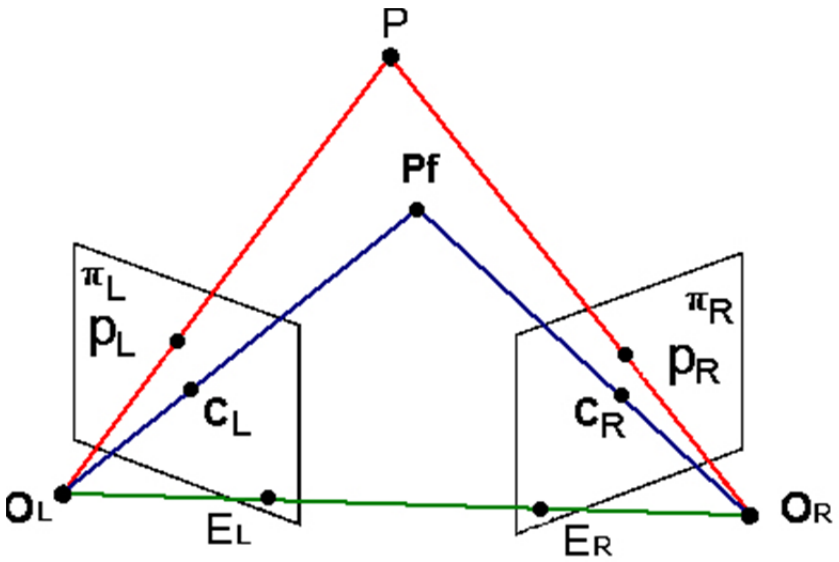


Figura 10 – L'immagine mostra un punto ripreso da due camere differenti in posizioni differenti.

La Figura 10 mostra un esempio di punto P ripreso da due camere differenti, da due prospettive diverse. Il punto O rappresenta il punto di messa a fuoco delle due camere e, per semplicità di rappresentazione, l'immagine viene mostrata antecedente a questo punto. La retta che congiunge i due punti di focalizzazione prende il nome di ***linea base*** (o "*baseline*" in inglese). Tale retta può appartenere o meno al piano immagini delle due camere. I punti O_R e O_L, essendo punti scena, possono essere proiettati, rispettivamente, nella camera di sinistra e nella camera di destra; eventualmente, essi possono essere anche esterni alle coordinate ammissibili dell'immagine. Il punto di proiezione del punto di fuoco dell'altra camera nel proprio piano immagini viene definito come ***epipolo***.

$$\overline{O_L O_R} \cap \pi_L = E_L$$
$$\overline{O_L O_R} \cap \pi_R = E_R$$

La geometria epipolare è quella branca della geometria che si applica alla stereo visione e che specifica le relazioni che sussistono tra i punti 3D del mondo proiettati sul piano immagine 2D.

Trovare il punto P partendo dalle sue proiezioni PL e PR viene definito come ***problema della ricostruzione*** ed è uno degli aspetti basilari della visione artificiale. Questo problema è differente da quello che viene definito come "*stereo-matching*". Infatti, nel primo caso si hanno a disposizione entrambe le proiezioni del punto sui piani immagine ed è necessario calcolare dove tale punto si trovava nella scena 3D; nel secondo caso, invece, l'obiettivo è quello di determinare quali punti di una immagine corrispondono a quelli dell'altra.

L'immagine da cui parte l'analisi dello *stereo-matching* prende il nome di ***immagine primaria,*** mentre quella su cui viene fatta la ricerca della corrispondenza, viene chiamata immagine secondaria. Una generica tecnica di *stereo-matching* parte da un punto dell'immagine primaria ed effettua una ricerca di corrispondenza dello stesso sull'immagine secondaria.

La geometria epipolare permette di semplificare questa ricerca (*matching*) introducendo alcuni vincoli. Il più importante viene chiamato vincolo epipolare e permette di limitare la ricerca della corrispondenza ad una singola retta di tutta l'immagine secondaria.

Si consideri il piano formato dai punti O_R , O_L e P, che è il medesimo piano che contiene anche i punti O_R, P_R e P, oppure O_L, P_L e P. Prendendo il piano formato da O_R , O_L e P_L, il punto P_R dovrà trovarsi sul medesimo piano e, quindi, la ricerca della corrispondenza viene limitata solamente a quel piano.

$$linea\ epipolare\ coniugata\ a\ M_1 = \pi_{epip} \cap \pi_2 = {}^{\pi_2}l_{\pi_1}$$

Nel caso in cui le rette che passano per il centro immagine siano parallele (ad esempio nel caso di camere parallele e non convergenti), anche le linee epipolari saranno tra loro parallele e coincideranno con le righe dell'immagine stessa.

L'ipotesi fondamentale su cui si basano i sistemi di visione stereoscopica è che l'intorno dei punti corrispondenti tra le due immagini deve essere simile. Più questa ipotesi sarà falsificata, maggiori saranno i problemi nel riuscire ad identificare le corrispondenze e sarà più arduo procedere alla ricostruzione dei punti nella scena 3D.

La scelta della baseline giusta

Alla luce dell'ipotesi fondamentale su cui si basano i sistemi di visione stereoscopica, è possibile fare qualche considerazione sulla linea base tra le due immagini (o tra le due camere).

Ricordiamo che, come visto in precedenza, si sta operando con dei sensori digitali, formati da una matrice di elementi fotosensibili. I punti della scena sono quindi soggetti ad un campionamento spaziale e la frequenza di questo campionamento dipende dalla dimensione e dalla distanza dei pixel del sensore stessi.

Quando ricostruiamo la scena, cioè quando retro-proiettiamo i punti della scena 2D verso il mondo 3D, in realtà stiamo proiettando dei punti di dimensione finita e non infinitesimale.

Questo fenomeno causa un'area di incertezza nella collocazione dei punti nel mondo 3D.

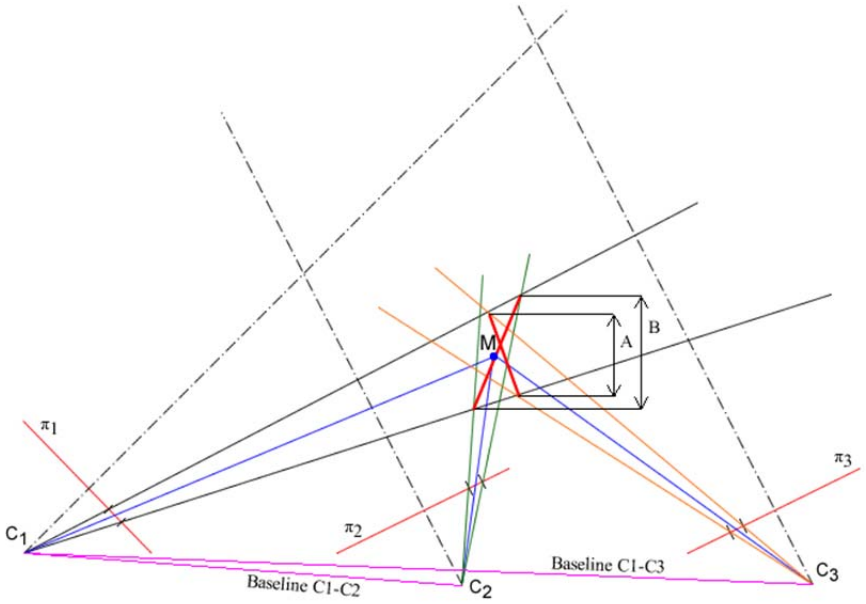


Figura 11 – L'immagine mostra un sistema di stereovisione con due baseline differenti. Viene rappresentata la proiezione dei bordi dei pixel a cui il punto M appartiene per ogni camera. Le aree di incertezza A e B si riferiscono, rispettivamente, alla baseline C1-C3 e alla baseline C1-C2. Le linee tratteggiate sono gli assi ottici delle camere.

Nella Figura 11 sono rappresentate due coppie di camere con diversa baseline (la camera C1 è in comune ad entrambi i sistemi stereo). Sono stati proiettati i bordi dei pixel che contengono il punto M della scena 3D. Questa proiezione determina l'area di incertezza della retro-proiezione del punto M. Come è possibile vedere, l'area di incertezza B creata dalla coppia di immagini π_1 e π_2 è nettamente maggiore dell'area di incertezza A formata dalla coppia di immagini π_1 e π_3. Si noti che la dimensione del pixel è stata rappresentata uguale in tutte e tre i piani immagine.

Viene quindi spontaneo pensare che tanto più è ampia la linea base, tanto migliore sarà il risultato della ricostruzione.

Tuttavia, l'ipotesi fondamentale della visione stereoscopica specifica che l'intorno immagine del punto M dell'immagine primaria deve essere simile all'intorno immagine del punto M dell'immagine secondaria. Se la baseline diventa sempre più grande, allora varieranno anche gli intorni dei punti immagine. Basti pensare ad una superficie non perfettamente opaca: con il cambiare dell'angolo di osservazione, cambierà anche la quantità di luce riflessa e, quindi, l'intorno del punto che stiamo osservando nelle due immagini.

E' importante, quindi, scegliere una baseline opportuna che permetta di avere un'area di incertezza della ricostruzione relativamente ridotta e che, tuttavia, non renda le immagini troppo dissimili.

Calibrazione e parametri della "testa stereo"

La calibrazione di una testa stereo parte dal presupposto che le due camere da cui è costituita siano calibrate. I loro parametri intrinseci, infatti, rimarranno invariati nella testa stereo, in quanto interni alle camere stesse. I parametri estrinseci, invece, verranno utilizzati per modellare la posizione di una camera rispetto all'altra. Come abbiamo visto nella sezione precedente, nel problema dello *stereo-matching* si parte da un'immagine, generalmente quella catturata dalla camera di sinistra, e si ricercano le corrispondenze sull'altra.

In un mondo ideale, sarebbe possibile posizionare le due camere esattamente nella posizione voluta, ad esempio, con assi ottici paralleli. Nel mondo reale questo risultato non è, però, raggiungibile e rimarrà sempre un certo grado di incertezza nel posizionamento delle due camere. I parametri estrinseci delle singole camere vengono utilizzati per computare precisamente dove è posizionata la camera secondaria rispetto alla camera di riferimento. Una volta identificata la rototraslazione che lega le due camere, il sistema di riferimento sarà quello della camera principale.

Elaborazione delle immagini a basso livello

Come abbiamo visto in precedenza, i sensori digitali sono soggetti ad alcuni problemi, come quello del rumore causato dalla black-current, un contrasto dinamico relativamente ridotto se paragonato a quello dell'uomo e quello della scena 3D che viene campionata spazialmente ad una frequenza dipendente dalla distanza e dalla dimensione dei pixel del sensore stesso.

In questo capitolo verranno presentate le tecniche di elaborazione a basso livello che consentono di ottenere delle immagini utilizzabili per svolgere delle elaborazioni di alto livello, come il calcolo della disparità.

Fondamenti: risoluzione, pixel deph, contrasto ed istogramma

Le immagini digitali sono una rappresentazione della quantità di luce incidente sui sensori. Così come i sensori sono delle matrici di elementi fotosensibili, le immagini sono delle matrici di valori d'intensità luminosa.

La risoluzione è la misura di quanti dettagli può contenere un'immagine o meglio, della qualità di un'immagine. La risoluzione può essere misurata in modi diversi, ma sostanzialmente quantifica quanto due linee possono essere vicine per essere ancora distinte visivamente. Essa può essere riferita alla dimensione fisica (pixel per mm^2) oppure alla dimensione totale dell'immagine (numero di pixel per l'altezza dell'immagine).

La più comune interpretazione di risoluzione è quella del conteggio dei pixel nell'immagine; ad esempio un'immagine con N righe e M colonne, avrà una risoluzione di NxM pixels. La risoluzione spaziale determina i più piccoli dettagli che sarà possibile distinguere nell'immagine che si sta osservando. Maggiore sarà la risoluzione, più piccoli saranno gli oggetti visibili e distinguibili gli uni dagli altri.

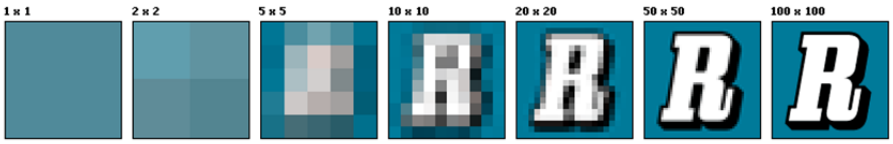


Figura 12 – La figura mostra un'immagine a diverse risoluzioni. In ordine, da sinistra a destra, abbiamo una immagine 1x1 (formata da un solo pixel), 2x2, 5x5, 10x10, 20x20, 50x50 e 100x100 come ultima. Come è possibile notare, i dettagli distinguibili aumentano man mano che la risoluzione stessa aumenta.

Con il termine risoluzione, tuttavia, non ci si riferisce, solamente a quella spaziale, cioè al numero di pixel, ma anche al numero di differenti livelli di luminosità che è possibile distinguere in un singolo pixel. Questo valore è molto soggettivo per l'occhio umano e dipende dalla sensibilità di ciascun individuo e dalle particolari condizioni di osservazione che si presentano. Nelle immagini questo valore è determinato dalla capacità del'ADC (Analog to Digital Converter) del sensore digitale di discriminare e convertire il livello di tensione accumulato nell'elemento fotosensibile in una sequenza di bit. Questo parametro è chiamato, generalmente, "*pixel-deph*" o "profondità colore".

Figura 13 – Le immagini sopra, da sinistra a destra e dall'alto in basso, rappresentano la stessa immagine in cui ogni pixel ha diverse profondità colore, rispettivamente 1bit (2 livelli), 2bit (4 livelli), 4bit (16 livelli) e 8bit (256 livelli).

Un altro aspetto importante delle immagini è il contrasto. Esso è definito come il rapporto tra il valore più elevato ed il valore più basso di luminosità dell'immagine. In altri termini si può affermare che il contrasto è la differenza visuale che rende possibile distinguere un oggetto (o la sua rappresentazione in un'immagine) dagli altri e dallo sfondo.

In visione artificiale, come in tutte le discipline che utilizzano le immagini per fare elaborazioni di alto livello, è essenziale che l'immagine sia ben contrastata in modo da potere distinguere gli oggetti dallo sfondo e potere discriminare i punti gli uni dagli altri.

Figura 14 – Le due immagini mostrano l'importanza del contrasto: quella di sinistra è scarsamente contrastata e per questo motivo è difficile riconoscere ed identificare i dettagli della scena. Quella di destra, invece, è ben contrastata ed possibile riconoscere chiaramente gli oggetti dallo sfondo ed i più piccoli dettagli

Il contrasto può essere misurato in diversi modi e con differenti tecniche, ma la più comune è quella del valore quadratico medio ***[37]***:

$$contrasto = \sqrt{\frac{1}{M \cdot N} \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} \left(I_{i,j} - \bar{I}\right)^2}$$

dove M e N sono le dimensioni dell'immagine, $I_{i,j}$ è il valore di intensità luminosa del pixel i,j e $\bar{I}$ è il valore medio di luminosità di tutta l'immagine. In questo caso, si suppone che l'intensità di ogni singolo pixel sia normalizzata nell'intervallo [0,1].

L'istogramma di un'immagine è una rappresentazione grafica della distribuzione dei diversi livelli di luminosità dei pixel dell'immagine.

L'asse orizzontale dell'istogramma rappresenta i vari livelli di luminosità ammissibili dai pixel e per ognuno di essi, sull'asse verticale, è rappresentato il numero di pixel che hanno quel particolare livello. Nella parte più a sinistra dell'istogramma sono presenti le tonalità più scure dell'immagine, mentre in quella più a destra le tonalità più chiare.

L'uso dell'istogramma è molto utile per capire come sono distribuite le intensità dei pixel e capire se sono uniformi come in un'immagine ideale, oppure se ci sono delle concentrazioni che potrebbero diminuire il contrasto dell'immagine.

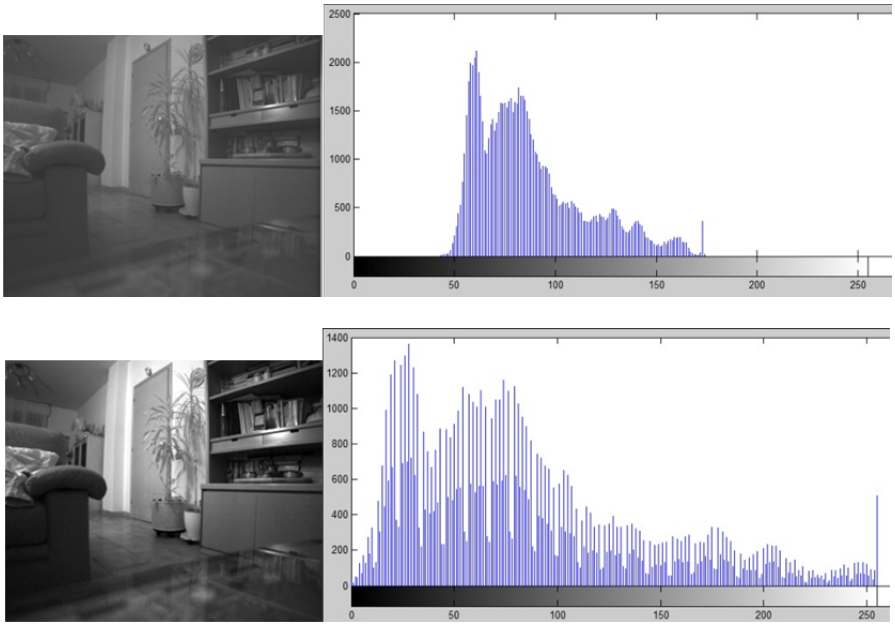


Figura 15 – Sopra sono mostrate le due immagini viste precedentemente con il relativo istogramma: come è possibile notare, l'istogramma dell'immagine che appare più contrastata copre un numero più alto di livelli di luminosità, mentre l'istogramma dell'immagine poco contrastata risulta molto compresso e non copre tutti i possibili valori.

Equalizzazione dell'istogramma

L'equalizzazione dell'istogramma è una tecnica di post-processing che mira ad aumentare il contrasto dell'immagine usando l'istogramma dell'immagine stessa.

L'idea che sta alla base di questa tecnica è quella di riuscire a distribuire le intensità dei pixel in modo uniforme lungo tutto il range dei possibili valori ottenendo, idealmente, un istogramma piatto.

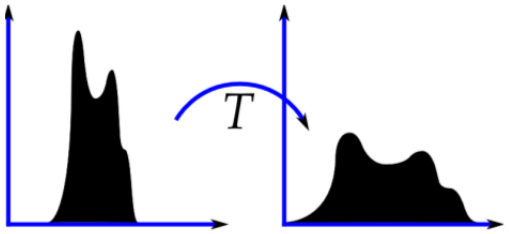


Figura 16 – L'immagine mostra l'idea che sta alla base della tecnica di equalizzazione dell'istogramma

Consideriamo un'immagine a scala di grigi *x* e chiamiamo con n_i il numero di occorrenze del livello *i* nell'immagine. La frequenza di un pixel con livello *i* nell'immagine è dato da:

$$p_x(i) = p(x = i) = \frac{n_i}{n} \text{ con } 0 \leq i \leq L$$

dove L rappresenta il numero totale di livelli (per un immagine di 8 bit, L=256), *n* rappresenta il numero totale di pixel dell'immagine e $p_x(i)$ è l'istogramma nell'immagine per il valore *i*, normalizzato tra [0,1].

Definiamo anche la funzione di distribuzione cumulativa come

$$cdf_x(i) = \sum_{j=0}^{i} p_x(j)$$

che corrisponde all'istogramma cumulativo dell'immagine. L'istogramma cumulativo dell'immagine è un istogramma in cui il valore del livello *i* è definito dalla somma delle occorrenze di tutti i livelli precedenti più le occorrenze del livello *i* attuale. Questo particolare grafico avrà un andamento sempre crescente e terminerà con l'ultima barra corrispondente al numero di pixel dell'immagine.

La forma ideale di un istogramma cumulativo, a differenza di quello tradizionale che deve essere idealmente il più piatto possibile, è quella di una funzione crescente il più possibile lineare.

Creiamo, ora, una trasformazione $y = T(x)$ per creare una nuova immagine *y* in cui l'istogramma cumulativo, o meglio la funzione di distribuzione cumulativa, risulti linearizzato nell'intervallo. Per esempio:

$$cdf_y(i) = i \cdot K$$

per una qualche costante *K*. Possiamo, quindi, avere la seguente trasformazione

$$y = T(x) = cdf_x(x)$$

Eliminare le distorsioni delle ottiche

Come abbiamo visto in precedenza, le ottiche utilizzate dalle camere introducono delle distorsioni. Le distorsioni radiali, tra cui troviamo quella a barile e quella a cuscino, sono tra quelle più semplici ed efficacemente corrette mediante l'uso di specifici algoritmi. Da notare che l'entità della distorsione è tanto maggiore quanto più ci si allontana dalla parte centrale dell'immagine ***[6]***.

Le distorsioni radiali possono essere corrette usando il modello di distorsione di Brown. Tale modello può essere utilizzato sia per la distorsione radiale che per quella tangenziale, causata dal non perfetto allineamento degli elementi delle ottiche.

$$x_u = x_d + (x_d - x_c)(K_1 \cdot r^2 + K_2 \cdot r^4 + \cdots) + (P_1(r^2 + 2(x_d - x_c)^2) + \\ +2P_2(x_d - x_c)(y_d - y_c)) \cdot (1 + P_3 r^2 + \cdots)$$
$$y_u = y_d + (y_d - y_c)(K_1 \cdot r^2 + K_2 \cdot r^4 + \cdots) + (P_1(r^2 + 2(y_d - y_c)^2) + \\ +2P_2(x_d - x_c)(y_d - y_c)) \cdot (1 + P_3 r^2 + \cdots)$$

dove:

- (x_u, y_u) = punto dell'immagine senza distorsione
- (x_d, y_d) = punto dell'immagine distorto
- (x_c, y_c) = punto centrale dell'immagine (punto principale)

- K_n = n-esimo coefficiente di distorsione radiale
- P_n = n-esimo coefficiente di distorsione tangenziale
- $r = \sqrt{(x_d - x_c)^2 + (x_d - x_c)^2}$

Le distorsioni a barile, di solito, hanno un valore positivo per il primo coefficiente delle distorsioni radiali K, mentre le distorsioni a cuscino hanno un valore negativo.

I sistemi di calibrazione delle camere esistenti sono in grado, durante il processo di calibrazione stesso, di effettuare una stima dei coefficienti di distorsione radiale e tangenziale presenti nelle immagini.

Esistono metodi consolidati per rimuovere le distorsioni delle immagini già integrati nelle librerie dei più importanti ambienti di sviluppo ad alto livello come Matlab ***[7] [8]*** o in librerie di programmazione come le OpenCV ***[9]***.

L'eliminazione delle distorsioni dell'immagine è un processo di fondamentale importanza nella visione artificiale: se non si eseguisse questo passaggio, durante la ricostruzione, cioè durante la retro-proiezione dei punti della scena 2D verso il mondo 3D, si farebbe riferimento ad una posizione in realtà distorta dalle lenti, ottenendo una tridimensionale della stessa errata.

Disparità e visione 3D

In precedenza abbiamo analizzato come calibrare le camere e quali sono i parametri che le caratterizzano, successivamente abbiamo visto come utilizzarne due per realizzare un sistema visivo binoculare che consente di ricostruire la posizione tridimensionale dei punti 2D rappresentati nella scena. Sono stati poi presentati i fondamenti dell'elaborazione delle immagini e delle tecniche di post-processing di basso livello e i metodi di correzione delle distorsioni introdotte dalle lenti. A questo punto, si hanno a disposizione immagini

utilizzabili (corrette ed ottimizzate) della scena presa da due prospettive differenti.

Vediamo ora come utilizzare tutto questo per creare la mappa di disparità ed ottenere una rappresentazione dello spazio libero del mondo.

Stereo matching e mappa della disparità

In precedenza abbiamo visto che utilizzando due immagini che rappresentano la stessa scena da due prospettive, è possibile eseguire la ricostruzione dei punti delle immagini nella scena 3D. Per eseguire questo procedimento, è necessario prendere i punti appartenenti all'immagine primaria ed identificarli nell'immagine secondaria; il vincolo epipolare ci aiuta a restringere la zona di ricerca, ma non ci dice esattamente dove ricercare il punto.

Le tecniche di stereo-matching permettono di risolvere questo problema. Ci sono due principali approcci per eseguire il matching dei punti dell'immagine primaria in quella secondaria:

- quelli basati sulle caratteristiche (*feature-based*), in cui dall'immagine primaria vengono estratte un certo numero di features che vengono confrontate, mediante appositi algoritmi, con quelle dell'immagine secondaria alla ricerca di una corrispondenza
- quelli basati sui pixel (*pixel-based*), dove l'analisi viene svolta senza alcuna conoscenza o ricerca del contenuto dell'immagine, ma analizzando i singoli pixel della stessa. Tra queste tecniche, la più importante è quella degli algoritmi a correlazione che si basa sull'analisi della somiglianza di aree intorno ad un pixel tra l'immagine primaria e quella secondaria. Tra i più popolari, c'è sicuramente l'algoritmo SSD (Square Sum of Differences) ***[38]***.

L'idea che sta alla base degli algoritmi a correlazione pixel-based è quella di sfruttare l'energia del pixel, ovvero la luminanza, per effettuare una stima della somiglianza. In questo processo è spesso

utilizzata la multi-risoluzione per raffinare progressivamente la ricerca delle corrispondenze. Si crea uno spazio delle scale (*scale-space*) applicando iterativamente alle immagini (sia quella primaria che quella secondaria) un filtro passa-basso o di smoothing in modo da eliminare i dettagli più piccoli. Una volta eseguita la ricerca sulle immagini a risoluzione più bassa, si passa, man mano, a quelle con maggiori dettagli usando, però, le informazioni precedentemente calcolate. In questo modo è possibile procedere ad una ricerca della corrispondenza in un'area più ristretta ed è più facile, quindi, individuare delle somiglianze. L'idea che sta alla base degli algoritmi a correlazione è quella di far "scorrere" una finestra di dimensione opportuna lungo tutta l'immagine secondaria al fine di trovare in quali posizioni la somiglianza con l'immagine primaria di una uguale finestra è massima. La multi-risoluzione aiuta limitando la dimensione della finestra di ricerca.

Uno dei più diffusi algoritmi a correlazione pixel-based è quello sviluppato da Mar, Poggio e Grimson. Esso basa il suo funzionamento sulla tecnica di rilassamento e sul vincolo di continuità ed unicità[2]. Prendiamo un pixel generico dell'immagine primaria e chiamiamolo n_i ed un pixel generico dell'immagine secondaria, chiamandolo m_j. Per ognuno di essi viene calcolata una misura di confidenza "*c*" che specifica quanto sono simili, o per meglio dire, quanto è probabile, che i due pixel siano associati.

---

[2] Oltre al vincolo epipolare visto in precedenza, è possibile definire altri vincoli. I principali sono:

- Unicità → Se si identifica una relazione, o vincolo, tra una feature dell'immagine primaria ed una feature dell'immagine secondaria, allora non dovrà esistere nessun altro vincolo tra queste due features
- Continuità → Si ipotizza che il mondo e la scena siano composti da oggetti con superfici semplici e continue
- Vincolo di ordinamento → il vincolo di ordinamento può essere applicato ad oggetti opachi: tale vincolo asserisce che punti che mantengono lo stesso ordinamento nell'immagine secondaria non sono ammessi.
- Vincolo del gradiente funzione disparità → Se si aggiunge un vincolo sul gradiente della disparità, si genera una zona proibita più ampia di quella creata dal vincolo dell'ordinamento

$$c(n_i, m_j) = \begin{cases} 1 \; se \; I(n_i) - I(m_j) < S \\ 0 \; altrimenti \end{cases}$$

Dove $I(n_i)$ e $I(m_j)$ sono rispettivamente le intensità dei pixel n_i e m_j e S è una certa soglia.

I pixel m_j con cui eseguire il confronto sono scelti tra quelli appartenenti alla linea epipolare del punto n_i nell'immagine secondaria. La misura di confidenza è quindi aggiornata mediante la formula:

$$c^{(n+1)}(n_i, m_j) = \begin{cases} 1 \; se \; |n_i'| \in V_i \mid c^n(n_i', m_j') = 1 \; con \; m_j' \in V_j \\ 0 \; altrimenti \end{cases}$$

dove V_i e V_j rappresentano, rispettivamente, un intorno del pixel n_i e m_j ***[24] [2]***.

Un altro algoritmo di stereo-matching molto diffuso, questa volta basato sulle features dell'immagine, è quello proposto da Pollard, Moyhew e Frisby. Questo algoritmo prevede una pre-elaborazione delle immagini da cui vengono estratti dei "*token*" che, preferibilmente, devono includere un certo numero di features. Quindi viene calcolato il valore c_{ij} che rappresenta la bontà di associazione del token t_i dell'immagine primaria al token t_j di quella secondaria.

Viene considerato un intorno dei token t_i e t_j (in particolare dei pixel m_i e n_j) che chiameremo, rispettivamente, pixel m_k *e* n_l. Tra tutti i pixel m_k *e* n_l ve ne sono solo alcuni che supportano l'associazione e sono quelli che rispettano il vincolo di disparità del gradiente:

$$DG(m_i, n_j, m_k, n_l) < \delta$$

Solamente i pixel m_k *e* n_l che supportano questa relazione supporteranno il match tra m_i e n_j.

Viene, quindi, calcolata la forza dell'associazione (*strenght match*) come segue:

$$SM(m_i, n_j) = c_{i,j} \cdot \sum_{t_k \in V_i} \frac{1}{dist(m_i, m_k)} \cdot \max_{t_l}\left(\frac{c_{k,l}}{dist(n_j, n_l)}\right) \cdot \sigma\big(DG(m_i, m_j, n_k, n_l)\big)$$

dove

$$\sigma(DG) = \begin{cases} 1 \; se \; DG < \Delta \\ 0 \; se \; DG > \Delta \end{cases}$$

Ad ogni iterazione, le associazioni per cui si ha una forza massima tra i token delle immagini vengono scelte come corrette. A causa del vincolo di unicità, tutte le altre associazioni legate a questi token sono eliminate e non vengono più prese in considerazione ***[2]***.

Mappa di disparità

Una volta trovate le associazioni tra i pixel, o le features, dell'immagine primaria verso l'immagine secondaria è possibile calcolare una mappa della diversità tra le stesse. Questa mappa viene chiamata ***mappa di disparità*** ed è di fondamentale importanza nella stereo visione.

Supponiamo di avere un punto $M_1(x_1, y_1)$ ed il suo punto equivalente nell'immagine secondaria che possiamo definire in relazione al pixel M_1stesso: $M_2(x_1 + \Delta x, y_1 + \Delta y)$:

$$M_1(x_1, y_1) \rightarrow M_2(x_2 = x_1 + \Delta x, y_2 = y_1 + \Delta y)$$

Δ è una funzione vettoriale:

$$\Delta = \begin{cases} \Delta x = \Delta x(x_1, y_1) \\ \Delta y = \Delta y(x_1, y_1) \end{cases}$$

Se si calcola questa funzione per ogni pixel dell'immagine principale, si otterrà un'immagine (mappa) che rappresenta le differenze tra le prospettive dell'immagine principale e di quella secondaria.

Figura 17 – Viene mostrato, da sinistra a destra, l'immagine catturata dalla camera di sinistra, l'immagine catturata dalla camera di destra e la mappa di disparità tra queste due immagini, opportunamente riscalata per essere visualizzata. Valori alti (colori chiari) corrispondono a grandi differenze tra le due immagini, mentre valori bassi (colori scuri) corrispondono a differenze minime. Più gli oggetti sono vicini alle camere, maggiore sarà la differenza che i due dispositivi osserveranno e più elevato sarà il valore della mappa di disparità.

I valori della mappa di disparità possono essere quindi riscalati in modo da essere visualizzati, come mostrato in Figura 17.

Gli algoritmi pixel-based, generalmente, consentono di ottenere delle mappe di disparità dense ovvero con un elevato numero di valori di disparità per superfice dell'immagine. Allo stesso tempo, però, tali mappe risulteranno più influenzate dal rumore presente nell'immagine che condiziona negativamente gli algoritmi per il calcolo della disparità che si basano, generalmente, sulle informazioni di luminanza dei singoli pixel. Gli algoritmi feature-based, al contrario, restituiscono delle mappe sparse, ovvero con poche informazioni, ma generalmente più precise perché si basano sull'analisi di dettagli di più alto livello, meno influenzati dal rumore. L'obiettivo della stereo visione è, ovviamente, quello di ottenere delle mappe della disparità dense e precise allo stesso tempo.

U-V disparity

La scena è spesso modellata mediante un insieme di piani appartenenti al mondo. La stima di piani 3D può essere ridotta ad un'identificazione di una linea retta in 2 dimensioni. La scena può essere modellata e semplificata come giacente su un piano orizzontale, che rappresenta la superficie stradale su cui il veicolo si muove, e con una serie di piani verticali che corrispondono agli oggetti presenti nella scena stessa. In questa condizione, la v-

disparity è una mappa che consente di stimare efficientemente il piano stradale.

La mappa v-disparity viene calcolata accumulando i pixel aventi lo stesso valore di disparità sull'asse "u" dell'immagine lungo le righe. Consideriamo un pixel P di coordinate (v_p, δ_p) della mappa v-disparity. L'intensità, o valore, del pixel P corrisponde al numero di pixel della riga v_p della mappa di disparità tradizionale che hanno valore di disparità δ_p. Grazie all'accumulo dei valori, la mappa v-disparity è robusta al rumore che è presente nella normale mappa di disparità. Questo risultato è estremamente utile per potere facilmente identificare il piano stradale e gli ostacoli posti sullo stesso. Il primo viene rappresentato come una linea obliqua che inizia nell'angolo in basso a destra dell'immagine e cresce verso l'angolo in alto a sinistra. Il piano rappresentato da un eventuale soffitto, invece, avrà il suo inizio nella parte in alto a destra della mappa di disparità per decrescere verso la parte in basso a sinistra ***[28] [27]***.

Gli ostacoli presenti sul piano stradale sono rappresentati da linee verticali. Utilizzando le tecniche per la rilevazione di linee nelle immagini, come ad esempio la trasformata di Hough ***[5] [10]***, è possibile determinare le equazioni della retta che rappresenta il piano stradale. I punti appartenenti a tale piano vengono quindi riportati sulla mappa di disparità e sono marcati in modo da non essere considerati come ostacoli. I punti rimasti vengono ulteriormente analizzati per definire se sono ostacoli o no.

Usando la mappa v-disparity è inoltre possibile effettuare delle stime sull'inclinazione delle camere rispetto al piano stradale e sull'altezza delle stesse.

La mappa u-disparity viene calcolata accumulando i pixel aventi lo stesso valore di disparità sull'asse "v" dell'immagine. Esattamente come per la v-disparity, anche la u-disparity, essendo una mappa ad accumulo, è robusta al rumore presente nella mappa di disparità tradizionale.

A differenza della v-disparity, però, la u-disparity non viene utilizzata per estrarre dei piani immagini, ma per identificare i pixel degli ostacoli. Un valore alto di un pixel della u-disparity, infatti, indica che molti pixel della mappa di disparità tradizionale hanno il medesimo valore di disparità lungo una singola colonna e quindi, non appartengono alla strada. Rappresentano, invece, allineamenti verticali nel mondo 3D e, quindi, vengono classificati come ostacoli ***[28] [27]***.

La Figura 18 mostra un esempio di mappa della disparità, di v-disparity e u-disparity. Come è possibile notare, nella mappa v-disparity è facilmente riconoscibile la retta corrispondente al piano stradale, mentre nella u-disparity sono ben visibili gli ostacoli di fronte alle camere.

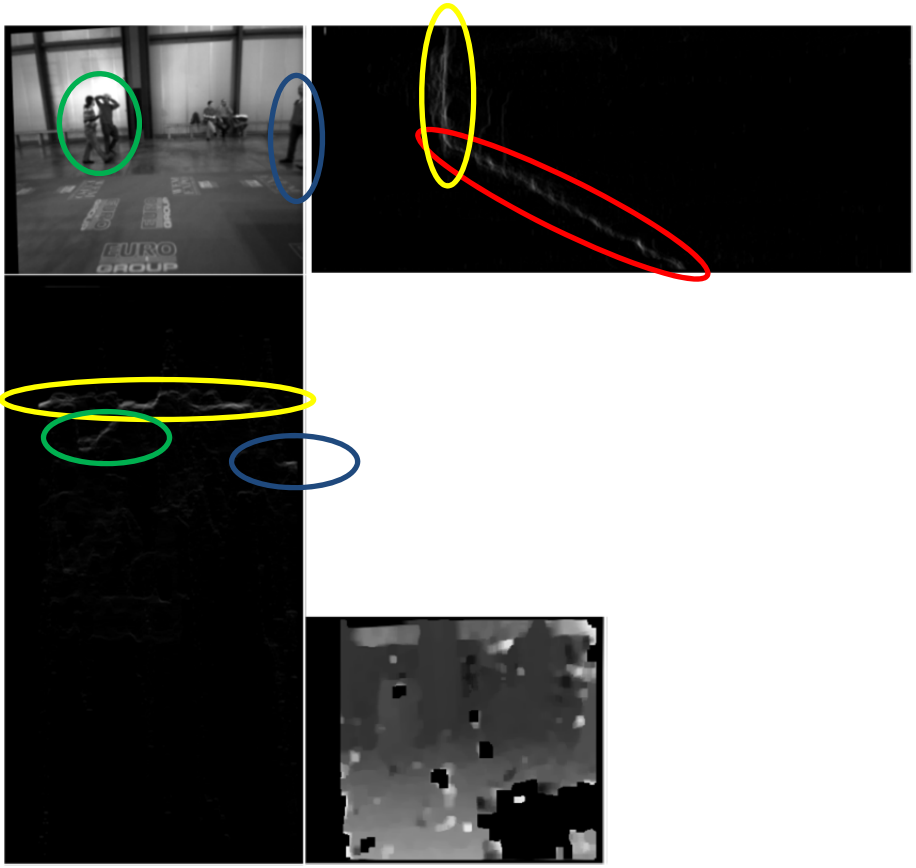


Figura 18 – L'immagine mostra, da sinistra a destra e dall'alto in basso, l'immagine della camera primaria opportunamente rettificata, la mappa v-disparity, la mappa u-disparity e la mappa di disparità tra l'immagine primaria e secondaria. L'ellisse rossa mostra sulla v-disparity la retta che rappresenta il piano stradale, mentre l'ellisse gialla, presente sia

nella v-disparity che nella u-disparity, rappresenta il muro che si vede in fondo alla scena. Nella mappa u-disparity è possibile notare due ellissi verde e blu a cui corrispondono due ostacoli identificati dai medesimi colori nell'immagine della camera principale.

Mappa dello spazio libero

Utilizzando le informazioni contenute nella mappa v-disparity è possibile classificare i pixel dell'immagine come appartenenti o meno al piano stradale. Tuttavia, per gli altri pixel non è possibile determinare con esattezza se siano ostacoli, rumore o qualcos'altro. Per potere eseguire questa valutazione viene usata la mappa u-disparity.

La classificazione degli ostacoli può essere fatta seguendo un semplice algoritmo: chiamiamo con P un pixel della mappa di disparità che non ha valore 0. Questo punto potrà essere identificato sia nella mappa v-disparity, che nella mappa u-disparity.

Se il punto rappresentato nella mappa u-disparity corrispondente al pixel P ha un valore di intensità alto, allora il pixel P viene considerato come ostacolo; altrimenti, se il punto rappresentato nella v-disparity ha un valore di intensità alto ed appartiene alla retta che rappresenta il piano stradale, allora viene marcato come spazio libero (in quanto appartiene alla strada stessa), altrimenti nulla si può dire su cosa sia il punto P (in generale viene classificato come rumore).

Il risultato della precedente elaborazione corrisponde ad una mappa con dei punti sparsi classificati come ostacoli, come piano stradale oppure senza una classificazione. Al fine di ottenere una mappa più densa e, quindi, maggiormente sfruttabile da algoritmi per la pianificazione del percorso o per la rilevazione degli ostacoli, devono essere applicate delle successive elaborazioni. I gruppi di pixel ottenuti vengono considerati come "semi" per applicare un'espansione delle relative aree: per ogni punto marcato come appartenente al piano stradale oppure come ostacolo, vengono analizzati i pixel contigui nelle mappe di v-disparity ed u-disparity. Se i pixel adiacenti hanno un valore superiore ad una certa soglia T determinata sperimentalmente, allora i corrispondenti valori nella

mappa dello spazio libero vengono marcati come appartenenti alle relative suddivisioni.

In questo modo, si ottengono delle aree più ampie di classificazione rispetto alla situazione iniziale, ma sarà ancora presente un grande numero di pixel non classificati. Per aumentare ulteriormente la densità della mappa dello spazio libero, per ogni pixel non classificato si analizzano i pixel del suo intorno. Assegnando, per esempio, a pixel appartenenti al piano stradale valori positivi e a pixel contrassegnati come ostacoli valori negativi, si sommano i valori dell'intorno del pixel non classificato considerato. Se il risultato sarà un valore positivo, allora verrà marcato come pixel appartenente al piano stradale, altrimenti, come ostacolo. In questo modo sarà possibile classificare ogni pixel della mappa dello spazio libero.

Conclusioni

In questo capitolo abbiamo analizzato lo stato dell'arte per le principali tecnologie utilizzate nell'ambito della visione artificiale e della guida autonoma di veicoli.

Nel capitolo successivo, vedremo come queste sono state implementate e quali scelte sono state fatte per ottenere un buon risultato del progetto.

Capitolo 2 - Implementazione

Nel precedente capitolo abbiamo analizzato qual è lo stato dell'arte delle tecniche e tecnologie utilizzate nella visione artificiale, quali sono i principali problemi e svantaggi e le possibili soluzioni.

Nel seguente capitolo verranno descritte le soluzioni progettuali adottate e le particolarità delle tecniche scelte.

Le camere

Abbiamo visto, nello stato dell'arte, che le camere e le loro caratteristiche rappresentano una scelta cruciale per la buona riuscita di un progetto di visione artificiale. Esaminiamo di seguito, le peculiarità e le caratteristiche delle camere adottate e le particolari tecnologie che implementano.

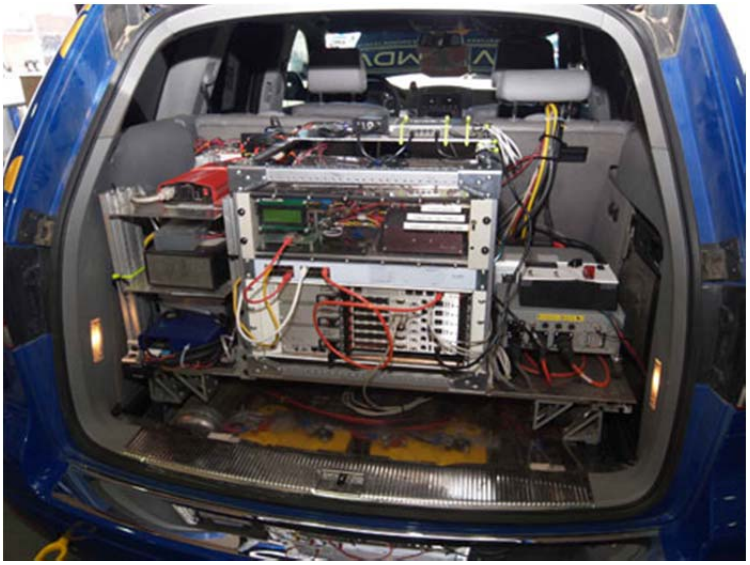
Le camere Photonfocus

Come abbiamo visto nel capitolo precedente, che le camere consumer attualmente disponibili sul mercato offrono delle caratteristiche troppo limitative per l'applicazione a veicoli a guida autonoma.

E' necessario, infatti, avere a disposizione un sensore che offra una buona sensibilità in condizioni di scarsa luce e che allo stesso tempo, abbia un rumore limitato. E' altresì importante che la camera disponga di una gamma dinamica elevata così da poter fornire immagini usabili anche in scene in cui la differenza di illuminazione è notevole.

Inoltre è necessario che le comunicazioni avvengano seguendo degli standard esistenti in modo da potere essere integrate in ambienti e sistemi pre-esistenti. In un veicolo a guida autonoma, infatti, saranno presenti un grande numero di dispositivi e sensori che dovranno comunicare sia tra loro che con un'unità centrale di elaborazione. In

applicazioni reali, la quantità d'informazioni da elaborare nell'unità di tempo è considerevole ed è difficilmente realizzabile un unico centro elaborativo per tutto il veicolo. Per questo motivo di solito il carico di computazione è suddiviso tra più calcolatori che, ovviamente, devono essere collegati a una stessa rete.



La figura mostra il bagagliaio del veicolo "Junior" creato dalla Stanford University. Com'è possibile osservare, la quantità di dispositivi e connessioni è molto elevata e per questo motivo è opportuno che ogni risorsa sia ottimizzata nel migliore dei modi.

E' di fondamentale importanza, quindi, che anche le camere si connettano alla stessa rete e che comunichino con essa in modo da ridurre le infrastrutture sul veicolo.

Le camere scelte sono il modello MV1-D1312-40-GB-12 prodotto dalla PhotonFocus.

Analizziamo ora le caratteristiche che hanno determinato la scelta di questo prodotto al posto di altri esistenti sul mercato.

Prima tra tutte la qualità del sensore: un sensore attivo CMOS A1312l con caratteristiche di tutto riguardo.

- Efficienza quantica[1] di oltre il 50%
- Pixel fill-factor[2] di oltre il 60%
- Dimensione dei singoli pixel di 8x8 µm, area totale del sensore di 10,48x8,64 mm con una risoluzione immagine massima di 1032x1082 pixel
- Profondità colore fino a 12bit (4096 livelli di luminosità identificabili)
- Shutter globale, con conseguente eliminazione del problema dello "*shuttering*"
- Range dinamico di 60dB in funzionamento lineare e di ben 120dB con tecnologia LinLog
- Dark-current a 27°C per pixel di soli 0,65fA

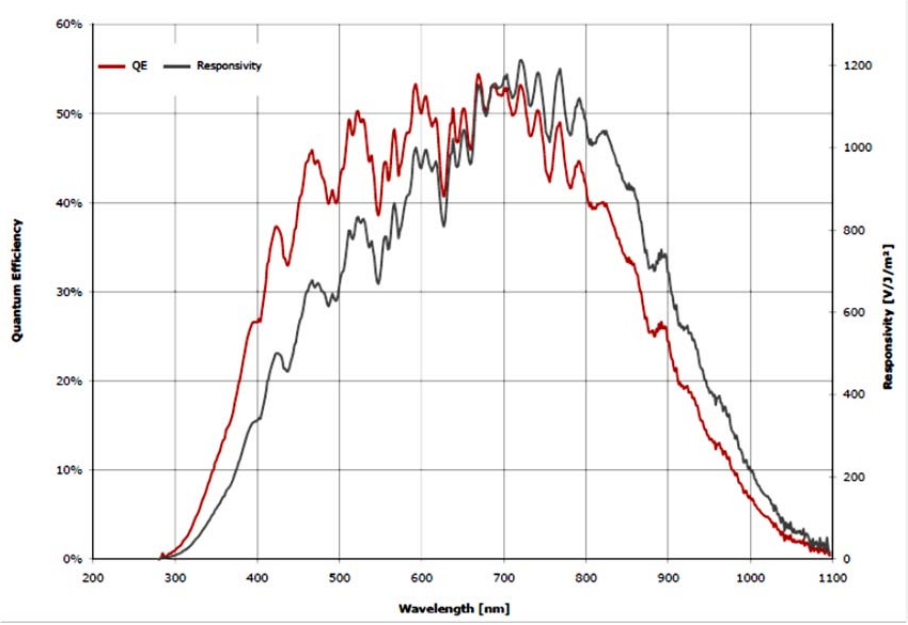


Figura 19 – L'immagine mostra il diagramma dell'efficienza quantica e della risposta del sensore utilizzato nella camera in relazione alle diverse lunghezze d'onda. Com'è possibile notare, il sensore risponde anche alle lunghezze d'onda appartenenti alla banda NIR (Near InfraRed) il che permette, ad esempio, di effettuare delle rilevazioni sui piani di scansione di sensori quali laser-scanner.

---

[1] Ricordiamo che l'efficienza quantica è la quantità di radiazione elettromagnetica incidente al sensore che viene effettivamente convertita in carica

[2] Il fill-factor, in un sensore digitale, è il rapporto tra l'area totale del sensore stesso e l'area della parte sensibile alla luce

L'ampia gamma dinamica di questo sensore (in modalità LinLog), ancorché inferiore a quella dell'occhio umano, permette di ottenere delle ottime immagini anche in scene con zone molto scure affiancate a zone molto chiare. Inoltre, il convertitore ADC da 12 bit consente di discriminare 4096 livelli di luminanza della scena, a differenza dei 256 livelli (8bit) delle normali camere diffuse sul mercato. Oltre a ciò, il sensore è stato realizzato in modo da presentare un'ampia resistenza al problema del *blooming* (o *smearing*) e ciò lo rende utilizzabile anche in condizioni di forte controluce.

La camera è dotata di una connessione Gigabit Ethernet (GigE) ed è quindi adatta all'interfacciamento con i calcolatori a bordo del veicolo a guida autonoma e agli altri dispositivi (che generalmente dispongono anch'essi di connessione Ethernet).

Un altro fattore da non sottovalutare nella scelta delle camere è il consumo energetico. In un veicolo a guida autonoma elettrico, infatti, tutto deve essere alimentato da batterie e il consumo dei calcolatori, dei sensori e dei dispositivi ne determina la velocità di scaricamento. E' importante, quindi, che tutto venga ottimizzato per un minore consumo.

Il consumo massimo dichiarato di questo modello è di 5.0 W anche se questo valore è stato sperimentalmente rilevato solo in fase di accensione ed inizializzazione. In condizioni di funzionamento normale, il consumo varia tra 3 e 4 W in relazione alla frequenza di cattura delle immagini.

Il corpo delle camere è costituito da alluminio con dei fori filettati utilizzabili per fissarle ad una struttura portante. Nel nostro caso, anche questo è un aspetto non trascurabile dato che le camere, per mantenere i valori di calibrazione, dovranno muoversi il meno possibile.

Ottiche grandangolari Pentax

In precedenza abbiamo già parlato delle lenti e delle ottiche e della loro importanza. Il modello teorico della camera a pin-hole è materialmente irrealizzabile se non con un foro di dimensione più grande. Per ovviare a questo problema, sono state introdotte le ottiche che hanno lo scopo di emulare il comportamento dei raggi luminosi che attraversano il foro di dimensione infinitesima.

Le ottiche Pentax B618(KA) C20616KA utilizzate in questo progetto sono delle lenti grandangolari con una distanza focale di 6.5mm ed un'apertura massima F/1.8.

L'apertura di queste lenti permette il passaggio di una notevole quantità di luce e la distanza focale relativamente ridotta consente di avere un campo di visione elevato.

Per calcolare esattamente gli angoli di visione che questa particolare lente ci permette di ottenere, consideriamo la seguente figura:

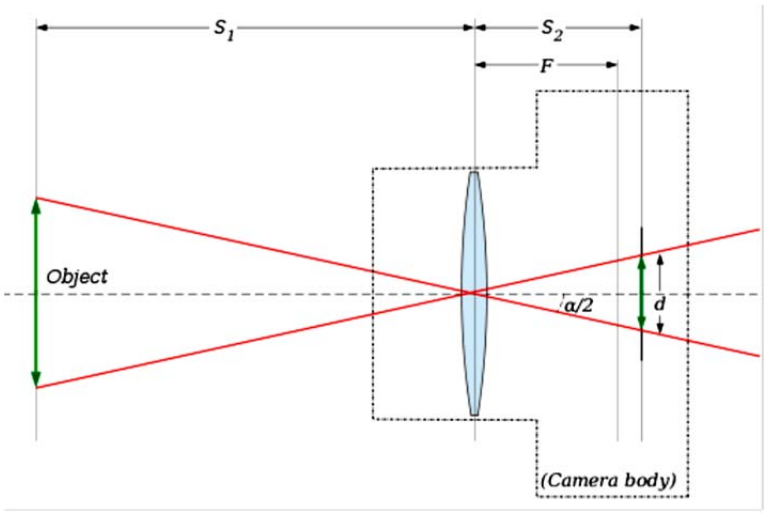


Figura 20 – La figura mostra la proiezione di un oggetto sul piano immagine (sensore), attraverso un'ottica

$\propto/2$ è l'angolo compreso tra l'asse ottico della lente e la retta congiungente il centro di proiezione con il bordo del sensore. $\propto$ è definito come l'angolo di visione, dato che rappresenta l'oggetto di dimensioni massime che può essere rappresentato sul sensore.

Usando la geometria fondamentale, possiamo scrivere che

$$\tan\frac{\propto}{2} = \frac{d/_2}{S_2}$$

Troviamo il valore di $\propto$ come:

$$\propto = 2 \cdot \tan^{-1}\frac{d}{2 \cdot S_2}$$

ma, dato che S_2 coincide proprio con la lunghezza focale (nel caso in cui la lente sia impostata per mettere a fuoco l'infinito o, comunque, un oggetto lontano), possiamo scrivere:

$$\propto = 2 \cdot \tan^{-1}\frac{d}{2 \cdot f}$$

dove *d* è la dimensione in millimetri del sensore nella direzione in cui si vuole determinare l'angolo di visuale e *f* è la lunghezza focale della lente.

Nel nostro caso il sensore è largo 10,48mm e alto 8,64mm e la lunghezza focale della lente è di 6,5mm, quindi si può calcolare che:

$$\propto_h = 77{,}8°$$

$$\propto_v = 67{,}2°$$

rispettivamente per l'angolo di visione orizzontale e verticale.

L'utilizzo di lenti grandangolari, tuttavia, porta con sè il problema delle distorsioni. Per rendere visibile una maggiore porzione della scena, infatti, l'immagine dovrà essere rimappata in modo non lineare dalla lente, introducendo delle distorsioni.

Le distorsioni di queste lenti si possono osservare nelle immagini sottostanti, ottenute durante il processo di calibrazione delle camere. E' possibile notare la presenza di una distorsione radiale molto accentuata che deve obbligatoriamente essere corretta.

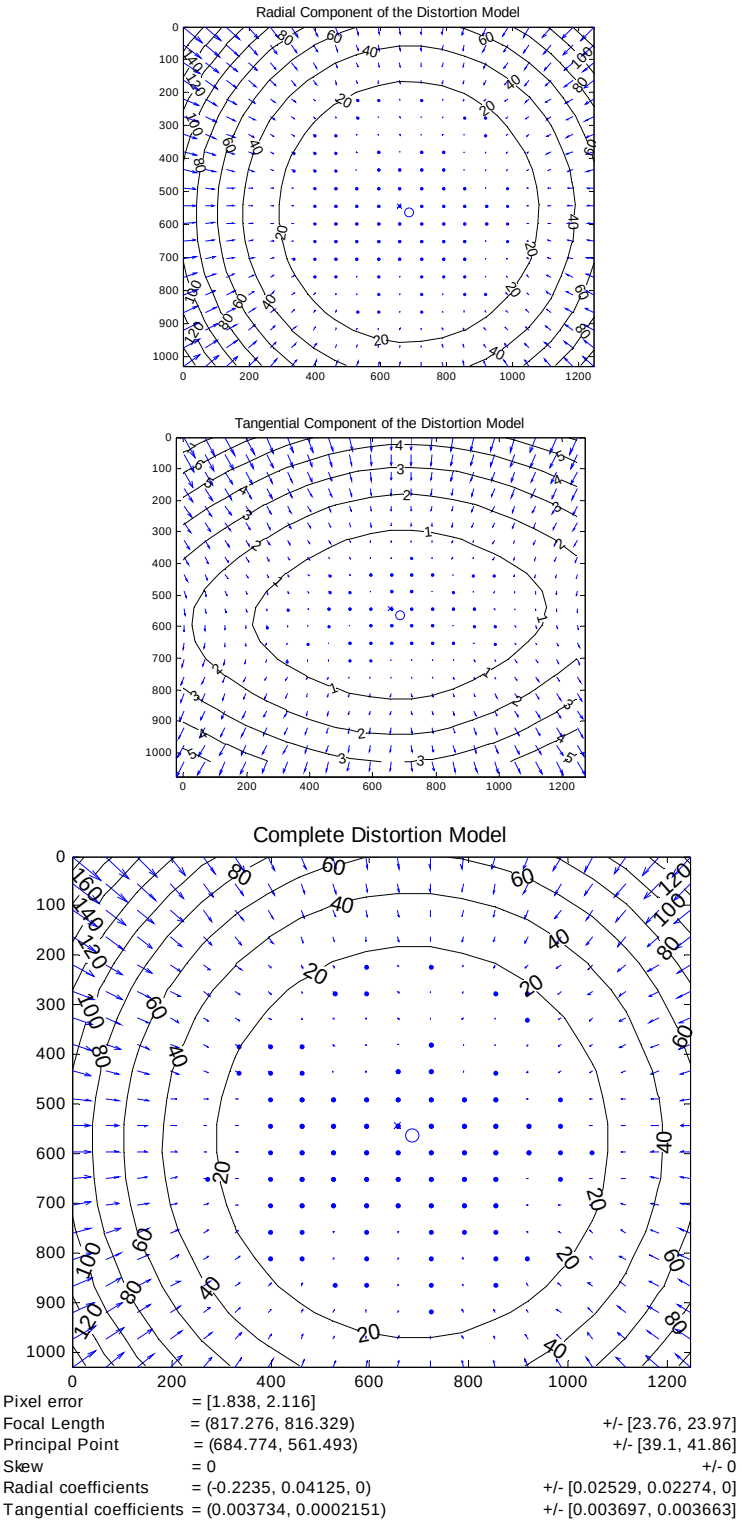


Figura 21 – Le immagini mostrano il modello di distorsione dell'ottica Pentax scelta. Dall'alto in basso, vediamo il modello di distorsione radiale, il modello di distorsione tangenziale ed infine, il modello di distorsione completo.

Nelle seguenti immagini è possibile osservare la stessa scena catturata dallo stesso punto, ma vista con due camere con ottiche differenti. E' evidente qual è il vantaggio introdotto da un'ottica grandangolare rispetto ad una tradizionale!

Figura 22 – Le immagini mostrano la stessa scena vista con due ottiche differenti. La prima (a sinistra) è l'ottica grandangolare pentax, mentre la seconda (a destra) è un'ottica non grandangolare con una distanza focale più ampia. E' possibile notare la grande differenza di campo visivo introdotta dalle due lenti.

LinLog di PhotonFocus

La tecnologia LinLog, di cui PhotonFocus è proprietaria, permette di aumentare la dinamica dell'immagine fino a 120dB senza, tuttavia, avere un'eccessiva compressione delle tonalità scure.

I sensori CMOS, avendo una circuiteria di amplificazione e conversione dedicata per ogni pixel, si prestano meglio dei CCD a tecniche hardware per l'aumento della dinamica, proprio come il LinLog. L'idea alla base di questa tecnologia è quella di "leggere" il contenuto dei singoli elementi fotosensibili prima che questi vadano in saturazione mantenendo traccia del numero di letture eseguite e del valore di tensione associato al singolo elemento fotosensibile per ogni lettura. In questo modo è come se il sensore fosse esposto per un tempo inferiore alla sorgente luminosa, ma per un certo numero di volte. Grazie a questo stratagemma, è possibile esporre il sensore a forti intensità di radiazione elettromagnetica senza che questo saturi.

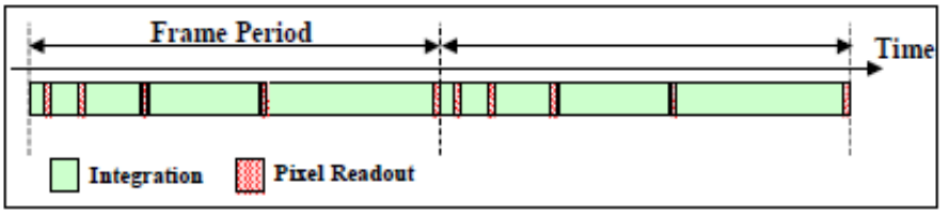


Figura 23 – L'immagine mostra la sequenza di letture (in rosso) del sensore durante il tempo di esposizione. Ogni lettura è preceduta da una fase di esposizione (verde) che è inferiore al tempo totale di esposizione dell'immagine

Il conseguimento di questo risultato avrà come conseguenza l'aumento della complessità costruttiva del circuito del sensore.

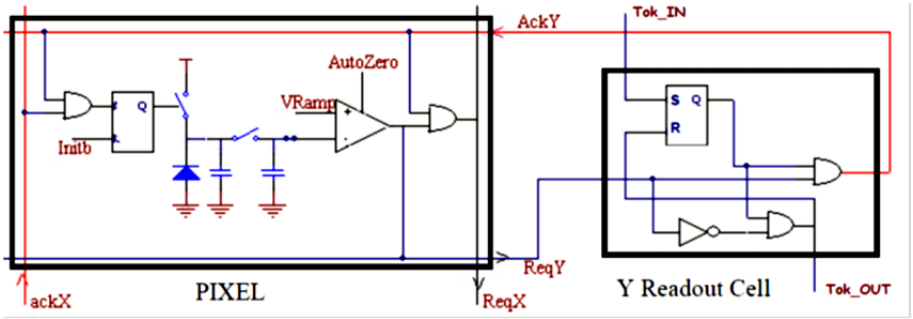


Figura 24 – La figura mostra uno schema concettuale della circuiteria di un pixel con tecnologia LinLog. La parte di acquisizione della luce rimane invariata, così come la struttura di accesso all'informazione, ma in uscita viene posto un apposito circuito che si occupa di effettuare la lettura intermedia del valore.

La Figura 24 mostra lo schema concettuale di funzionamento di un pixel di un sensore CMOS dotato di LinLog. I segnali VRamp e AutoZero fanno parte del comparatore auto-zero con segnale di comparazione a "rampa" globale. Questo tipo di segnale viene generato da un apposito DAC (Digital to Analog Converter) che definisce la precisione di lettura del livello di tensione dell'elemento fotosensibile. Quando il valore di tensione accumulato nell'elemento fotosensibile supera quello raggiunto dal segnale a rampa, il comparatore cambia di stato ed invia una richiesta di lettura (ReqY) del pixel. A questo punto, le informazioni sul valore della rampa generata dal DAC e le coordinate del pixel vengono memorizzate, in modo da tenere traccia del numero di letture per ciascun pixel.

Questa tecnica è molto funzionale perché evita la saturazione anche di un solo pixel esposto ad un intenso fascio luminoso lasciando, tuttavia, invariata la lettura di quelli circostanti.

Tale comportamento permette di rielaborare le letture riferite ai singoli pixel, ad esempio è possibile mantenere una risposta lineare per basse intensità luminose, mentre per le alte si può ottenere una compressione logaritmica.

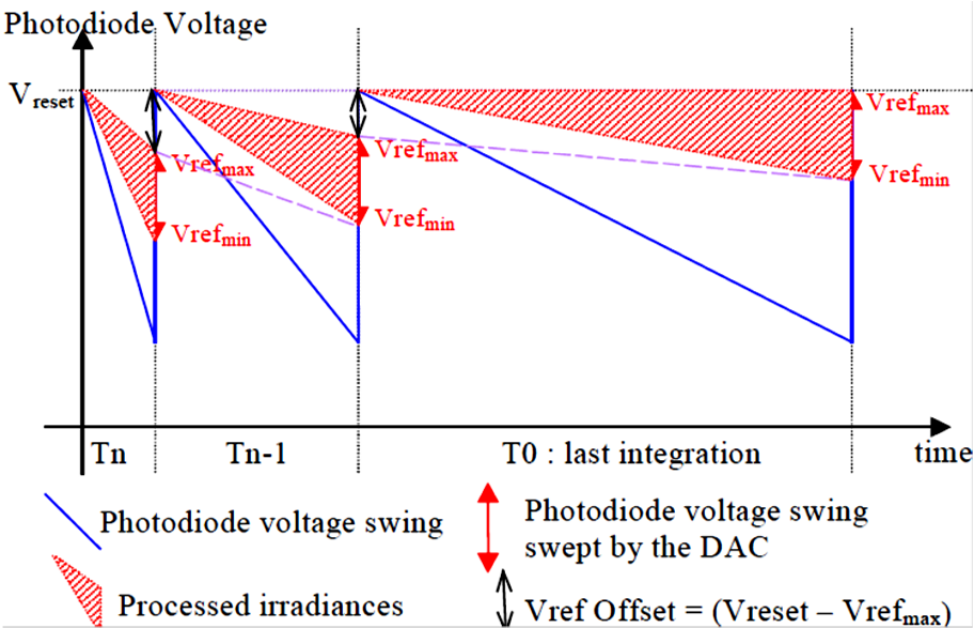


Figura 25 – La figura mostra l'andamento della tensione sul foto recettore (in rosso) rispetto alla capienza massima del pixel (in blu). Le linee verticali corrispondono ad una lettura del contenuto del pixel.

La tecnologia LinLog applicata al sensore CMOS della camera PhotonFocus permette di scegliere tra 4 impostazioni predefinite che variano la risposta e la compressione applicata alla gamma.

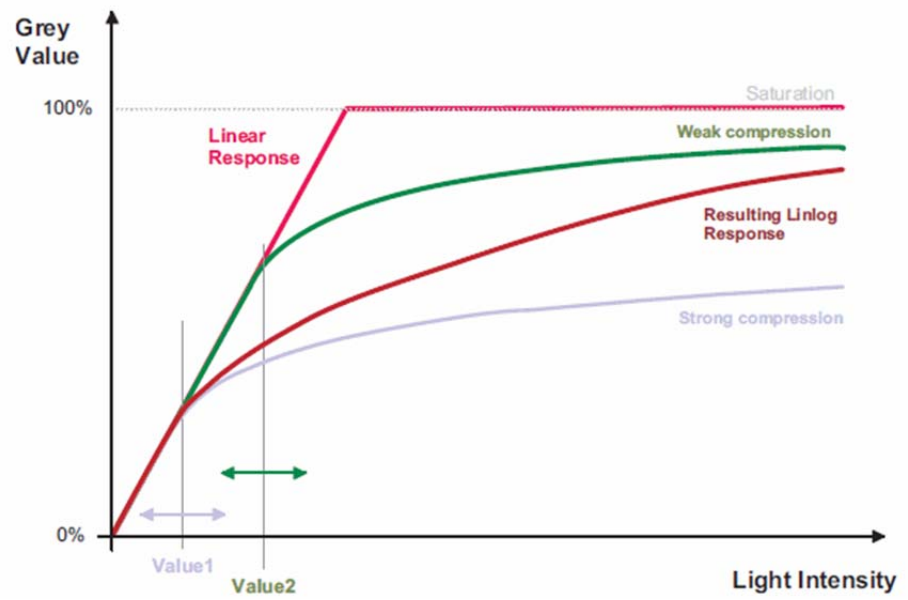


Figura 26 – L'immagine mostra l'uscita del sensore corrispondente ai diversi settings predefiniti offerti dalla camera.

Nella Figura 26 vengono mostrati i settings predefiniti resi disponibili dalla camera. La risposta lineare non offre alcuna compressione degli alti livelli e corrisponde alla disattivazione della tecnologia LinLog. Le diverse altre impostazioni, compressione debole, normale o forte, aumentano la dinamica del sensore eseguendo lo svuotamento (o lettura) frequente dei singoli pixel che stanno raggiungendo la tensione di saturazione.

La principale caratteristica di questa tecnologia è di mantenere lineare la risposta nelle aree scure dell'immagine e di comprimere esclusivamente le alte intensità luminose e ciò si rivela molto utile per mantenere i dettagli nelle aree con poca illuminazione.

I parametri personalizzabili per la funzione LinLog sono 4:

- Time1 e Time2 sono valori normalizzati secondo il tempo di esposizione e possono variare da 0 a 1000
- Value1 e Value2 possono variare da 0 a 200 e corrispondono al voltaggio applicato al sensore, ovvero alla soglia in cui fare intervenire la lettura dei dati acquisiti dal pixel.

Combinando opportunamente tali parametri è possibile individuare tre modalità di funzionamento di questa tecnologia:

- Se sono impostati solamente i parametri Value1 e Value2 allo stesso valore e s'imposta Time1 e Time2 al valore massimo, corrispondente al tempo di esposizione, allora Value1=Value2 determina il punto di transizione tra la parte lineare della risposta del sensore e la parte logaritmica. Maggiore è il valore, tanto più le alte intensità luminose saranno compresse e tanto più sarà ridotta la parte lineare di risposta del sensore. In questo caso, la risposta cambia improvvisamente da lineare a logaritmica, causando una scarsa qualità e risoluzione delle tonalità di grigio dell'immagine.

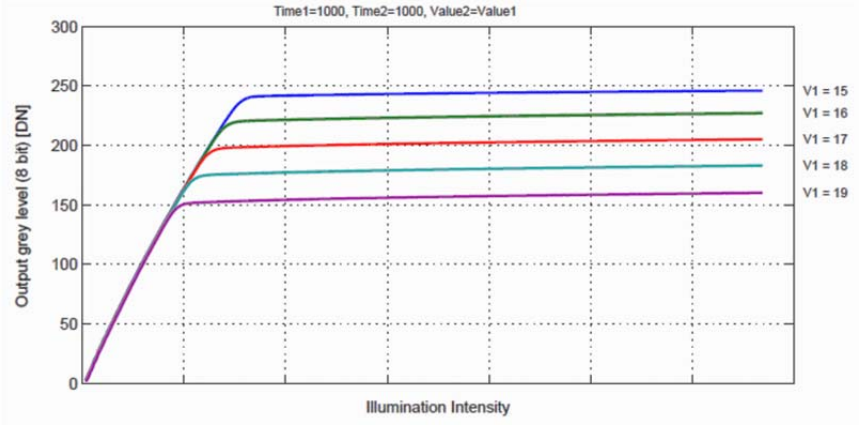


Figura 27 – L'immagine mostra la curva di risposta del sensore quando il parametro Value1=Value2 ed i parametri Time1 e Time2 corrispondono al tempo di esposizione. Il passaggio dalla zona di risposta lineare a quello di risposta logaritmica è molto brusco.

- Nella seconda modalità di utilizzo, vengono impostati i parametri Value1, Value2 e Time1, mentre Time2 viene impostato al massimo valore possibile, corrispondente al tempo di esposizione.

  Time1 determina il tempo, normalizzato secondo il tempo di esposizione della camera, in cui avviene il passaggio da Value1 a Value2. Questo significa che durante il tempo di esposizione è possibile avere due differenti compressioni logaritmiche: la prima, più forte, durante la fase iniziale (ovvero fino a Time1)

e la seconda, più debole nella fase finale (dopo Time1). Questo determina una transizione più smussata tra la zona lineare e quella logaritmica e porta ad avere una migliore risoluzione delle tonalità di grigio.

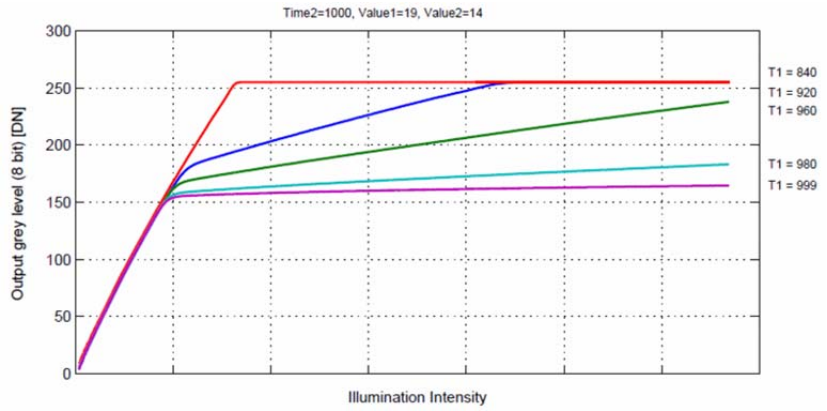


Figura 28 – L'immagine mostra la curva di risposta del sensore quando i parametri Value1 , Value2 e Time1 sono settati ed il parametro Time2 viene lasciato al suo massimo valore. La transizione tra la zona lineare e quella logaritmica è così più graduale.

- Nell'ultima modalità di funzionamento, tutti i parametri vengono utilizzati per definire il comportamento della tecnologia LinLog. Il parametro Time2 specifica il tempo, normalizzato rispetto all'intervallo di esposizione, in cui disattivare la risposta logaritmica. Si avranno, quindi, tre fasi: una prima fase con una forte compressione, quindi una fase con compressione minore ed, infine, un ultimo intervallo di tempo in cui la compressione logaritmica sarà nulla.

  Questo porta ad avere la migliore risoluzione di grigio possibile consentendo di giungere ad una dinamica fino a 120dB nel caso in cui si impostino al meglio tali parametri.

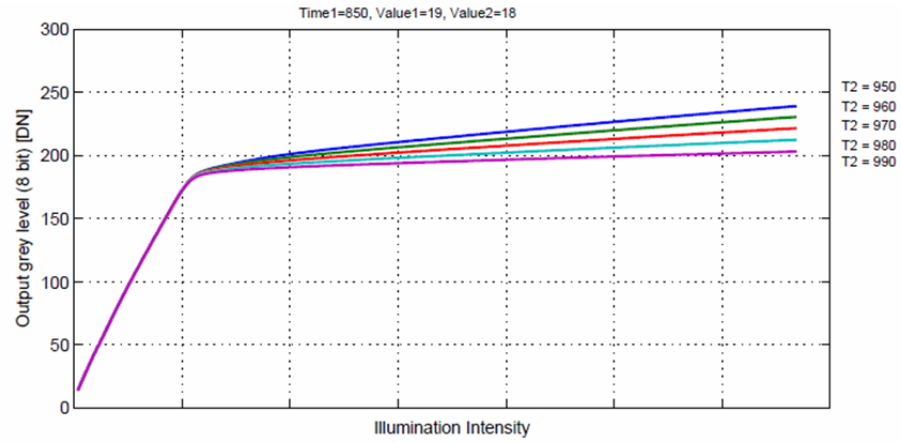


Figura 29 - L'immagine mostra la curva di risposta del sensore quando vengono utilizzati tutti i parametri Value1, Value2, Time1 e Time2. Time2 specifica il tempo in cui la risposta logaritmica diventa nulla.

Ovviamente la non linearità nella compressione sulla luminanza avrà delle conseguenze sulle varie soglie usate nelle elaborazioni successive, ad esempio nello stereo matching. Questa influenza, tuttavia, non è stata oggetto di questo lavoro di tesi.

La testa stereo per USAD

U.S.A.D. è il progetto Urban Shuttle Autonomously Driven nato con l'intento di creare un veicolo a guida autonoma in grado di guidare in normali condizioni di traffico cittadino.

Il lavoro svolto, presentato in questa tesi, è stato applicato direttamente nel progetto in oggetto. Nel capitolo corrente viene esaminata la fase di realizzazione fisica della testa stereo installata in seguito su USAD.

Progettazione della testa stereo

Il primo problema affrontato è stato quello di determinare quale porzione del mondo fosse possibile vedere e cosa, effettivamente, si voleva vedere utilizzando le camere presentate in precedenza. Una

volta montate sul veicolo, le camere saranno i suoi "occhi" ed è impensabile guidare in sicurezza se si ha una visione troppo ristretta del mondo.

Parte del progetto oggetto di questa tesi è stato dedicato a realizzare una visualizzazione che riuscisse a rendere l'idea del campo visivo delle camere montate su USAD e a determinare quale fosse l'angolo di inclinazione ottimale a cui fissarle.

L'obiettivo primario è quello di potere osservare le persone e gli oggetti che sono nell'immediata vicinanza di USAD, di avere una discreta visione laterale che permetta di accorgersi dell'entrata in scena di oggetti in movimento, prima del loro arrivo sulla traiettoria stessa e di potere distinguere una persona nella sua altezza intera.

Quest'ultimo punto è richiesto dalle applicazioni che eseguono il riconoscimento (*object-detection*) e l'inseguimento degli oggetti (*object-tracking*). In questi casi, per riuscire a marcare (o meglio categorizzare) correttamente le varie componenti del mondo è necessario vederle nella loro completezza e, di conseguenza, è importante che le persone vengano viste interamente.

Per ottenere questo obiettivo è stato scelto di montare la testa stereo sulle barre del telaio situate nella parte in alto ed anteriormente al veicolo, che si trovano ad un'altezza di circa 175 cm. Tale posizionamento è stato scelto in virtù della possibilità di un solido fissaggio alla struttura di USAD stesso (il telaio, infatti, è solidale con l'intera struttura del veicolo), della posizione riparata rispetto agli agenti atmosferici, immediatamente sotto il tetto e dietro il parabrezza, e della visuale libera ottimale disponibile da quella particolare collocazione.

E' importante che la posizione di montaggio della testa stereo non comporti la vista di parti di USAD stesso che, per ovvie ragioni, andrebbero scartate dalle elaborazioni con una conseguente riduzione dell'area utilizzabile delle immagini.

Si è realizzato in Matlab uno strumento che permette di osservare il campo visivo della testa stereo a diverse posizioni ed angolazioni di

montaggio, e di misurare con precisione le distanze di visione. E' stata inoltre aggiunta la possibilità di mostrare delle "sagome" di persone in modo da assicurarsi di poterle vedere nella loro interezza.

Si sottolinea che ogni camera dispone, come illustrato in precedenza, di un angolo visivo di 77,8° sull'asse verticale e di 67,2° sull'asse orizzontale.

Dalle misurazioni teoriche eseguite, la testa stereo montata ad un'altezza di 175 cm dal piano stradale e con un'inclinazione rispetto all'orizzontale di 25° verso la strada stessa permette di:

- vedere il piano stradale ad una distanza di 1 metro dal punto focale. Le camere sono montate a circa 50 cm dalla parte più esterna di USAD e, quindi, è possibile osservare quello che accade 50 cm di fronte al veicolo stesso,
- vedere interamente una persona alta 180 cm (circa la media) posta a 1 metro di fronte alle camere (50 cm dalla parte anteriore del veicolo),
- vedere una persona posta a 1 metro di fronte alle camere ed a 1 metro lateralmente alle stesse,
- osservare per una distanza di circa 5 metri a destra ed altrettanti a sinistra ad una distanza dal piano focale di circa 6 metri.

Questi risultati si possono definire più che accettabili come campo visivo della testa stereo perché permettono di osservare gli oggetti nella loro interezza, di avere la cognizione completa di quello che avviene vicino e lontano dal veicolo e di accorgersi con discreto anticipo dell'entrata nel campo visivo di oggetti prima che influiscano sulla traiettoria di USAD.

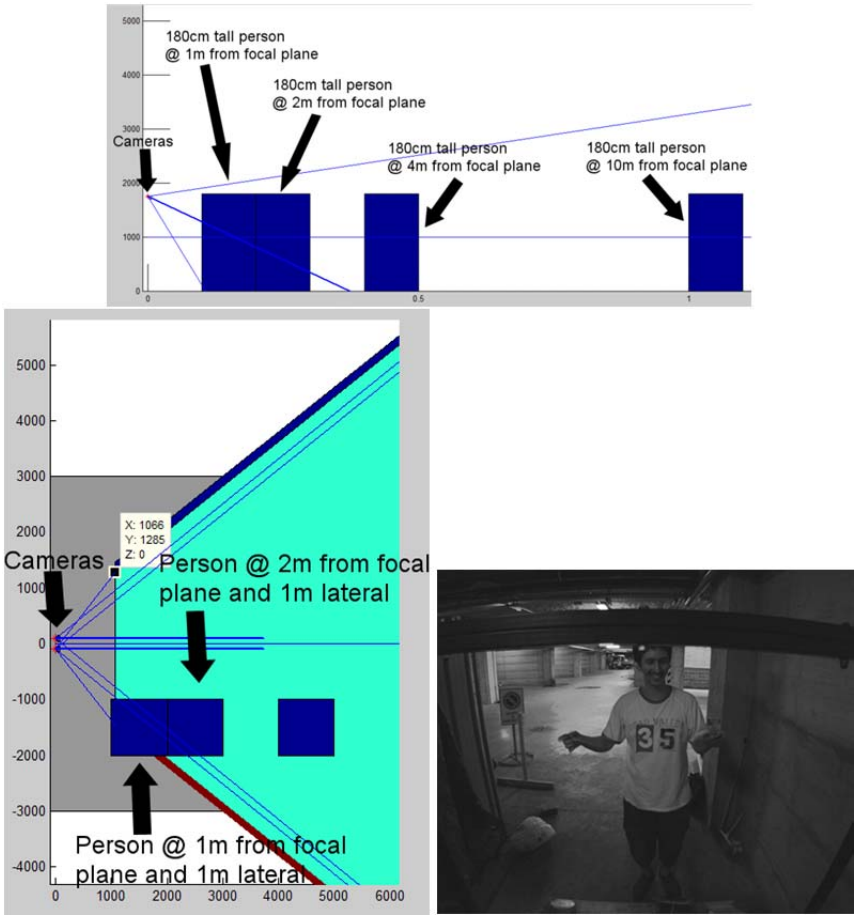


Figura 30 – Le immagini mostrano il risultato della visualizzazione del campo visivo delle camere. In alto la visione laterale in cui è possibile notare che le persone ad 1 metro di distanza vengono viste nella loro interezza, mentre sotto a sinistra, la vista dall'alto mostra quanto lateralmente sono visibili gli oggetti. Le persone sono state modellizzate con dei parallelepipedi di 1 m^2 di base e 1,8 metri di altezza. Nell'immagine in basso a destra è possibile vedere un esempio di campo visivo reale delle camere (Grazie al Dott. F. Sacchi per la comparsa!)

Realizzazione e montaggio

Come abbiamo visto nei capitoli precedenti, la calibrazione è un processo fondamentale ed essenziale al fine della buona riuscita della ricostruzione della scena 3D a partire dai punti immagine 2D. E' altresì importante che, una volta completata la calibrazione, i parametri estrinseci delle camere (che sono i più soggetti a cambiamenti) rimangano gli stessi.

La testa stereo deve, quindi, essere realizzata tenendo conto di questa esigenza: una volta assemblate, le camere dovranno essere solidamente ancorate alla struttura che forma la testa stereo che, a sua volta, dovrà essere ben fissata alla struttura del veicolo.

Le camere hanno un peso di 0,5Kg ciascuna, esclusi i supporti ed i cablaggi ed è perciò importante che la struttura di sostegno sia adatta a sostenere tale peso e non si torca quando sottoposta alle sollecitazioni della guida.

In considerazione di tale necessità si è scelto un supporto principale formato da un parallelepipedo di alluminio cavo all'interno, con lo spessore di 4 mm. Dopo aver praticato dei fori in corrispondenza di quelli di fissaggio delle camere, esse sono state avvitate sul supporto con delle viti M4. Dato il discreto peso delle camere e le possibili sollecitazioni a cui potrebbero essere sottoposte durante la guida (buche, sobbalzi del piano stradale, brusche frenate e/o accelerate) si è ritenuto poco sicuro ancorare le camere usando solamente 2 viti.

E' stato realizzato quindi un supporto ad "U" che calzasse perfettamente sopra ciascuna camera, praticando dei fori in corrispondenza della predisposizione dei fissaggi delle camere stesse. I supporti sono stati poi fissati mediante 6 viti M4 per ciascuna camera e, a loro volta, sono stati ancorati al parallelepipedo di alluminio cavo con 6 viti ciascuno. In questo modo, il posizionamento delle camere si può dire stabile e permanente.

La Figura 31 mostra la proiezione ortogonale del supporto della testa stereo in cui è possibile vedere le soluzioni di fissaggio e montaggio descritte in precedenza.

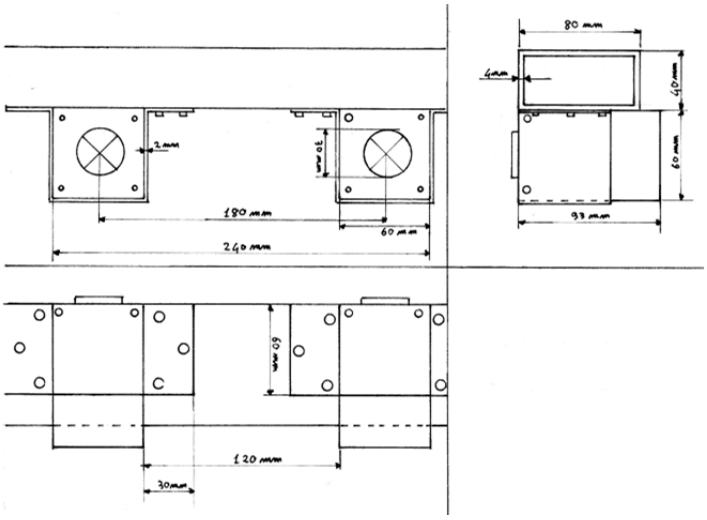


Figura 31 – L'immagine mostra la proiezione ortogonale del supporto della testa stereo e delle camere montate su di essa. Il parallelepipedo di alluminio ha uno spessore di 4mm che lo rende particolarmente rigido alle torsioni ed i supporti ad "U" che fissano le camere al supporto hanno uno spessore di 2mm. Le camere sono fissate con 6 viti alla "U" di alluminio e con 2 viti al supporto stesso. La "U" che avvolge le camere è ancorata al supporto in alluminio mediante 6 viti ciascuno. In questo modo, le camere risultano particolarmente stabili ai movimenti e alla torsione.

Per permettere la variazione dell'inclinazione delle camere rispetto al piano orizzontale, sono stati usati tre giunti snodati della ITEM ancorati al telaio del veicolo nella parte anteriore in alto. Ogni giunto è in grado di sorreggere un peso di 35 Kg e quindi, i 3 giunti possono sopportare una forza di 1050 Nm (105 Kg), più che sufficiente a mantenere la testa stereo stabile rispetto alle sollecitazioni della guida.

Figura 32 – Le fotografie mostrano due particolari del montaggio della testa stereo: a sinistra i giunti snodati utilizzati per fissarla, mentre a destra la posizione di montaggio su USAD.

Durante la realizzazione della testa stereo ed il suo utilizzo iniziale abbiamo incontrato un insolito problema relativo al protocollo di comunicazione Gigabit Ethernet utilizzato dalle camere. A seguito dell'interfacciamento delle stesse, infatti, riscontravamo un elevato numero di errori di comunicazione. Dopo numerose ricerche sul codice e sul protocollo utilizzato, il problema è stato rilevato come "*Missing frames*".

Le camere utilizzano un protocollo di comunicazione UDP (User Data Protocol) che non presenta alcuna funzionalità di controllo di flusso e non supporta il rinvio di pacchetti "persi" o danneggiati, ma la scelta di tale protocollo è stata necessaria per rispondere alle esigenze di ottenere un'alta velocità di comunicazione. Se durante una comunicazione si verifica un errore di "*Missing frames*", significa che uno o più pacchetti della sequenza che compone l'immagine non sono arrivati a destinazione e non è stato quindi possibile ricostruire in modo completo l'immagine stessa.

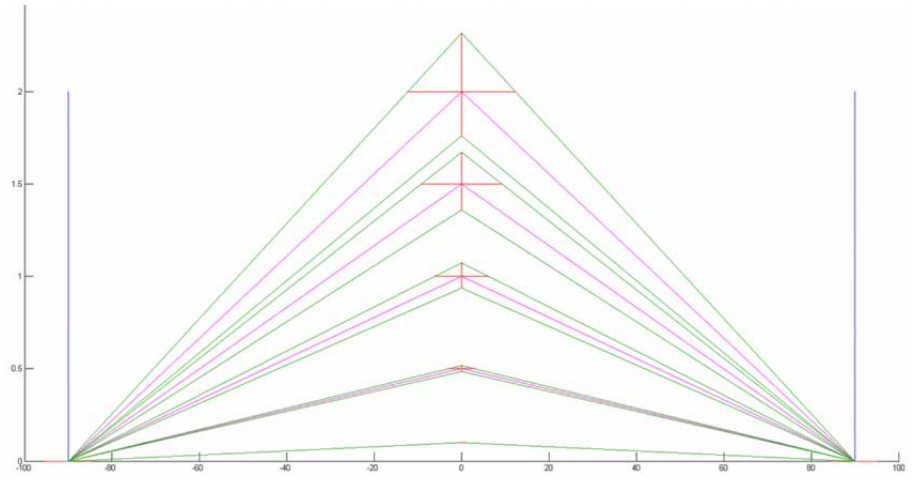
Analizzando le possibili cause del suddetto errore si è arrivati alla conclusione che il problema potesse essere ricercato nelle interferenze elettromagnetiche a cui i cavi ethernet erano sottoposti quando il motore elettrico di USAD era in funzione o, più in generale, quando correnti elevate fluivano nei conduttori prossimi ai cavi di rete stessi. Perciò sono stati utilizzati dei cavi Cat.6 STP al fine di garantire allo stesso tempo la velocità di connessione Gigabit, supportata dalle camere, e l'immunità alle interferenze elettromagnetiche presenti nei pressi di USAD stesso. Tuttavia tale sostituzione non ha portato alla soluzione del problema.

Dopo avere svolto un'approfondita ricerca, abbiamo optato per l'utilizzo di cavi Cat.7 che presentano una doppia schermatura (una sulle singole coppie di cavi ed una sull'intero cavo), dei connettori schermati con impedenza calibrata rispetto a quella del cavo stesso e una frequenza di trasmissione massima di 600 MHz, a differenza dei tradizionali 250 MHz per la categoria Cat.6 ed i 100 MHz di Cat.5

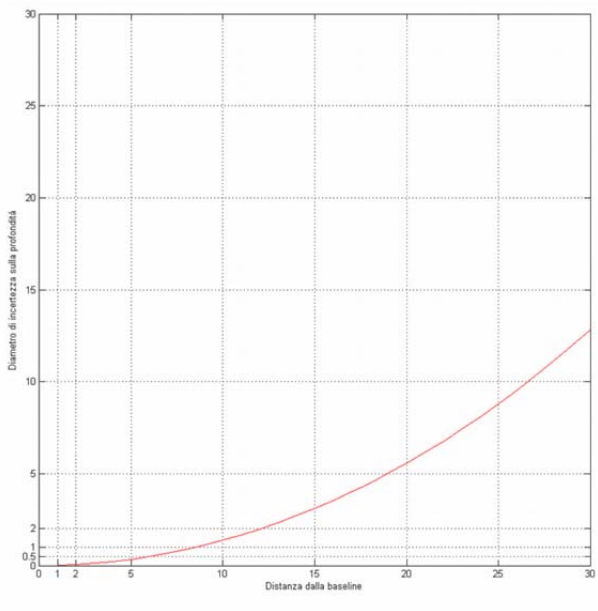
La scelta di realizzare il cablaggio mediante cavi di categoria Cat.7 ci ha permesso di risolvere efficacemente il problema degli errori di comunicazione con "*Missing frames*".

Calibrazione e baseline

Come illustrato in precedenza, la baseline ricopre un ruolo di fondamentale importanza per la buona riuscita della ricostruzione dei punti immagine 2D nella scena. Un valore troppo elevato di questo parametro rischierebbe di portare alla falsificazione dell'ipotesi fondamentale della stereo visione e, quindi, non sarebbe più possibile identificare con successo i punti dell'immagine primaria in quella secondaria. Partendo da questa considerazione, osservando quanto disponibile in letteratura sul problema e basandosi sull'esperienza del D.R. Marzorati, è stato scelto di fissare le camere con una baseline di 18 cm.



L'immagine mostra gli assi ottici delle due camere (rappresentati in blu) ed i relativi sensori (perpendicolari agli assi ottici e coincidenti con l'asse delle x). Le rette viola rappresentano la proiezione di punti a 1, 5, 10, 15 e 20 metri di distanza dalla baseline sul piano immagine. Le rette verdi rappresentano la retroproiezione dell'area di incertezza provocata dalla dimensione degli elementi fotosensibili del sensore. Le croci rosse centrate nel punto proiettato rappresentano i diametri di incertezza sulla larghezza (segmento orizzontale) e sulla profondità (segmento verticale). In particolare, il diametro di incertezza sulla larghezza riferito alle differenti distanze del punto è: 1.22, 6.15, 12.30, 18.45 e 24.61 millimetri, mentre il diametro di incertezza relativo alla profondità è: 13.50, 341.39, 1372, 3107, 5571 millimetri. Com'è possibile notare, il diametro di incertezza sulla profondità aumenta molto in funzione della distanza del punto stesso, mentre quello relativo alla larghezza rimane contenuto. N.B. Gli assi del grafico hanno scale differenti!



L'immagine mostra il variare della dimensione del diametro di incertezza sulla profondità in funzione della distanza del punto dalla baseline delle camere

Completato il montaggio della testa stereo, il passo successivo è stato quello della calibrazione. Esistono in letteratura diverse tecniche per procedere alla calibrazione delle camere e, successivamente, della testa stereo. Quella adottata in questo caso è Camera Calibration Toolbox for Matlab ***[29]*** che consente l'acquisizione in un pattern predefinito (scacchiera) in diversi piani dello spazio. In seguito è necessario individuare sul pattern stesso (in Matlab nel nostro caso) gli estremi della scacchiera ed indicare quanti quadrati sono contenuti in ogni asse (orizzontalmente e verticalmente). Tale procedura va ripetuta sulle diverse immagini di calibrazione catturate (generalmente se ne usano più di 20 per ogni camera).

Come risultato, si ottengono le stime dei parametri intrinseci ed estrinseci della camera, oltre ad una stima della distorsione che l'ottica introduce. Si ripete la medesima procedura anche per l'altra camera in modo da acquisire i parametri di calibrazione intrinseci ed estrinseci di entrambe. I valori così ottenuti, vengono utilizzati per determinare la calibrazione della testa stereo complessiva.

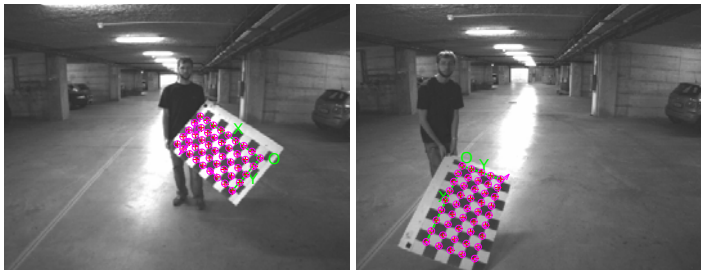


Figura 33 – Le immagini mostrano la procedura di calibrazione con il pattern di riconoscimento posto in due posizioni diverse e con gli angoli dei quadrati marcati correttamente (Si ringrazia il Dott. A. Galbiati per l'aiuto).

I parametri intrinseci ed estrinseci delle camere ottenuti al termine del processo di calibrazione sono i seguenti:

Camera sinistra (principale)

Focal Length: fc_left = [830.41826 830.63978] ± [1.42736 1.43712]
Principal point: cc_left = [669.64934 511.90074] ± [2.71602 2.86062]
Skew: alpha_c_left = [0.00000] ± [0.00000]
Distortion: kc_left = [-0.21831 0.06742 -0.00143 -0.00025 0.00000]

Camera destra (secondaria)

Focal Length: fc_left = [834.86578 834.46106] ± [3.17416 3.20841]
Principal point: cc_left = [660.66259 518.15022] ± [5.97033 6.41949]
Skew: alpha_c_left = [0.00000] ± [0.00000]
Distortion: kc_left = [-0.22028 0.06839 -0.00189 0.00146 0.00000]

I parametri di calibrazione della testa stereo, cioè la traslazione e la rotazione che portano dall'immagine primaria a quella secondaria, sono i seguenti:

Rotation vector: om = [0.00843 -0.00040 -0.00499]
Translation vector; T = [-183.19299 -10.26640 2.70762]

Elaborazione delle immagini a basso livello

Analizziamo ora la parte di progetto che si è occupata di rendere le immagini catturate dalla testa stereo utilizzabili ai fini delle elaborazioni di alto livello. Come visto in precedenza, infatti, molto spesso le immagini restituite dal sensore sono imperfette, hanno un contrasto non ottimale, sono troppo scure oppure troppo chiare.

Prima di procedere alla fase di ricostruzione è necessario, quindi, operare a basso livello sulle immagini.

Pixel binning: quando la quantità di luce è scarsa

Una delle situazioni più comuni quando si opera con parametri fissi della camera è quella di ottenere delle immagini troppo scure, soprattutto quando si inquadrano spazi chiusi, si viaggia di notte oppure si utilizzano tempi di esposizione bassi per evitare l'effetto di mosso delle immagini.

La tecnica del pixel binning serve proprio a risolvere queste problematiche a scapito, però, della risoluzione finale dell'immagine stessa.

Con il termine ***pixel binning*** ci si riferisce ad una tecnica volta a creare dei "*super pixel*" dall'unione di più pixel dell'immagine originale. Ad esempio, se diciamo che stiamo utilizzando un binning 1x1, intendiamo che ogni pixel dell'immagine è elaborato singolarmente. Se utilizziamo un binning 2x2, significa che stiamo aggregando le informazioni di due pixel sulle righe e due pixel sulle colonne per ottenere un "super pixel" con una quantità di informazioni pari alla somma dei 4 pixel reali. In questo modo, tuttavia, l'immagine risultante avrà una risoluzione di ¼ rispetto a quella originale.

Questa tecnologia può trovare applicazione direttamente nell'hardware delle camere, oppure può essere eseguita come tecnica di post elaborazione una volta acquisita l'immagine. Nel

primo caso, otterremo già direttamente dalla camera un'immagine di risoluzione minore, ma con una maggiore quantità di informazioni per ogni pixel. Tendenzialmente, questo tipo di operazioni vengono svolte con circuiteria dedicata e sono, quindi, molto veloci.

Nel secondo caso, sarà necessario un apposito algoritmo che tratti l'immagine ricevuta dalla camera ed effettui la riduzione di risoluzione, unendo le informazioni dei pixel necessari.

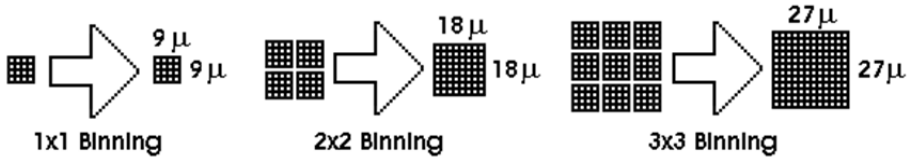


Figura 34 – L'immagine mostra l'idea della tecnica di pixel-binning. A partire da un certo numero di pixel dell'immagine originale, viene creato un "super pixel" che contiene le informazioni di tutti i pixel.

Questa tecnica è particolarmente utile quando le riprese sono molto scure e l'uso delle normali tecniche di elaborazione delle immagini porterebbe ad un eccessivo aumento del rumore. Se si è disposti ad avere una certa perdita di risoluzione, il pixel-binning permette di sommare la poca quantità di luce catturata da ogni elemento fotosensibile in gruppi di super-pixel in modo da ottenere un'immagine nettamente più utilizzabile.

A titolo di esempio della potenzialità di questa tecnica, nella Figura 35 è possibile osservarne un esempio di applicazione. Si tratta di un caso "estremo", nel quale è possibile apprezzare un certo numero di dettagli prima del tutto invisibili. L'immagine è stata catturata a notte fonda (senza luna) con una delle camere usate in questo progetto a un tempo di esposizione di 10 ms (1/100 di secondo), apertura F/1.8 e nessun guadagno digitale. Nell'immagine risultante è presente una grande quantità di rumore (causato dalla scarsa illuminazione della scena e dall'amplificazione introdotta dal pixel-binning), ma sono comunque distinguibili dettagli importanti come le sagome delle vetture, la strada e gli alberi.

Figura 35 – Le fotografie mostrano, a sinistra, l'immagine catturata dal sensore della camera e, a destra, la stessa immagine a cui è stato applicato l'algoritmo di binning 4x4, software sviluppato per questa tesi. Come appare evidente, l'immagine è "estrema" come condizioni di acquisizione, ma nell'immagine risultante è possibile distinguere un elevato numero di dettagli.

Le camere PhotonFocus scelte per questo progetto non integrano nella propria circuiteria questa tecnica. E' stato quindi necessario realizzare un algoritmo che svolgesse questo compito. Di seguito vediamo qual è il principio di funzionamento:

1. Input ($\mathcal{I}$,$\boldsymbol{bx}$,$\boldsymbol{by}$)
2. $\mathcal{O} = \emptyset$
3. **foreach** pixel *in* row($\mathcal{I}$) *step* $\boldsymbol{bx}$ **do**
4. **for** i:=1 *to* $\boldsymbol{bx}$ **do**
5. $\mathcal{O}$(pixel) += $\mathcal{I}$(pixel+i)
6.
7. **foreach** pixel *in* col($\mathcal{I}$) *step* $\boldsymbol{by}$ **do**
8. **for** j:=1 *to* $\boldsymbol{by}$ **do**
9. $\mathcal{O}$(pixel) += $\mathcal{I}$(pixel+j)
10.
11. ritorna $\mathcal{O}$

Algoritmo 1 – Lo pseudo codice sopra riportato rappresenta l'esecuzione del pixel-binning

Il risultato del pixel binning dovrà essere riscalato nell'intervallo desiderato: sommando i valori di luminanza di ogni pixel, infatti, si ottengono dei numeri sicuramente più elevati del limite massimo riferito al pixel-deph dell'immagine. Per raggiungere dei valori utilizzabili e visualizzabili è possibile eseguire il seguente rescaling di valori:

$$y = x * \frac{\max \text{dest}}{\max(\text{x}) - \min(\text{x})} - \min(x)$$

Dove *y* è il valore risultante, *x* è il valore originale, *maxdest* è il valore massimo del nuovo intervallo, *max(x)* e *min(x)* sono il massimo ed il minimo valore della funzione *x*. *y* sarà riscalato tra 0 e *maxdest*.

Equalizzazione dell'istogramma con stretchlim: saturare volontariamente una parte dei dati

Una tecnica molto utilizzata nell'elaborazione delle immagini è l'equalizzazione dell'istogramma che permette di ridistribuire i valori di luminanza dei pixel dell'immagine lungo tutto il range possibile aumentando il contrasto dell'immagine stessa.

Questa tecnica è molto efficace quando l'istogramma dell'immagine che si sta considerando è compresso e l'immagine risulta scura, chiara o poco contrastata. Nel caso in cui siano presenti degli spot luminosi oppure delle aree molto scure, questa tecnica restituisce un risultato mediocre e, in alcuni casi più estremi, non produce alcuna variazione nell'immagine finale.

Se osserviamo l'istogramma a sinistra nella Figura 36, possiamo notare che le intensità dei pixel sono concentrate completamente nei bassi livelli. Nell'istogramma di destra, invece, si osserva che i pixel dell'immagine sono distribuiti in un range più ampio e che è presente una componente di alte intensità luminose. Applicando una tradizionale tecnica di equalizzazione dell'istogramma, da quello di destra si otterrà un istogramma identico, come mostrato nella Figura 37 a sinistra. Con la tecnica stretchlim, invece, l'istogramma è ben contrastato e i valori sono distribuiti uniformemente lungo tutto il range, a scapito, ovviamente, della saturazione di alcuni di essi.

In visione artificiale, a differenza di altre discipline dove è importante non avere parti sovraesposte o sottoesposte dell'immagine, è accettabile saturare una parte di dati. In questo modo, vengono recuperati dei dettagli che gli algoritmi di alto livello possono utilizzare per calcolare in maniera più precisa, ad esempio, la

disparità immagine. Se vi è uno spot luminoso o una piccola area molto scura, è lecito pensare che in tali zone non siano contenuti dettagli importanti, molto probabilmente collocati nella parte restante dell'immagine. Tentare di leggere correttamente queste aree rischierebbe di compromettere ulteriormente la parte di scena che ci interessa ed è quindi preferibile optare per una saturazione di queste piccole zone.

La tecnica di stretchlim si basa proprio su questo concetto, individuando nell'immagine, quali valori in percentuale, corrispondono al livello di saturazione ammesso. A questo punto, si procede all'equalizzazione dell'istogramma non più rispetto ai limiti del range identificato dal valore di *pixel-deph*, bensì utilizzando quelli calcolati dalla funzione stretchlim.

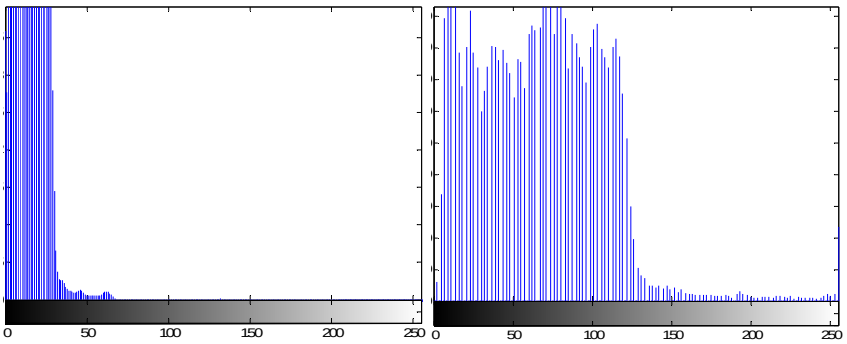


Figura 36 – L'immagine mostra, a sinistra, l'istogramma di un'immagine concentrato nella parte scura senza, tuttavia, avere un contributo nella parte chiara e, a destra, l'istogramma di un'immagine in cui i livelli di luminanza sono distribuiti in un range più ampio e, soprattutto, con un contributo anche nelle alte intensità.

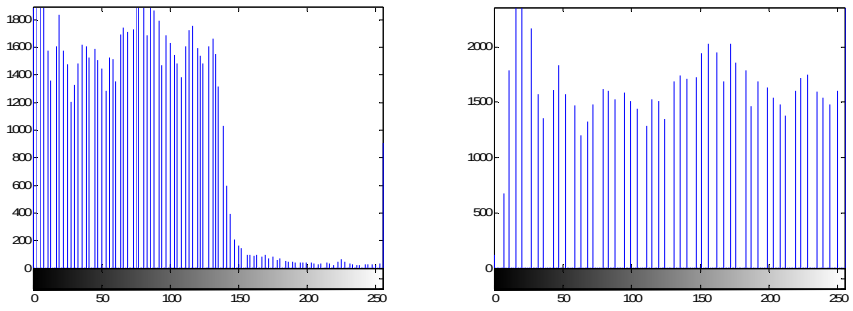


Figura 37 – La figura mostra l'applicazione della procedura tradizionale di equalizzazione dell'istogramma, a sinistra, e quella con stretchlim, a destra. L'istogramma di riferimento, è quello presentato nella Figura 36 a destra.

L'idea alla base di questo algoritmo è quella di calcolare la probabilità cumulativa per i livelli di intensità luminosa che corrisponde all'istogramma cumulativo. Viene richiesto come input il valore di saturazione ammissibile per le basse luminanze e per gli alti valori di intensità luminosa. Tali valori devono essere stabiliti sperimentalmente in relazione all'applicazione e alla risoluzione dell'immagine, arrivando a dei compromessi sulla dimensione della parte "sacrificabile".

```
12. Input (ℑ, lim, lim_min, lim_max)
13. 𝒪 = ∅
14. cumulative_hist = ∅
15.
16. foreach pixel in ℑ do
17.       cumulative_hist (ℑ (pixel)) += 1
18.
19.       if cumulative_hist (ℑ (pixel))< lim_min then
20.             new_low_limit = cumulative_hist (ℑ (pixel))
21.       if cumulative_hist (ℑ (pixel))< lim_ max then
22.             new_high_limit = cumulative_hist (ℑ (pixel))
23.
24. foreach pixel in ℑ do
25.       𝒪(pixel) = (ℑ (pixel)- new_low_limit)*limit/
26.                   /(new_high_limit- new_ low _limit)
27.
28.       if 𝒪(pixel)> limit then
29.             𝒪(pixel) = limit
30.       else if 𝒪(pixel)<0 then
31.             𝒪(pixel) = 0
32.
33. ritorna 𝒪
```

Algoritmo 2 – L'algoritmo sopra riportato rappresenta il codice di stretchlim presentato in questo capitolo.

Impostazioni della camera costanti per ogni situazione

Un ambizioso obiettivo della visione artificiale è quello di utilizzare dei parametri fissi delle camere per ogni condizione di esposizione e

per ogni tipo di scena. Questo eviterebbe di dovere compiere delle elaborazioni in backgound sulle immagini catturate dalle camere e di dovere comandare le stesse per adattarle alle nuove condizioni di acquisizione.

Ogni elaborazione su delle immagini comporta un certo grado di calcolo computazionale. Tutte le analisi di alto livello svolte nell'ambito della visione artificiale, come ad esempio il calcolo della disparità, della u-v disparity e della mappa dello spazio libero, hanno degli impatti non indifferenti sulle risorse del/i calcolatore/i utilizzati. Aggiungere altro carico computazionale a queste macchine significherebbe rallentare i calcoli di alto livello ottenendo una frequenza minore della stima dello spazio libero e un maggiore consumo di energia causato dal crescente sfruttamento delle risorse elaborative.

In questa tesi sono state utilizzate tutte le tecniche presentate in precedenza per raggiungere l'obiettivo esplicitato all'inizio di questa sezione.

Dopo avere valutato la velocità massima a cui USAD può muoversi e dopo avere svolto numerosi test, è stato determinato empiricamente un tempo di esposizione che permette di non avere immagini mosse, anche con movimento alla massima velocità, e che, allo stesso tempo, consente di ottenere con la post elaborazione dei risultati utilizzabili per le analisi di alto livello.

Nel capitolo dei risultati sperimentali, saranno esposti in dettaglio i risultati di questa tecnica.

Sistema di retroazione sui parametri delle camere

Nella sezione precedente abbiamo analizzato un approccio che mira ad avere i parametri delle camere fissi in ogni condizione di ripresa. Questo approccio, se pur ambizioso ed adatto per USAD, trova delle

limitazioni nell'applicazione in veicoli a guida autonoma più generici. Se, infatti, USAD ha una velocità massima piuttosto limitata, un veicolo generico potrebbe avere una velocità maggiore ed il tempo di esposizione scelto genererebbe delle immagini potenzialmente "mosse".

Per ovviare a questo problema è possibile agire su un singolo parametro della camera, ovvero il tempo di esposizione, lasciando invariati gli altri.

Analisi dell'immagine: luminosità e contrasto

Quando il tempo di esposizione dell'immagine è troppo basso rispetto alla luminosità della scena, si avrà come risultato un'immagine troppo scura. Viceversa, se il tempo di esposizione è troppo alto, l'immagine sarà mediamente sovraesposta.

Per effettuare un calcolo esatto del tempo di esposizione ideale per un'immagine, è necessario misurare, con un esposimetro, la luminanza media della scena. In ottica e in fotografia, il valore di esposizione (o valore esposimetrico) EV è definito come l'intensità luminosa a cui corrispondono coppie di diaframma (apertura della lente) e tempi di esposizione. Ad un valore di EV pari a 0, corrisponde un tempo di esposizione di 1 secondo ad un'apertura di F/1.0. La formula che lega queste grandezze è:

$$EV = \log_2 \frac{A^2}{T}$$

dove A è l'apertura del diaframma e T indica il tempo di esposizione.

Nelle tradizionali macchine fotografiche, l'esposimetro è integrato nella circuiteria interna e, prima di scattare la fotografia, viene valutato il valore di EV della scena e vengono quindi scelti i valori di tempo di esposizione ed apertura del diaframma più opportuni. Nel nostro caso, tuttavia, la camera PhotonFocus non ha un esposimetro integrato e, quindi questo tipo di analisi è impossibile.

Il tempo di esposizione, tuttavia, ha un effetto diretto sulla luminosità media dell'immagine. E' quindi possibile eseguire delle analisi su tale parametro per valutare se un certo valore è appropriato o meno alla scena.

Tuttavia l'obiettivo delle elaborazioni di basso livello esaminate in precedenza è quello di riuscire a portare un'immagine con un istogramma fortemente compresso ad una situazione in cui i livelli di luminanza siano distribuiti in modo omogeneo nell'intervallo. Queste tecniche hanno, però, dei limiti e, in situazioni estreme non riescono ad espandere completamente l'istogramma.

Nell'analisi della luminosità media dell'immagine è importante quindi tenere in considerazione sia l'immagine originale catturata dal sensore sia l'immagine elaborata con le tecniche già viste. Nel primo caso si avrà un'indicazione di quanta luce stia effettivamente arrivando al sensore, mentre nel secondo caso si potrà misurare la "difficoltà" che le tecniche di post elaborazione stanno incontrando nel riportare l'istogramma ad una distribuzione uniforme. Idealmente, la situazione ottimale sarebbe quella di ottenere un'immagine perfettamente contrastata già a livello del sensore in modo da evitare le tecniche di post-processing che sono generalmente onerose per quanto riguarda il tempo di computazione ed introducono, quindi, dei ritardi, ma questo è in contrasto con la necessità di utilizzare dei tempi di esposizione ridotti per diminuire la quantità di "mosso" presente nell'immagine.

Per controllare correttamente il tempo di esposizione è quindi opportuno valutare entrambe le informazioni di luminosità media dell'immagine catturata dal sensore e luminosità media dell'immagine elaborata. Analizzando la prima si tenderà ad aumentare il tempo di esposizione dell'immagine, mentre considerando la seconda, si può limitare questo fenomeno nel momento in cui l'immagine elaborata possiede una buona media ed un istogramma correttamente equalizzato.

Queste due valutazioni, tuttavia, sono limitative in quanto non tengono conto delle reali necessità del tempo di esposizione della

scena. Per scene molto movimentate oppure per movimenti del veicolo ad alte velocità potrebbe, infatti, essere necessario utilizzare un tempo di esposizione che produca un'immagine leggermente più scura della media a vantaggio della nitidezza della stessa. Per questo motivo, è opportuno introdurre una metrica per valutare e definire la quantità di "mosso" presente nell'immagine ed intervenire in modo opportuno variando il tempo di esposizione della camera.

Analisi dell'immagine: sfocatura e "mosso"

L'analisi di quanto l'immagine catturata dalla camera sia "mossa" è in generale essenziale in visione artificiale, ma ancor di più per un veicolo a guida autonoma. Con l'aumentare della velocità, infatti, dovrà diminuire il tempo di esposizione usato, a causa del moto relativo crescente del mondo.

Esistono differenti tipologie di approccio all'analisi del motion blur delle immagini e si possono suddividere in tre principali categorie: full-reference, reduced-reference e no-reference. Nel primo caso, l'immagine analizzata viene comparata con una di riferimento, ad esempio l'originale. Questo tipo di approccio richiede che sia disponibile l'immagine priva di effetto mosso e, quindi, è generalmente poco utilizzabile in ambiti di visione artificiale in quanto non è possibile disporre di un immagine di riferimento priva di sfocatura.

Nel secondo caso vengono utilizzate delle features dell'immagine da analizzare che vengono comparate con alcune features dell'immagine originale. Anche in questo caso, nonostante siano necessarie porzioni minori dell'immagine originale, bisogna averne a disposizione alcune parti e, quindi, queste tecniche sono ancora una volta poco impiegabili nell'ambito della visione artificiale.

Nelle tecniche di motion blur detection "no-reference", invece, la metrica non è relativa ad un'immagine di riferimento, ma viene calcolata in relazione a specifiche caratteristiche dell'immagine a disposizione. L'approccio senza riferimento è piuttosto impegnativo

in quanto la distinzione tra features dell'immagine ed effetti di mosso è molte volte ambigua e difficilmente individuabile.

Esistono numerose tecniche "no-reference" per la rilevazione e quantificazione della quantità di blur presente nell'immagine. Le principali e più diffuse sono ***[39] [40]***:

- Metrica della varianza. Si basa sul calcolo della varianza dell'intera immagine. Dato che un aumento della sfocatura dell'immagine corrisponde ad un aumento dello smussamento degli edge, si avranno delle transizioni sempre più graduali dei livelli di grigio. Questo si traduce in una diminuzione della varianza dell'immagine stessa.
- Metrica basata sull'autocorrelazione. Viene calcolata una funzione di autocorrelazione su due finestre di analisi dell'immagine. Nel caso in cui l'immagine sia sfocata, i dettagli presenti in due finestre di analisi vicine saranno simili e, quindi, il valore di autocorrelazione sarà elevato. Maggiore sarà tale valore, tanto più l'immagine stessa presenterà l'effetto mosso.
- Metrica basata sulla derivata. Questo approccio utilizza le metriche calcolate utilizzando il gradiente (derivata del prim'ordine) ed il laplaciano (derivata del second'ordine). Questa metrica può essere vista come l'applicazione di un filtro passa-alto nel dominio delle frequenze dell'immagine. Minori saranno i dettagli presenti nelle alte frequenze (ovvero tanto più basso sarà il valore della metrica), tanto più l'immagine sarà sfocata. Tuttavia, le derivate, soprattutto il laplaciano, sono molto sensibili al rumore e sono quindi molto suscettibili alle condizioni di illuminazione della scena e alla qualità del sensore che può introdurre del rumore
- Metrica percettiva. Per prima cosa vengono rilevati gli edge presenti nell'immagine quindi, ciascuna riga viene scansionata e viene calcolata la larghezza di ogni edge presente sottraendo la posizione di fine a quella di inizio dell'edge stesso. Tanto più l'immagine sarà sfocata, tanto più

aumenterà la larghezza media degli edge presenti nell'immagine.

- Metrica sulla soglia di frequenze. Si basa sulla somma dell'ampiezza delle frequenze dell'immagine sopra ad una certa soglia T che viene scelta sperimentalmente: maggiore sarà il valore di tale soglia, tanti più edge riusciremo a rilevare, ma nello stesso tempo aumenterà anche la suscettibilità al rumore.

Le tecniche di blur-detection appena esaminate soffrono di due principali problematiche: sono fortemente dipendenti dal contenuto della scena che abbiamo catturato nell'immagine e si basano su soglie difficilmente impostabili a priori con dei valori utilizzabili in ogni circostanza. Va ricordato, infatti, che in applicazioni di guida autonoma possiamo incorrere in situazioni estreme di illuminazione e di oggetti rappresentati ed è possibile, quindi, che il contenuto in termini di edge dell'immagine o di frequenze sia molto variabile. In questo caso, le metriche descritte in precedenza restituiscono dei valori fuorvianti.

Per ottenere una metrica il più possibile indipendente dalle caratteristiche della scena e da particolari soglie da impostare, è stato scelto di utilizzare un differente approccio nell'identificazione del mosso di un'immagine.

La metrica CPBD (Cumulative Probability of Blur Detection) si basa sul concetto base di "*Just Noticeable Blur*" (JNB) che può essere definito come la minima quantità di sfocatura percepita dato un livello di contrasto superiore al valore JND ***[30]***.

In psicofisica la "*Just Noticeable Difference*" (JND) viene anche chiamata soglia differenziale e corrisponde al minimo valore di energia con cui uno stimolo fisico deve essere variato perché la variazione sia avvertita dall'uomo. In particolare in visione artificiale si è interessati allo stimolo sensoriale della vista e, quindi, possiamo dire che il JND è la minima differenza rilevabile tra due livelli di luminanza immagine.

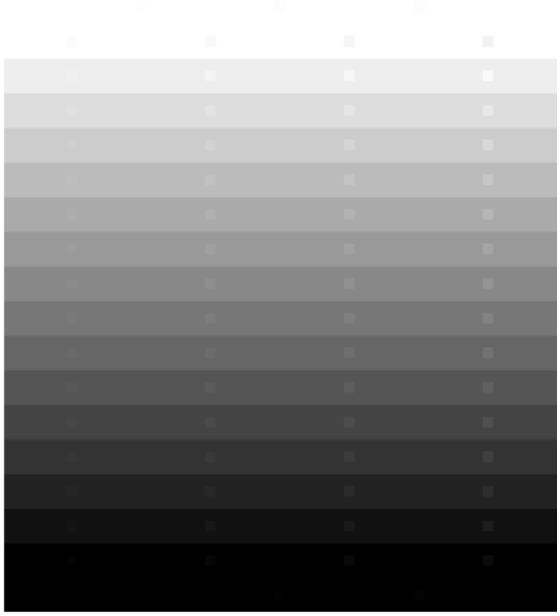


Figura 38 – L'immagine mostra delle bande di contrasto. La differenza di intensità luminosa tra le diverse bande ed i quadrati in esse contenuti rappresenta la minima differenza rilevabile dall'uomo tra due livelli di luminanza.

Dalla definizione di JND si può affermare, quindi, che il JNB corrisponde alla quantità di sfocatura minima percepita intorno ad un edge, dato un contrasto superiore a JND.

La probabilità di una rilevazione di una sfocatura su di un dato edge con contrasto C può essere modellizzata da una funzione psicometrica data da:

$$P_{BLUR} = 1 - e^{-\left|\frac{w(e_i)}{w_{JNB}(e_i)}\right|}$$

dove $w_{JNB}(e_i)$ è la larghezza della sfocatura JNB riferita all'edge e_i, e $w(e_i)$ è la larghezza misurata dell'edge e_i.

L'immagine di partenza viene suddivisa in blocchi di 64x64 pixel ed ognuno di essi è classificato come "blocco di edge" oppure "blocco non di edge" in relazione alla quantità di contorni presenti. I blocchi marcati come "non edge" vengono scartati e non elaborati.

La dimensione di 64x64 pixel scelta per la suddivisione in blocchi dell'immagine corrisponde, all'incirca, alla dimensione della regione foveale dell'occhio umano.

Gli edge che vengono considerati ai fini di questa tecnica sono solamente quelli verticali in quanto da studi condotti ***[31] [30] [20] [21]***, l'effetto di motion blur è presente in maniera meno rilevante sugli edge orizzontali. I contorni sono calcolati utilizzando un opportuno edge detector che permetta di discriminare tra la direzione degli edge stessi, come, ad esempio, il filtro di Sobel verticale.

Per la classificazione dei blocchi di edge, da elaborare, e di non edge, da scartare, viene, tuttavia, utilizzato un edge-detector più diffuso ed efficiente: il filtro di Canny **[41] [42]**. L'edge detector di Canny si basa su un approccio multi stadio al fine di ottenere il migliore algoritmo possibile di rilevazione dei contorni. Nel contesto della selezione, infatti, è importante assicurarsi di rilevare gli edge del blocco stesso nel modo più efficiente e corretto possibile.

I blocchi classificati come "di edge" vengono ulteriormente elaborati, calcolando la larghezza di ogni singolo contorno presente in ciascun blocco. Tale misura viene fatta scorrendo le righe presenti nel blocco analizzato e calcolando la differenza tra la posizione di fine ed il punto di inizio di ciascun edge. La probabilità di rilevazione della sfocatura in ogni edge è calcolata utilizzando la precedente formula, dove il valore w_{JNB} dipende dal contrasto locale C del blocco a cui l'edge analizzato appartiene. Il contrasto di ogni blocco dell'immagine analizzato viene calcolato sottraendo il valore massimo di luminanza a quello minimo presenti nel blocco stesso.

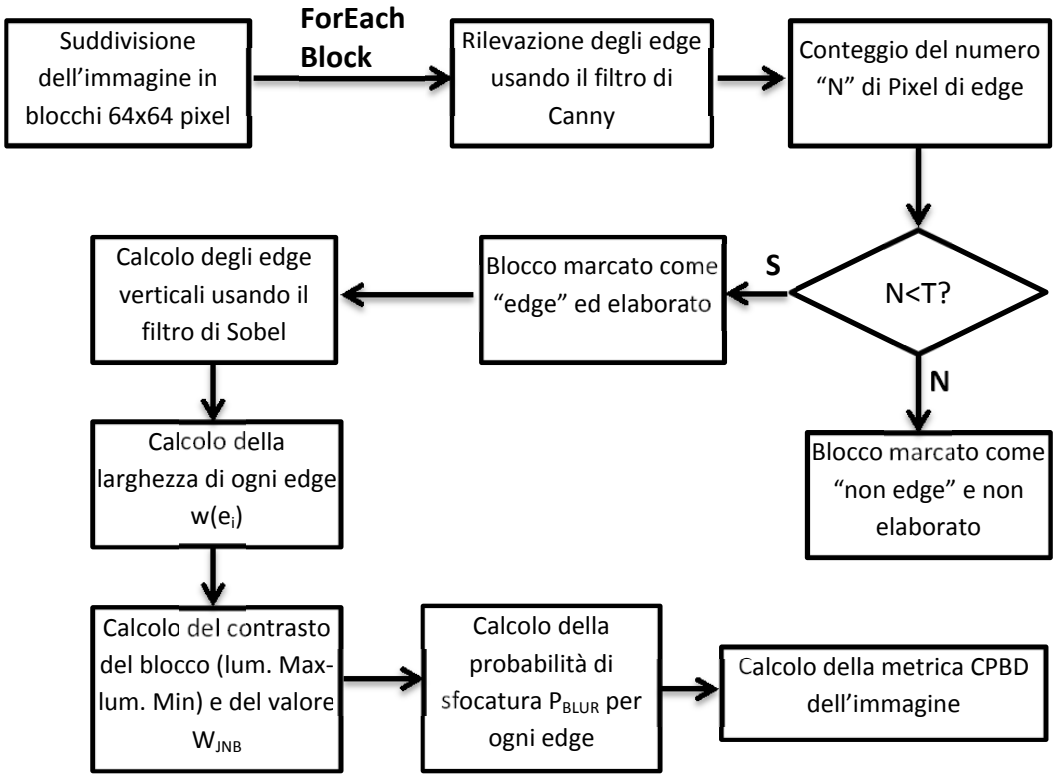
Da test condotti empiricamente ***[31] [30] [20] [21]***, è stato rilevato che $P_{JNB} = 63\% = P_{BLUR}$, nel caso in cui $w(e_i) = w_{JNB}(e_i)$. Ne consegue che la sfocatura non viene rilevata se per un edge $P_{BLUR} \leq P_{JNB}$.

La probabilità cumulativa della rilevazione della sfocatura, CPBD, è calcolata come

$$CPBD = P(P_{BLUR} \leq P_{JNB}) = \sum_{P_{BLUR}=0}^{P_{BLUR}=P_{JNB}} P(P_{BLUR})$$

dove $P(P_{BLUR})$ rappresenta la distribuzione di probabilità dato un particolare valore di P_{BLUR}.

Di seguito è possibile vedere il diagramma di flusso del funzionamento dell'algoritmo analizzato:



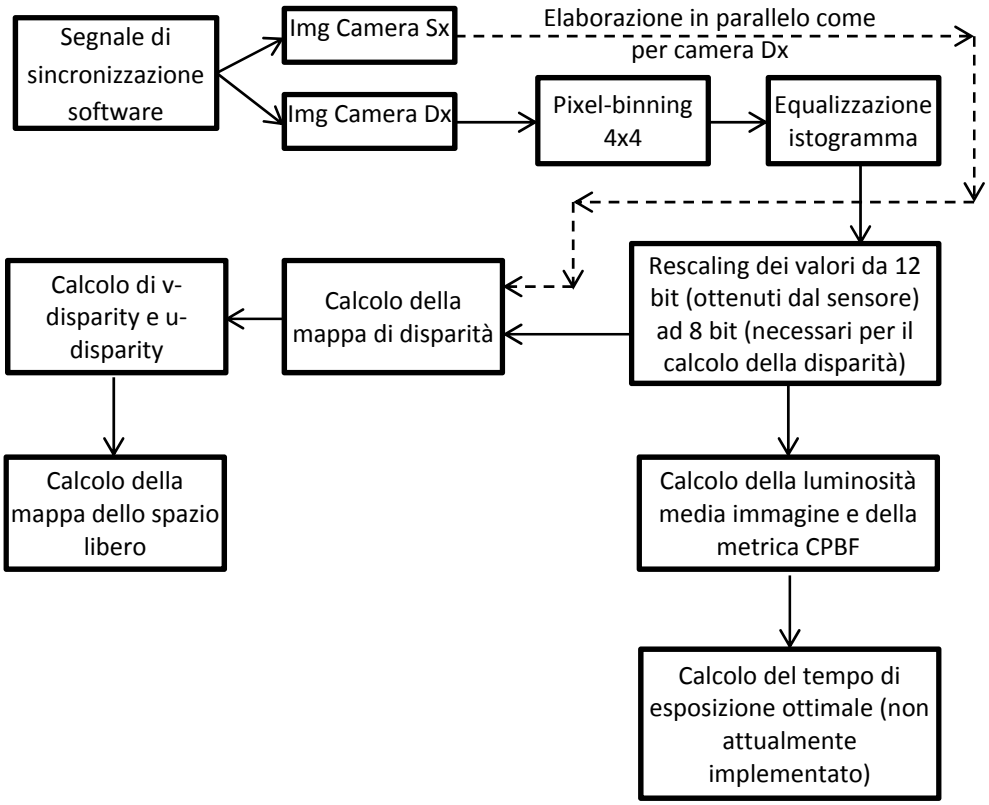
La soglia T presentata nel diagramma di flusso è determinata sperimentalmente, così come il valore w_{JNB} . E' opportuno sottolineare che questa metrica è più rappresentativa del grado di nitidezza dell'immagine, piuttosto che di quello di sfocatura. Infatti, avremo valori prossimi a 1 per immagini con completa assenza di mosso o sfocatura, mentre rileveremo valori prossimi a 0 per immagini molto mosse o sfocate.

Grazie a questa metrica è possibile stabilire quando il tempo di esposizione delle camere è adeguato alla velocità del veicolo ed ai movimenti presenti nelle immagini della scena. Se il valore CPBD delle immagini tende a scendere, allora si dovrà diminuire il tempo di esposizione per ottenere delle immagini più nitide. Se, invece, il valore CPBD delle immagini è soddisfacente, allora è possibile aumentare il tempo di esposizione in funzione della luminosità media dell'immagine.

La variazione di tempo d'esposizione, quindi, è direttamente legata alla luminosità media dell'immagine, ma supervisionata dalla metrica CPBD: se l'immagine è mediamente scura, ma con un valore di blur alto non si potrà aumentare il tempo di esposizione e sarà necessario sopperire a questa mancanza con le tecniche di post-elaborazione.

Schema complessivo processo acquisizione immagini

Vediamo, di seguito, lo schema complessivo del processo di acquisizione ed elaborazione delle immagini



Disparità e visione 3D

Le diverse tecniche per il calcolo della mappa di disparità, che è una rappresentazione della diversa prospettiva della scena esistente tra le due camere della testa stereo, determinano la qualità di rilevazione e la velocità computazionale della stessa.

In questo capitolo, verrà analizzata la tecnica proposta da Mattoccia ***[13]*** per la determinazione della mappa di disparità.

A tale riguardo, segnalo che il mio contributo si è limitato all'integrazione di software sviluppato dal D. R. Marzorati.

Calcolo della disparità secondo Mattoccia

Gli algoritmi per il calcolo di mappe di disparità dense vengono eseguiti in 4 principali step:

1. calcolo dei "costi" di ogni associazione
2. aggregazione dei costi precedentemente calcolati
3. ottimizzazione della disparità
4. raffinamento della disparità (rimozione delle associazioni errate)

Questi algoritmi sono suddivisi in due principali categorie: quelli ad approccio locale (che generalmente si compongono dei punti 1,2 e 4) e quelli ad approccio globale (che generalmente si compongono dei punti 1,3 e 4).

L'algoritmo proposto da Mattoccia mira ad integrare queste due tipologie di algoritmi cercando di procedere ad un'esecuzione locale (generalmente più veloce a livello di computazione rispetto ad una globale), ma usando una sorta di ragionamento globale sui pixel dell'intorno del pixel considerato.

Consideriamo la situazione mostrata in Figura 39 dove viene rappresentato, data una certa ipotesi di disparità *d*, il supporto dell'immagine di riferimento S_f e dell'immagine secondaria $S_{f'}$.

Assumiamo che la scena non abbia cambiamenti bruschi e che sia osservata per mezzo di camere con gli assi ottici paralleli tra loro, come nel nostro caso.

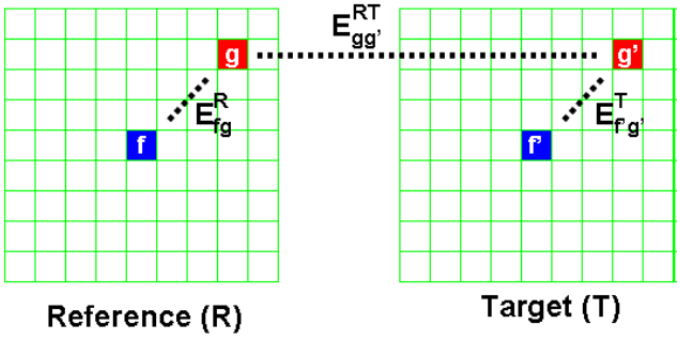


Figura 39 – L'immagine mostra il supporto Sf e Sf' per i punti $f \in R$ e $f' \in T$ per un'ipotesi di disparità "d".

La plausibilità dei punti g e g' facenti parte del supporto, rispettivamente S_f, centrato in f, e $S_{f'}$, centrato in f', può essere modellizzato come:

- $E^R_{f\,g}$: il punto $g \in S_f$ appartiene a S_f. Questo corrisponde al fatto che il punto *g* appartiene al supporto centrato in *f*. Se la scena non ha cambiamenti bruschi, allora la probabilità di questo evento è relativa alla somiglianza cromatica (o di luminanza) tra f e g. I punti più vicini al centro del supporto sono più rilevanti di quelli posti sul bordo dello stesso.
- $E^T_{f',\,g'}$: il punt $g' \in S'_f$o appartiene a S'_f. Questo corrisponde al fatto che il punto *g'* appartiene al supporto centrato in *f'*. E' modellato similarmente a $E^R_{f\,g}$.
- $E^{RT}_{g\,g'}(d)$: i punti $g \in S_f$ e $g' \in S'_f$ hanno, rispettivamente, disparità *d* e *-d*. Questo denota il fatto che i punti *g* e *g'* sono omologhi ed è rappresentato dalla prossimità del colore tra *g* e *g'*.

Dato un certo spazio colore, chiamiamo con Δ^{φ} una funzione che denota la somiglianza in termini cromatici tra i punti *f,g* e *f',g'* e con Δ^{ω} una funzione che denota la somiglianza cromatica tra i punti *g,g'*. Utilizzando la regola di Bayes è possibile scrivere la priorità a posteriori degli eventi mostrati in precedenza come:

$$P\left(E^R_{f\,g}, E^T_{f'\,g'}, E^{RT}_{g\,g'}(d) \middle|\, \Delta^{\varphi}_{f\,g}, \Delta^{\varphi}_{f'\,g'}, \Delta^{\omega}_{g\,g'}\right) \propto P_P\left(E^R_{f\,g}, E^T_{f'\,g'}, E^{RT}_{g\,g'}(d)\right) \cdot P_L\left(\Delta^{\varphi}_{f\,g}, \Delta^{\varphi}_{f'\,g'}, \Delta^{\omega}_{g\,g'} \middle|\, E^R_{f\,g}, E^T_{f'\,g'}, E^{RT}_{g\,g'}(d)\right)$$

Dove P_P e P_L sono rispettivamente la probabilità a priori e la probabilità di somiglianza. Assumiamo, per semplicità, che gli eventi $E^R_{f\,g}, E^T_{f'\,g'}$ *e* $E^{RT}_{g\,g'}(d)$ siano indipendenti. La precedente equazione può quindi essere riscritta come segue:

$$P\left(E^R_{f\,g}, E^T_{f'\,g'}, E^{RT}_{g\,g'}(d) \middle|\, \Delta^{\varphi}_{f\,g}, \Delta^{\varphi}_{f'\,g'}, \Delta^{\omega}_{g\,g'}\right)$$
$$\propto P_P\left(E^R_{f\,g}\right) \cdot P_L\left(\Delta^{\varphi}_{f\,g} \middle|\, E^R_{f\,g}\right)$$
$$\cdot\ P_P\left(E^T_{f'\,g'}\right) \cdot P_L\left(\Delta^{\varphi}_{f',\,g'} \middle|\, E^T_{f'\,g'}\right)$$
$$\cdot P_P\left(E^{RT}_{g\,g'}(d)\right) \cdot P_L\left(\Delta^{\omega}_{g\,g'} \middle| E^{RT}_{g\,g'}(d)\right)$$

Fissiamo la probabilità a priori $P_P(E^R_{f\,g})$ e $P_P(E^R_{f',\,g'})$ secondo il seguente vincolo:

$$P_P\left(E^R_{f\,g}\right) = e^{-\frac{\Delta_{f\,g}}{\gamma_s}}$$

dove Δ è la distanza euclidea tra f, g mentre γ_s è un parametro che controlla il vincolo di vicinanza spaziale. Assumiamo, inoltre, nessuna conoscenza a priori di $P_P(E^{RT}_{g\,g'}(d))$ per quanto riguarda l'aspetto dei punti *g* e *g'*.

Facciamo un'ulteriore assunzione: supponiamo che la scena sia formata solo da materiali lambertiani, ovvero materiali perfettamente opachi, e che le immagini abbiano un rumore bianco gaussiano di tipo additivo.

Denotiamo con $I_{(p)}$ un vettore che identifica l'intensità colore del punto *p*. Modellizziamo la prossimità cromatica dei punti *f,g* e *f',g'* mediante la seguente formula:

$$\Delta^{\varphi}_{f\,g} = \sqrt{\sum_{c \in R,G,B} (I_c(f) - I_c(g))^2}$$

e, similarmente, definiamo

$$\Delta^{\omega}_{g\,g'} = \sqrt{\sum_{c \in R,G,B} (I_c(g) - I_c(g'))^2}$$

Dalle formule precedenti è possibile ottenere la plausibilità che i punti *g* e *g'*, appartenenti ai supporti S_f e S'_f, abbiano la disparità *d* come segue:

$$P\left(E^{R}_{f\,g}, E^{T}_{f'\,g'}, E^{RT}_{g\,g'}(d)\middle|\ \Delta^{\varphi}_{f\,g}, \Delta^{\varphi}_{f'\,g'}, \Delta^{\omega}_{g\,g'}\right)$$

$$\propto e^{-\frac{\Delta_{f\,g}}{\gamma_s}} \cdot e^{-\frac{\Delta_{f\,g}}{\gamma_c}} \cdot e^{-\frac{\Delta_{f'\,g'}}{\gamma_s}} \cdot e^{-\frac{\Delta_{f'\,g'}}{\gamma_c}} \cdot e^{-\frac{\Delta_{g\,g'}}{\gamma_t}}$$

dove γ_c e γ_t sono due parametri che controllano il comportamento di $P_L(\ \Delta^{\varphi}_{f',\,g'}\middle|\ E^{T}_{f'\,g'})$ e $P_L\left(\ \Delta^{\omega}_{g\,g'}\middle|E^{RT}_{g\,g'}(d)\right)$.

I parametri γ_s, γ_ce γ_t sono determinati empiricamente .

L'equazione appena vista, quantifica la plausibilità dell'ipotesi di disparità *d* assunta per i due supporti S_f e S'_f, una volta che è stata individuata una corrispondenza tra *f* e *f'* con valore di disparità *d*.

E' da notare che la plausibilità include anche il termine $e^{-\frac{\Delta_{g\,g'}}{\gamma_t}}$ che quantifica la somiglianza cromatica tra i punti *g* e *g'* a differenza degli approcci tradizionali che utilizzano il termine $\Delta^{\omega}_{g\,g'}$. Questo accorgimento permette di rendere il contributo di $E^{R}_{f\,g}, E^{T}_{f'\,g'}, E^{RT}_{g\,g'}(d)$ omogeneo.

Conclusioni

In questo capitolo sono state presentate le tecniche e le scelte interpretative ed implementative adottate per la realizzazione di questa tesi. La realizzazione di questa parte di progetto ha comportato il superamento di diversi problemi.

In primo luogo, la scelta delle tecnologie più idonee da utilizzare, cioè camere, sensori e mezzi di comunicazione. Inoltre, l'ambiziosa idea di creare un sistema con parametri fissi delle camere per ogni condizione di scena, ha richiesto un'approfondita ricerca delle tecniche esistenti e del funzionamento della tecnologia LinLog di PhotonFocus per sfruttare al meglio queste tecnologie. Al fine di identificare i limiti di questo approccio ed individuare sperimentalmente i parametri da utilizzare implementativamente negli algoritmi, è stato inoltre necessario effettuare un ampio

numero di test sperimentali nelle più svariate condizioni di illuminazione e metereologiche.

Successivamente si sono dovute convogliare tutte queste elaborazioni ed informazioni verso le analisi di alto livello (calcolo della disparità, u-v disparity e free space map) in maniera efficace.

Si è poi analizzato il problema delle immagini mosse, che si presenta con velocità di movimento superiori a quelle possibili con USAD, in modo da rendere il sistema più robusto possibile a qualsiasi applicazione.

Nel seguente capitolo, verranno illustrati i risultati sperimentali di queste scelte implementative.

Capitolo 3 – Valutazioni sperimentali

Dopo avere mostrato qual è lo stato dell'arte per le tecniche applicate in visione artificiale e quali scelte implementative sono state adottate per il progetto presentato in questa tesi, verranno ora esposti i risultati sperimentali ottenuti.

Risultati per LinLog di PhotonFocus

La tecnologia presentata da PhotonFocus è effettivamente molto promettente e, sotto alcuni punti di vista, i risultati ottenuti sono stati sorprendenti.

Tuttavia, usarla correttamente e, soprattutto, riuscire ad identificare un set di parametri che permetta di ottenere buoni risultati, ha richiesto sia un notevole sforzo, al fine di capirne il funzionamento, sia un considerevole impegno pratico, per testare questa tecnologia in svariate condizioni.

Dopo avere realizzato un codice di lettura dei parametri predefiniti del LinLog contenuti all'interno della camere (parametri non forniti da PhotonFocus anche dietro esplicita richiesta), è stato più semplice capirne il funzionamento ed identificare un insieme di parametri ottimale per l' applicazione in esame.

Com'è possibile notare nelle immagini proposte in Figura 40, la tecnologia LinLog impedisce effettivamente la saturazione in condizione di forte luce. Nella prima immagine, i raggi visibili sono causati dall'ottica ed è possibile notare che l'unico punto sovraesposto è il sole stesso. Nella seconda immagine, si può osservare un'immagine dell'IraLab catturata con il medesimo tempo di esposizione della precedente, con una forte luce proveniente dalle finestre ed una parte molto scura sulla sinistra. Nella terza immagine, ancora una volta è possibile osservare i dettagli riconoscibili nel cielo

nonostante il forte controluce ed i dettagli nella parte scura in ombra sotto la finestra stessa.

Figura 40 – Le fotografie mostrano il risultato della tecnologia LinLog in una situazione molto sfavorevole. Sono state catturate in IraLab in condizione di forte controluce con sole estivo diretto attraverso la finestra. La rimanente parte di laboratorio non era illuminata da luce artificiale ed era quindi mediamente buia. Le condizioni di acquisizione sono le medesime.

I parametri della tecnologia LinLog visti in precedenza sono Value1, Value2 che rappresentano il valore di compressione logaritmica applicata alla parte luminosa dell'immagine, e Time1, Time2 che corrispondono al tempo, normalizzato rispetto a quello di esposizione, in cui fare intervenire i diversi livelli di compressione Value1 e Value2.

Questi valori sono stati determinati sperimentalmente analizzando il risultato delle immagini catturate in diverse condizioni di esposizione:

- Value1 = 170
- Value2 = 145
- Time1 = 960
- Time2 = 999

In questo modo si avrà una forte compressione per il 96% del tempo di esposizione, una media compressione per il 3,9% del tempo di esposizione e nessuna compressione per lo 0,1%.

Di seguito è possibile osservare un esempio della stessa scena catturata dalle camere PhotonFocus con LinLog impostato come indicato sopra e da una camera senza questa tecnologia.

Figura 41 – Le immagini mostrano la stessa scena ripresa da due camere differenti: una dotata di LinLog, a sinistra, e l'altra senza questa tecnologia, a destra. Il diverso campo visivo è causato dalla differenza lunghezza di focale delle ottiche utilizzate.

Risultati per pixel binning

Al fine di ottenere il massimo risultato dalle immagini catturate, si è scelto di sacrificare la risoluzione delle stesse per ottenere delle riprese chiare anche in ambienti con scarsissima illuminazione. La perdita di risoluzione, in questa particolare applicazione è tollerabile e a volte addirittura necessaria dato lo stato dell'arte sul fronte del computing power. Infatti, a causa dei tempi di calcolo che servono per eseguire le mappe di disparità, è stato indispensabile accettare dei compromessi sulla dimensione delle immagini.

In questo caso, al posto di effettuare un semplice sotto-campionamento, l'utilizzo del pixel-binning genera risultati migliori.

Il pixel-binning da noi utilizzato è di 4x4 pixel inglobando, quindi, le informazioni di 16 elementi fotosensibili in un unico "super-pixel". La risoluzione immagine passerà, dunque, da 1312x1082 a 328x270 pixel.

Figura 42 – Le coppie di immagini sopra riportate mostrano il risultato della sola operazione di pixel-binning. A sinistra è mostrata l'immagine del sensore (cioè con binning 1x1) e a destra il risultato dell'operazione (con binning 4x4).

Com'è possibile vedere dalle immagini riportate nella Figura 42, la luminosità media dell'immagine aumenta notevolmente e il numero di dettagli riconoscibili è nettamente superiore.

Risultati dell'equalizzazione dell'istogramma con stretchlim

L'equalizzazione dell'istogramma utilizzando la funzione di stretchlim ha permesso di ottenere ottimi risultati nelle immagini finali. Durante i test condotti per il pixel binning e per il funzionamento della tecnologia LinLog, è capitato sovente di trovarsi ad inquadrare una scena con illuminazione piuttosto uniforme ad eccezione di alcuni spot luminosi. La tecnica di equalizzazione dell'istogramma standard in queste circostanze, forniva dei pessimi risultati e, alcune volte, non portava alla variazione dell'immagine stessa.

Invece, applicando alle stesse immagini la tecnica di stretchlim, si sono ottenuti dei risultati più che soddisfacenti ed in alcuni casi l'unione del pixel-binning con lo stretchlim ha fornito esiti sorprendenti.

Osserviamo di seguito alcuni esempi significativi dell'utilizzo della tecnica di binning in unione a quella di equalizzazione dell'istogramma con stretchlim:

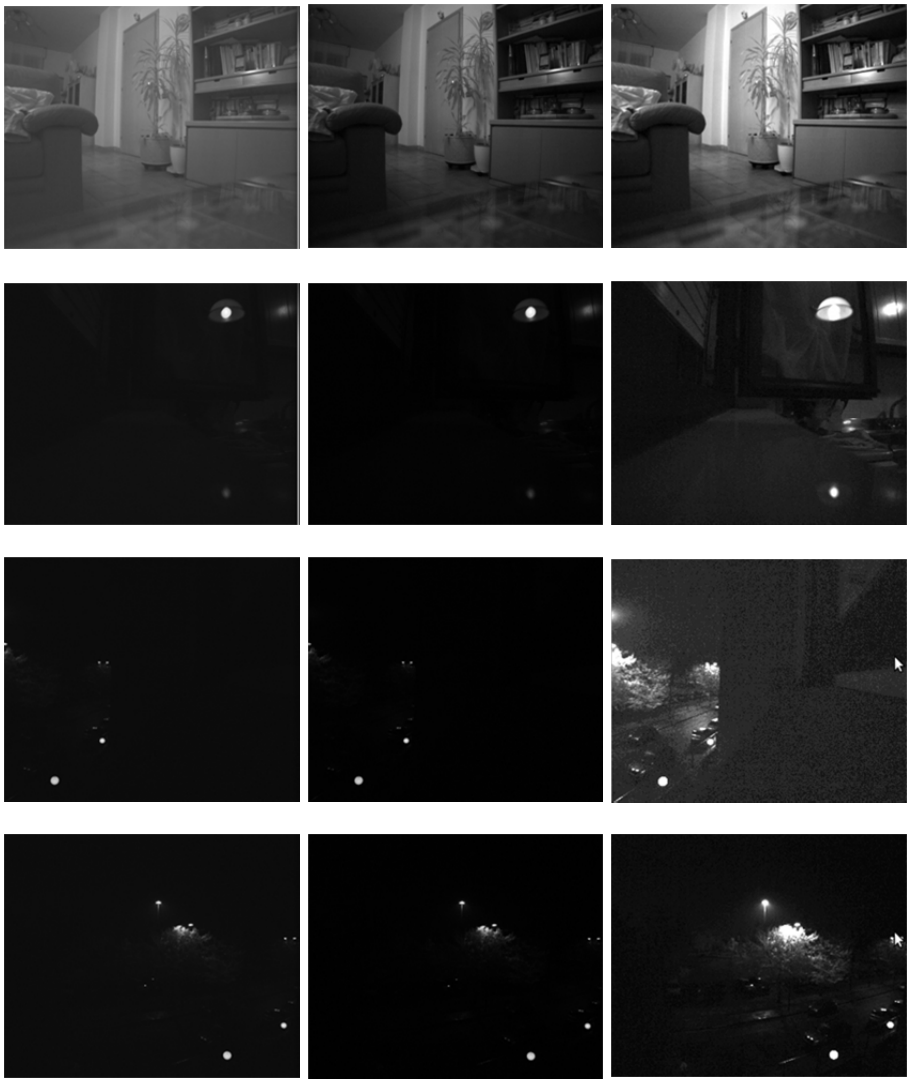


Figura 43 – Le triplette di figure sopra riportate mostrano: a sinistra l'immagine originale del sensore, in centro l'esecuzione dell'equalizzazione dell'istogramma "tradizionale" e a destra l'equalizzazione dell'istogramma utilizzando stretchlim.

Le immagini rappresentate nella seconda, terza e quarta riga sono delle immagini "estreme". Sono state catturate a notte fonda (senza luna) con un tempo di esposizione di 10ms e senza l'ausilio del guadagno digitale. Si nota che nelle immagini originali la quantità di dettagli riconoscibile è veramente scarsa, per non dire nulla, ad

eccezione degli spot luminosi. La tecnica dell'equalizzazione dell'istogramma, in questi casi, non migliora la foto, mentre quella proposta in questa tesi permette di osservare un numero di dettagli particolarmente elevato. In questa condizione specifica è possibile anche osservare una certa quantità di rumore, causata dalla bassa illuminazione della scena e dell'amplificazione introdotta dal pixel-binning. Il rumore presente nell'immagine verrà poi eleminato con degli appositi filtri di smoothing durante il calcolo della mappa di disparità.

Le percentuali di dati saturabili con questa tecnica sono state determinate sperimentalmente e fissate allo 0,1% per le tonalità scure e allo 0,5% per le tonalità chiare.

L'immagine finale, a cui è stato applicato il pixel binning 4x4, ha una risoluzione di 328x270 e quindi un totale di 88.560 pixel. Con le percentuali riportate, si è disposti a perdere 88 pixel che verranno saturati al livello di minore intensità luminosa, e 440 pixel che verranno saturati a quello con maggiore luminanza.

Risultati per impostazioni della camera costanti per ogni situazione

Presentiamo ora il risultato dell'ambiziosa idea di riuscire ad utilizzare dei parametri fissi delle camere per ogni condizione della scena. Pur avendo raggiunto dei compromessi per quanto riguarda la risoluzione delle immagini, i risultati ottenuti sono ottimi.

E' stato scelto un tempo di esposizione di 10ms che è adatto alle condizioni di movimento di USAD. Per identificare questo valore sono stati condotti una serie di test in movimento in diverse condizioni, considerando la velocità di crociera del veicolo, la sua velocità massima e variando le situazioni di controllo con muri nelle vicinanze (e quindi moto relativo maggiore) oppure veicoli/persone in movimento con verso opposto (con conseguente percezione di movimento con somma delle velocità).

In tutte le circostanze, il tempo di 10ms non ha causato il ***blur*** delle immagini e si è rivelato adatto alle condizioni d'illuminazione più disparate.

Al fine di rendere il più esaustivo possibile il test del sistema a parametri fissi delle camere si è provveduto ad effettuare numerosi test in condizioni di ripresa molto differenti in modo da considerare il maggior numero di situazioni a cui si deve far fronte in un ambiente di guida non controllato.

I test sono stati eseguiti nelle seguenti condizioni:

- Interno: fiera EIV a Rho-Pero. Illuminazione interna con luci a spot, molte zone sottoesposte e sovraesposte e sole in ingresso dalle finestre alte del padiglione. Questo ambiente è particolarmente impegnativo per le differenti e repentine variazioni di illuminazione della scena e per la presenza di luce molto forte proveniente dalle finestre e localizzata in determinati punti (Immagine nella seconda riga a sinistra della Figura 44).
- Interno: parcheggio interrato edificio U5. Illuminazione mediante luci al neon e grate laterali da cui entra luce diurna. L'illuminazione di questo ambiente è mediamente molto scarsa; vi è, tuttavia, la presenza di forti sorgenti di illuminazione dalle grate nelle giornate soleggiate.
- Esterno: parcheggio esterno edificio U5 in condizione di sole. I test sono stati eseguiti sia con il sole a perpendicolo, sia al tramonto in condizioni di forte controluce
- Esterno: parcheggio esterno edificio U5 in condizione di neve. E' stato anche condotto un test durante una nevicata a Milano. Le acquisizioni sono state eseguite sia in presenza di illuminazione diurna, sia con l'illuminazione pubblica (lampioni stradali).
- Passaggio da esterno ad interno. Questa condizione è particolarmente significativa per simulare, ad esempio, l'entrata in una galleria.

In tutte le condizioni di acquisizione sopra descritte, le immagini ottenute con i parametri fissi delle camere sono risultate perfettamente utilizzabili per le successive elaborazioni di alto livello.

Come precedentemente detto, tuttavia, il tempo di 10ms di esposizione scelto potrebbe non essere adatto a velocità di guida superiori a quelle raggiunte da USAD.

Di seguito, viene mostrata una selezione di casi significativi, tutti catturati con le medesime impostazioni delle camere.

Figura 44 – Nelle immagini sopra sono rappresentati alcuni casi significativi acquisiti durante differenti orari del giorno e diverse condizioni metereologiche e di luce. Com'è possibile notare, tutte le immagini sono ben contrastate e sufficientemente luminose. Nell'ultima immagine è possibile osservare un tipico problema delle ottiche quando una "lama" di luce colpisce il vetro, rendendo cieca la camera.

Nell'ultima immagine è possibile notare il più insidioso problema riscontrato durante tutti i test, ovvero la luce del sole incidente sulla lente dell'ottica che causa l'accecamento della camera. Questo inconveniente è piuttosto raro e limitato a poche combinazioni di condizioni d'acquisizione. Per evitarlo è possibile installare sulle ottiche un paraluce che impedisca ai raggi della luce con particolari incidenze di colpire il vetro della lente esterna creando il fenomeno di accecamento.

Risultati controllo della metrica per la rilevazione del "mosso" nelle immagini

Nonostante gli ottimi risultati ottenuti con i parametri fissi delle camere, si è scelto in ogni caso di testare la metrica per l'identificazione del mosso delle immagini presentata nei capitoli precedenti.

Utilizzando la letteratura disponibile ***[31] [30] [20] [21]*** e svolgendo dei test su alcuni set di immagini con differenti livelli di sfocatura, è stato possibile definire delle classi di blur. Nella tabella di seguito, riportiamo i centroidi di tali classi:

Centroide della classe	Quantità di blur/qualità immagine
0.02	Molto elevato / pessima
0.16	Elevato / scarsa
0.34	Poco / mediocre
0.54	Molto poco / buona
0.73	Assente / ottima

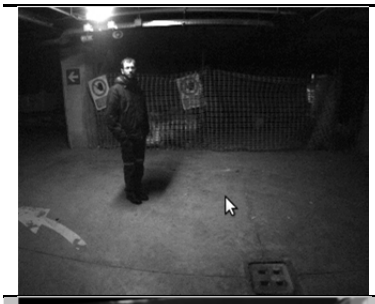
Sono stati condotti dei test specifici per questa metrica su immagini direttamente rilevate dalle camere di USAD variandone opportunamente i tempi di esposizione in modo da ricreare l'effetto mosso, oppure per eliminarlo del tutto.

Tutte le immagini catturate con i parametri fissi delle camere visti in precedenza hanno riscosso un valore appartenente alle classi di qualità immagine "buona" e "ottima", mentre quelle catturate aumentando volontariamente il tempo di esposizione, si sono attestate tra le classi "scarsa" e "pessima". Quest'ultima osservazione permette di fare le seguenti considerazioni:

- il tempo di esposizione di 10ms scelto come parametro fisso per le camere è effettivamente adatto al movimento di USAD stesso e la metrica CPBD conferma il risultato visivo di assenza di mosso nelle immagini catturate.
- La metrica CPBD è in grado di rilevare con buona efficienza la quantità di sfocamento presente nelle immagini ed è possibile

utilizzarla per identificare quando il tempo di esposizione è troppo elevato in relazione alla scena che stiamo osservando.

Di seguito riportiamo alcuni esempi di immagini con il risultato ottenuto dal calcolo della metrica CPBD.

	Scena statica, con nessun movimento. **CPBD = 0,91** L'immagine non presenta traccia di sfocatura o mosso
	USAD statico, ma movimento delle persone davanti. **CPBD = 0,66** L'immagine non presenta sfocatura o è comunque molto poco mossa
	USAD in movimento alla massima velocità con presenza di veicoli e piante nelle immediate vicinanze. **CPBD = 0,69** L'immagine non presenta sfocatura o è comunque molto poco mossa
	USAD in movimento alla massima velocità con presenza di veicoli e piante nelle immediate vicinanze. **CPBD = 0,75** L'immagine non presenta sfocatura o mosso

	USAD in movimento alla massima velocità e tempo di esposizione molto alto (100ms) **CPBD = 0,17** L'immagine è completamente mossa ed è fortemente sfocata
	USAD in movimento a bassa velocità e tempo di esposizione molto alto (100ms) **CPBD = 0,26** L'immagine a vista sembra poco mossa, ma in realtà la metrica rileva che il tempo di esposizione non è adeguato.
	USAD in movimento a media velocità e tempo di esposizione molto alto (100ms) **CPBD = 0,19** L'immagine è sfocata e mossa a causa dell'alto tempo di esposizione

Com'è possibile notare dai casi sopra riportati, la metrica CPBD permette di identificare in modo preciso il grado di sfocatura presente in un'immagine. E' possibile, dunque, utilizzarla per supervisionare la variazione del tempo di esposizione operata dall'analisi della luminosità media delle immagini, descritta precedentemente.

Risultati per il calcolo della mappa di disparità, U-V disparity e mappa dello spazio libero

Come abbiamo visto in precedenza, avere una mappa della disparità densa è molto importante quando si opera nel campo della visione artificiale e della guida autonoma. Più dettagli riusciremo ad avere

nella mappa di disparità, tanto più ne riusciremo a ricostruire nel mondo e maggiori saranno gli ostacoli che potremo identificare.

Un altro aspetto fondamentale del calcolo della mappa di disparità è il tempo di computazione. Queste elaborazioni di alto livello, infatti, utilizzano una grande quantità di risorse del calcolatore.

I risultati delle immagini ottenuti mediante le tecniche presentate in questa tesi sono stati utilizzati per eseguire il codice prodotto dal D.R. Marzorati che ha effettivamente implementato l'algoritmo presentato da Mattoccia.

Le immagini utilizzate per la computazione hanno una risoluzione di 328x270 pixel ed i risultati qualitativi del calcolo della mappa di disparità sono, come vedremo di seguito, più che soddisfacenti a dimostrazione della validità dei risultati ottenuti in questa tesi.

Abbiamo riscontrato un problema piuttosto grave nei tempi di calcolo della disparità in quanto la frequenza ottenuta è scarsamente utilizzabile in applicazioni reali. Con un computer dotato di processore Core2Duo da 2,4GHz di frequenza, 4 GB di RAM e un FSB di 800MHz, i tempi rilevati per il calcolo di una singola mappa di disparità su single core erano superiori al secondo. Per questo motivo, abbiamo deciso di provare l'implementazione multi-core sfruttando tutta la potenzialità del processore. Sempre riferendosi alla precedente configurazione (questa volta usando entrambi i core disponibili) il tempo di computazione è sceso a 400-500 ms. Questo tempo è tuttavia ancora troppo elevato per consentire la navigazione in sicurezza.

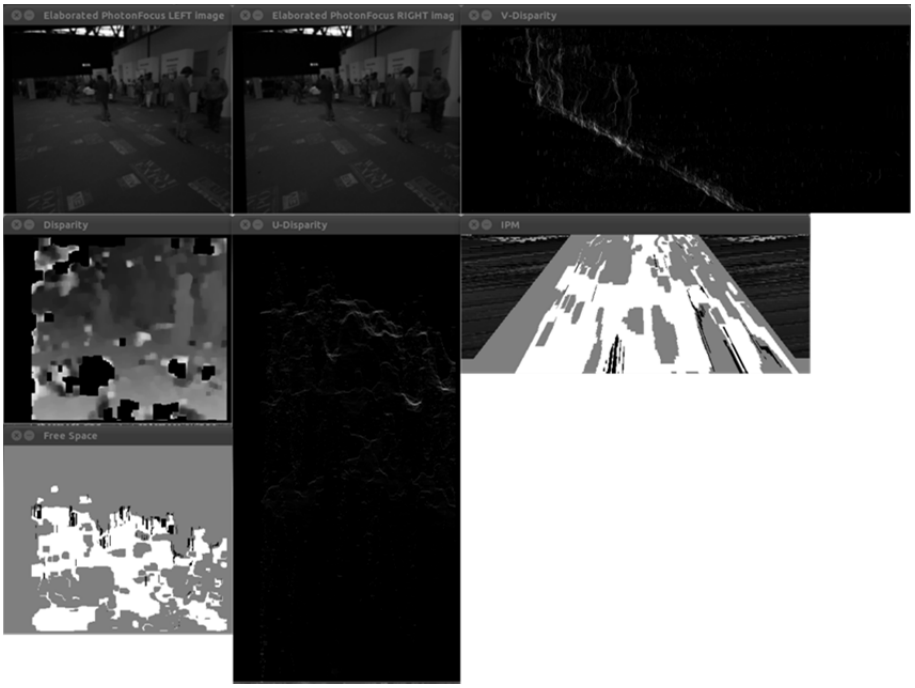
Lo stesso algoritmo è stato anche testato su una macchina con processore Core I7 a 3,06 GHz, 12 GB di RAM e QPI a 4,8GT/s ottenendo dei risultati migliori. Il tempo di computazione è sceso a 250-300 ms per ogni mappa della disparità. Anche se migliori dei precedenti, questi risultati sono ancora troppo elevati per riuscire a navigare in modo autonomo. Abbiamo tentato un ulteriore test nel quale abbiamo overcloccato il precedente processore alla frequenza di 3,7 GHz. In questo caso abbiamo ottenuto dei tempi di calcolo

nettamente più soddisfacenti, attestati tra 150 e 180ms che permettono di ottenere circa sei mappe di disparità della scena ogni secondo.

L'obiettivo ottimale, tuttavia, sarebbe quello di avere una frequenza di mappe della disparità superiore a 10Hz, traguardo ad oggi ancora difficilmente raggiungibile.

Di seguito vediamo alcuni risultati del calcolo della mappa di disparità, della u-v disparity e della rappresentazione dello spazio libero.

Nelle immagini sono riportate le acquisizioni della camera destra e della camera sinistra elaborate con le tecniche presentate in questa tesi, l'immagine che rappresenta la mappa di disparità (posta sotto l'immagine della camera di sinistra), la mappa v-disparity (a destra dell'immagine della camera di destra), la mappa u-disparity (sotto l'immagine della camera di destra) e la mappa dello spazio libero (sotto la mappa di disparità). Viene, inoltre, rappresentata la IPM, ovvero la prospettiva della mappa dello spazio libero come se fosse proiettata dalle camere sulla superficie stradale.



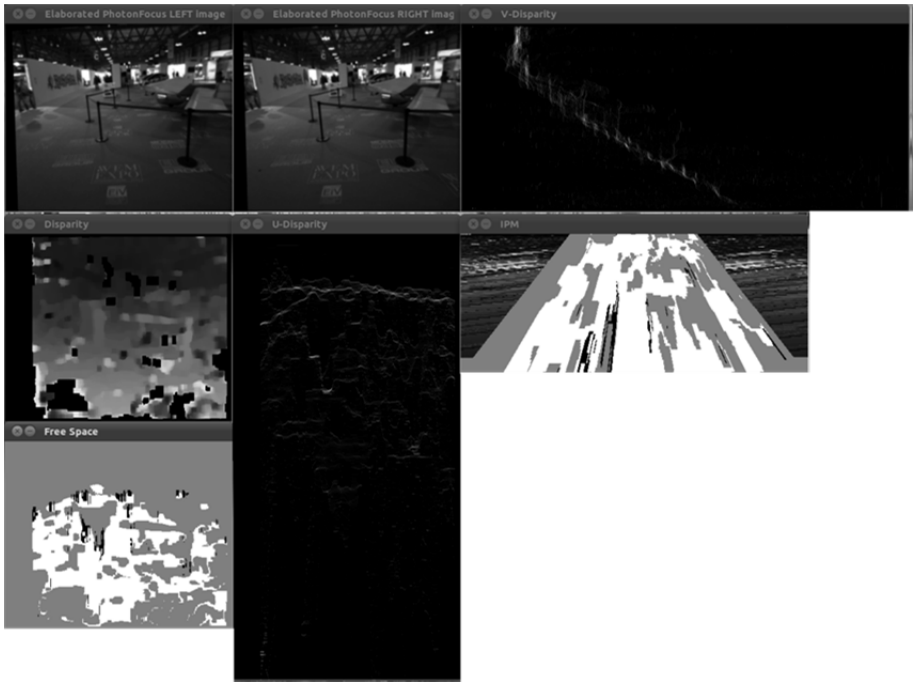


Figura 45 – Le immagini mostrano il risultato dell'elaborazione delle immagini della camera di sinistra e della camera di destra, l'immagine che rappresenta la mappa di disparità (posta sotto l'immagine della camera di sinistra), la mappa v-disparity (a destra dell'immagine della camera di destra), la mappa u-disparity (sotto l'immagine della camera di destra) e la mappa dello spazio libero (sotto la mappa di disparità). Viene, inoltre, rappresentata la IPM, ovvero la prospettiva della mappa dello spazio libero come se fosse proiettata dalle camere sulla superficie stradale

Capitolo 4 – Conclusioni e sviluppi futuri

Conclusioni

Con questo progetto ci siamo posti come obiettivo di implementare un sistema di stereo visione da applicare ai veicoli con guida autonoma che fornisse delle immagini sempre utilizzabili, in ogni condizione di acquisizione e movimento del veicolo, per le elaborazioni di alto livello, come il calcolo della mappa di disparità e dello spazio libero.

Il lavoro si è articolato come segue:

- **analisi dello stato dell'arte.** Questa prima fase iniziale del progetto ha richiesto la ricerca e l'analisi delle tecnologie presenti sul mercato al fine di identificare quelle più adatte al nostro progetto.

- **realizzazione della testa stereo.** Durante questa fase ci siamo occupati della creazione fisica della testa stereo. Si è reso necessario identificare quale fosse la migliore posizione di montaggio trovando un modo per rendere le camere più solidali possibili ad USAD. Abbiamo quindi proceduto alla calibrazione della testa stereo stessa.

- **analisi della tecnologia LinLog.** E' stato necessario realizzare un programma specifico per leggere i parametri interni non pubblici delle camere riguardanti le impostazioni del LinLog. Successivamente sono stati svolti test sperimentali approfonditi per identificare quali fossero i parametri ottimali da utilizzare.

- **identificazione delle tecniche da usare per l'elaborazione di basso livello delle immagini.** In questa fase si è scelto quali

tecniche utilizzare per l'elaborazione delle immagini catturate dalle camere optando per il pixel-binning ed la normalizzazione dell'istogramma con stretchlim.

- **implementazione nel linguaggio C.** La fase successiva ha riguardato la scrittura del codice C per le tecniche viste precedentemente. E' stato poi necessario integrare il tutto nella libreria DAFNE sviluppata dal D.R. Marzorati che include, tra l'altro, il calcolo della mappa della disparità.

- **identificazione dei parametri fissi per le camere.** Terminata la fase implementativa, ci siamo dedicati ad identificare sperimentalmente quali fossero i migliori parametri a cui impostare le camere per ottenere il risultato di utilizzare parametri costanti in ogni situazione

- **verifica sperimentale dei risultati.** Successivamente all'integrazione, si è proceduto ad un'esaustiva fase di verifica sperimentale dei risultati ottenuti. I test sono stati condotti in diverse condizioni di acquisizione, in ambienti chiusi ed all'aperto. Si è scelto, inoltre, di verificare l'affidabilità del sistema anche sotto diverse condizioni metereologiche, come neve, notte e durante il tramonto con sole radente.

- **analisi delle tecniche per il blur-detection.** Nonostante i più che soddisfacenti risultati ottenuti con i parametri costanti delle camere, abbiamo svolto un'approfondita ricerca per identificare quali fossero le tecniche esistenti per la rilevazione della sfocatura (o effetto mosso) nelle immagini.

- **scelta e test della metrica per il blur-detection.** In questa fase abbiamo valutato l'efficacia delle varie tecniche di blur detection individuate in precedenza ed abbiamo scelto la promettente metrica CPBD, implementandone una prima versione in Matlab

- **verifica sperimentale dei risultati per il blur-detection.** Sono stati condotti dei test sperimentali basati sui dataset immagine, catturati nelle fasi precedenti, volti a verificare

l'efficacia della metrica CPBM nel rilevare ed identificare le immagini mosse.

Sviluppi futuri

Analizziamo ora quali sono i possibili sviluppi futuri di questo progetto:

- la tecnica del pixel-binning può essere resa dinamica ricorrendo ad esempio, ad un binning meno incisivo (come ad esempio un 3x3 oppure un 2x2) al fine di ottenere delle immagini a più alta risoluzione quando l'illuminazione della scena è sufficiente ed incrementare il valore di binning con il diminuire della luminosità;
- l'ottimizzazione algoritmica del calcolo della disparity map può essere migliorata cercando di individuare tecniche che riducano il tempo di computazione;
- è necessario implementare in linguaggio C l'algoritmo per il blur-detection presentato in questa tesi;
- è opportuno implementare un processo in background che effettui la retroazione sul tempo di esposizione delle camere in relazione alla luminosità media dell'immagine ed alla metrica di blur-detection presentata in questa tesi
- le informazioni fornite dalla mappa dello spazio libero possono essere utilizzate per riuscire a guidare in modo autonomo, interfacciandosi con il ciclo di controllo di USAD stesso.

Bibliografia

[1] Trucco E., Verri A. - Introductory techniques for 3D computer vision (1998). Prentice Hall Editore

[2] O. Faugeras - Three dimensional computer vision: a geometric viewpoint, Mit press, 1993.

[3] D. A. Forsyth, J. Ponce - Computer vision, a modern approach, Pearson Prentice-Hall, 2003

[4] B. K. P. Horn - Robot vision, Mit press, 1986

[5] Tinku Acharya, Ajoy K. Ray - Image Processing: Principles and Applications, Wiley, 2005

[6] Frédréric Devernay and Olivier Faugeras - Straight lines have to be straight: automatic calibration and removal of distortion from scenes of structured enviroments. Mach. Vision Appl., 13(1):14–24, 2001.

[7] Janne Heikkila and Olli Silven - A four-step camera calibration procedure with implicit image correction. In CVPR '97: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97), page 1106, Washington, DC, USA, 1997. IEEE Computer Society.

[8] J. Heikkilti, O. Silv´e, and n. Calibration procedure for short focal length off-the-shelf ccd-cameras. In ICPR '96: Proceedings of the 1996 International Conference on Pattern Recognition (ICPR '96) Volume I, page 166, Washington, DC, USA, 1996. IEEE Computer Society.

[9] http://sourceforge.net/projects/opencvlibrary/ - The open computer vision library.

[10] Rafael C. Gonzalez, Richard E. Woods – Digital Image processing, Prentice Hall, 2002

[11] Mortimer Abramowitz, Michael W. Davidson - Concepts in Digital Imaging Technology: Pixel binning

[12] http://itee.uq.edu.au/~iris/CVsource/OpenCVreferencemanual.pdf - OpenCV reference manual

[13] L. Di Stefano, M. Marchionni, S. Mattoccia – A PC-based Real-Time stereo vision system. Machine GRAPHICS & VISION vol. 13, no. 3, 2004, pp. 197-220

[14] Motilal Agrawal, Kurt Konolige, Morten Rufus Blas – CenSurE: Center Surround Extremas for Realtime Feature Detection and Matching. European Conference on Computer Vision ECCV08 (2008) Volume**:** 5305, Publisher: Springer, Pages: 102-115

[15] Alexandre Guilavard, Pierre Magnan, Josep Segura, Philippe Martin-Gonthier – A High dynamic range digital LinLog CMOS image sensor architecture based on Event Readout of pixel and suitable for low voltage operation. 2007 International Image Sensor Workshop, 07-10 Jun 2007, Ogunquit Maine, United States.

[16] Zhencheng Hu, Francisco Lamosa, Heiichi Uchimura – A Complete U-V-Disparity Study for stereovision based 3D driving environment analysis. 3DIM '05 Proceedings of the Fifth International Conference on 3-D Digital Imaging and Modeling. 13-16 June 2005 IEEE Computer Society Washington, DC, pp.: 204 – 211

[17] Javier Civera, Andrew J. Davison, J.M.M Montiel – Inverse Depth to depth conversion for monocular SLAM. Robotics and Automation, 2007 IEEE International Conference, April 2007, pp. 2778-2783.

[18] Raphal Labayrade, Dider Aubert – A single framework for Vehicle Roll, Pitch, Yaw estimation and obstacle detection by stereovision. Intelligent Vehicles Symposium, 2003. Proceedings. IEEE, 9-11 June 2003, pp. 31 - 36

[19] Stefano Mattoccia – A locally global approach to stereo correspondence. IEEE Workshop on 3D Digital Imaging and Modeling (3DIM2009), October 3-4, 2009, Kyoto, Japan, pp. 1763 - 1770

[20] Pina Marziliano, Frederic Dufaux, Stefan Winkler and Touradj Ebrahimi - A No-Reference Perceptual Blur Metric. International Conference on Image Processing, Rochester, NY, September 22--25, 2002, vol. 3, 2002, p. 57-60

[21] Niranjan D. Narvekar and Lina J. Karam - An Improved No-Reference Sharpness Metric Based On The Probability Of Blur Detection. International Workshop on Quality of Multimedia Experience (2009). Pages: 87-91

[22] Dave Litwiller – CCD vs. CMOS: Facts and Fiction. January 2001 issue of PHOTONICS SPECTRA , Laurin Publishing Co. Inc.

[23] W. S. Boyle and G. E. Smith - Charge Coupled Semiconductor Devices. Bell Syst. Tech. J. 49:587-93, 1970. [Bell Laboratories, Murray H i l l , NJ]. (April 1970)

[24] Tomaso Poggio, David Marr. - Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. New York, W. H. Freeman and Company 1982.

[25] Stan Birchfield and Carlo Tomasi - Depth Discontinuities by Pixel-to-Pixel Stereo. Proceedings of the Sixth IEEE International Conference on Computer Vision, Mumbai, India, pages 1073-1080, January 1998

[26] Stan Birchfield and Carlo Tomasi - A Pixel Dissimilarity Measure that is Insensitive To Image Sampling. IEEE Transactions on Pattern Analysis and Machine Intelligence,20(4):401-406, April 1998

[27] Nicolas Soquet, Mathiaas Perrollaz, Raphael Labayrade, Dedier Aubert – Free space estimation for autonomous navigation. The 5th International Conference on Computer Vision Systems, 2007 Bielefeld, 21. - 24. March 2007

[28] Labayrade, R.; Aubert, D.; Tarel, J.-P. - Real time obstacle detection in stereovision on non flat road geometry through "v-disparity" representation. Intelligent Vehicle Symposium, 2002. IEEE 17-21 June 2002

[29] http://www.vision.caltech.edu/bouguetj/calib_doc/ - Camera Calibration Toolbox for Matlab

[30] R. Ferzli and L. J. Karam, “A no-reference objective image sharpness metric based on the notion of Just Noticeable Blur (JNB),” IEEE Transactions on Image Processing, vol. 18, pp. 717-728, Apr. 2009.

[31] http://people.usd.edu/~schieber/coglab/WebersLaw.html - Fondamenti sulla legge di Weber e la Just Noticeable Difference

[32] http://www.ccd.com/ccd103.html - Pixel-binning. AIS (Apogee Imagin Systems) – CCD University

[33] R. Dyck and G. Weckler. "Integrated arrays of silicon photodetectors for image sensing". IEEE Trans. Electron Devices ED-15 (4): 196–201.

[34] Hanghang Tong, Mingjing Li, Hongjiang Zhang, Changshui Zhang - Blur Detection for Digital Images Using Wavelet Transform.

[35] Renting Liu. Zhaorong Li. Jiaya Jia. - Image Partial Blur Detection and Classification. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008

[36] James H. Elder, Steven W. Zucker - Local Scale Control for Edge Detection and Blur Estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence archive Volume 20 Issue 7, July 1998

[37] E. Peli - Contrast in Complex Images. Journal of the Optical Society of America A 7 (10): 2032–2040.doi:10.1364/JOSAA.7.002032 (Oct. 1990).

[38] E. Trucco, V. Roberto, S. Tinonin, M. Corbatto - SSD Disparity Estimation for Dynamic Stereo. Proceedings of the British Machine Vision Conference (1996)

[39] Xiaogang Chen, Jie Yang, Qiang Wu, Jiajia Zhao - Motion Blur Detection Based On Lowest Directional High-Frequency Energy.

Proceedings of 2010 IEEE 17th International Conference on Image Processing, September 26-29, 2010, Hong Kong

[40] R. Ferzli, and Lina J. Karam - No-Reference Objective Wavelet Based Noise Immune Image Sharpness Metric. Image Processing, 2005. ICIP 2005. IEEE International Conference, pp. I - 405-8

[41] Canny, J., *A Computational Approach To Edge Detection*, IEEE Trans. Pattern Analysis and Machine Intelligence, 8(6):679–698, 1986

[42] R. Deriche, *Using Canny's criteria to derive a recursively implemented optimal edge detector*, Int. J. Computer Vision, Vol. 1, pp. 167–187, April 1987.

Ringraziamenti

Un grazie di cuore ai miei genitori e a mia sorella Barbara che mi hanno sempre supportato, e hanno reso possibile tutto il percorso della mia carriera universitaria sostenendomi anche nei momenti difficili.

Un sincero ringraziamento al Prof. Sorrenti per la fiducia accordatami e per la sua infinita disponibilità.

Un sentito ringraziamento anche ai miei due correlatori, Axel e Daniele che mi hanno sostenuto e saggiamente consigliato nel cammino di questo elaborato, e che hanno condiviso con me splendidi momenti in Iralab.

Uno speciale ringraziamento anche a mio zio Gianpaolo, grazie al quale ho iniziato il mio percorso di studi in informatica e che mi ha sempre consigliato saggiamente e supportato pazientemente. Senza il suo prezioso contributo ed i suoi consigli, probabilmente ora starei facendo tutt'altro.

Un grazie anche ai miei amici più cari, con i quali ho condiviso il mio percorso accademico.

Un particolare ringraziamento ad Ignazio e Teresa per le lunghe chiacchierate, per il loro sostegno e per avere sempre creduto in me.

Grazie anche a Giuseppe per i consigli, le nuotate estive nello splendido mare di Tropea e il continuo supporto morale e materiale!

Non posso non nominare tutti i compagni "iralabbiani" (per non fare torno a nessuno, in ordine alfabetico :-P) Andrea, Augusto, Francesco S., Francesco V. con i quali ho condiviso splendidi momenti (di "follia") in Iralab e momenti difficili nei "piani interrati".

Un grazie a mia nonna Clelia per la sua disponibilità e generosità.

Un grazie anche ai miei nonni, Maddalena, Antonio e Dario che, anche se non più vicini fisicamente, sono sempre presenti per me ed il loro ricordo ed il loro esempio mi hanno sempre dato la forza e l'energia per continuare il mio percorso e superare anche i momenti più difficili.

Un sincero grazie a tutti quelli che non ho esplicitamente citato, ma che hanno partecipato a questa mia carriera universitaria e che hanno contribuito a fare di me la persona che sono.