

UNIVERSITÀ DEGLI STUDI DI NAPOLI “PARTHENOPE”

SCUOLA INTERDIPARTIMENTALE DELLE SCIENZE,
DELL'INGEGNERIA E DELLA SALUTE

DIPARTIMENTO DI SCIENZE E TECNOLOGIE

CORSO DI LAUREA IN INFORMATICA

Tesi di Laurea Triennale in Informatica

Multi-person 3D skeleton tracking using Intel RealSense D435 and OpenPose with Kafka support

Relatore

Ch.mo prof. Ferone Alessio

Candidato

Denny Caruso

matr. 0124002062

Anno Accademico 2021/2022

Learn from other people's mistakes and do that before they do.

Sommario

Sempre più spesso e in sempre più contesti la sicurezza è messa a rischio: la sicurezza delle persone, la sicurezza di oggetti e di ambienti. Al fine di garantire il massimo grado di sicurezza vengono adottate misure sempre più avanzate e in alcuni casi anche costose e stringenti. Uno dei metodi più convenzionali per sorvegliare una scena o un ambiente all'interno del quale si vuole garantire la sicurezza, è quella d'installare dei dispositivi di video-sorveglianza che catturano video 2D. Questo tipo di approccio necessita di sempre più operatori al crescere del numero di videocamere.

Oggi la tecnologia permette di automatizzare parte di questo lavoro, riuscendo a rilevare eventuali anomalie, scene sospette ed eventi che potrebbero mettere a rischio la sicurezza della scena osservata con un elevato grado di precisione. Una forte limitazione a tale approccio però, è l'assenza dell'informazione sulla profondità all'interno della scena e di conseguenza di oggetti e persone in essa presenti, o dell'imprecisione sull'informazione profondità a causa del suo "retrieving" manuale.

Si vuole dimostrare come sia possibile realizzare una soluzione di video-sorveglianza 3D che tenga conto dell'informazione sulla profondità grazie a un'appropriata sensoristica, che sia in grado di rilevare più persone all'interno della scena osservata, che sia in grado di rilevare anomalie e comportamenti anomali, che sia di facile implementazione ed economicamente conveniente.

Indice

1	Introduzione	1
1.1	Contesto	1
1.2	Concetti preliminari	4
1.3	Obiettivo	9
1.4	Organizzazione della tesi	12
2	Stato dell'arte	13
2.1	Aspetti teorici	13
2.1.1	Profondità e videocamere RGBD	13
2.1.2	Cenni di geometria epipolare	16
2.1.3	Intelligenza artificiale	20
2.1.4	Il percepitrone	21
2.1.5	Artificial Neural Network	23
2.1.6	Convolutional Neural Network	23
2.2	Intel RealSense	26
2.2.1	Intel RealSense D435	26
2.2.2	Intel RealSense D435 e Intel RealSense T265	30
2.2.3	Tecnologia infrarossi, laser, lidar	31
2.3	SLAM - Simultaneous localization and mapping	34
2.4	Ottenerne la profondità in un'immagine	36
2.5	Pose Estimation	37
2.5.1	OpenPose	37
2.6	Trasformazione di spazi di coordinate	37
2.7	Streaming di eventi	37
3	Progettazione	38
4	Implementazione	39
5	Risultati sperimentali	40

Elenco delle figure

1.1	Telecamere Intel RealSense [2]	2
1.2	Casi d'uso videocamere [4]	3
1.3	Rappresentazione immagine digitale	4
1.4	Spettro elettromagnetico [6]	5
1.5	Spazio colore RGB	6
1.6	Immagine RGB	7
1.7	Immagine RGBD	7
1.8	Skeleton [34]	9
1.9	Acquisizione e scissione frame video, rilevazione ed estrazione skeleton, conversione e rendering delle coordinate	10
2.1	Due fotocamere che acquisiscono un'immagine idealmente nel- lo stesso istante della stessa scena [11]	17
2.2	Determinazione della profondità per ogni pixel dell'immagine acquisita [11]	19
2.3	Un neurone biologico confrontato con una rete neurale artifi- ciale: (a) neurone umano; (b) neurone artificiale; (c) sinapsi biologica; (d) sinapsi ANN [24]	21
2.4	Rete neurale convoluzionale [24]	24
2.5	Viste differenti della telecamera Intel RealSense D435 [2] . . .	27
2.6	Architettura interna telecamera Intel RealSense D430 [2] . . .	27
2.7	Effetto Rolling Shutter [36]	28
2.8	Telecamera Intel RealSense T265 [2]	30
2.9	Pattern circolari Intel RealSense D415 e D435 [25]	31
2.10	Immagine lidar [25]	33
2.11	Rappresentazione del problema SLAM [26]	34
2.12	Tecnologia Raytrix [31]	37

Elenco delle tabelle

2.1 Vantaggi e svantaggi delle telecamere orientate al Field of View oppure al Light Spectrum	29
--	----

xi

Capitolo 1

Introduzione

Sicurezza: la condizione che rende e fa sentire di essere esente da pericoli, o che dà la possibilità di prevenire, eliminare o rendere meno gravi danni, rischi, difficoltà, evenienze spiacevoli, e simili. [1]

1.1 Contesto

Oggigiorno si sente parlare sempre più spesso di sicurezza: sicurezza sul lavoro, in mare, in volo, in casa, di persone, di oggetti, di ambienti e così via. Per preservare e mantenere la sicurezza è possibile adottare diversi approcci, alcuni più costosi, altri più invadenti e stringenti; altri ancora più innovativi. Un approccio classico per mantenere la sicurezza è l'utilizzo di videocamere di sorveglianza. Una videocamera è un apparecchio portatile che permette

di registrare su uno o più supporti segnali corrispondenti a immagini e suoni; comprende un sistema ottico (in genere uno zoom con regolazione automatica dell'esposizione), un microfono, un sensore capace di trasformare i segnali luminosi in segnali elettronici, un piccolo videoregistratore e solitamente viene collegata a un monitor più o meno grande su cui si possono osservare le immagini registrate. Una videocamera può avere anche funzionalità tecnologicamente avanzate o di rilievo, come la capacità di operare in maniera wireless, la capacità di acquisire immagini ad alta risoluzione e suoni limpidi, la capacità d'interfacciarsi con smartphone, tablet, computer e cloud, la capacità di acquisire o dedurre la profondità all'interno dell'immagine e così via.



Figura 1.1: Telecamere Intel RealSense [2]

Una videocamera di sorveglianza, o alternativamente telecamera di sorveglianza, ha tutte le caratteristiche di una videocamera. Il fine per il quale viene utilizzata però, è specificamente rivolto alla sicurezza. Anche questa tipologia di videocamere possono avere funzioni più o meno avanzate. Per esempio in commercio sono presenti delle videocamere di sorveglianza che permettono il rilevamento di anomalie, di comportamenti inconsuete, di persone e di oggetti grazie all'utilizzo di tecniche avanzate di Machine Learning e intelligenza artificiale. Inoltre, esistono apposite telecamere che permettono l'acquisizione dell'informazione sulla profondità, mentre altre ancora permettono l'elaborazione d'immagini di diversa natura. Un esempio di telecamere in grado di acquisire e gestire l'informazione sulla profondità della scena inquadrata è dato dalle telecamere Intel RealSense. [2] In figura 1.1 si riporta un'immagine che mostra questa famiglia di telecamere.

Con i progressi tecnologici degli ultimi anni, si è reso possibile utilizzare delle videocamere per la visione artificiale, nella robotica, per svolgere compiti e attività di magazzino, di logistica, in ambito agricolo, medico, industriale, per la collision avoidance, gesture recognition, skeleton tracking, in ambito videoludico e in tanti altri domini applicativi. In figura 1.2 si riportano delle immagini che mostrano alcuni dei domini applicativi citati e la cui fonte è consultabile al riferimento seguente. [4] Con il veloce avanzamento della tecnologia è possibile impiegare telecamere, ormai sempre più avanzate,

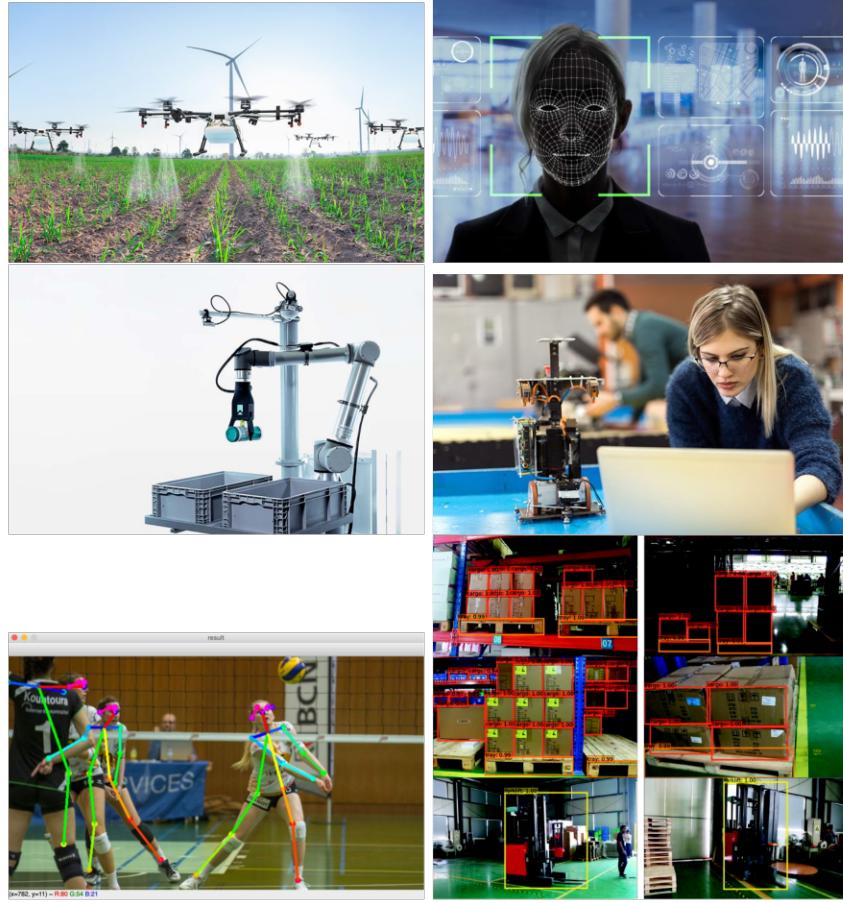


Figura 1.2: Casi d'uso videocamere [4]

anche per il *perceptual computing*. Con questa espressione si fa riferimento alla capacità di un calcolatore di riuscire a percepire e/o analizzare l'ambiente circostante che lo circonda e agire di conseguenza. L'idea del *perceptual computing* è quella di sfruttare voce, gesti e altri input forniti mediante apposita sensoristica per utilizzare il calcolatore, anziché usare mouse e tastiera. Quindi si tratta di un approccio del tutto *touch-less*. Molti se non tutti questi domini applicativi necessitano dell'informazione sulla profondità degli oggetti all'interno della scena osservata. Come si vedrà in seguito, esistono oggi diversi approcci per ricavare quest'informazione. Uno dei più importanti è quello che fa uso di telecamere con capacità di rilevamento della profondità.

1.2 Concetti preliminari

Si consideri il video in quanto elemento multimediale. Un video acquisito da una videocamera può essere visto in prima approssimazione e in termini di rappresentazione digitale come un insieme d'immagini e suoni. Tralasciando l'aspetto sonoro, un video può essere più o meno fluido a seconda di quanti frame per secondo, e quindi immagini, vengono acquisiti dal dispositivo incaricato.

Consideriamo dapprima un'immagine in scala di grigi. Una singola immagine di questo tipo, può essere definita come una funzione bidimensionale, $f(x, y)$, dove x e y sono le coordinate spaziali (sul piano), e l'ampiezza di f in ogni coppia di coordinate (x, y) viene chiamata *intensità* o livello di grigio dell'immagine in quel determinato punto. Quando x , y e i valori dell'ampiezza assunti da f sono tutti finiti, dunque quantità discrete, possiamo definire l'immagine come un'immagine digitale. Ogni elemento di un'immagine è detto *picture element* o *pixel* ed è caratterizzato dall'avere una posizione e un valore all'interno dell'immagine. In figura 1.3 si schematizza quanto detto poc'anzi.

$$f(x, y) = \begin{bmatrix} f(0,0) & f(0,1) & \cdots & f(0, N - 1) \\ f(1,0) & f(1,1) & \cdots & f(1, N - 1) \\ \vdots & \vdots & & \vdots \\ f(M - 1,0) & f(M - 1,1) & \cdots & f(M - 1, N - 1) \end{bmatrix}$$

Figura 1.3: Rappresentazione immagine digitale

Esistono diversi dispositivi d'imaging in grado di trattare e acquisire immagini a partire da componenti differenti dello spettro elettromagnetico e non. Questa trattazione si soffermerà alle immagini acquisite a partire dalla luce visibile dello spettro elettromagnetico. In figura 1.4 si riporta l'intero spettro elettromagnetico, le lunghezze d'onda espresse in metri e la locazione della banda della luce visibile con le lunghezze d'onda espresse in nanometri.

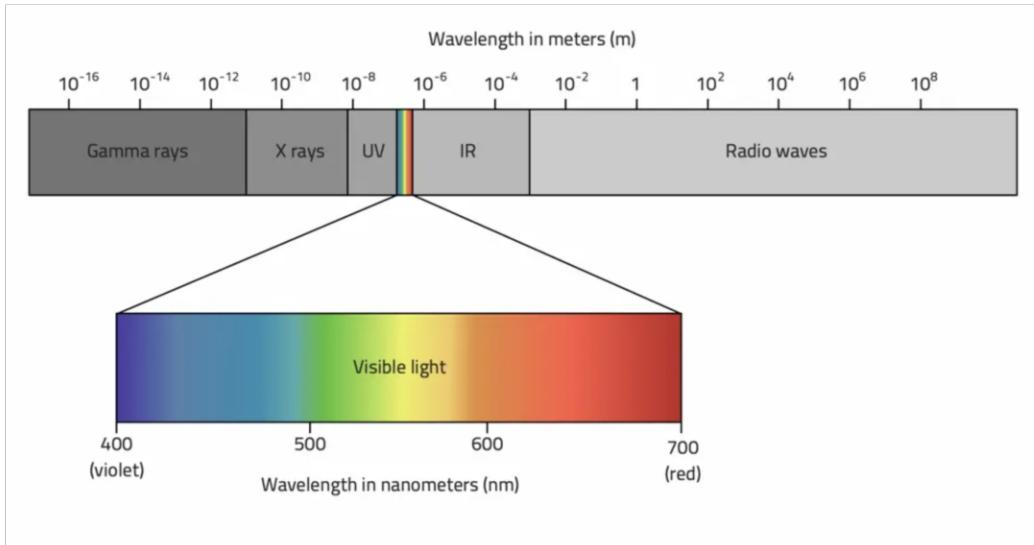


Figura 1.4: Spettro elettromagnetico [6]

Durante questo lavoro di tesi sono state adoperate tecniche dell'elaborazione delle immagini sia di basso, sia di medio, che di alto livello; integrando così sia operazioni di base, che avanzate nell'elaborazione e nel trattamento delle immagini. Al fine di alleggerire il documento, si tralasciano i dettagli relativi alla struttura e al funzionamento di un dispositivo d'imaging, all'acquisizione e alla formazione di un'immagine digitale, alla risoluzione spaziale e d'intensità.

Nell'elaborazione dell'immagini non ci si limita soltanto all'uso di determinati livelli di grigio. L'uso del colore nell'elaborazione delle immagini è motivato da due fattori principali. Per prima cosa, il colore è un descrittore che semplifica l'identificazione di un oggetto e la sua estrazione da una scena. In secondo luogo, gli uomini sono in grado di distinguere migliaia di gradazioni di colore e d'intensità, in confronto a solo due dozzine di tonalità di grigio. Ciò è di particolare importanza nell'analisi manuale dell'immagine. Per rappresentare le immagini a colori si sfrutta un modello colore (o spazio colore). Il suo scopo è quello di facilitare e standardizzare la specifica dei colori. In sostanza, un modello colore è un sistema di coordinate e di un sottospazio all'interno di quel sistema dove ogni colore viene rappresentato da un singolo punto. I modelli colore maggiormente noti e utilizzati sono RGB, CMY, CMYK e HSI. Il modello colore usato durante questo lavoro di tesi è quello RGB, dove il modello si basa su un sistema di coordinate cartesiane, mentre il sottospazio d'interesse è un cubo unitario. Il modello RGB viene

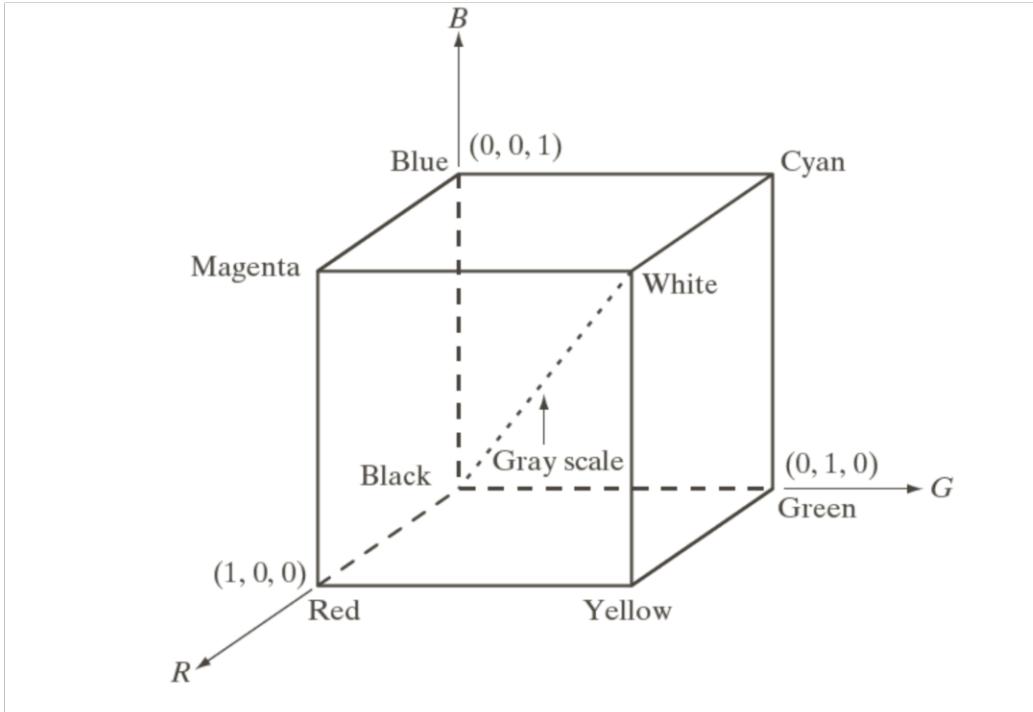


Figura 1.5: Spazio colore RGB

mostrato in figura 1.5

Le immagini rappresentate nel modello RGB sono formate da tre immagini, una per ogni colore primario; come mostrato in figura 1.6. Quando vengono visualizzate in un monitor RGB, queste tre immagini si combinano per produrre un'immagine a colori composta. Dal punto di vista dell'accesso in memoria, è possibile sia accedere a ognuna delle tre immagini, che all'immagine a colori nel suo complesso. Invece, dal punto di vista dell'elaborazione delle immagini e dei risultati ottenuti, in alcuni casi l'output di operazioni ripartite sui singoli canali differiscono dall'output dell'operazione effettuata sull'immagine intera. [7]

Un'immagine RGBD è semplicemente un'immagine RGB combinata con una corrispondente immagine di profondità. Un'immagine di profondità è un'immagine a un canale singolo dove per ogni pixel si conserva l'informazione sulla distanza tra il piano immagine e il corrispondente oggetto nella realtà. Uno dei dispositivi storici utilizzati per catturare immagini RGBD è il Kinect [14]. Al fine di ottenere l'informazione sulla profondità in maniera accurata e senza fare ricorso a tecniche di Machine Learning e intelligen-

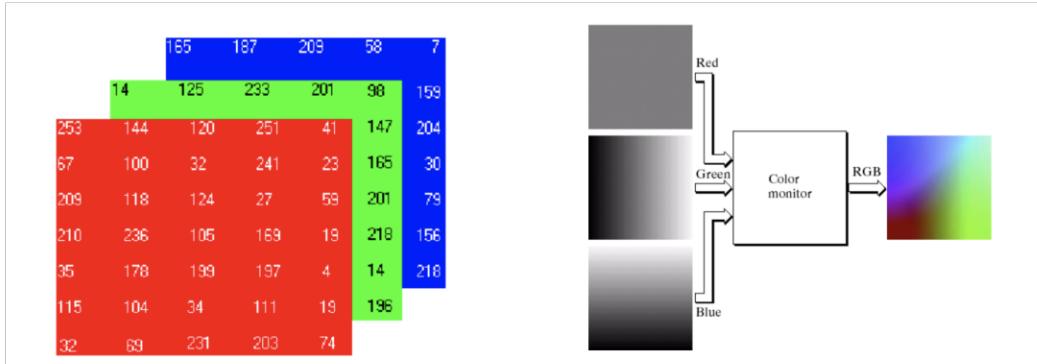


Figura 1.6: Immagine RGB

za artificiale, sono necessarie almeno due videocamere. A partire da queste videocamere, è possibile fare ricorso a nozioni di geometria epipolare e trigonometria in modo da ricavare la profondità per ogni pixel dell’immagine acquisita. In figura 1.7, è mostrata un’immagine RGBD acquisita durante alcuni test condotti presso il laboratorio “*Alfredo Petrosino - CVPR Lab*”.

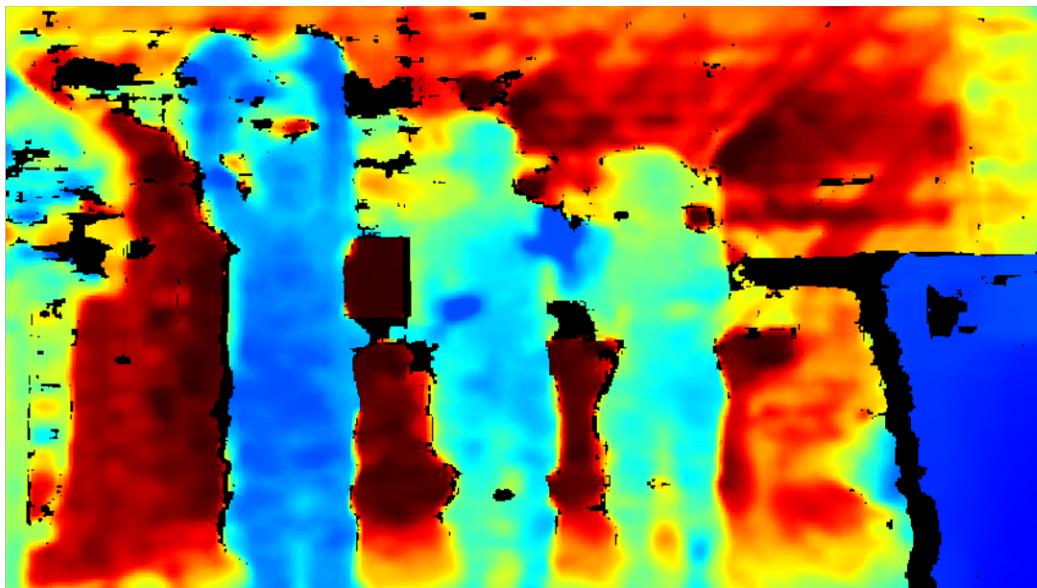


Figura 1.7: Immagine RGBD

In un’immagine contenente informazioni sulla profondità degli elementi inquadrati nella scena, la visualizzazione è semplificata dal particolare uso del colore. Infatti nel caso delle telecamere Intel RealSense D435, come si vede nella figura 1.7, si usa la cosiddetta mappa di profondità tale per cui

colori caldi (e.g. giallo, arancione, rosso) indicano elementi lontani; viceversa colori freddi (e.g. azzurro, blu, viola) indicano elementi vicini. Elementi contraddistinti dal colore verde sono invece a una distanza media, dove il concetto di medio è da intendersi rispetto alla calibrazione iniziale effettuata dalla telecamera. Infine, laddove sono visibili zone di colore nero, queste intendono indicare zone ove non è stato possibile valutare la distanza o zone di colore nero ai fini interpretativi della profondità (tipicamente in presenza dei bordi di persone e oggetti). Si noti come la persona più a sinistra sia caratterizzata da un colore più freddo, mentre quella più a destra sia caratterizzata dall'avere un colore più caldo. Questo indica semplicemente che la prima è più vicina alla telecamera, a differenza della seconda che è più lontana.

1.3 Obiettivo

Si descrive ora l'idea del lavoro di tesi in maniera sintetica ma completa, per poi approfondirne maggiormente i dettagli nei capitoli che seguono. Il progetto completo si è svolto in modalità ibrida: da casa e presso il laboratorio “*Alfredo Petrosino - CVPR Lab*”. Il laboratorio è visibile nel video [5]. Il progetto è articolato in tre moduli operativi e il candidato si è occupato del primo dei tre moduli poc’anzi menzionati.

L’idea del primo modulo è la seguente. Data una videocamera RGBD Intel RealSense D435 [2] in grado di acquisire informazioni sia sul colore, che sulla profondità, si estrae e si scinde l’informazione relativa al colore da quella relativa alla profondità. Una volta fatto ciò, si sfruttano tecniche avanzate di Machine Learning e intelligenza artificiale, in modo da individuare e ricavare i cosiddetti “skeleton”, ovvero gli scheletri delle persone individuate all’interno dell’immagine. Ogni scheletro è composto banalmente da punti e linee. I punti fanno riferimento a punti di giuntura fondamentali, mentre i segmenti si occupano di unire opportunamente tali punti di giuntura in modo da formare lo skeleton finale. Esempi di punti di giuntura di uno skeleton sono l’orecchio destro, l’orecchio sinistro, il naso, l’occhio destro, l’occhio sinistro e così via. In figura 1.8 è mostrato uno skeleton con una possibile annotazione per identificare i diversi punti di giuntura.

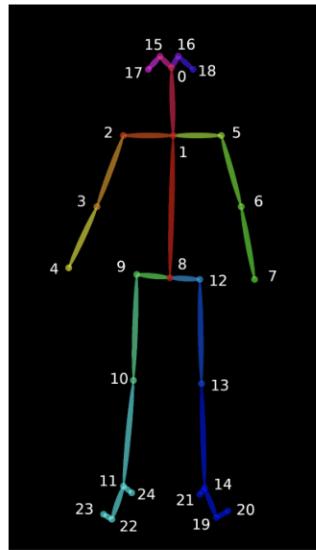


Figura 1.8: Skeleton [34]

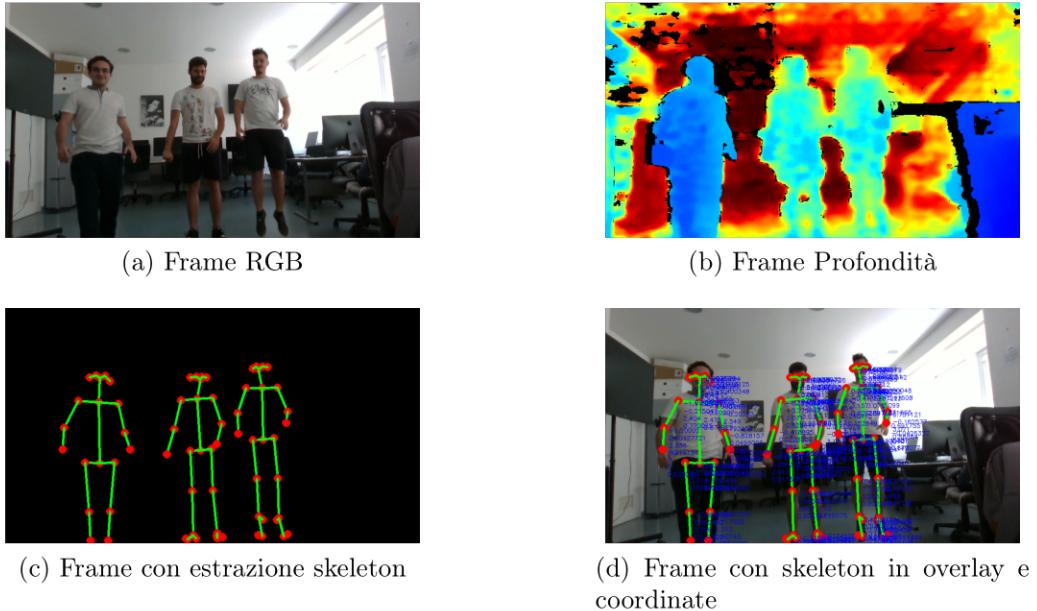


Figura 1.9: Acquisizione e scissione frame video, rilevazione ed estrazione skeleton, conversione e rendering delle coordinate

Una volta ottenuti gli skeleton di tutte le persone presenti all'interno dell'immagine, si passa alla trasformazione del sistema di coordinate. In particolare si passa dal sistema di coordinate usato dalla videocamera RGBD a un sistema di coordinate opportunamente fissato e basato su parametri di configurazione della scena inquadrata. A questo punto si hanno a disposizione le coordinate di tutti i punti di giuntura di ogni skeleton individuato espresse in un formato adottato per convenzione anche dai successivi moduli realizzati e che indicano per ogni punto di giuntura la sua posizione all'interno della stanza ove è stata installata la telecamera. Tutte le informazioni finora ottenute, vengono inviate al modulo successivo che si occuperà di eseguire un controllo sulla presenza di eventuali anomalie nei movimenti eseguiti da uno o più skeleton. Per fare ciò, si fa uso ancora una volta di tecniche di Machine Learning e intelligenza artificiale. [8]

Una volta fatto ciò, il secondo modulo si occupa d'inviare tutte le informazioni finora prodotte e le informazioni relative alle anomalie al terzo modulo, che si occuperà di effettuare il rendering digitale (seguendo il modello dei digital twin) in ambiente Unity [10]. Inoltre, in tale ambiente, si mostra a video in maniera dettagliata se un determinato skeleton sta compiendo un'azione anomala o meno. [9]

In figura 1.9a e 1.9b si mostra come viene effettuata la scissione del frame RGB da quello relativo alla profondità successivamente alla fase di acquisizione da parte della telecamera. In figura 1.9c si mostrano gli skeleton estratti dopo l'opportuna fase di rilevamento degli stessi, mentre in figura 1.9d si ricostruisce lo skeleton appena estratto sull'immagine 1.9a e si aggiunge l'informazione per ogni punto di giuntura sulle coordinate spaziali facenti riferimento alla stanza ove installata la telecamera.

L'integrazione dei tre moduli appena menzionati permette di mettere in funzione un sistema di tracciamento delle anomalie efficiente e multi-persona con conseguente rendering video in ambiente Unity. Nell'attuale documento si discuterà in maniera approfondita del primo modulo, in quanto è quello relativo al lavoro di tesi del candidato. Si precisa che nel documento si farà riferimento a questo modulo come "AI Watch A1", o più semplicemente modulo "A1". I dettagli relativi ai successivi due moduli sono rintracciabili nei documenti citati nei riferimenti. [8] [9] Invece, al riferimento che segue, è possibile visionare un video dimostrativo che mostra in esecuzione il primo modulo del sistema nelle sue fasi primordiali. [5]

1.4 Organizzazione della tesi

Di seguito si riporta una panoramica della tesi, con una breve descrizione per ogni capitolo. Il capitolo 1 riassume il contesto e l'obiettivo finale delle tesi. Inoltre, fornisce alcune nozioni di base per gli aspetti teorici affrontati nei successivi capitoli. Il capitolo 2 introduce il problema della stima della profondità, il problema del rilevamento di uno skeleton, il problema della conversione fra sistemi di coordinate e gli approcci utilizzati per affrontare tutti questi problemi.

Il capitolo 3 contiene tutte le scelte eseguite in fase di progettazione al fine di realizzare il modulo A1: l'approccio usato per ricavare le informazioni sulla profondità, quello usato per estrarre gli skeleton, quello usato per la conversione fra sistemi di coordinate, per l'invio delle informazioni prodotte e altri dettagli tecnici.

Il capitolo 4 indica le modalità d'implementazione del modulo A1. Il capitolo 5 mostra i risultati sperimentali ottenuti dall'esecuzione del modulo A1 e dell'esecuzione di tutti i moduli in fase d'integrazione del sistema. Infine, il capitolo 6 riassume brevemente la tesi, trae delle deduzioni sul lavoro svolto e si conclude proponendo sviluppi e miglioramenti futuri.

Capitolo 2

Stato dell'arte

Questo capitolo presenterà attività simili di ricerca e sviluppo condotte da altri e/o presenti sul mercato, dopo aver introdotto gli aspetti teorici fondamentali propri di questo dominio applicativo.

2.1 Aspetti teorici

2.1.1 Profondità e videocamere RGBD

Le videocamere capaci di determinare la profondità degli oggetti e delle persone presenti all'interno della scena inquadrata hanno avuto negli ultimi tempi sempre più importanza in quanto forniscono la possibilità ai dispositivi di vedere, capire, interagire e imparare dall'ambiente osservato. L'output prodotto da questa tipologia di telecamere è dato da immagini a colori in cui elementi, e più in generale regioni, con una stessa profondità all'interno dell'immagine hanno un colore simile, indipendentemente dal colore che possiedono all'interno della scena. È possibile migliorare la resa delle immagini catturate andando a calibrare la telecamera all'avvio rispetto alla scena inquadrata e impostando determinati parametri in maniera sperimentale in modo da avere dei miglioramenti sui singoli frame (e.g. riempimento buchi, riduzione rumore, potenza emissione laser, e così via). Inoltre, è possibile applicare tecniche di post-processing delle singole immagini acquisite, così da migliorare ulteriormente il risultato finale.

È utile fare una prima distinzione tra il concetto di “depth”, rispetto a quello di “range”. Il primo vocabolo fa riferimento alla distanza memorizzata all'interno dell'immagine, mentre il secondo fa riferimento alla distanza effettiva nella scena. Questi due valori possono differire a seconda di vari fattori:

telecamera utilizzata, versione firmware telecamera, stabilità installazione supporto fisico, luci, ombre e altre specifiche della scena inquadrata.

La riflettanza misura, in ottica, la capacità di riflettere parte della luce incidente su una data superficie o materiale. Essendo quindi il rapporto tra intensità del flusso radiante riflesso e intensità del flusso radiante incidente, è una grandezza adimensionale. Un oggetto che ha una bassa riflettanza produce immagini più rumorose e con maggiore incertezza sul valore della distanza dalla telecamera, anche detto “depth confidence”. Viceversa oggetti con una riflettanza più elevata, permettono di avere meno rumore, qualità più elevata e informazioni sulla distanza più accurate. Questo può essere un problema, in quanto spesso si sfruttano tecnologie laser o infrarossi al fine di dedurre la distanza. Una delle cause delle zone nere visibili nei frame contenenti informazioni sulla distanza, è proprio quella citata poc’anzi. Inoltre, a parità di distanza dalla telecamera che inquadra la scena, un oggetto con maggiore riflettanza sarà maggiormente definito e limpido nell’immagine risultante, a differenza di un oggetto con minore riflettanza. Di conseguenza all’aumentare della distanza dalla telecamera, gli oggetti con riflettanza maggiore riescono a conservare meglio il grado di chiarezza col quale verranno rappresentati nelle immagini risultanti. Chiaramente, quando si parla di questi aspetti di basso livello, si fa riferimento a immagini reali e non a immagini sintetiche prodotte al computer.

L’informazione sulla profondità all’interno di un’immagine può tornare utile anche ai fini della segmentazione, infatti questo tipo di operazione è di gran lunga facilitato avendo a disposizione anche quest’ulteriore informazione. In linea generale alcune problematiche dell’elaborazione delle immagini sono più semplici se affrontate in RGBD, anziché in RGB; oppure è possibile affrontarle in maniera più efficace. Per esempio applicare il Deep Learning su immagini RGBD tendenzialmente porta a risultati migliori rispetto a usare semplici immagini RGB, perché la rete neurale non deve sforzarsi più di tanto per segmentare i singoli oggetti: basta trovare i pixel che fra di loro sono più vicini in termini di distanza e considerarli come appartenenti allo stesso oggetto.

Un aspetto importante da precisare è che le immagini catturate con una telecamera di profondità non saranno mai perfette, nemmeno con la visione stereoscopica di una potenziale telecamera. La motivazione è che non è possibile vedere “tutto” con i soli due “occhi” della telecamera di profondità. Per esempio le regioni della scena ai lati non saranno visibili a seconda dell’ampiezza del campo visivo della telecamera. L’errore che si commette è lo stesso

errore di quando si pone un dito davanti agli occhi e si cerca di vedere oltre. L'effetto prodotto di norma sarebbe vedere due dita, anziché una, ma la visione umana aiutata dal cervello unisce le due viste e ne vede uno soltanto. Una telecamera digitale, benché si possano usare algoritmi molto avanzati, non è in grado di farlo completamente e in ogni scenario. Di conseguenza, in alcuni casi si generano degli artefatti.

Inoltre, è possibile ricavare l'informazione sulla profondità per ogni pixel solitamente grazie alla tecnologia laser o infrarossi. In entrambi casi in caso di luce diretta del Sole, tali tecnologie potrebbero essere leggermente inibite.

2.1.2 Cenni di geometria epipolare

La geometria epipolare è la geometria che lega due immagini acquisite da due punti di vista differenti. Le relazioni che intercorrono tra le immagini tuttavia non dipendono dalla scena osservata, ma dipendono solamente dai parametri intrinseci delle camere e dalle posizioni relative. Per ogni punto osservato, il piano epipolare è il piano formato dal punto in coordinate del mondo e dai due centri ottici. La geometria epipolare è quindi la geometria che descrive la visione stereoscopica (o binoculare). Descrive le relazione e i vincoli geometrici che legano due immagini bidimensionali della stessa scena tridimensionale, catturata da due fotocamere con posizione e orientamento distinto.

L'essere umano riesce a percepire la distanza grazie al fatto che ha a disposizione due occhi. Con un solo occhio possiamo ricavare qualche indizio, capire approssimativamente la distanza in base all'esperienza umana; ma la vera visione tridimensionale e la percezione della profondità avviene con entrambi gli occhi. Il cervello sa dove sono fisicamente situati gli occhi sul viso e quindi riesce a elaborare e capire in qualche modo come interpretare la scena osservata dagli occhi. Invece, nel caso di un dispositivo elettronico stereoscopico, questi deve essere messo al corrente di dove sono situate le due telecamere. Si precisa che si parlerà di due telecamere, ma in realtà si fa riferimento a una sola telecamera con due sensori fotosensibili separati. I concetti sono applicabili anche nel caso di telecamere del tutto distinte con le opportune osservazioni.

Supponiamo di avere due telecamere situate in un cubo unitario, rivolte verso una delle pareti con una determinata angolazione e posizionate a una certa altezza. La prima telecamera acquisisce un'immagine idealmente nello stesso istante in cui la seconda telecamera acquisisce un'altra immagine. Questo scenario è mostrato in figura 2.1. A questo punto è necessario trovare i punti corrispondenti tra le due immagini. Qualora non conoscessimo dove sono situate le telecamere all'interno del cubo, la complessità computazionale e lo spazio di ricerca in generale aumenterebbero vertiginosamente (e.g. angoli che appaiono più volte, punti confusi, sovrapposti e così via). Per esempio un angolo appare più volte quando si inquadra un libro.

Un'ulteriore problematica è che ci saranno sezioni della scena visibili a una telecamera, ma non all'altra, magari sezioni appartenenti a uno stesso oggetto. Quello che si fa inizialmente in linea generale è calibrare le due telecamere. Questo permette, date le due telecamere supposte poc'anzi, di

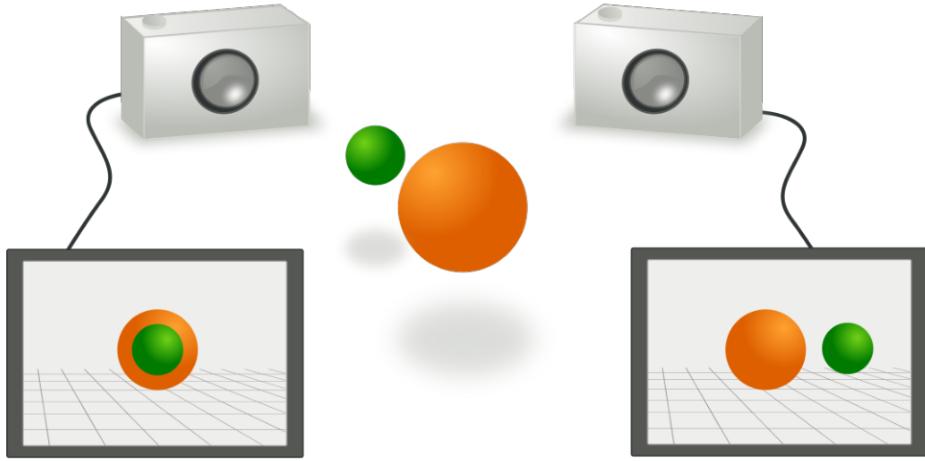


Figura 2.1: Due fotocamere che acquisiscono un’immagine idealmente nello stesso istante della stessa scena [11]

capire le loro angolazioni e posizioni nello spazio. Per fare ciò si acquisisce un’immagine da entrambe idealmente nello stesso istante, sebbene il concetto di contemporaneità esatta non sempre è possibile. In linea generale, è necessario che il tempo che intercorre tra le due acquisizioni sia il più contenuto possibile, altrimenti la scena cambierebbe e non sarebbe possibile procedere con i passaggi successivi. A questo punto si hanno a disposizione due immagini che sono dette “left view” e “right view”, rispettivamente l’immagine acquisita dalla telecamera sinistra (che chiameremo telecamera A) e quella acquisita dalla telecamera destra (che chiameremo telecamera B).

Come determinare la distanza per ogni pixel di un’immagine? Si consideri la figura 2.2: l’oggetto che immaginiamo di fotografare è quello che viene rappresentato dal punto X . Supponiamo di avere a disposizione due fotocamere pin-hole A e B . Asserire ciò, presuppone che tutti i raggi ottici che concorrono alla formazione dell’immagine siano “transitati”, a un certo tempo del loro cammino, per un unico punto, detto centro di proiezione. Il termine pin-hole nasce proprio dal fatto che si suppone che la luce che impressiona la lastra o i sensori di una fotocamera digitale, attraversi un foro di piccole dimensioni, tanto piccole da potersi considerare un punto nell’accezione geometrica del termine.

Le due fotocamere A e B , sono centrate rispettivamente nei punti O_L e O_R . Il punto X in questione verrà rispettivamente proiettato sul piano

immagine della fotocamera sinistra in x_L e in x_R nel piano immagine della fotocamera destra. Si denoti con a il segmento che congiunge il punto O_L con X e con b il segmento che congiunge il punto O_R con X . Il valore della distanza fornito dalla fotocamera A , potrebbe essere un valore qualsiasi lungo il segmento a od oltre. Analogamente per la fotocamera B con il segmento b . Si tratta in un problema non da poco. Supponiamo quindi di essere a conoscenza del fatto che un pixel nell'immagine acquisita da A è equivalente a un pixel nell'immagine acquisita da B . In tal caso, effettuando delle proiezioni e con l'aiuto della trigonometria si ricava la profondità del punto nella scena. Purtroppo, non si ha l'informazione relativa al pixel di equivalenza tra le due immagini acquisite da A e B . Infatti, questo pixel può differire nel tempo, può non essere visibile a entrambe le fotocamere e così via. Per risolvere questo problema piuttosto complesso e trovare i punti di corrispondenza tra le immagini acquisite, estendere il discorso a ogni pixel e ricavare la profondità, si sfrutta la geometria epipolare.

Si consideri il punto O_L di A e il punto X nella scena inquadrata. Per rendere più semplice la ricerca nell'immagine di B e al fine di dedurre la profondità a cui si trova X , si prendono in considerazione questi due punti poc'anzi menzionati e il punto O_R di B , come appartenenti a un unico grande triangolo. Si immagini che dalla fotocamera B escano tanti segmenti che vanno a intersecarsi con il segmento a generando una serie di punti, che nell'immagine acquisita da B rivelano essere presenti su un'unica linea. Questa retta è detta retta epipolare. Modellando il problema in questo modo, la difficoltà di risoluzione cala vertiginosamente. Infatti, sapendo dove sono situate le fotocamere e con l'informazione relativa alla retta epipolare, si deduce che la corrispondenza tra x_L e x_R , è possibile trovarla solo e unicamente sulla retta epipolare poc'anzi descritta. C'è quindi un insieme limitato di pixel da dover controllare e fra questi determinare quello più simile a x_L , triangolarizzare come nell'approccio ideale e determinare di conseguenza l'informazione sulla profondità dell'oggetto nella scena. Questo tipo di approccio è possibile solo perché si è a conoscenza di dove sono situate le fotocamere; altrimenti, se così non fosse, sarebbe necessario cercare nell'intera immagine e ciò sarebbe costosissimo. Questo problema è definito come *problema di corrispondenza*.

Siccome nel caso di una telecamera Intel RealSense, di cui si discuterà nella sottosezione 2.2.1, le due telecamere sono allineate e ben calibrate all'avvio, il lavoro di triangolarizzazione per ricavare la profondità è semplificato. Inoltre, si precisa che durante il procedimento di triangolarizzazione, avvengono anche alcuni passaggi di smoothing per eliminare il rumore e altri

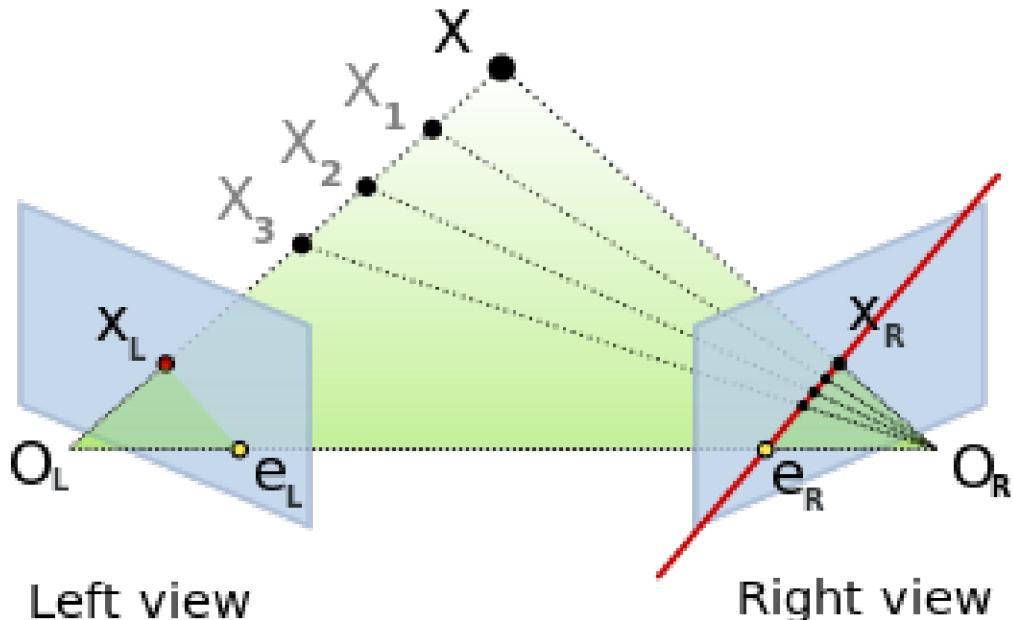


Figura 2.2: Determinazione della profondità per ogni pixel dell’immagine acquisita [11]

dettagli meno rilevanti. In molti casi vengono applicate anche tecniche di post-processing per migliorare la qualità dell’immagine finale.

Quando si hanno due o più telecamere gli approcci maggiormente noti sono due: quello della geometria epipolare appena descritto, che permette una precisione e un’efficienza maggiore; oppure l’approccio che fa uso della minimizzazione delle somme delle distanze al quadrato (anche noto come SSD). Quest’ultimo approccio però risulta essere abbastanza costoso e poco preciso. Di conseguenza non verrà considerato in questa trattazione. Infatti, l’aspetto fondamentale da tenere a mente quando è necessario lavorare con due telecamere è quello di considerare le disparità tra i pixel delle due immagini acquisite e abbinarli in modo che corrispondano alla stessa regione, oggetto, persona. Tra tutte le possibili scelte, va fatta ovviamente quella che permette di avere le disparità più piccole tra tutti i pixel dell’immagine. Solo agendo in questo modo si riesce a ottenere un’immagine di profondità qualitativamente soddisfacente.

2.1.3 Intelligenza artificiale

L’Intelligenza Artificiale si propone di sviluppare delle macchine dotate di capacità autonome di apprendimento e adattamento che siano ispirate ai modelli di apprendimento umani. Diversi anni fa ci furono delle scoperte che, nonostante un periodo di scoraggiamento, hanno rivoluzionato il mondo della tecnologia. Da sempre l'uomo ha cercato di costruire macchine pensanti, prendendo spunto dal funzionamento del cervello umano. Le unità responsabili del passaggio d’informazioni nel cervello sono i neuroni e in uno studio del 1943, Warren Sturgis McCulloch e Walter Pitts, introdussero un neurone artificiale [17]. Tale studio schematizzò un combinatore lineare a soglia con dati binari in entrata e un singolo dato binario in uscita. Le prime ipotesi di apprendimento furono introdotte dallo psicologo canadese Donald Olding Hebb, che propose un modello di neurone basato sul funzionamento complesso dei neuroni del cervello umano [18]. Il primo schema di rete neurale per il riconoscimento e classificazione di forme, antesignano delle attuali reti neurali, è da attribuire a Frank Rosenblatt [19]. Esso fornì un’interpretazione dell’organizzazione generale dei sistemi biologici ed era in grado di apprendere. Inoltre, utilizzava funzioni booleane linearmente separabili. L’interesse e l’euforia di tali studi, tuttavia, nel 1969 furono notevolmente ridimensionate da Marvin Minsky e Seymour A. Papert, i quali mostrarono i limiti operativi dell’apprendimento delle semplici reti a due strati basate sul percettrone [20].

A causa di queste limitazioni, seguì un periodo di diffidenza durante il quale tutte le ricerche in questo campo, non ricevettero più alcun finanziamento dai governi. Ciò portò a far ristagnare la ricerca per oltre un decennio. Nonostante molti anni prima del 1982, il matematico Paul Werbos nella sua tesi dimostrò come addestrare le reti MLP (Multi-Layers Perceptron), solo l’intervento di John Hopfield, nel suo studio sui modelli di riconoscimento di pattern molto generali, riaprì degli spiragli per la ricerca nel campo dell’intelligenza artificiale. Pochi anni più tardi, David E. Rumelhart, Geoffrey Hinton e Ronald J. Williams introdussero uno dei metodi più noti ed efficaci per l’addestramento delle reti neurali: il cosiddetto algoritmo di retropropagazione dell’errore (error backpropagation). Esso modifica sistematicamente i pesi delle connessioni tra i nodi, così che la risposta della rete si avvicini sempre di più a quella desiderata. Tale lavoro fu prodotto riprendendo il modello creato da Werbos [21]. L’addestramento di une rete neurale avviene in due diversi fasi: forward-pass e backward-pass. Nella prima fase, i vettori in input sono elaborati dai nodi in ingresso con una propagazione in avanti dei segnali attraverso ciascun livello della rete. Durante questa fase i valori

dei pesi della rete sono tutti fissati. Nella seconda fase la risposta della rete viene confrontata con l'output desiderato ottenendo l'errore di classificazione. Questo errore è propagato nella direzione inversa rispetto a quella del passo forward-pass. Infine, i pesi vengono modificati in modo da minimizzare la differenza tra l'uscita attuale e l'uscita desiderata. Attraverso questo algoritmo si consente di superare le limitazioni del modello introdotto da McCulloch e Pitts. Infatti, viene risolto il problema della separabilità non lineare.

2.1.4 Il percepitrone

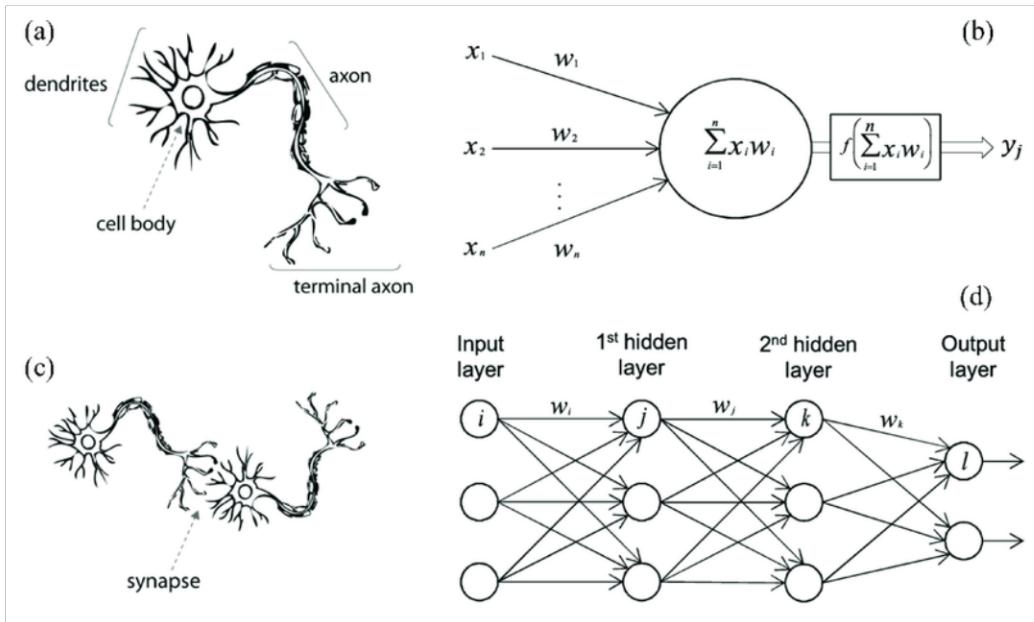


Figura 2.3: Un neurone biologico confrontato con una rete neurale artificiale: (a) neurone umano; (b) neurone artificiale; (c) sinapsi biologica; (d) sinapsi ANN [24]

I neuroni artificiali sono le unità di base delle reti neurali artificiali. La loro struttura cerca di somigliare quanto più possibile ai neuroni biologici. Questi neuroni artificiali ricevono in input degli stimoli e durante l'elaborazione vengono moltiplicati per un opportuno valore detto *peso*. Il risultato delle moltiplicazioni viene sommato e se la somma supera una certa soglia, il neurone si attiva restituendo un output. Questo peso serve a quantificare l'importanza di uno stimolo rispetto agli altri. Per esempio, se più neu-

roni comunicano fra loro e solo alcune connessioni vengono maggiormente utilizzate, allora tali connessioni avranno un peso maggiore.

$$f(x) = k \left(\sum_{i=1}^m w_i x_i \right) \quad (2.1)$$

L'equazione 2.1 si riferisce al neurone introdotto da McCulloch e Pitts e si può osservare la funzione di attivazione k , la quale permette di eccitare o inibire l'informazione in transito all'altro neurone [17]. Le quattro funzioni di attivazione più utilizzate sono le seguenti. Si precisa che w_i rappresenta il peso i -esimo, mentre x_i l'input i -esimo.

- La funzione soglia.

$$Y = \begin{cases} 0, & \text{se } \sum_{i=1}^m w_i x_i < 0 \\ 1, & \text{se } \sum_{i=1}^m w_i x_i \geq 0 \end{cases} \quad (2.2)$$

- La funzione sigmoide.

$$Y = \frac{1}{1 + e^{(-\sum_{i=1}^m w_i x_i)}} \quad (2.3)$$

- La funzione ReLU.

$$Y = \max \left\{ 0, \sum_{i=1}^m w_i x_i \right\} \quad (2.4)$$

- La funzione Tangente Iperbolica.

$$Y = \frac{1 - e^{-2 \sum_{i=1}^m w_i x_i}}{1 + e^{-2 \sum_{i=1}^m w_i x_i}} \quad (2.5)$$

Se si considerano più livelli di neuroni, si crea la rete multilayer perceptron (MLP) [21]. La retropropagazione dell'errore, in inglese *back-propagation of error*, ma solitamente abbreviato in backpropagation, è un algoritmo per l'addestramento delle reti neurali artificiali, usato in combinazione con un metodo di ottimizzazione, per esempio il gradiente discendente. Gli output desiderati per ogni input della rete, sono contenuti solitamente in una *ground truth* e in base agli output ottenuti dalla rete, si calcola il gradiente. Questa operazione viene maggiormente utilizzata nei metodi di apprendimento supervisionato, sebbene venga anche utilizzata in metodi non supervisionati.

Inoltre, la retropropagazione richiede che la funzione d'attivazione usata dai neuroni artificiali sia differenziabile. Una delle principali difficoltà quando si utilizza il gradiente, è il problema noto come *scomparsa del gradiente*. Esso è dovuto alla presenza di funzioni di attivazione non lineari che causano una diminuzione esponenziale del valore del gradiente all'aumentare della profondità della rete neurale. Il funzionamento del gradiente discendente è molto semplice: esso permette di trovare un minimo locale di una funzione in uno spazio a N dimensioni. Ottimizzando i parametri del modello, esso minimizza la funzione *costo* e aumenta l'accuratezza del modello apprendendo le combinazioni non lineari delle caratteristiche in input.

2.1.5 Artificial Neural Network

Una rete neurale artificiale imita la struttura del cervello umano basandosi su neuroni artificiali disposti a strati e collegati tra di loro mediante l'invio di segnali. Dato un neurone artificiale dello strato i -esimo, vanno definiti un peso, una soglia, un certo numero d'input e un certo numero di output. L'approccio classico è quello di una rete *feed forward*, in cui l'output dei neuroni di uno strato della rete costituisce l'input dei neuroni dello strato successivo. Le ANN sono caratterizzate dall'avere un'elevata accuratezza, migliorano col tempo e sono ideali quando il fattore tempo è fondamentale nel dominio applicativo considerato.

Nel caso classico i neuroni di una ANN sono i percetroni, mentre in altri contesti si possono adottare i neuroni sigmoidi, ovvero dei neuroni che hanno valori compresi fra zero ed uno, anziché assumere soltanto valore zero oppure uno. Tale tipo di neuroni sono utili quando il problema neurale è non lineare. La funzione di accuratezza misura quanto si adatta (o generalizza) il modello scelto. La convergenza della funzione di accuratezza è raggiunta quando si raggiunge un punto di minimo, generalmente mediante il metodo del gradiente discendente. Un'ANN fa ampio uso della back-propagation analizzata alla sottosezione precedente.

2.1.6 Convolutional Neural Network

Le reti neurali convoluzionali (o convolutive) dette CNN (Convolutional Neural Network), sono identificate dai loro loop di feedback. Ci si avvale di questo approccio di apprendimento soprattutto quando si utilizzano dati di serie temporali per fare delle previsioni su risultati futuri (e.g. mercato azionario o previsioni prezzo di vendita). Alla base di una CNN vi è il concetto di convoluzione tra funzioni. Si può immaginare una CNN di base come formata

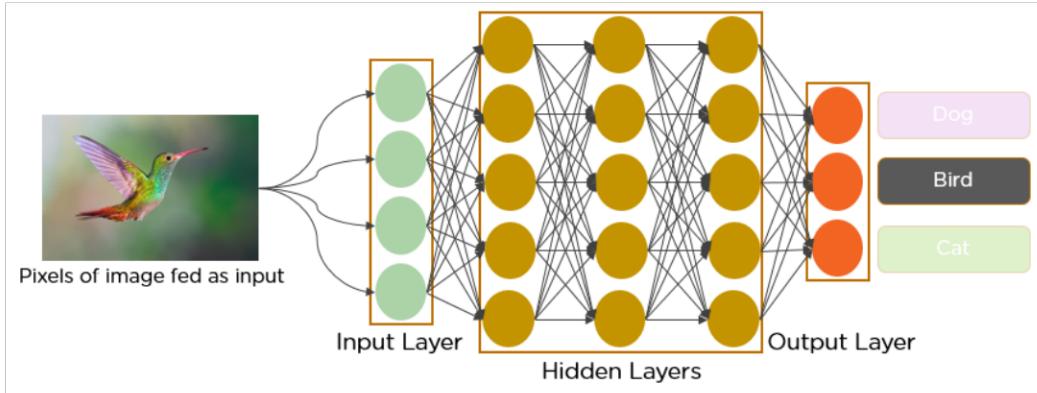


Figura 2.4: Rete neurale convoluzionale [24]

da un livello convoluzionale, un livello di pooling e un livello completamente connesso. In una CNN si fa ampio uso di conversioni e filtri per analizzare gli input forniti di volta in volta.

Il campo che ha reso importante il deep learning è quello della computer vision. Tale successo è dovuto all'introduzione, da parte di Lecun nel 1998, della rete neurale convoluzionale [23]. Questo studio ha permesso di ottenere risultati ottimali, perché consente alle macchine di vedere o percepire il mondo come fanno gli esseri umani in una moltitudine di attività. Nella figura 2.4 viene mostrato un esempio dell'architettura di una rete neurale convoluzionale. L'input della CNN è un'immagine da cui si apprendono i pesi e i bias (o threshold), differenziando vari aspetti degli oggetti contenuti in essa, al fine ultimo di classificare l'intera immagine. Il preprocessing richiesto in un CNN è molto inferiore rispetto a quello che viene eseguito negli algoritmi di machine learning. Infatti, in questi algoritmi i filtri sulle immagini sono progettati a mano per il caso specifico, mentre con un buon addestramento la rete in esame riesce ad apprendere filtri in grado di estrarre determinate caratteristiche discriminanti.

L'immagine è una matrice di valori di pixel e la CNN è in grado di catturare con successo le dipendenze spaziali e temporali in un'immagine attraverso l'applicazione di filtri pertinenti. Una componente principale di questa rete è il layer di convoluzione. Esso permette di eseguire sull'immagine l'operazione di convoluzione attraverso un kernel, ovvero una matrice. Nel caso l'immagine fosse a più canali, per esempio RGB, il kernel ha la stessa profondità dell'immagine e tutti i risultati vengono poi sommati con il bias. Conventionalmente, il primo layer di convoluzione è responsabile dell'acquisizione delle

caratteristiche di basso livello, ad esempio: i bordi, il colore, l'orientamento del gradiente e così via. Nei successivi layer, l'architettura si adatta anche alle caratteristiche di alto livello. Oltre al layer di convoluzione, un altro layer ha molta importanza nella rete CNN: il layer di pooling. Esso è responsabile della riduzione delle dimensioni spaziali dell'immagine con l'obiettivo di diminuire la computazione richiesta per elaborare i dati (in altri casi si va a ridurre il numero di *feature*). Inoltre, esso è utile anche per estrarre le caratteristiche dominanti che sono invarianti sia per rotazione, che per posizione. Questi due layer insieme formano l' i -esimo strato di una rete neurale convoluzionale e il loro numero può essere aumentato a seconda della complessità delle immagini acquisendo così ulteriori dettagli di alto livello, ma bisogna pagare il costo di una maggiore complessità di calcolo in quanto aumentano i parametri. Dopo aver eseguito i vari livelli del modello CNN, l'output prodotto si appiattisce e viene inserito in una normale rete neurale MLP per scopi di classificazione. [22]

In una rete neurale tipica del Deep Learning, le immagini RGB vengono immesse nella rete a strati singoli. A questo punto nel primo layer della rete si possono effettuare alcune operazioni quali: una convoluzione per unire le informazioni, operazioni di filtraggio, segmentazione, determinazione di fattori decisionali e così via. Alcune di queste operazioni, alternativamente, vengono effettuate negli strati successivi della rete. Nel caso d'immagini RGBD l'approccio è analogo, ma ora vi è un altro strato relativo alla profondità dei pixel nell'immagine. Alla rete non interessa che ci siano tre o quattro canali: semplicemente si rimodula il modo di agire del primo livello della rete che prenderà in input quattro canali, anziché tre. In altri casi la rete neurale ha due *primi livelli*: in uno dei si prendono in input le informazioni relative a RGB e nell'altro solo l'informazione relativa alla profondità. L'approccio da scegliere dipende dalle performance ottenute, dalle risorse a disposizione, dal tempo impiegato per un determinato dominio applicativo e così via.

2.2 Intel RealSense

2.2.1 Intel RealSense D435

La telecamera Intel RealSense D435 è una telecamera della famiglia D400 RealSense e offre il campo visivo più ampio di tutte le fotocamere Intel RealSense. Questa telecamera dispone di un otturatore globale sul sensore di profondità che la rende ideale anche per applicazioni in rapido movimento e rappresenta una soluzione che offre misure della profondità di qualità per diversi campi applicativi. Infatti, il suo ampio campo visivo la rende perfetta per applicazioni come la robotica o la realtà aumentata e virtuale, dove vedere quanto più possibile della scena è di vitale importanza. Con una distanza di visuale fino a dieci metri e con la sua forma compatta, questa telecamera può essere anche integrata in soluzioni già esistenti. Inoltre, grazie al supporto fornito con Intel RealSense SDK 2.0 e al supporto multipiattaforma, lo sviluppo viene semplificato e velocizzato.

Un altro vantaggio di questa telecamera è dato dal fatto che rappresenta una soluzione a basso costo, leggera e potente, che consente lo sviluppo di soluzioni innovative e in grado di comprendere e interagire con l'ambiente circostante. I sensori dell'otturatore globale forniscono una grande sensibilità alla luce bassa che consente ai robot sui quali potenzialmente può essere installata di navigare negli spazi con le luci spente.

In figura 2.5 si mostrano diverse viste della telecamera D435, mentre in figura 2.6 si mostra com'è strutturata internamente la telecamera D430.

Come si nota in figura 2.6 e 2.5, la telecamera Intel RealSense D435 è formata da diverse componenti. Quelle principali sulle quali è importante soffermarsi sono il *Left Imager*, il *Right Imager*, il modulo RGB e l'emettitore di raggi infrarossi.

Ora si prenderanno in considerazione le principali specifiche tecniche della telecamera Intel RealSense D435, così da chiarire campi e contesti applicativi. La telecamera, come riportato dal sito ufficiale d'Intel [2], fornisce validi risultati sia in ambienti interni che esterni; il range di utilizzo ideale è compreso fra i 0.3 metri e i 3 metri. La tecnologia adottata per la profondità è di tipo stereoscopico, ovvero una tecnica di realizzazione e visione d'immagini, disegni, fotografie e filmati, atta a trasmettere una illusione di tridimensionalità, analoga a quella generata dalla visione binoculare del sistema visivo umano. Il campo visivo è di $87^\circ \times 58^\circ$, la minima distanza rilevabile è fissata



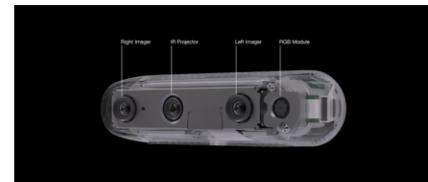
(a) Fronte



(b) Retro



(c) Lato



(d) Componenti fondamentali

Figura 2.5: Viste differenti della telecamera Intel RealSense D435 [2]

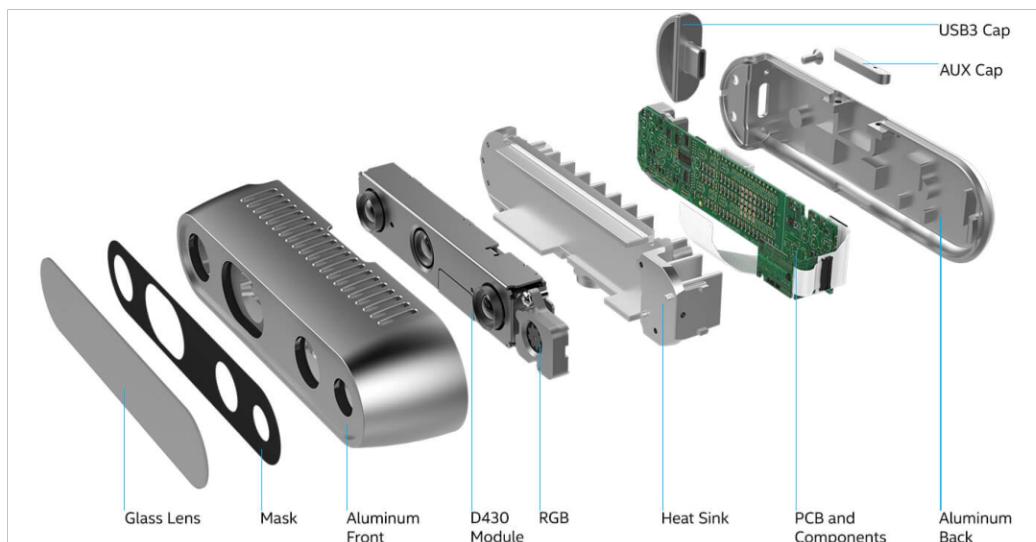


Figura 2.6: Architettura interna telecamera Intel RealSense D430 [2]

a 0.28 metri, la risoluzione massima raggiungibile è 1280x720, il massimo frame rate è 90 FPS, mentre la precisione è inferiore al 2% a una distanza di due metri.

Per quanto riguarda il canale RGB, la risoluzione massima del frame è 1920x1080, il campo visivo è di 69°x42° (H x V), il frame rate è di 30 FPS, la risoluzione del sensore RGB è di 2 MP, mentre l'otturatore è di tipo “rolling shutter”. Purtroppo con un otturatore di questo tipo e non di tipo globale si ha l'effetto rolling shutter, che assume la forma di una distorsione nelle foto in cui un oggetto si muove rapidamente. Questo fenomeno si verifica quando si scatta una foto dal finestrino di un treno: gli oggetti con forti forme verticali appariranno piegati nella parte superiore o inferiore dell'immagine. Questo può capitare anche se i soggetti sono fermi ma si esegue una rapida panoramica della fotocamera. La distorsione del rolling shutter influisce anche sulla registrazione video.



Figura 2.7: Effetto Rolling Shutter [36]

In figura 2.7 si riporta un esempio dell'effetto rolling shutter. Ulteriori dettagli e specifiche tecniche riguardanti la telecamera Intel RealSense D435

sono consultabili al seguente riferimento. [3]

Le telecamere Intel RealSense sono dispositivi *plug-and-play*, ovvero una volta collegate al calcolatore, è già possibile iniziare a utilizzarle sviluppando i primi programmi con l'ausilio del SDK 2.0 fornito da Intel [12] o testando il funzionamento con programmi di esempio come Intel RealSense Viewer [13] e simili.

Quando si parla d'immagini contenenti informazione sulla profondità degli oggetti nella scena inquadrata, non si può non menzionare il dislivello esistente tra *qualità della profondità rilevata* e *tracciamento di persone e/o oggetti* che è possibile realizzare. Si consideri la tabella 2.1. Una telecamera orientata al *Field of View* (FOV) ha vantaggi che una telecamera orientata al *Light Spectrum* non ha e viceversa. Una soluzione ottimale è quella di adottare una telecamera con un ampio FOV, ma con filtro infrarossi. Le telecamere Intel RealSense adottano esattamente questo approccio.

	Tracking	Depth Quality
Field of View	Wide	Narrow
Light Spectrum	Visible Light	Infrared

Tabella 2.1: Vantaggi e svantaggi delle telecamere orientate al Field of View oppure al Light Spectrum

La tecnologia delle telecamere Intel RealSense D435 è integrabile con il processore *Movidius*. Quest'ultimo permette di adempiere task di computer vision avanzati risparmiando sull'utilizzo altrimenti eccessivo delle risorse del sistema. In ultima analisi si precisa che la telecamera Intel RealSense D435 è in grado di utilizzare sia la tecnologia laser che quella a infrarossi al fine di ottenere l'informazione sulla profondità. Tale aspetto sarà trattato nella sottosezione 2.2.3.

2.2.2 Intel RealSense D435 e Intel RealSense T265

La telecamera Intel RealSense T265 è visibile in figura 2.8 e si tratta di un'altra telecamera di casa Intel appositamente progettata per attività di tracking. Infatti, ha a disposizione un Field of View molto esteso e ciò rende le operazioni di tracking molto precise. Viceversa, la telecamera Intel RealSense D435 ha un Field of View più ristretto e di conseguenza è maggiormente utilizzata per domini applicativi ove l'informazione sulla profondità è di vitale importanza da ricavare.

Inoltre, il modello T265 non è una telecamera propriamente pensata per ricavare informazioni sulla profondità e non fornisce dati a essa relativi. Una possibile applicazione allo stato dell'arte potrebbe essere quella di abbinare una telecamera Intel RealSense D435 (anche modelli precedenti o successivi) con una telecamera Intel RealSense T265. Così facendo è possibile rilevare la profondità dalla scena con molta precisione e al tempo stesso realizzare un tracking dettagliato e molto preciso. La telecamera T265 è dotata di sensore IMU, argomento della sottosezione 2.3.



Figura 2.8: Telecamera Intel RealSense T265 [2]

2.2.3 Tecnologia infrarossi, laser, lidar

La procedura di determinazione dei punti di equivalenza al fine di ricavare la profondità per ogni pixel dell’immagine acquisita della scena inquadrata, diventa più complicata quando la scena è a “tinta unita”, per esempio una parete monocromatica. In tal caso si sfruttano le potenzialità dei raggi infrarossi: la telecamera trasmette dei pattern solitamente circolari in maniera pseudo-casuale sulla scena e in base alla deformazione subita dai pattern circolari nella scena. Queste deformazioni faranno sì che a un pattern circolare più piccolo in termini di raggio, sarà associata una profondità maggiore, mentre a un pattern circolare più grande in termini di raggio, sarà associata una profondità inferiore. Si precisa che tutti i pattern circolari proiettati sono dello stesso raggio. Questo approccio è usato dalla famiglia di telecamere Intel RealSense e permette di ottenere risultati migliori anche in situazioni particolari in termini di luce, cromaticità e quant’altro. In generale i raggi infrarossi vengono sfruttati per avere informazioni sulle tessiture (o maglia) degli oggetti. In figura 2.9 si mostrano i pattern circolari emessi dalle telecamere Intel RealSense D415 e D435.

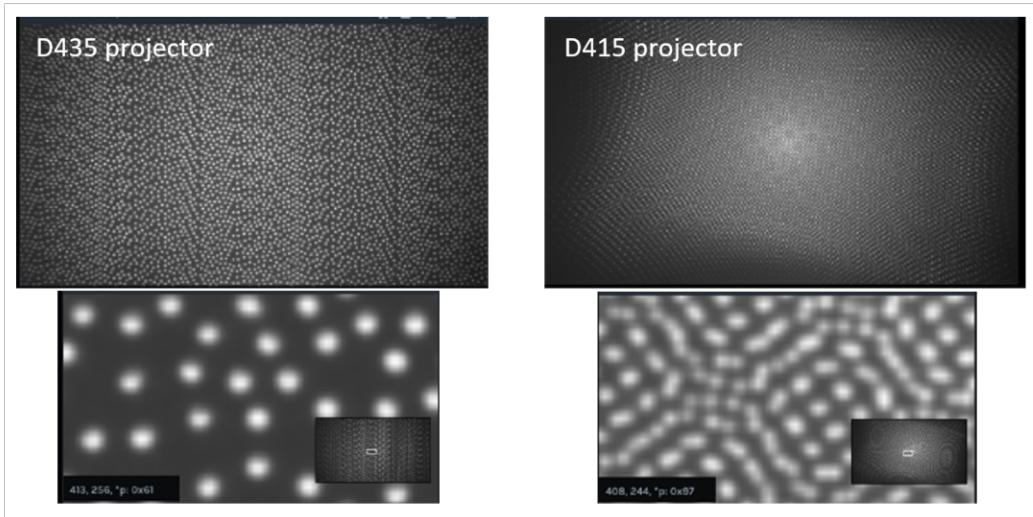


Figura 2.9: Pattern circolari Intel RealSense D415 e D435 [25]

Si supponga di coprire una delle due telecamere RGB d’imaging. Ciò che si otterrà è la perdita dell’informazione sulla profondità ovviamente. Se invece si copre il sensore infrarossi, allora l’immagine continua a essere prodotta e visibile, ma la qualità dell’immagine cala in maniera drastica. Questo perché si perdono alcune informazioni fondamentali sulle tessiture, per esempio

le pareti monocromatiche risulteranno molto rumorose. Inoltre, si precisa che il dispositivo delegato ai raggi infrarossi non è soltanto un emettitore di raggi infrarossi, ma anche di un sensore di raggi infrarossi. Così facendo è in grado di emettere i pattern circolari di cui prima, e successivamente dedurne le deformazioni una volta che questi vengono riflessi.

In ambienti esterni il contributo dei raggi infrarossi, come già accennato in precedenza, è abbastanza limitato, se non del tutto assente se la telecamera inquadra il Sole. Quindi in ambienti esterni la telecamera solitamente fa affidamento maggiormente sulla visione stereoscopica.

Tutta la computazione relativa allo stereo-matching, ovvero attività come la triangolarizzazione per corrispondenza tra le due immagini e conseguente ottenimento della profondità, l'emissione e il rilevamento dei raggi infrarossi, il miglioramento delle texture, l'abbinamento del frame RGB con quello della profondità, il pre-processing e il post-processing, sono tutte effettuate interamente sulla telecamera Intel RealSense, permettendo così al calcolatore alla quale viene collegata di realizzare solo la fase di rendering del video. Questo aspetto è molto importante. Infatti, qualora dovessimo usare parte delle risorse del calcolatore per effettuare parte o tutte le operazioni poc'anzi descritte, l'intero processo sarebbe di gran lunga rallentato e questo soprattutto se il calcolatore, nel mentre, è utilizzato anche per altre attività. Inoltre, questo aspetto è utile anche per applicativi in campo IoT ed embedded. Approcci alternativi all'utilizzo della tecnologia a infrarossi sono il laser e il lidar [15]. Si tratta di tecnologie che permettono di raggiungere un grado di precisione maggiore, ma sono anche di gran lunga più costose.

La riflessione acquisita dal sensore infrarossi di tipo CMOS [16] della telecamera Intel RealSense, per calcolare la profondità finale dell'oggetto nella scena dalla telecamera e quindi assegnarla al pixel di riferimento, si confronta ciò che è stato proiettato dall'emettitore a infrarossi con ciò che è stato acquisito dal sensore infrarossi. A tal punto si sfruttano i metodi trigonometrici e di triangolarizzazione visti in precedenza. Questo è anche il principio di funzionamento del Kinect nella sua versione *v1*. Nella versione *v2* si sfrutta invece il *ToF*, ovvero il *Time of Flight*. Il concetto è simile a quello delle telecamere Intel RealSense. Una telecamera che fa affidamento sul ToF, è composta da alcune componenti principali: lenti, sorgente luminosa integrata, sensore che cattura tutte le informazioni dell'immagine e un'interfaccia (e.g. USB) per il collegamento esterno. Questo approccio con queste componenti permette di ottenere informazioni sia sulla profondità degli oggetti nella scena, che sull'intensità da memorizzare nel pixel relativo al corrispet-

tivo oggetto; questo per ogni pixel dell’immagine. L’idea di funzionamento è ancora una volta quella di emettere un fascio luminoso sulla scena e poi osservare e determinare la luce riflessa dagli oggetti nella scena. Siccome la velocità della luce è nota, allora la distanza degli oggetti nella scena è facilmente ricavabile per ogni punto della scena. L’uso del ToF comporta alcuni vantaggi, ma anche svantaggi come la sensibilità molto maggiore alla luce del Sole rispetto alla tecnologia infrarossi o laser, riflessioni multiple e conseguenti artefatti nell’immagine risultato.

Infine, per quanto riguarda la tecnologia laser, questa permette di raggiungere livelli di precisione molto maggiori rispetto alla tecnologia infrarossi. Alcune tecnologie laser possono far uso a loro volta di tecnologia infrarossi, oppure basarsi sull’ultravioletto o sulla luce visibile. Invece, per quanto riguarda la tecnologia lidar (Light Detection And Ranging), questa identifica la tecnologia che misura la distanza da un oggetto illuminandolo con una luce laser e che al contempo è in grado di restituire informazioni tridimensionali ad alta risoluzione sull’ambiente circostante. Un lidar utilizza tipicamente diversi componenti: laser, fotorilevatori e circuiti integrati di lettura (ROIC) con capacità di tempo di volo (TOF) per misurare la distanza illuminando un bersaglio e analizzando la luce riflessa. Di base il lidar è una tecnica simile a un radar basata sul principio dell’eco. Lo stesso principio utilizzato dai radar, che utilizza come “segnaletica” la luce (pulsata) anziché un segnale radio.

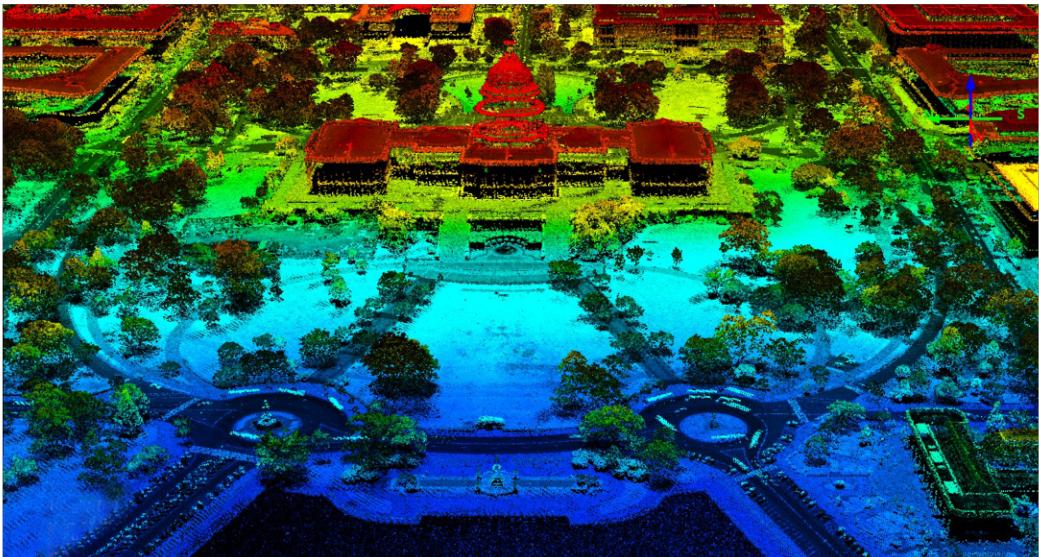


Figura 2.10: Immagine lidar [25]

2.3 SLAM - Simultaneous localization and mapping

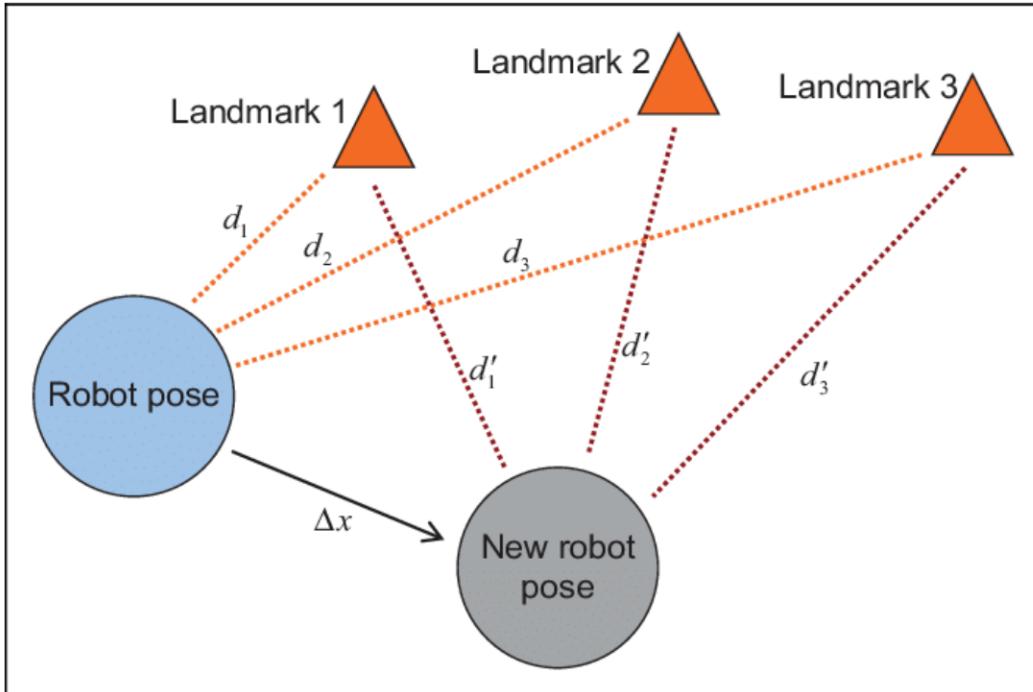


Figura 2.11: Rappresentazione del problema SLAM [26]

Quando si parla di SLAM, si sta parlando del problema computazionale *Simultaneous localization and mapping*, ovvero come fa un dispositivo a ricostruire e/o aggiornare una mappa tridimensionale di un ambiente sconosciuto e contemporaneamente tenere traccia della sua propria posizione all'interno dell'ambiente in cui si muove? Sebbene sia un problema proprio del mondo della robotica, ha conseguenza anche in questo dominio applicativo. Prima dell'avvento del GPS (Global Positioning System), anche noto come NAVSTAR GPS (NAVigation Satellite Timing And Ranging Global Positioning System), i marinai navigavano con le stelle, sfruttando il loro movimento e la loro posizione per trovare la rotta da seguire nell'oceano. Per risolvere la problematica, si sfrutta l'approccio V-SLAM (Visual-SLAM) dove l'idea è quella di usare una combinazione di telecamere e un dispositivo elettronico chiamato IMU (Inertial Measurements Units) per navigare in maniera analoga a quanto detto prima nell'approccio dei marinai, ovvero sfruttando caratteristiche visuali nell'ambiente circostante al fine di tracciarlo con precisione, benché non sia noto a priori.

Di fronte a pareti a tinta unita, questo approccio risulta avere ancora qualche problema pratico come inesattezza e artefatti. Il sensore IMU serve a fornire misurazioni accurate anche con rotazioni e/o grandi velocità intraprese. Si tratta di un dispositivo elettronico utilizzato per calcolare e segnalare una forza esatta del corpo, velocità angolare e direzione del corpo, che può essere ottenuta utilizzando una miscela di tre sensori come giroscopio, magnetometro e accelerometro. Questi sensori sono normalmente utilizzati per pianificare aerei inclusi UAV (veicoli aerei senza pilota), tra molti altri, e veicoli spaziali, compresi lander e satelliti. Gli sviluppi moderni consentono la produzione di dispositivi GPS basati su IMU.

In diversi modelli di telecamere Intel RealSense è presente un sensore IMU che permette di ricavare anche informazioni in merito all'orientazione della telecamera all'interno dell'ambiente circostante, così come di ciò che viene osservato nella scena. Il modello di telecamera Intel RealSense D435 non dispone di questo ulteriore modulo, a differenza di altri modelli come D435i e il modello T265 menzionato nella sottosezione 2.2.2. La figura 2.11 mostra graficamente il problema SLAM.

2.4 Ottener la profondità in un'immagine

Esistono diversi approcci per ricavare la profondità degli oggetti e quindi dei pixel presenti in ogni immagine. Un primo modo è quello di utilizzare una telecamera in grado di determinare la profondità per ogni pixel (tali telecamere sono anche dette nel gergo *telecamere 3D*). A questa categoria appartengono le telecamere con tecnologia Intel RealSense [2]. Vale la pena menzionare la presenza di altre telecamere in grado di adempiere compiti analoghi, come le telecamere Zivid [27], Revopoint [28], Orbbec [29] e Stereolabs [30].

Al fine di ricavare la profondità da un'immagine però, quello di utilizzare una telecamera dedicata non è l'unico approccio. Quelle elencate finora sono tutte famiglie di telecamere e in generale tecnologie che fanno uso di una telecamera in cui sono presenti poi due *sotto-telecamere*, ovvero si basano sull'emulazione dell'apparato visivo umano. Alternativamente, è possibile utilizzare una singola telecamera con un sistema complesso e avanzato di lenti esagonali detto *Hexagonal Microarray Lens* che permette di raggiungere lo stesso risultato. Questo tipo di approccio è utilizzato dalle telecamere Raytrix [31]. Questo tipo di telecamere sono anche dette telecamere *lightfield*. Le fotocamere Light Field sono un nuovo tipo di telecamere 3D che acquisiscono un'immagine standard insieme alle informazioni di profondità di una scena. Le informazioni metriche 3D possono essere acquisite con una singola telecamera di campo luminoso attraverso un singolo obiettivo in un singolo scatto utilizzando solo la luce disponibile. Raytrix si è specializzata nello sviluppo di telecamere leggere e pratiche per applicazioni industriali. Usando la tecnologia *Hexagonal Microarray Lens* visibile in figura 2.12, consente un compromesso ottimale tra alta risoluzione efficace e grande profondità di campo. L'idea è che ogni microlente esagonale dell'array vede l'immagine da un'angolazione leggermente differente per poi utilizzare lo stesso approccio visto nelle telecamere stereoscopiche: lo stereo-matching mediante geometria epipolare.

Un terzo approccio utile per ricavare la profondità da un'immagine è quello di sfruttare sempre una sola telecamera, ma questa volta la si integra con l'ausilio dell'intelligenza artificiale e del Machine Learning. Quello che si fa è utilizzare un modello specializzato a determinare la profondità d'immagini grazie a una corposa fase di addestramento, per poi ottenere la profondità da immagini completamente nuove. Questo approccio è quello che viene utilizzato da PoseNet [32], da MoveNet [33], da OpenPose [34] e dal noto Mediapipe [35].

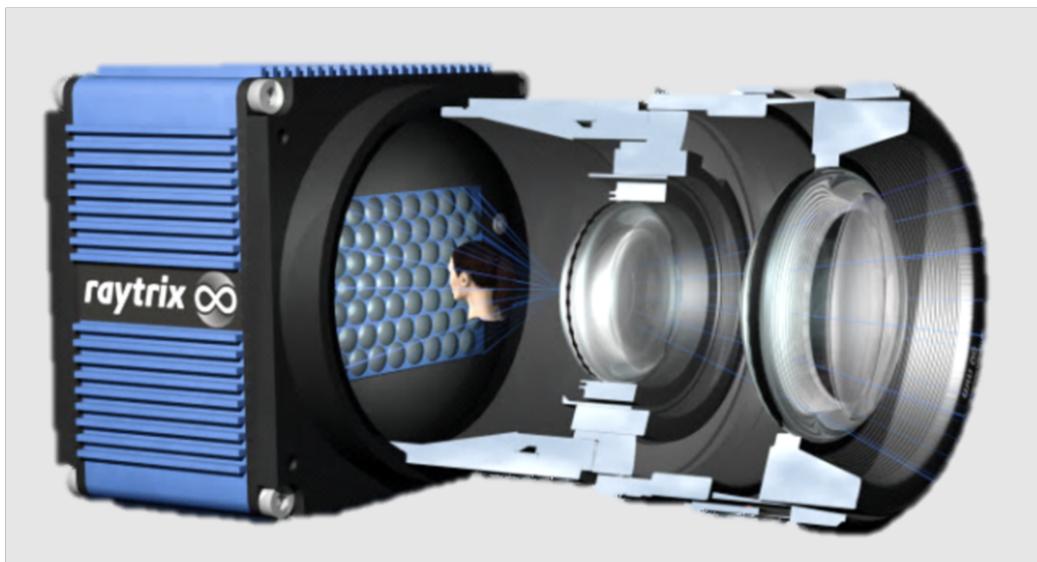


Figura 2.12: Tecnologia Raytrix [31]

2.5 Pose Estimation

2.5.1 OpenPose

2.6 Trasformazione di spazi di coordinate

2.7 Streaming di eventi

Capitolo 3

Progettazione

Progettazione interessante

Capitolo 4

Implementazione

Implementazione interessante

Capitolo 5

Risultati sperimentali

Risultati sperimentali

Capitolo 6

Conclusione

Conclusione interessante

Appendice

Riferimenti

- [1] Treccani - Enciclopedia Online,
Definizione “Sicurezza”,
www.treccani.it/enciclopedia/sicurezza/
- [2] Tecnologia Intel RealSense e fonti immagini relative,
Prodotti e casi d'uso RealSense,
www.intelrealsense.com,
www.intelrealsense.com/use-cases/,
<https://dev.intelrealsense.com/docs/projectors>,
<https://www.intelrealsense.com/tracking-camera-t265/>,
<https://www.intelrealsense.com/depth-camera-d435/>,
<https://www.intelrealsense.com/depth-camera-d435i/>
- [3] Specifiche tecniche Intel RealSense D435,
[https://www.intelrealsense.com/wp-content/uploads/2022/05/Intel-RealSense-D400-Series-Datasheet-April-2022.pdf/](https://www.intelrealsense.com/wp-content/uploads/2022/05/Intel-RealSense-D400-Series-Datasheet-April-2022.pdf)
- [4] Fonti Immagini figura 1.2,
www.width.ai/services/object-recognition-software-services,
www.biometricupdate.com/202004/morocco-extends-facial-recognition-moratorium-to-year-end-proposes-biometric-authentication-service,
www.therobotreport.com/righthands-yaro-tenzer-on-robotic-piece-picking-waymos-sf-expansion/,
<https://maelfabien.github.io/tutorials/open-pose/>
- [5] Demo YouTube,
OpenPose + RealSense D435 - 3D Body Tracking - Demo CVPRLAB,
[www.youtu.be/Ac0V8Dj0FbI](https://www.youtube.com/watch?v=Ac0V8Dj0FbI)

- [6] Fonti Immagini sottosezione Concetti preliminari,
<https://www.radio2space.com/it/le-componenti-dello-spettro-elettromagnetico/>
- [7] Rafael C. Gonzalez, Richard E. Woods,
Elaborazione delle Immagini, terza edizione, Pearson, 2008
- [8] Luca Rubino,
Sviluppo di un'architettura di comunicazione distribuita con digital twins per la videosorveglianza, 2022
- [9] Renato Esposito,
Sviluppo di un ambiente virtuale con digital twins per la videosorveglianza, 2022
- [10] Unity,
www.unity.com
- [11] Geometria epipolare - Wikipedia,
https://it.wikipedia.org/wiki/Geometria_epipolare
- [12] Intel RealSense SDK 2.0,
<https://www.intelrealsense.com/sdk-2/>
- [13] Intel RealSense Viewer,
<https://www.intelrealsense.com/download/7144/>
- [14] Microsoft Kinect,
https://it.wikipedia.org/wiki/Microsoft_Kinect
- [15] Tecnologia Lidar,
<https://consystem.it/faq/tecnologia-lidar-che-cosa-e-come-funziona/>
- [16] Sensore CMOS,
https://it.wikipedia.org/wiki/Sensore_a_pixel_attivi
- [17] W.S. McCulloch, W. Pitts,
A logical calculus of the ideas immanent in nervous activity - The bulletin of mathematical biophysics, 1943
- [18] D. O. Hebb,
The organization of behavior: a neuropsychological theory, 1949

- [19] F. Rosenblatt *The perceptron: a probabilistic model for information storage and organization in the brain*, *Psychological review*, 1958
- [20] M. Minsky, S. Papert,
An introduction to computational geometry, Cambridge tiass., HIT, 1969
- [21] P. J. Werbos,
An overview of neural networks for control, *IEEE Control Systems Magazine*, 1991
- [22] Pasquale Auriemma,
Classificazione di segnali EEG mediante tecniche di deep learning, 2019
- [23] Y. LeCun, Y. Bengio,
Convolutional networks for images, speech, and time series, of brain theory and neural networks, 1995
- [24] Fonti Immagini sottosezioni 2.1.3, 2.1.4, 2.1.6,
<https://www.analyticsvidhya.com/blog/2021/05/convolutional-neural-networks-cnn/>
https://www.researchgate.net/publication/339446790_Using_a_Data_Driven_Approach_to_Predict_Waves_Generated_by_Gravity
- [25] Fonti Immagini sottosezioni 2.2.3,
<https://gisgeography.com/lidar-light-detection-and-ranging/>,
<https://dev.intelrealsense.com/docs/projectors>
- [26] Fonti Immagini sottosezione 2.3,
https://www.researchgate.net/publication/341348773_A_survey_of_image_semantics-based_visual_simultaneous_localization_and_mapping_Application-oriented_solutions_to_autonomous_navigation_of_mobile_robots
- [27] Telecamere Zivid,
<https://www.zivid.com/>
- [28] Telecamere Revopoint,
<https://3dcamera.revopoint3d.com/>
- [29] Telecamere Orbbec,
<https://orbbec3d.com/>

- [30] Telecamere Stereolabs,
<https://www.stereolabs.com/>
- [31] Telecamere Raytrix,
<https://raytrix.de/>
- [32] PoseNet,
<https://github.com/tensorflow/tfjs-models/tree/master/pose-detection>
- [33] MoveNet,
<https://www.tensorflow.org/hub/tutorials/movenet>
- [34] OpenPose,
<https://github.com/CMU-Perceptual-Computing-Lab/openpose>
- [35] Mediapipe,
<https://mediapipe.dev/>
- [36] Fonte Immagine 2.7 https://en.wikipedia.org/wiki/Rolling_shutter

Ringraziamenti