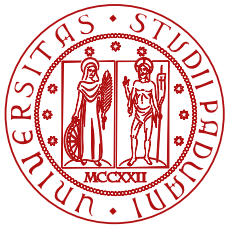


Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione

Laurea Triennale in Ingegneria Informatica

Ricerca sistematica e comparazione delle metodologie per la stima della posa 6D

Laureando

Matteo Tonello

Matricola 2008596

Relatore

Prof. Stefano Ghidoni

Università di Padova

Correlatore

Dott. Davide Allegro

Università di Padova

Anno Accademico
2022/2023

Sommario

Nel campo della Computer Vision, la capacità di stimare la posa 6D degli oggetti, cioè determinare la loro posizione e il loro orientamento rispetto a un sistema di riferimento, sta diventando sempre più importante: trova applicazioni in diversi settori, tra cui la robotica, dove è essenziale per la manipolazione degli oggetti, e nella guida autonoma, dove permette di tracciare i veicoli circostanti. Tuttavia, determinare la posa 6D degli oggetti è una sfida complessa, complicata dalla presenza di occlusioni nella scena o oggetti parzialmente visibili. Per superare questa sfida, negli ultimi anni, grazie anche alla diffusione del Deep Learning, sono stati sviluppati numerosi metodi che utilizzano strategie e approcci diversi. Questa tesi si propone di fornire una panoramica completa dello stato dell'arte delle metodologie per la stima della posa 6D attraverso una ricerca sistematica, focalizzata in particolare sui metodi che affrontano il problema delle occlusioni. Il processo di selezione e analisi degli studi è stato condotto basandosi sul protocollo PRISMA, che fornisce un insieme di linee guida fondamentali per la conduzione di revisioni sistematiche con l'obiettivo di garantire una metodologia chiara, trasparente e affidabile. Attraverso questa revisione, è stato possibile analizzare e confrontare i vari metodi, mettendo in luce le loro caratteristiche principali e identificando le categorie più adatte a risolvere il problema considerato. Vengono infine discusse le limitazioni di queste tecniche, fornendo un punto di partenza per ulteriori lavori futuri.

Indice

Elenco delle Figure

Elenco delle Tabelle

1 Introduzione

Grazie ai progressi tecnologici recenti, ad esempio nei settori della robotica, dei veicoli autonomi e della realtà aumentata, la capacità di determinare con precisione la posizione e l'orientamento degli oggetti sta diventando fondamentale.
In questo capitolo, al fine di comprendere al meglio l'argomento della tesi, è fornita una panoramica dettagliata sul problema della stima della posa 6D, definendo in particolare il concetto di posa 6D, esplorando le sue applicazioni e mettendo in evidenza una delle sfide più significative che si manifesta in questo contesto, ovvero la presenza delle occlusioni.

1.1 Problema della stima della posa 6D

La stima della posa 6D degli oggetti è una delle sfide recenti più interessanti nel campo della Computer Vision. Si propone infatti di risolvere il problema di determinare con precisione la posizione e l'orientamento di un oggetto tridimensionale nello spazio, in relazione a un sistema di riferimento.
L'applicabilità della stima della posa 6D trova spazio in svariati ambiti di studio che, con l'evoluzione tecnologica, sono destinati a espandersi ulteriormente. Alcuni esempi di applicazioni li troviamo nel campo della robotica [1], della realtà aumentata [2] e della guida autonoma [3].
Tuttavia, lo sviluppo di questa tecnologia è tutt'altro che privo di sfide. Numerosi sono infatti i problemi e le difficoltà che emergono durante il processo di sviluppo, come ad esempio la presenza di oggetti deformabili, simmetrici o che si sovrappongono.

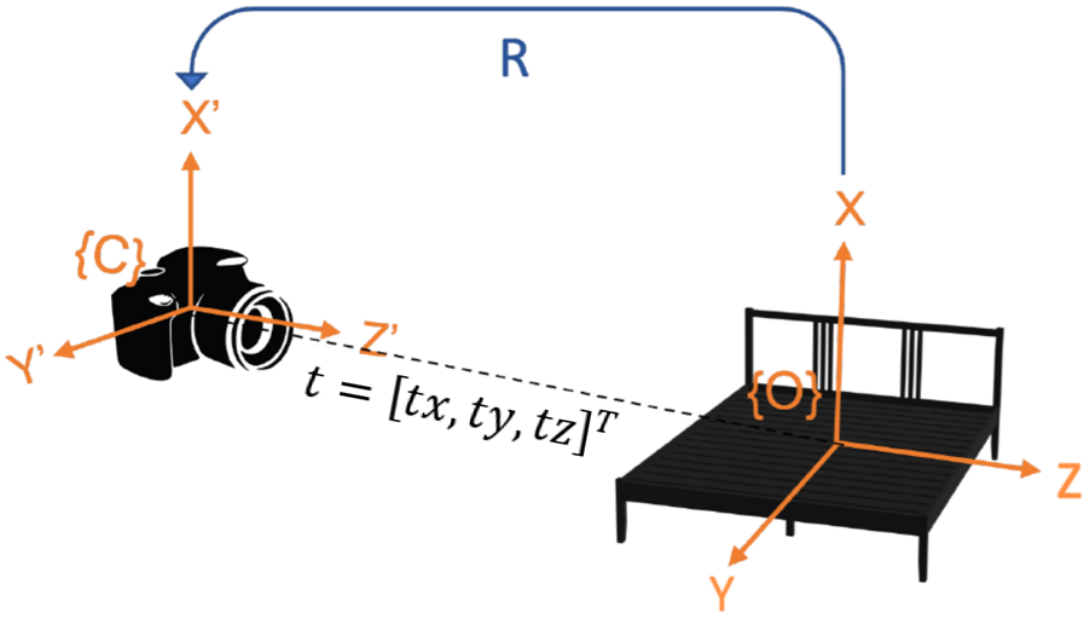


Figura 1.1: Illustrazione della definizione di posa 6D. Sono raffigurati i sistemi di riferimento {O} e {C}, rispettivamente quello dell'oggetto e quello della fotocamera. R e t rappresentano rispettivamente la rotazione e la traslazione dell'oggetto rispetto alla telecamera.

1.1.1 Cos'è la posa 6D

La stima della posa 6D di un oggetto è una tecnica di Computer Vision che determina la posizione e l'orientamento di un oggetto nello spazio rispetto a un sistema di coordinate di riferimento. I sei gradi di libertà (6D) si riferiscono alle traslazioni lungo gli assi x, y e z (posizione) e le tre rispettive rotazioni attorno agli assi stessi (orientamento).
In particolare la posa 6D di un oggetto può essere definita attraverso una generica matrice di rototraslazione della seguente forma $G = \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix}$, dove R identifica la matrice di rotazione, $R \in SO(3)$, e T rappresenta il vettore di traslazione $t = [tx, ty, tz]^T$. Questa matrice di rototraslazione descrive la trasformazione geometrica (posizione e orientamento) di un oggetto rispetto a un sistema di riferimento preciso. Facendo riferimento alla figura 1.1 , la matrice di rototraslazione G mette in relazione ogni punto $Xo = [xo, yo, zo]^T$ del sistema di riferimento dell'oggetto {O} col punto corrispondente $Xc = [xc, yc, zc]^T$ del sistema di riferimento della fotocamera {C}. La traslazione $t = [tx, ty, tz]^T$ può essere interpretata come la posizione dell'origine del sistema di coordinate dell'oggetto nel sistema di coordinate della telecamera. Mentre, come illustrato alla figura 1.2, la rotazione

R in tre dimensioni può essere espressa come una composizione di 3 rotazioni intorno ai tre assi x,y e z: $R = Rz(\alpha)Ry(\beta)Rx(\gamma)$.

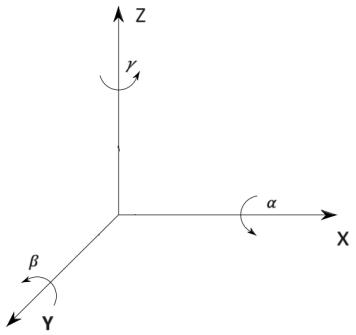


Figura 1.2: Rappresentazione della rotazione degli assi e i relativi angoli.

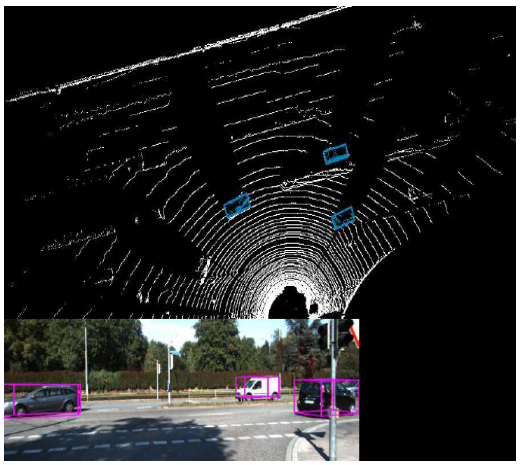
La relazione tra un qualsiasi punto Xo nel sistema {O} e il punto corrispondente Xc nel sistema {C} può essere espressa come segue: $Xc = [R|t]Xo$.
Normalmente vengono considerati solamente oggetti rigidi, quelli in cui le caratteristiche strutturali rimangono invariate rispetto al sistema di riferimento o al passare del tempo. Di conseguenza tutti i punti di tali oggetti condividono la stessa matrice di rototraslazione.

1.1.2 Applicazioni

L'importanza della stima della posa 6D degli oggetti deriva dalla sua capacità di essere applicata in una vasta gamma di scenari.
Un primo esempio di applicazione possiamo trovarlo nel campo della realtà aumentata (AR) [2]. Questa tecnologia consente di mettere in relazione oggetti reali e virtuali in modo interattivo e real-time. Grazie alle informazioni della posizione e dell'orientamento di un oggetto, molti metodi per la stima della posa 6D consentono di far interagire elementi virtuali in modo coerente con il mondo reale (figura 1.3a).
Un altro campo di applicazione è quello della robotica, dove la stima della posa 6D di un oggetto rispetto al sistema di riferimento del robot (end-effector o base) permette al braccio robotico di comprendere la posizione e l'orientamento degli oggetti nel suo spazio di lavoro. Questa tecnologia consente quindi ai robot di identificare, afferrare e manipolare oggetti in modo molto preciso, ad esempio per applicazioni di bin-picking, ovvero l'abilità dei robot di maneggiare oggetti da

(a) Realtà aumentata.

(b) Bin-picking.



(c) Veicoli a guida autonoma.

Figura 1.3: Esempi di applicazioni della stima della posa 6D.

contenitori (figura 1.3b). Un notevole sviluppo in questo settore è stato evidenziato dalla competizione "Amazon Picking Challenge"[1] [1].
Un ulteriore esempio di applicazione in cui la stima della posa 6D dimostra la sua importanza è rappresentato dai veicoli a guida autonoma. In questo contesto, emerge la necessità di individuare con precisione la posizione e il comportamento non solo dei veicoli circostanti [3], ma anche di altre entità [4] , quali pedoni, ciclisti e ostacoli vari (figura 1.3c). La capacità di effettuare una stima accurata della posizione 6D consente alle vetture autonome di acquisire una comprensione approfondita dell'ambiente circostante, affinando la loro abilità di prevedere le azioni degli altri attori sulla strada.

[1]https://arc.cs.princeton.edu

1.1.3 Occlusioni

Una delle principali sfide che sorge nella stima della posizione 6D di un oggetto è la presenza di occlusioni nella scena, come nell'esempio riportato nella figura 1.4. Questo accade quando l'oggetto di interesse è parzialmente o completamente nascosto da altri oggetti nel campo visivo. In altre parole, una o più parti dell'oggetto sono oscurate da elementi circostanti. Questo fenomeno, tipico delle complesse scene industriali, può verificarsi in molte situazioni, come ad esempio quando gli oggetti si sovrappongono l'uno all'altro o in ambienti disordinati.
Le occlusioni rappresentano una sfida significativa per la stima della posa 6D, poiché influenzano negativamente la capacità di percepire e riconoscere correttamente l'oggetto. Quando una parte dell'oggetto è nascosta, le caratteristiche visive utilizzate per riconoscerlo o per stimare la sua posizione e orientamento potrebbero non essere completamente visibili. Questo può portare a errori nella stima della posa, poiché le informazioni incomplete o ambigue possono portare a risultati imprecisi o addirittura a un fallimento nell'individuare l'oggetto. Molti metodi tentano di risolvere questo problema, attuando vari tipi di strategie, che verranno analizzate più approfonditamente in questa tesi.

Figura 1.4: Esempio di scena con occlusioni, dal dataset Occlusion Linemod [5].

1.2 Scopo e struttura della tesi

Questa tesi ha l'obbiettivo di studiare lo stato dell'arte dei metodi per la stima della posa 6D, in particolare nelle scene caratterizzate da occlusioni. Vengono analizzati e confrontati tutti i metodi presenti nella letteratura che tentano di risolvere questo problema, in modo tale da fornire una visione più generale possibile. La parte successiva della tesi è strutturata nel seguente modo: nel secondo capitolo, viene fornita una panoramica più specifica sui principali metodi per la stima della posa 6D. Nel terzo capitolo viene delineata la metodologia adottata per condurre la revisione sistematica, fornendo una panoramica sulle strategie e gli strumenti utilizzati per l'analisi. Il quarto capitolo è dedicato alla presentazione dei risultati ottenuti dalla ricerca. Nel quinto capitolo vengono discussi e analizzati i risultati conseguiti. Infine, nell'ultimo capitolo, la conclusione, vengono offerti possibili spunti per sviluppi futuri.

2

Metodi per la stima della posa 6D

Per la stima della posa 6D, sono state sviluppate diverse metodologie in letteratura, che tipicamente prendono in input immagini 2D o 3D. Successivamente, attraverso algoritmi dedicati e reti neurali, si stima la posa dell'oggetto, che viene poi utilizzata per lo specifico scopo.
Questi approcci differiscono notevolmente tra loro, poiché la loro scelta è influenzata dal contesto e dall'applicazione specifica. Sebbene esistano numerose tecniche applicabili in vari scenari, è importante notare che alcune metodologie sono appositamente ottimizzate per contesti particolari, come la robotica industriale o la realtà aumentata.
All'interno di questo capitolo, vengono descritte le diverse caratteristiche dei metodi utilizzati per la stima della posa 6D e le categorie principali in cui tali metodi possono essere suddivisi.

2.1 Caratteristiche

La scelta del tipo di metodo da utilizzare è guidata anche dalla disponibilità tecnologica. Sensori di input, capacità di calcolo e disponibilità di dati variano, influenzando la selezione dell'approccio ottimale.
Pertanto, si possono individuare numerose caratteristiche distintive tra i diversi metodi, tra cui le seguenti:

- **Single/Multi view**: Gli approcci "single view" si basano su un singolo punto di vista o immagine, mentre gli approcci "multi view" utilizzano più punti di vista o immagini della stessa scena per ottenere stime più accurate e robuste della posa.

- **Instance/Category level**: Gli approcci a livello di istanza mirano a stimare la posa esatta di oggetti specifici. Gli approcci a livello di categoria catturano le caratteristiche generali di una categoria di oggetti e stimano la posa basandosi sulla conoscenza generale delle proprietà della categoria.
- **Stima individuale/multi-istanza**: Gli approcci per la "stima individuale" si concentrano sulla determinazione della posa 6D di singoli oggetti, isolatamente dalla scena circostante. Gli approcci per la "stima multi-istanza" affrontano la sfida di stimare la posa 6D di più istanze di oggetti contemporaneamente.
- **Asincrono / Tempo reale**: Gli approcci "asincroni" non richiedono una risposta immediata, consentendo maggiore complessità computazionale. Gli approcci "in tempo reale" richiedono risposte istantanee e sono ottimizzati per eseguire calcoli rapidi, adatti a scenari dove la tempestività è essenziale.

In generale quindi, la stima della posa 6D si adatta alle esigenze specifiche, basandosi sull'ambiente di applicazione, sul livello di accuratezza desiderato e sulle risorse tecnologiche disponibili.

2.2 Categorie

Ciò che costituisce una differenza significativa tra i diversi metodi, oltre alle caratteristiche precedentemente menzionate, risiede negli approcci metodologici adottati per risolvere il problema della stima della posa 6D. Si possono quindi suddividere i metodi in tre principali categorie: ***template-based***, ***feature-based*** e ***learning-based***.

2.2.1 Template-based

I metodi template-based includono una prima fase offline in cui viene costruito un database di template da un modello 3D dell'oggetto. Questo database è ottenuto variando posizione e orientamento dell'oggetto, così da avere una sua prospettiva da diverse angolazioni. La seconda fase è una fase di test, eseguita online per determinare la posizione 6D. L'immagine in input viene confrontata con tutti i template del database generati nella fase precedente. Questi algoritmi scelgono la posa che ha la migliore corrispondenza. Lo schema di funzionamento è mostrato alla figura 2.1.

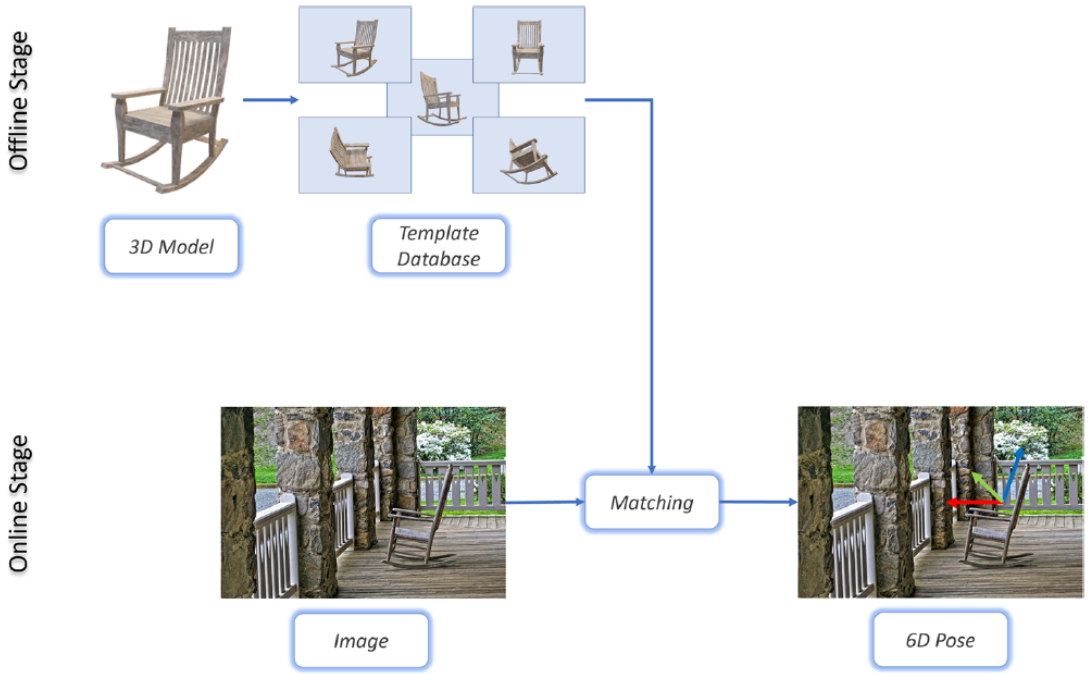


Figura 2.1: Schema dei metodi Template-based.

Vantaggi

- Funzionano bene nel caso di oggetti privi di texture.
- Se il database è vasto e completo, possono raggiungere un'alta precisione.

Svantaggi

- Sono molto sensibili alle occlusioni, poiché influenzano negativamente la comparazione dell'oggetto, visibile parzialmente, con i template.
- La velocità di esecuzione è inversamente proporzionale al numero di elementi appartenenti al database. Tuttavia, questo numero è direttamente proporzionale alla precisione del metodo. Pertanto, è necessario trovare un compromesso.

2.2.2 Feature-based

Altri metodi di stima della posa 6D si basano su approcci geometrici, come ad esempio i metodi feature-based. I metodi appartenenti a questa categoria sfruttano le caratteristiche locali (keypoints o edges) estratte dalle regioni di interesse o da tutti i pixel nell'immagine per poi confrontarle con le caratteristiche di un modello 3D dell'oggetto (ad esempio un modello CAD) al fine di stabilire corrispondenze 2D-3D. Pertanto, il processo si articola in due fasi: la prima fase estrae le carat-

teristiche locali e le confronta con i punti chiave tridimensionali; la seconda fase coinvolge le corrispondenze 2D-3D per ottenere la posa 6D, ad esempio tramite l'algoritmo PnP (Perspective-n-Point) [6]. Queste tecniche possono comprendere anche reti neurali convoluzionali (CNN), le quali vengono utilizzate in diverse fasi del processo al fine di migliorare le prestazioni complessive del sistema. Alla figura 2.2 è illustrato il funzionamento di questi metodi.

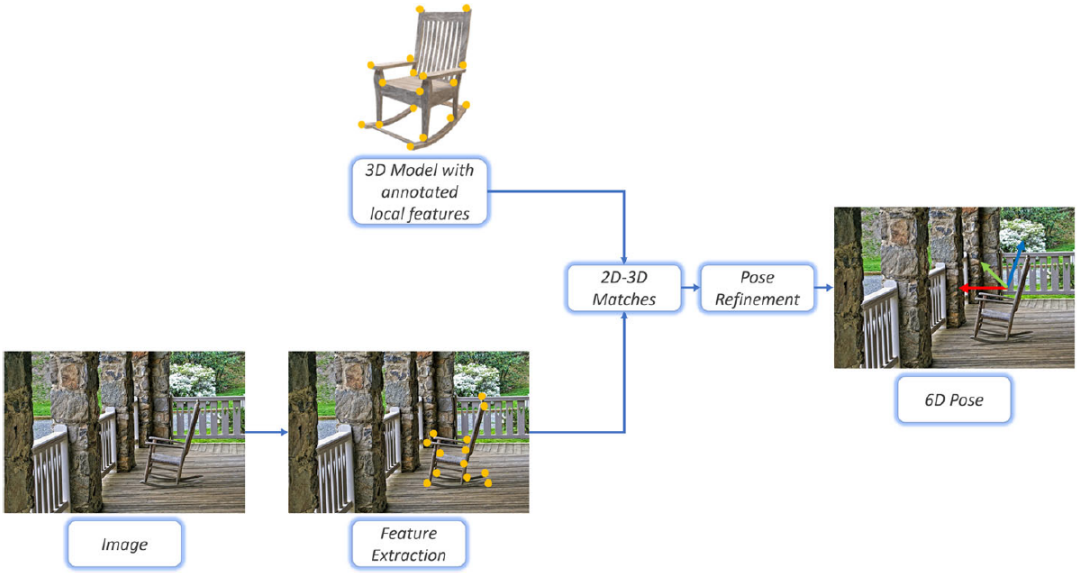


Figura 2.2: Schema dei metodi Feature-based.

Vantaggi

- Sono veloci e robusti alle occlusioni tra oggetti e scene affollate.

Svantaggi

- Gli oggetti dovrebbero avere texture ricche, ben definite e distinte per il calcolo delle caratteristiche locali.
- Non funzionano bene con oggetti simmetrici.
- Di solito, questi metodi richiedono molto tempo per eseguire la stima perché le corrispondenze 2D-3D generano una posizione 6D approssimativa, quindi di solito necessitano di un ulteriore fase per ottenere la posa finale.

2.2.3 Learning-based

Con la diffusione del Deep Learning, i ricercatori hanno migliorato i metodi tradizionali introducendo metodi basati sull'apprendimento, ossia i metodi

learning-based, rendendoli più efficienti e performanti. Questi metodi prevedono la stima della posa 6D utilizzando reti neurali convoluzionali (CNN), richiedendo quindi una fase di addestramento che impiega una quantità considerevole di dati, ma consente alle CNN di apportare miglioramenti significativi sia per la stima della posizione 3D che della rotazione. Possono essere di tipo "one-stage" e "two-stage", a seconda se viene utilizzato un ulteriore algoritmo Pnp per perfezionare i parametri di posa. Lo schema di funzionamento è mostrato alla figura 2.3.

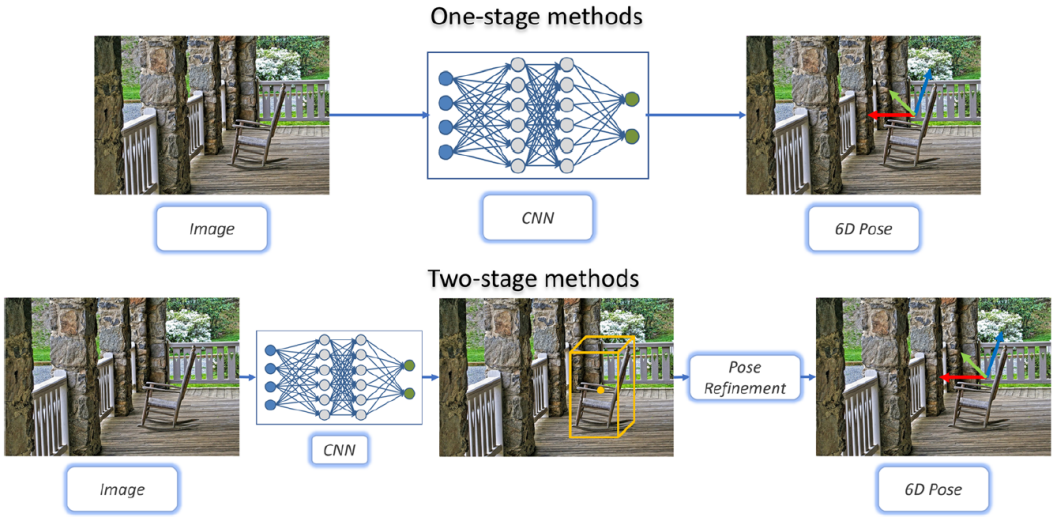


Figura 2.3: Schema dei metodi Learning-based "one-stage" e "two-stage".

Vantaggi

- Sono potenti e possono fornire risultati eccellenti.
- Presentano alte prestazioni anche se l'oggetto è parzialmente occluso o in presenza di scene disordinate.

Svantaggi

- Richiedono un processo di addestramento che necessita di molto tempo.
- Non sono particolarmente robusti quando gli oggetti sono poco visibili poiché è difficile gestire tutte le possibili occlusioni con immagini reali.

3

Analisi PRISMA

In questo capitolo sarà esaminato con attenzione il metodo di ricerca adottato per selezionare i metodi più idonei alla stima della posa 6D in ambienti in cui sono presenti occlusioni, in vista della revisione sistematica.

3.1 PRISMA statement

Per selezionare gli articoli più adatti tra le centinaia di essi presenti nella base di dati riguardanti l'argomento in questione, sono state utilizzate diverse strategie. In particolare sono state seguite le direttive fornite da *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) [1]. Il PRISMA statement rappresenta un insieme di linee guida fondamentali per il reporting accurato di revisioni sistematiche, con l'obiettivo di garantire trasparenza, affidabilità e la possibilità di replicazione della ricerca.
La versione più recente del PRISMA statement [7], distribuita nel 2020, consiste in una checklist composta da 27 punti chiave, progettati per guidare gli autori attraverso un processo rigoroso di descrizione del lavoro svolto.
I punti chiave si suddividono in varie categorie: titolo, abstract, introduzione, metodi, risultati, discussione e altre informazioni. In particolare, nella parte dei risultati è fondamentale fornire il *PRISMA Statement Flowchart*, un diagramma di flusso che viene utilizzato per rappresentare graficamente il processo di selezione degli studi. Per questa revisione sistematica sui metodi per la stima della posa 6D

[1] http://www.prisma-statement.org/

in ambienti con occlusioni, il *PRISMA Statement Flowchart* è illustrato alla figura 3.1.

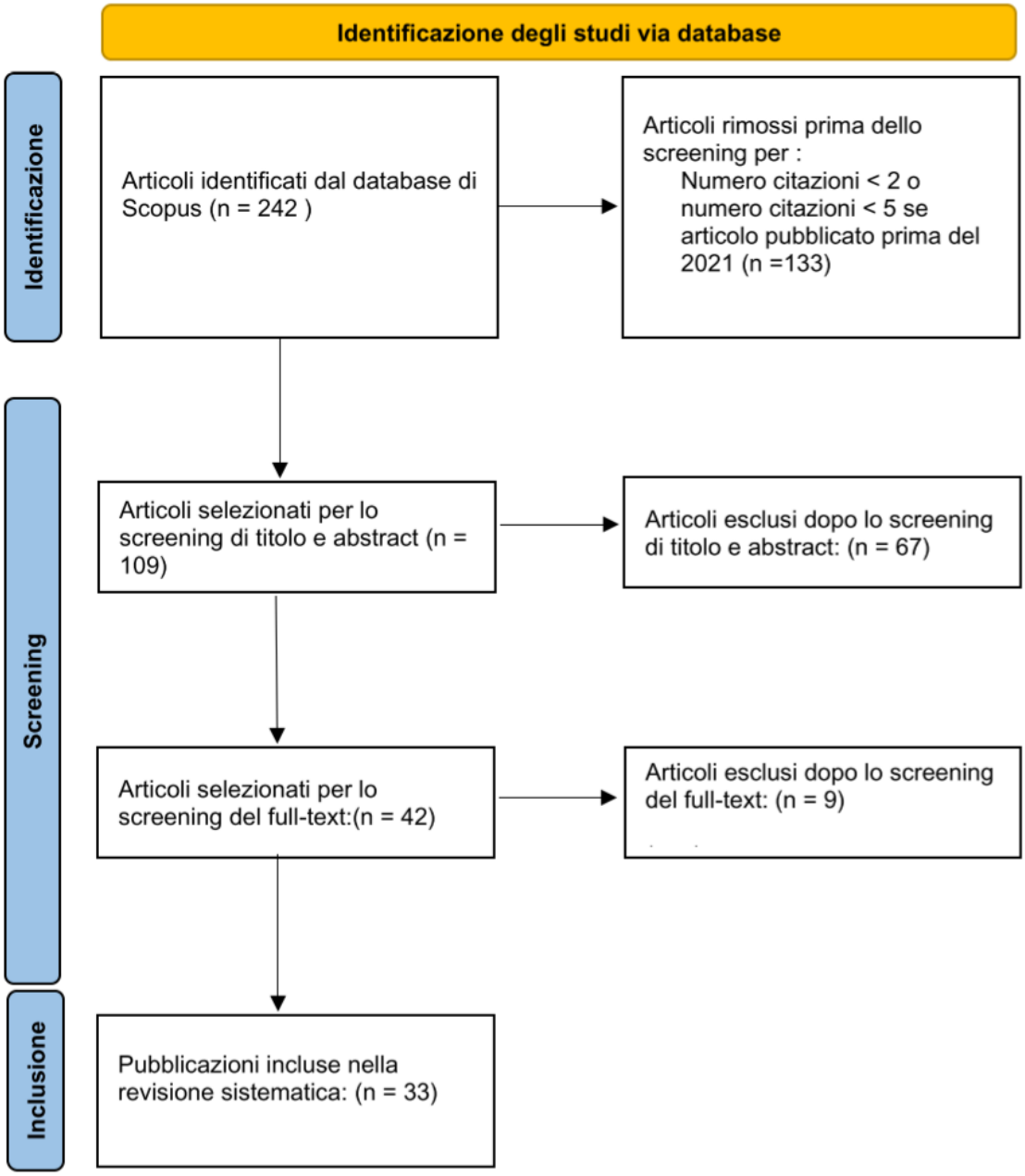


Figura 3.1: Diagramma di flusso PRISMA.

3.2 Database

La ricerca è stata condotta utilizzando la base di dati di Scopus[2], una piattaforma contenente un'ampia gamma di articoli scientifici. A differenza di altri database simili, Scopus offre una vasta gamma di filtri e opzioni di ricerca, rendendola un'opzione ideale per ottenere un'analisi completa delle fonti disponibili. La flessibilità di Scopus ha consentito di raffinare la ricerca e applicare diversi criteri di selezione, quali intervallo temporale delle pubblicazioni, lingua, tipo di documento, autori e riviste specifiche.

3.3 Criterio di ricerca

Il processo di selezione degli articoli è iniziato mediante la scelta della stringa di ricerca ottimale. Dopo aver verificato la sua adeguatezza, è stato valutato il numero di citazioni presenti nei vari documenti individuati. Infine, è stata condotta un'analisi approfondita, iniziando dall'esame dei titoli e degli abstract, per poi approfondire con l'analisi dei full-text disponibili. Solo gli articoli che hanno soddisfatto pienamente tutti i criteri sono stati inclusi nella revisione sistematica.

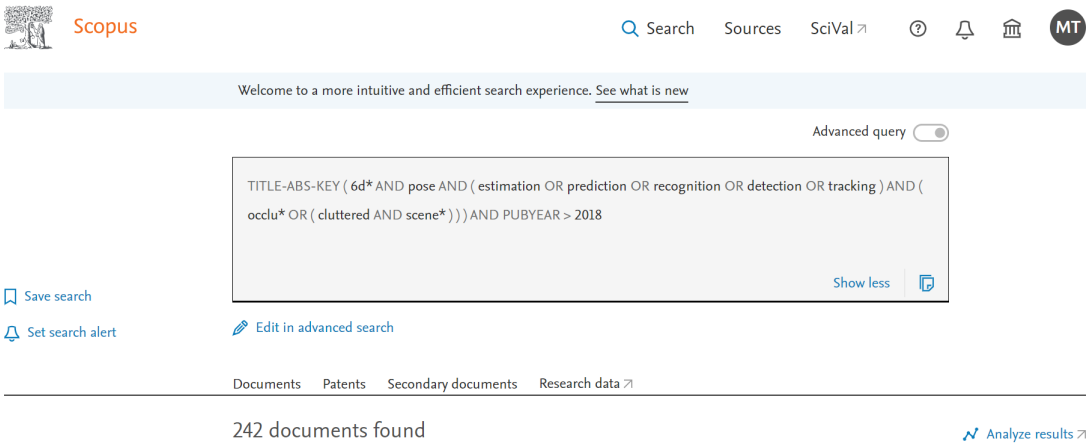


Figura 3.2: Pagina di ricerca di Scopus.

3.3.1 Stringa di ricerca

Dopo svariati tentativi, è stata applicata la seguente stringa di ricerca: "TITLE-ABS-KEY (6d* AND pose AND (estimation OR prediction OR recognition OR

[2] https://www.scopus.com

detection OR tracking) AND (occlu* OR (cluttered AND scene*))) AND PUBYEAR > 2018"[3][4][5], come illustrato nella Figura 3.2. Questa ricerca ha prodotto un totale di 242 risultati (dati aggiornati al 1 agosto 2023). Solamente gli articoli più recenti, dal 2018, escluso, ad oggi, sono stati inclusi nella revisione, in quanto tali pubblicazioni presentano metodologie più all'avanguardia, tenendo conto dell'evoluzione delle tecnologie.

3.3.2 Citazioni

Nella fase successiva del processo di selezione degli articoli, sono stati selezionati solamente gli articoli con un numero di citazioni pari o superiore a 5. Questo criterio di selezione ha permesso di filtrare gli articoli che si riferiscono a contesti di nicchia o applicabili solo a scenari specifici.
Al fine di fornire una panoramica completa e più flessibile, sono stati esaminati anche gli articoli a partire dall'anno 2021, con un numero di citazioni pari o superiore a 2. Questa scelta si è rivelata utile per catturare ulteriori sviluppi recenti nel campo.
Seguendo i criteri precedentemente delineati, sono stati eliminati un totale di 133 articoli, portando alla selezione di 109 articoli.

3.3.3 Abstract e titolo

In seguito, è stato analizzato il contenuto dei titoli e degli abstract dei documenti individuati. Durante questa fase, è emersa la necessità di escludere ulteriori 67 articoli dalla selezione.
Le motivazioni, riassunte nel grafico presente alla figura 3.3, sono state diversificate:

- Dataset: Alcuni articoli descrivevano e fornivano solamente i dataset da utilizzare per la valutazione dei metodi per la stima della posa 6D.
- Duplicati: Sono stati trovati duplicati o metodi che sono stati successivamente aggiornati alla versione più recente.
- Revisioni o report: Altri articoli erano invece revisioni su tematiche simili.

[3]TITLE-ABS-KEY indica che la ricerca è stata effettuata nel titolo, nell'abstract e nelle parole chiave

[4]Il simbolo * rappresenta qualsiasi sequenza di caratteri ed è utile per cercare parole che condividono la stessa radice, ad esempio 6d* include le parole 6d e 6dof.

[5]PUBYEAR indica l'anno di pubblicazione dell'articolo.

- Applicazioni: In alcuni casi, alcuni studi non si sono dedicati in maniera diretta alla sfida della stima della posa 6D. Piuttosto, si sono concentrati sull'impiego degli algoritmi più adeguati per affrontare in modo ottimale tale problema all'interno del contesto applicativo specifico.
- Mancanza di approfondimento sul problema delle occlusioni: Alcuni articoli non erano adeguatamente focalizzati sul problema delle occlusioni.

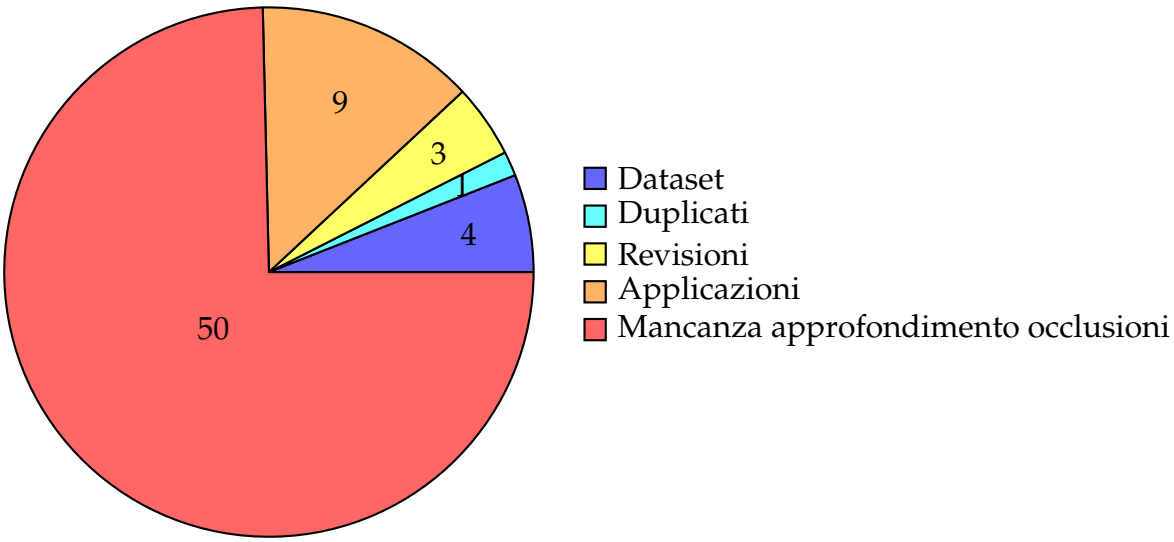


Figura 3.3: Numero di articoli per categoria eliminati dopo l'analisi di abstract e titolo.

3.3.4 Lettura full-text

I rimanenti 42 articoli sono stati sottoposti a una lettura completa. Durante questa fase, 9 articoli sono stati esclusi per varie ragioni:

- Il metodo non è focalizzato nel risolvere il problema delle occlusioni. In particolare o si sofferma solamente su scene con oggetti chiaramente visibili e non sovrapposti, o ha una bassa accuratezza con oggetti occlusi.
- L'articolo propone una applicazione di un metodo per la stima della posa 6D che è efficacie contro le occlusioni. Non viene presentata nessuna nuova tecnica o miglioramento di altri approcci.

Alla fine, nella revisione sistematica sono stati inclusi un totale di 33 articoli rilevanti.

Nel grafico presente alla figura 3.4, è possibile osservare la distribuzione cronologica delle pubblicazioni, suddivise per singolo anno.

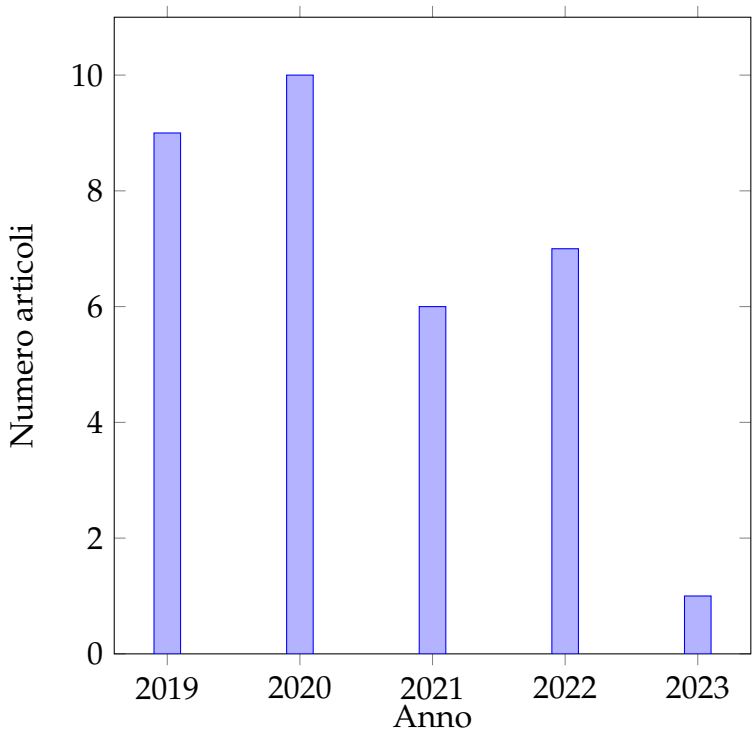


Figura 3.4: Numero di articoli divisi per anno inclusi nella revisione.

4 Risultati

In questo capitolo, vengono esposti i risultati emersi dalla ricerca sistematica che è stata condotta seguendo i criteri definiti nel capitolo precedente. Dei metodi delineati nei documenti presi in esame vengono illustrate le differenti caratteristiche che contraddistinguono tali metodi, allo scopo di agevolarne la comparazione. In conclusione, viene fornita una breve descrizione del funzionamento di queste tecniche.

4.1 Caratteristiche principali dei metodi

Nella Tabella 4.1 sono riportati i vari metodi che sono stati discussi negli articoli selezionati.La disposizione degli articoli nella tabella segue un ordine cronologico, partendo dal più recente, ovvero l'anno 2023, fino al più datato, che risale al 2019. Un aspetto rilevante fornito riguarda il numero di citazioni, che offre una metrica indicativa per valutare l'impatto e il contributo di ciascun articolo all'avanzamento della ricerca nell'ambito. Inoltre, ogni metodo è stato classificato secondo la categoria di appartenenza (vedi 2.2). Inoltre viene fornita un'analisi dettagliata delle caratteristiche più rilevanti che definiscono i diversi metodi presi in considerazione. In particolare, al fine di agevolare un confronto approfondito tra tali approcci, è stata posta l'attenzione sui seguenti parametri chiave:

- **Input**: Indica il tipo di dato che il metodo elabora come input. Generalmente consiste in immagini RGB o RGB-D (composte da un'immagine RGB e una mappa di profondità), modelli CAD, mappe di profondità o Point Cloud.

- **Rete neurale**: Indica se il metodo utilizza una rete neurale e, nel caso affermativo, specifica il tipo di rete neurale impiegata.
- **Dataset**: Indica sui dati su cui il metodo è stato testato. Questa informazione risulta preziosa per valutare l'accuratezza del metodo in diversi contesti e per consentire il confronto tra diversi approcci nello stesso contesto. I principali dataset sono descritti nella tabella 4.2.
- **Real-time**: Specifica se il calcolo della posa avviene in tempo reale o meno. Questo dettaglio può avere un impatto significativo sulle possibili applicazioni dell'algoritmo in vari settori.
- **Applicazione**: Indica il tipo di applicazione per il quale il metodo è stato sviluppato. L'attributo "Generale" indica che l'algoritmo può essere applicato in diversi campi.

Tabella 4.1: Caratteristiche metodi analizzati.

Metodo	Anno	Citazioni	Categoria	Input	Rete neurale	Dataset	Real-time	Applicazione
[8]	2023	3	Learning-based	Point cloud	CNN(include PointNet++ [9]), CNN(include Resnet101 e Resnet50)	S3DIS	NA	Robotica
[10]	2022	2	Feature-based	Point cloud	No	Cvlabs kinect, Princetons red-kitchen e UWA T-rex	No	Generale
[11]	2022	5	Learning-based	Immagine RGB	Rete encoder-decoder, Darknet-53 [12]	Occlusion Line-mod	Sì	Generale
[13]	2022	6	Learning-based	Immagine RGB/RGB-D	Rete encoder-decoder, SphereCNN, GCN	YCB-Video, Li-nemod	NA	Generale

Tabella 4.1 – continua

Metodo	Anno	Citazioni	Categoria	Input	Rete neurale	Dataset	Real-time	Applicazione
[14]	2022	2	Learning-based	Immagine RGB-D	Faster-RCNN	Shelf-Tote, Extended Rutgers RGBD, Linemod, Occlusion Linemod	No	Generale
[15]	2022	27	Feature-based	Immagine RGB	Pixel-wise Voting Network	Linemod, Occlusion Linemod, Truncated Linemod, T-LESS, YCB-Video	Sì	Generale
[16]	2022	3	Feature-based	Immagine RGB-D	Pixel-level Attention CNN, RandLA-like network	YCB-Video, Linemod, Occlusion Linemod	Sì	Generale
[17]	2022	2	Feature-based	Immagine RGB	PVNet	NA	Sì	Robotica
[18]	2021	30	Learning-based	Immagine RGB	Rete encoder-decoder	YCB-Video, Linemod, Occlusion Linemod	No	Generale

Tabella 4.1 – continua

Metodo	Anno	Citazioni	Categoria	Input	Rete neurale	Dataset	Real-time	Applicazione
[19]	2021	11	Feature-based	Immagine RGB, Point Cloud	RetinaNet, PPF-MEAM	NA	NA	Robotica
[20]	2021	15	Feature-based	Immagine RGB-D	CNN con 4 sub-networks PoitNet++	T-LESS, Occlusion Linemod, Nocs-real275, ShapeNetPose	NA	Generale
[21]	2021	83	Learning-based	Immagine RGB	GDR-Net	YCB-Video, Linemod, Occlusion Linemod	Sì	Generale
[22]	2021	28	Feature-based	Immagine RGB-D	Transductive-VOS network, LF-Net	NOCS, YCBInEOAT	Sì	Generale
[23]	2021	9	Feature-based	Immagine RGB	Rete encoder-decoder, PointNet++, ResNet18, PSPNet	Linemod, Occlusion Linemod	No	Generale
[24]	2020	18	Learning-based	Immagine RGB-D	YOLO V3, PointPoseNet	Linemod, Occlusion Linemod	NA	Generale

Tabella 4.1 – continua

Metodo	Anno	Citazioni	Categoria	Input	Rete neurale	Dataset	Real-time	Applicazione
[25]	2020	7	Feature-based	Immagine RGB	No	OPT, RBOT	Sì	Realtà aumentata
[26]	2020	126	Learning-based	Immagine RGB	ResNet	Linemod, Occlusion Linemod	Sì	Generale
[27]	2020	7	Learning-based	Immagine RGB-D	Mask R-CNN, DenseFusion	NA	NA	Robotica
[28]	2020	44	Learning-based	Immagine RGB-D, modello CAD	Mask-RCNN, ResNet18 Point-wise encoder, 3D-CNN, DenseFusion	YCB-Video, Cluttered YCB	Sì	Robotica
[29]	2020	16	Feature-based	Immagine RGB-D, modello CAD	No	NA	No	Robotica
[30]	2020	42	Learning-based	Immagine RGB-D, modello CAD	2 encoder e CNN per la stima	YCB-Video, YC-BInEOAT	Sì	Robotica

Tabella 4.1 – continua

Metodo	Anno	Citazioni	Categoria	Input	Rete neurale	Dataset	Real-time	Applicazione
[31]	2020	7	Learning-based	Immagine RGB	CNN (include Mask R-CNN)	YCB-Video, Occlusion Linemod	Sì	Robotica
[32]	2020	14	Template-based	Immagine RGB-D	No	Doumanoglou	NA	Generale
[33]	2020	10	Feature-based	Point Cloud, modello CAD	No	NA	No	Generale
[34]	2019	29	Learning-based	Point Cloud	Point-wise Pose Regression Network (include PointNet++)	Siléane	Sì	Robotica
[35]	2019	5	Learning-based	Immagine RGB-D, maschera di segmentazione	2 versioni: PointNet o rete Dynamic Nearest Neighbour Graph (DG), PoseCNN	YCB-Video	NA	Generale
[36]	2019	9	Learning-based	Immagine RGB/RGB-D	Mask R-CNN, rete encoder-decoder	Occlusion Linemod	NA	Generale

Tabella 4.1 – continua

Metodo	Anno	Citazioni	Categoria	Input	Rete neurale	Dataset	Real-time	Applicazione
[37]	2019	175	Learning-based	Immagine RGB	Rete encoder-decoder	YCB-Video, Occlusion Linemod	Sì	Generale
[38]	2019	6	Learning-based	Immagine RGB-D	ResNet-50, classificatore WILDCAT	YCB-video, Occlusion Linemod	Sì	Generale
[39]	2019	228	Learning-based	Immagine RGB	Architettura auto-encoder	Linemod, Occlusion Linemod, T-LESS	No	Generale
[40]	2019	17	Learning-based	Mappa di profondità	CNN, Faster RCNN, GossipNet	Doumanoglou, Siléane	NA	Robotica
[41]	2019	52	Feature-based	Immagine RGB, modello 3D	No	OPT, RBOT	Sì	Realtà aumentata
[42]	2019	268	Learning-based	Immagine RGB-D	CNN con una Mask R-CNN modificata	CAMERA 25, REAL275, Occlusion Linemod	Sì	Generale

Di seguito viene fornita una breve descrizione dei principali dataset utilizzati dai metodi della tabella 4.1.

Tabella 4.2: Descrizione dataset principali.

Nome	Descrizione
Linemod [43]	Include 15 oggetti domestici di vari colori, forme e dimensioni. Presenta molte sfide per la stima della posa: scene disordinate, occlusioni, oggetti senza texture e poca luminosità.
YCB-Video [44]	Contiene 21 oggetti con un totale di 133,827 immagini.
Occlusion Linemod [5]	Creato basandosi sulle immagini del dataset Linemod. Ogni immagine contiene più oggetti in un ambiente fortemente occluso.
Truncated Linemod [15]	Ottenuto ritagliando in modo casuale le immagini nel set di dati Linemod. La risoluzione di ogni immagine ritagliata è 256x256.
T-LESS [45]	Consiste di trenta oggetti, che non hanno segni significativi e nessun colore discriminante. Fornisce 39k immagini di addestramento e 10k immagini di prova.
Shelf-Tote [46]	Contiene oltre 7.000 immagini RGB-D che coprono 477 scene a una risoluzione 640x480. Offre 148 diverse configurazioni di posizionamento degli oggetti, collocati su uno scaffale. Le configurazioni includono occlusioni tra oggetti.
Siléane [47]	Rappresenta tipici scenari industriali di bin-picking dove oggetti di un singolo tipo si sovrappongono casualmente.
OPT [48]	Contiene 6 oggetti e 552 sequenze di immagini con un numero totale di 79.968 fotogrammi a una risoluzione di 1920x1080 . Il set di dati copre diversi movimenti della telecamera e condizioni di illuminazione. Tuttavia, non include occlusioni.
RBOT [41]	Contiene 18 oggetti e 72 sequenze di immagini. Gli oggetti variano in forma e dimensioni. Ogni sequenza di immagini contiene 1000 fotogrammi con una risoluzione di 640x512. Per ogni oggetto il dataset compone diverse varianti di sequenze di immagini includendo casi con occlusioni.

Tabella 4.2 – continua

Nome	Descrizione
Doumanoglou [49]	Contiene un set di addestramento, formato da 4740 immagini, e un set di test, formato da 177 immagini.
YCBInEOAT [30]	Contiene 9 sequenze video catturate da una telecamera RGB-D statica, mentre gli oggetti sono mossi. Coinvolge un totale di 5 oggetti.
Cluttered YCB [28]	Abbiamo utilizzato un simulatore per posizionare gli oggetti nelle possibili pose in modo casuale. Questo set di dati ha 1200 scene e 15 frame per ciascuna.
Nocs-real275 [50]	Il set di test contiene oggetti, mai visti nella precedente fare, di cinque categorie con forma e dimensioni diverse da quelle del training set.
ShapeNetPose [20]	Contiene immagini RGB-D di oggetti appartenenti a 22 categorie.
S3DIS [51]	Contiene punti da 6 diverse aree e 272 stanze. Ogni punto presenta una etichetta che rappresenta un istanza, appartenente a una delle 13 categorie.

4.2 Descrizione dei metodi

Nella tabelle 4.3 è stata presentata, seguendo anche in questo caso un ordine cronologico decrescente, una breve descrizione del funzionamento di ciascun metodo, concentrandosi in particolare sul processo impiegato per stimare la posa 6D.

Tabella 4.3: Descrizione dei metodi.

Metodo	Descrizione
[8]	Formato da due moduli, il primo per la segmentazione delle istanze, il secondo per la stima della posa 6D. Infine quest'ultima è perfezionata con un algoritmo ICP.
[10]	Un fase offline dove le caratteristiche delle coppie di punti sono salvate in una tabella hash, e una fase online dove si ottengono le più probabili corrispondenze attraverso un sistema di voto *Hough-like voting scheme*. Viene infine usato un algoritmo ICP.

Tabella 4.3 – continua

Metodo	Descrizione
[11]	Composto da un encoder comune e da due decoder, per la segmentazione e per la regressione. I due output sono utilizzati per stimare la posa finali in un ultimo modulo.
[13]	Diviso in 4 parti: segmentazione, predizione della rotazione, GCN per la predizione della struttura 3D, la stima della posa.
[14]	Dalla rete neurale viene estratta la segmentazione degli oggetti nella scena. Vengono successivamente formate delle ipotesi per la posa. La migliore viene infine scelta come output. Include successivamente un auto allenamento della rete.
[15]	Utilizza la rete PVNet per la localizzazione dei punti chiave e per etichettare gli oggetti. Usa poi un algoritmo Pnp modificato per determinare la posa finale. Può includere un perfezionamento con l'algoritmo ICP.
[16]	Utilizza due reti neurali per estrarre le caratteristiche geometriche e dell'aspetto dell'oggetto. Poi sono combinate per prevedere il vettore *point-wise* con gli oggetti etichettati.Successivamente viene stimata la posa.
[17]	Utilizza una rete PVNet e successivamente un algoritmo Pnp per la stima della posa. Confronta infine il risultato con una mappa di profondità se disponibile.
[18]	L'encoder estrae le caratteristiche principali, processate da due decoder (inclusa l'analisi del self-occlusion). I due output sono poi utilizzati per stimare la posa.
[19]	Utilizza RetinaNet e l'algoritmo Canny edge per determinare le caratteristiche 3D dell'oggetto e i bordi. Il risultato viene passato come input, insieme al Point Cloud, alla rete PPF-MEAM che ne stima la posa finale.
[20]	All'inizio gli oggetti vengono rilevati nell'immagine e successivamente categorizzati in base alla loro simmetria. In base a quest'ultima viene infine stimata la posa finale, grazie a un algoritmo che analizza le caratteristiche geometriche dell'oggetto.
[21]	Dall'immagine di input la rete prevede le caratteristiche geometriche degli oggetti. Poi Patch-PnP ne stima direttamente la posa.

Tabella 4.3 – continua

Metodo	Descrizione
[22]	La prima rete segmenta gli oggetti, la seconda ne estrae le caratteristiche che vengono poi vengono utilizzate per determinare la posizione e l'orientamento dell'oggetto grazie a un grafo per l'ottimizzazione della posa.
[23]	Combina le caratteristiche del colore e la Point Cloud degli oggetti per poi utilizzare PointNet++ per determinare le caratteristiche globali e locali, poi fuse insieme per stimare la posa.
[24]	Attraverso la rete neurale convoluzionale vengono determinate la segmentazione degli oggetti e la posizione 3D dei punti chiave. Da questi vengono generate ipotesi della posa 6D, dalle quali viene alla fine selezionata la migliore.
[25]	Inizialmente vengono determinati i bordi degli oggetti attraverso l'algoritmo Canny Edge Detector. Poi, considerando solo quelli attendibili, viene determinata la posa con un algoritmo IRLS (Iteratively reweighted least squares).
[26]	Grazie alla rete neurale ResNet determina i punti chiave, i bordi e le simmetrie. Vengono utilizzati poi altri due moduli(uno per il perfezionamento) per definire la posa.
[27]	Ottiene inizialmente, dalle reti neurali convoluzionali, la stima della posa per poi migliorarla confrontandola con le caratteristiche geometriche nella scena. Grazie poi a un grafo determina la posa più attendibile.
[28]	Diviso in 4 parti: creazione di una mappa volumetrica, determinazione posa iniziale, ridefinizione controllando eventuali incongruenze con il modello CAD, sostituzione della posa con il modello CAD.
[29]	Prima viene stimata la configurazione della mano robotica e vengono generate una serie di ipotesi di posa dell'oggetto dal modello CAD. Viene poi valutata la posa più adatta basandosi sulla posizione della mano robotica.
[30]	L'immagine RGB-D e il modello CAD vengono elaborati separatamente in una nuova rete neurale formata da due encoder. Il loro output viene concatenato per la stima finale della transazione e della rotazione.

Tabella 4.3 – continua

Metodo	Descrizione
[31]	Una CNN estrae le principali caratteristiche che vengono analizzate separatamente per calcolare rotazione, traslazione, maschera, classe e box degli oggetti. Con questi dati viene calcolata la posa e applicato un algoritmo ICP.
[32]	L'algoritmo migliora il modello LineMod, ispirandosi all'illusione Muller-Lyer, per gestire le occlusioni. Vengono creati template nella fase di training. L'immagine di input poi viene divisa in template e comparata con quelli di training per la stima della posa.
[33]	Inizialmente vengono estratte le caratteristiche dei bordi dai modelli CAD e memorizzati in una tabella hash. Dall'immagine di input si trovano le migliori corrispondenze nella tabella e infine viene determinata la posa dall'algoritmo SVD.
[34]	Per ogni punto dell'input determina la posa dell'oggetto a cui appartiene. Successivamente vengono comparate e determinata la posa finale per ogni oggetto.
[35]	Viene stimata la transazione dell'oggetto attraverso una rete PoseCNN e crea un Point Cloud, che poi è utilizzato come input a un'altra rete neurale convoluzionale che ne stima la rotazione.
[36]	Inizialmente vengono segmentati gli oggetti dell'immagine con una rete Mask R-CNN. Di questi vengono poi determinate le coordinate 3D dei punti attraverso una CNN. Infine è stimata la posa e perfezionata attraverso una ottimizzazione geometrica.
[37]	La rete è formata da un encoder in comune e due decoder, uno per la segmentazione e uno per determinare le coordinate 2D dei punti chiave. La posa viene determinata tramite l'algoritmo EPnP.
[38]	Le caratteristiche dell'immagine RGB-D di input vengono estratte nella prima parte. Vengono poi create delle heatmap per ogni classe. La posa viene infine determinata grazie al metodo Stochastic Congruent Sets [52].
[39]	Viene inizialmente determinata la maschera e il box di selezione degli oggetti, poi le sue coordinate 3D e la predizione degli errori, e infine la posa grazie all'algoritmo PnP.

Tabella 4.3 – continua

Metodo	Descrizione
[40]	Formato da tre moduli, eseguiti sequenzialmente: rilevamento bounding box 2D degli oggetti, stima posa 6D e registrazione congiunta. In particolare per la posa 6D viene stimata la coordinata 2D del centro, la posa 3D e la profondità.
[41]	Prevede una segmentazione statistica degli oggetti nelle immagini. La posa viene stimata elaborando e confrontando i dati con uno modello 3D conosciuto e con la posa stimata precedentemente, attuando uno schema di ottimizzazione Gauss-Newton.
[42]	La CNN stima l'etichetta della classe, la maschera delle istanze e la mappa NOCS (Normalized Object Coordinate Space) direttamente dall'immagine RGB. La mappa di profondità e la mappa NOCS sono utilizzati per stimare la posa 6D.

5

Discussione

In questo capitolo vengono discussi i risultati ottenuti dalla ricerca sistematica, presentati nel capitolo precedente. Grazie a questi dati siamo in grado di comparare i vari metodi e valutare quali si comportano meglio in situazioni in cui nella scena sono presenti occlusioni. In particolare vengono discusse le caratteristiche principali dei metodi e le categorie che meglio si adattano a questo contesto, fornendo un esempio esplicativo per ognuna.
Per confrontare le varie tecniche tra di loro è fondamentale testarle con un dataset comune. A questo scopo molti metodi sono stati valutati utilizzando i dataset più importanti (vedi figura 5.1). In questo modo è possibile confrontare le loro performance in modo più accurato.

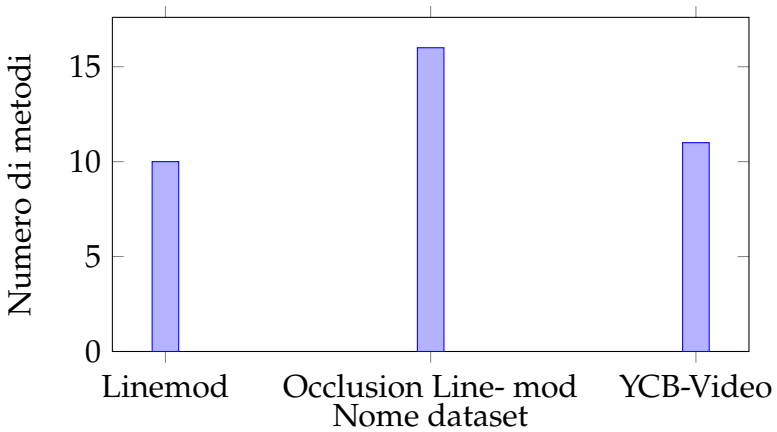


Figura 5.1: Dataset più utilizzati.

5.1 Analisi delle caratteristiche dei metodi

Le varie caratteristiche presentate alla tabella 4.1 ci permettono di trovare elementi condivisi dei metodi presi in considerazione e quindi capire quali sono i principali attributi che permettono a una tecnica di ottenere prestazioni superiori in presenza di occlusioni. In questo senso è importante focalizzarsi in particolare sui dati di input e sull'utilizzo delle reti neurali.

5.1.1 Input

Innanzitutto va sottolineato che la maggioranza di queste richiede l'utilizzo di dati spaziali tridimensionali. Questi dati, spesso ottenuti da fonti come immagini RGB-D o Point Cloud, sono essenziali per ottenere stime di posizione e orientamento accurate degli oggetti in uno spazio tridimensionale. Rispetto alle immagini bidimensionali, i dati 3D apportano un notevole incremento di precisione nelle stime di posa. Essi catturano le relazioni spaziali tra gli oggetti e permettono ai vari algoritmi di considerare non solo l'aspetto superficiale, ma anche caratteristiche come la profondità. Questo maggiore livello di informazione è particolarmente vantaggioso in scenari caratterizzati da occlusioni parziali o complete, poiché consente una migliore discriminazione degli oggetti anche quando sono parzialmente nascosti dalla vista.

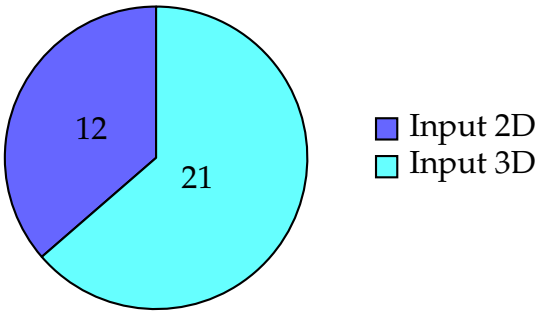


Figura 5.2: Numero di metodi per categoria.

Tuttavia, è fondamentale riconoscere che l'introduzione di dati 3D porta con sé una complessità aggiuntiva. La necessità di acquisire, elaborare e gestire informazioni tridimensionali richiede tecnologie più sofisticate rispetto alle classiche immagini 2D. Questo può tradursi in una fase di preparazione dei dati più impegnativa e in algoritmi di stima più complessi da implementare.

5.1.2 Reti neurali

Esaminando i risultati ottenuti, emerge chiaramente come l'applicazione delle reti neurali stia acquisendo un ruolo sempre più cruciale. Infatti solamente 5 metodi sui 33 considerati non ne fanno uso.
Con l'avvento del Deep Learning, l'impiego delle reti neurali, specialmente delle reti neurali convoluzionali (CNN), sta aumentando in modo significativo. Alcuni metodi utilizzano le CNN per la previsione diretta della posa 6D, come ad esempio in se(3)-TrackNet [30]. Altri, invece, le utilizzano per risolvere specifici problemi all'interno del processo di stima della posa: ci sono fasi che, attraverso l'uso di algoritmi classici, risulterebbero più complesse da affrontare (un esempio è PANet [16] che inizialmente estrae dai dati di input le caratteristiche degli oggetti attraverso una CNN).

5.1.3 Real-time

Nonostante non sia un aspetto intrinseco che contribuisce a un miglioramento dell'accuratezza nella stima della posa 6D, è comunque interessante osservare come quasi la metà dei metodi, 16 sui 33 considerati, consentano il rilevamento della posa in tempo reale. Questo aspetto riveste una notevole importanza poiché dimostra che, nonostante la sfida rappresentata dalla stima della posa in presenza di occlusioni, è possibile ottenere risultati di elevata qualità senza compromettere la velocità di esecuzione. Questo comporta una maggior possibilità di impiego di tali metodi in svariati contesti, anche in quelli in cui la tempestività del rilevamento gioca un ruolo fondamentale.

5.2 Analisi delle categorie dei metodi

Dalla tabella 4.1 è possibile osservare quali sono le categorie di metodi più utilizzati in questo contesto. I risultati di tali osservazioni sono illustrati alla figura 5.3. Possiamo constatare, come previsto dalle analisi preliminari presentati al capitolo 2.2, che quasi la totalità dei metodi che affrontano in modo efficiente il problema delle occlusioni appartiene alle categorie Learning-based e Feature-based. Tuttavia, è interessante notare che anche un metodo Template-based ha ugualmente conseguito risultati promettenti.

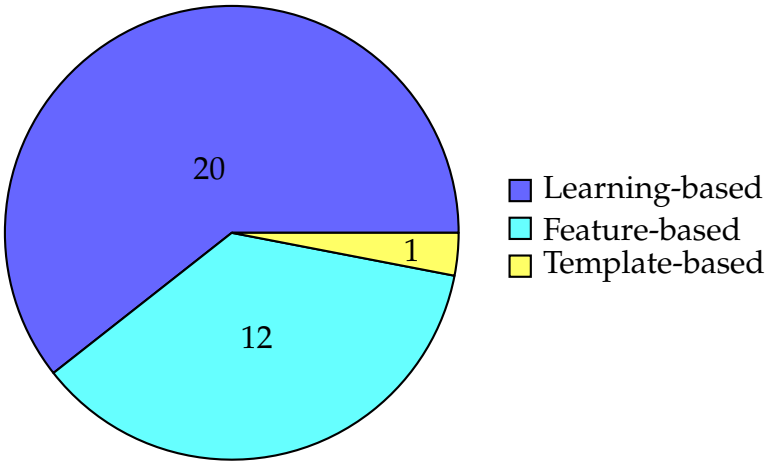


Figura 5.3: Numero di metodi per categoria.

5.2.1 Learning-based: Segmentation-driven

I metodi Learning-based sono quelli più utilizzati in questo contesto. Sono infatti in grado, dopo un accurato addestramento della rete neurale con varie scene differenti, di apprendere caratteristiche che permettono di identificare un oggetto anche in presenza di occlusioni. Sono inoltre in grado di generalizzare: possono rilevare occlusioni diverse rispetto a quelle presenti nelle immagini utilizzate per il training.
Tra quelli presentati nei risultati della ricerca sistematica, un esempio di metodo learning-based, in particolare di tipo "two-stage" (vedi paragrafo 2.2.3), è il metodo Segmentation-driven [37]. In questo approccio è stata sviluppata un'architettura di rete neurale convoluzionale (CNN) composta da due flussi distinti: uno dedicato alla segmentazione, che prevede l'etichettatura dell'oggetto osservato in ciascuna posizione dell'immagine, e l'altro focalizzato sulla stima della posizione bidimensionale dei punti chiave (come illustrato alla figuare 5.4).
I due flussi condividono un'unità di codifica (encoder), ma presentano due unità di decodifica separate. A differenza di altri metodi che effettuano previsioni globali per ciascun oggetto, in questo caso specifico, porzioni individuali dell'immagine prevedono a quale oggetto appartengono e dove sono localizzati i punti chiave. In seguito, le previsioni di tutte le porzioni vengono integrate per ottenere una stima più accurata, completando il processo con l'utilizzo di un algoritmo PnP. Questa tecnica si dimostra particolarmente efficace nell'affrontare occlusioni, conferendo una maggiore robustezza nell'estrazione delle informazioni utilizzate per stimare la posa.

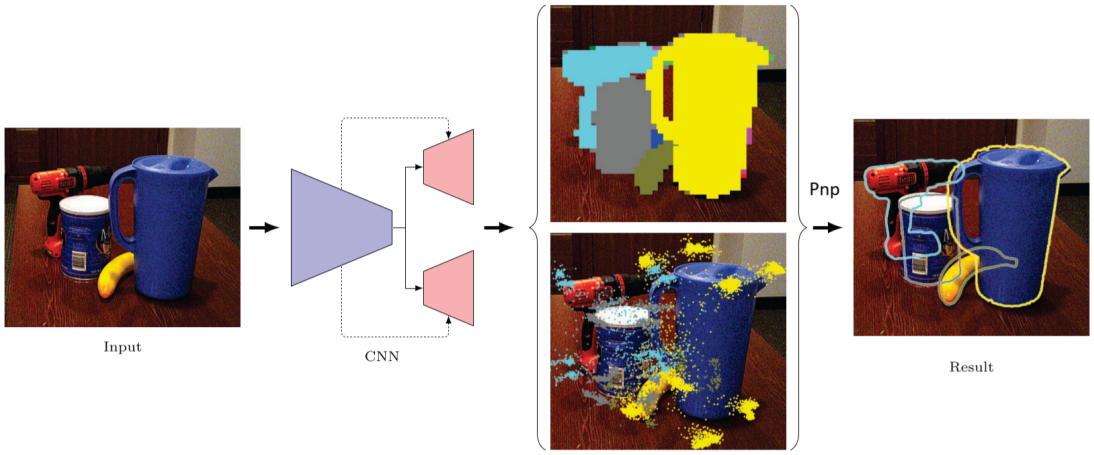


Figura 5.4: Schema di funzionamento del metodo Segmentation-driven [37].

5.2.2 Feature-based: PVNet

Anche le tecniche feature-based hanno ottime performance in questo scenario. Questi metodi infatti sfruttano caratteristiche specifiche e locali, spesso elementi geometrici come contorni o bordi, che permettono di ricostruire la posa dell'oggetto anche senza una sua visione completa. Riescono quindi a estrarre le informazioni necessarie solamente analizzando la parte visibile.
Un esempio di metodo feature-based è PVNet [15]. Si tratta di un framework per la stima della posa 6D che utilizza una Pixel-wise Voting Network, seguita da un algoritmo Pnp, come illustrato alla figura 5.5.

Invece di stimare direttamente le coordinate nell'immagine dei punti chiave, la

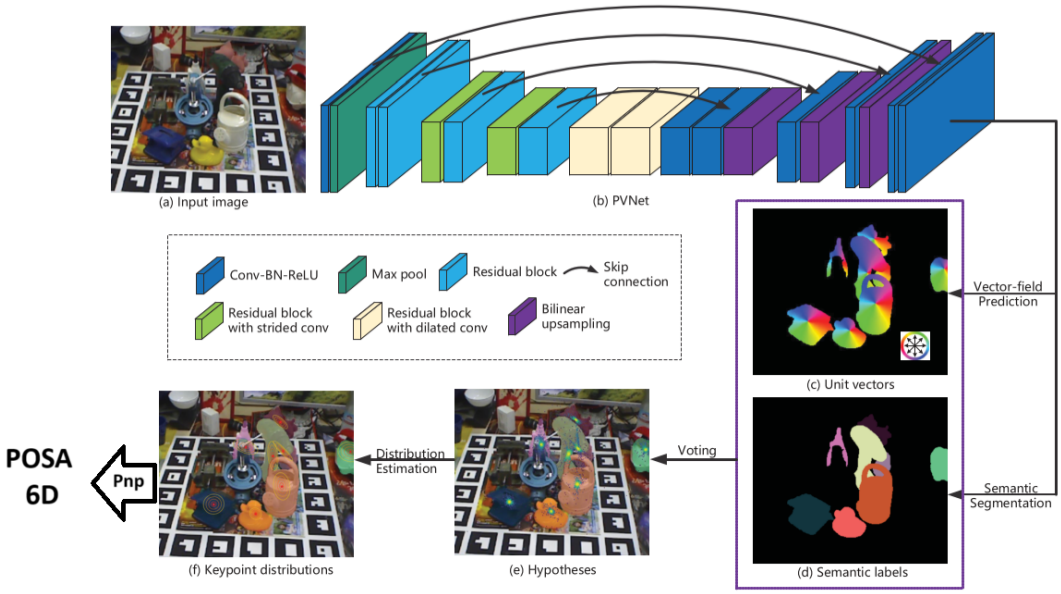


Figura 5.5: Schema di funzionamento del metodo PVNet [15].

PVNet determina i vettori che rappresentano le direzioni da ciascun pixel dell'oggetto verso i keypoints. Queste direzioni votano quindi per le posizioni dei punti chiave basandosi su RANSAC [53]. Questo schema di voto è basato su una proprietà degli oggetti rigidi: una volta che osserviamo alcune parti locali, siamo in grado di dedurre le direzioni relative verso altre parti. Questo approccio crea essenzialmente una rappresentazione vettoriale per la localizzazione dei punti chiave. In seguito, grazie ad un algoritmo Pnp, in particolare l'algoritmo EPnp [6], viene stimata la posa 6D finale degli oggetti.
A differenza delle rappresentazioni basate su coordinate, imparare tale rappresentazione costringe la rete a concentrarsi sulle caratteristiche locali degli oggetti e sulle relazioni spaziali tra le parti dell'oggetto. Di conseguenza, la posizione di una parte invisibile può essere dedotta dalle parti visibili. Inoltre, questa rappresentazione è in grado di rappresentare anche punti chiave dell'oggetto nascosti o che si trovano al di fuori dell'immagine. Questo permette una grande accuratezza anche quando si hanno occlusioni tra oggetti o dove una parte di essi non è visibile.

5.2.3 Template-based: CT-LineMod

È importante sottolineare che anche un metodo template-based ha ottenuto buoni risutati con le occlusioni, nonostante queste tecniche non siano generalmente considerati adatti a questo scopo.
In particolare CT-LineMod [32] ha ottenuto una buona accuratezza, migliorando il metodo LineMod [43]. È stato utilizzando un approccio locale: l'immagine è divisa in più porzioni e per ognuna di queste è determinato il template più adatto, tra quelli nel batabase iniziale. La posa finale è quindi la composizione di tutti i template locali. Il processo che porta alla stima della posa è illustrato alla figura 5.6. Questa tecnica di analizzare gli oggetti localmente rende il metodo più accurato quando nella scena sono presenti occlusioni.

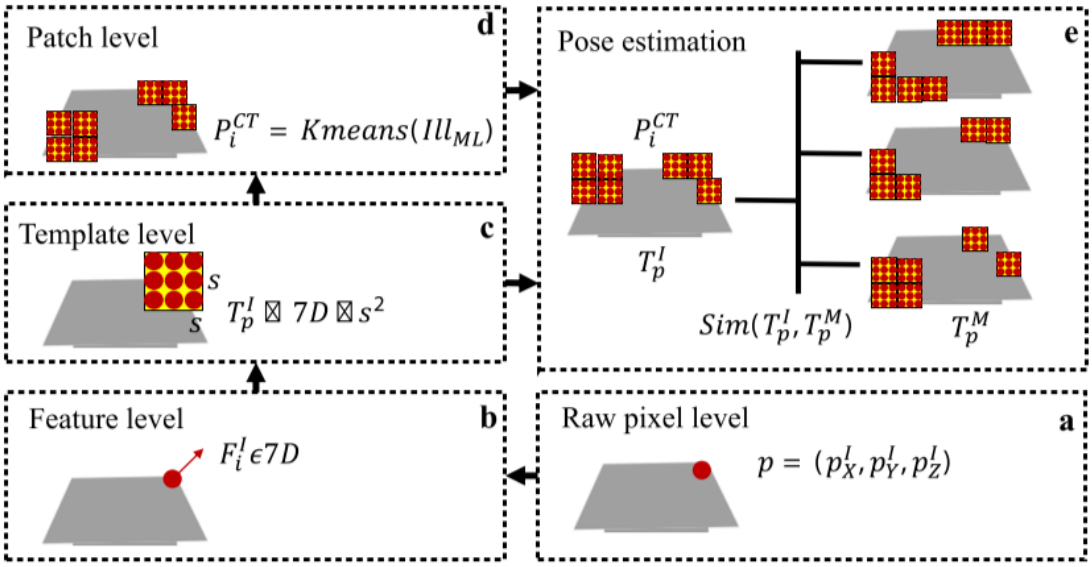


Figura 5.6: Schema di funzionamento del metodo CT-LineMod [32].

6

Conclusioni e lavori futuri

Recentemente l'importanza dello sviluppo di tecniche per la stima della posa 6D degli oggetti è cresciuta in modo significativo all'interno del campo della Computer Vision. Nel corso degli ultimi anni, sono stati sviluppati molti metodi al fine di affrontare questa sfida e la diffusione del Deep Learning ha contribuito in maniera sostanziale a rendere questi metodi sempre più accurati e precisi.
Molte sono però le sfide che si incontrano in questo ambito, come ad esempio la forma degli oggetti, la composizione della scena o la disponibilità delle tecnologie impiegabili. In particolare questa tesi si è focalizzata nel problema delle occlusioni.
Attraverso la revisione sistematica proposta in questa tesi, effettuata secondo i criteri del protocollo PRISMA, è stata proposta una panoramica sullo stato dell'arte di questi metodi. L'analisi ha permesso di comprendere quali sono i metodi con le prestazioni migliori in questo ambito e quali sono le loro principali caratteristiche. In questo modo è possibile individuare le strategie più performanti e che ottengono risultati più precisi, le quali possono servire come base per nuovi metodi che verranno sviluppati in futuro. In particolare è emerso che i metodi learning-based e quelli feature-based sono i più adatti in questo contesto: quasi la totalità dei metodi analizzati appartengono a queste categorie. Inoltre si può notare che la presenza di reti neurali può drasticamente migliorare le performance, anche se richiedono un notevole volume di dati per l'addestramento.
Nonostante le tecnologie attuali siano in grado di effettuare una stima accurata, da questa revisione sono emerse anche diverse limitazioni che possono essere un punto di partenza per ulteriori sviluppi futuri in questo ambito.
Una di queste è l'input utilizzato per l'elaborazione: in svariati casi è necessario

un vasto numero di dati per ottenere una buona precisione. Tuttavia, questo non è sempre possibile in tutti i contesti applicativi. Infatti ci possono essere casi in cui la disponibilità di dati è limitata.
Un'altra limitazione è il tempo di esecuzione della stima: non tutti i metodi sono in grado di determinare la posa in tempo reale. Questo è molto limitante in molte applicazioni, in particolare nella campo della realtà aumentata e della guida autonoma, dove la rapidità di esecuzione svolge un ruolo fondamentale.
Sarà quindi importante affrontare queste sfide per sviluppare dei metodi più performanti, accurati e versatili, applicabili quindi in più ambiti, che con il rapido sviluppo della tecnologia saranno destinati ad aumentare.

Bibliografia

[1] C Eppner, S Höfer e R Jonschkowski. «Lessons From The Amazon Picking Challenge: Four Aspects of Building Robotic System». In: *IJCAI. 2017:4831-4835.* (2017).

[2] R. Hachiuma e H. Saito. «Recognition and pose estimation of primitive shapes from depth images for spatial augmented reality.» In: *In Proceedings of the 2016 IEEE 2ndWorkshop on Everyday Virtual Reality (WEVR), Greenville, SC, USA, 2020 March 2016; pp. 3235.* (2016).

[3] Chen X, Ma H e Wan J. «Multi-view 3d object detection network for autonomous driving.» In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017* (2017).

[4] R. Gu, G. Wang e J. Hwang. «Efficient Multi-person Hierarchical 3D Pose Estimation for Autonomous Driving.» In: *n Proceedings of the 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), San Jose, CA, USA, 2830 March 2019; pp. 163168.* (2019).

[5] Brachmann E., Krull A. e Michel F. «Learning 6D object pose estimation using 3D object coordinates». In: *Proc. Eur. Conf. Comput. Vis., 2014, pp. 536551* (2014).

[6] V. Lepetit, M. Moreno-Noguer e P. Fua. «EPnP: An Accurate O(n) Solution to the PnP Problem». In: *International Journal of Computer Vision. 81 (2): 155166* (2008).

[7] Page Matthew J, McKenzie Joanne E e Bossuyt Patrick M. «The PRISMA 2020 statement: an updated guideline for reporting systematic reviews». In: *BMJ 2021;372:n71* (2020).

[8] C. Zhuang, S. Li e H. Ding. «Instance segmentation based 6D pose estimation of industrial objects using point clouds for robotic bin-picking». In: *Robotics and Computer-Integrated Manufacturing 82,102541* (2023).

[9] Qi C.R. et al. «PointNet++: deep hierarchical feature learning on point sets in a metric space». In: *Conference on Neural Information Processing Systems, Long Beach, USA, 2017, pp. 50995108* (2017).

[10] Z. Ge et al. «A Fast Point Cloud Recognition Algorithm Based on Keypoint Pair Feature». In: *Sensors 22(16),6289* (2022).

[11] W.-L. Huang, C.-Y. Hung e I.-C. Lin. «Confidence-Based 6D Object Pose Estimation». In: *IEEE Transactions on Multimedia 24, pp. 3025-3035* (2022).

[12] Redmon J. e Farhadi A. «YOLOV3: An incremental improvement». In: *arXiv:1804.02767* (2018).

[13] J. Mei, X. Jiang e H. Ding. «Spatial feature mapping for 6DoF object pose estimation». In: *Pattern Recognition 131,108835* (2022).

[14] C. Mitash, A. Boularias e K. Bekris. «Physics-based scene-level reasoning for object pose estimation in clutter». In: *International Journal of Robotics Research 41(6), pp. 615-636* (2022).

[15] S. Peng et al. «PVNet: Pixel-Wise Voting Network for 6DoF Object Pose Estimation». In: *IEEE Transactions on Pattern Analysis and Machine Intelligence 44(6), pp. 3212-3223* (2022).

[16] T. Xie et al. «PANet: A Pixel-Level Attention Network for 6D Pose Estimation with Embedding Vector Features». In: *IEEE Robotics and Automation Letters 7(2), pp. 1840-1847* (2022).

[17] Z. Zhang et al. «Single RGB Image 6D Object Grasping System Using Pixel-Wise Voting Network». In: *Micromachines 13(2),293* (2022).

[18] Y. Di et al. «SO-Pose: Exploiting Self-Occlusion for Direct 6D Pose Estimation». In: *Proceedings of the IEEE International Conference on Computer Vision pp. 12376-12385* (2021).

[19] D. Liu et al. «6D Pose Estimation of Occlusion-Free Objects for Robotic Bin-Picking Using PPF-MEAM with 2D Images (Occlusion-Free PPF-MEAM)». In: *IEEE Access 9,9385060, pp. 50857-50871* (2021).

[20] Y. Shi et al. «StablePose: Learning 6D object poses from geometrically stable patches». In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition pp. 15217-15226* (2021).

[21] G. Wang et al. «GDR-Net: Geometry-guided direct regression network for monocular 6D object pose estimation». In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1660616616* (2021).

[22] B. Wen e K. Bekris. «BundleTrack: 6D Pose Tracking for Novel Objects without Instance or Category-Level 3D Models». In: *IEEE International Conference on Intelligent Robots and Systems pp. 8067-8074* (2021).

[23] G. Zhou et al. «A Novel Depth and Color Feature Fusion Framework for 6D Object Pose Estimation». In: *IEEE Transactions on Multimedia 23,9115222, pp. 1630-1639* (2021).

[24] W. Chen et al. «PointPoseNet: Point pose network for robust 6D object pose estimation». In: *Proceedings - 2020 IEEE Winter Conference on Applications of Computer Vision, WACV 2020 9093272, pp. 2813-2822* (2020).

[25] H. Huang et al. «An Occlusion-aware Edge-Based Method for Monocular 3D Object Tracking using Edge Confidence». In: *Computer Graphics Forum 39(7), pp. 399-409* (2020).

[26] C. Song, J. Song e Q. Huang. «HybridPose: 6D Object Pose Estimation under Hybrid Representations». In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 9157758, pp. 428-437* (2020).

[27] Z. Sui et al. «GeoFusion: Geometric Consistency Informed Scene Estimation in Dense Clutter». In: *IEEE Robotics and Automation Letters 5(4),9144435, pp. 5913-5920* (2020).

[28] K. Wada et al. «Morefusion: Multi-object reasoning for 6d pose estimation from volumetric fusion». In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 9157179, pp. 14528-14537* (2020).

[29] B. Wen et al. «Robust, Occlusion-aware Pose Estimation for Objects Grasped by Adaptive Hands». In: *Proceedings - IEEE International Conference on Robotics and Automation 9197350, pp. 6210-6217* (2020).

[30] B. Wen et al. «Se(3)-TrackNet: Data-driven 6D pose tracking by calibrating image residuals in synthetic domains». In: *IEEE International Conference on Intelligent Robots and Systems 9341314, pp. 10367-10373* (2020).

[31] Y. Wu, Y. Fu e S. Wang. «Deep instance segmentation and 6D object pose estimation in cluttered scenes for robotic autonomous grasping». In: *Industrial Robot 47(4), pp. 593-606* (2020).

[32] T. Zhang et al. «Cognitive Template-Clustering Improved LineMod for Efficient Multi-object Pose Estimation». In: *Cognitive Computation 12(4), pp. 834-843* (2020).

[33] J. Zhou et al. «BOLD3D: A 3D BOLD descriptor for 6Dof pose estimation». In: *Computers and Graphics (Pergamon) 89, pp. 94-104* (2020).

[34] Z. Dong et al. «PPR-Net:Point-wise Pose Regression Network for Instance Segmentation and 6D Pose Estimation in Bin-picking Scenarios». In: *IEEE International Conference on Intelligent Robots and Systems 8967895, pp. 1773-1780* (2019).

[35] G. Gao et al. «Occlusion resistant object rotation regression from point cloud segments». In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11129 LNCS, pp. 716-729* (2019).

[36] O. Hosseini Jafari et al. «iPose: Instance-Aware 6D Pose Estimation of Partly Occluded Objects». In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11363 LNCS, pp. 477-492* (2019).

[37] Y. Hu et al. «Segmentation-driven 6D object pose estimation». In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2019-June,8953567, pp. 3380-3389* (2019).

[38] J.-P. Mercier, P. Mitash C. adn Giguere e A. Boularias. «Learning object localization and 6d pose estimation from simulation and weakly labeled real images». In: *Proceedings - IEEE International Conference on Robotics and Automation 2019-May,8794112, pp. 3500-3506* (2019).

[39] K. Park, T. Patten e M. Vincze. «Pix2pose: Pixel-wise coordinate regression of objects for 6D pose estimation». In: *Proceedings of the IEEE International Conference on Computer Vision 2019-October,9008819, pp. 7667-7676* (2019).

[40] J. Sock et al. «Multi-task deep networks for depth-based 6D object pose and joint registration in crowd scenarios». In: *British Machine Vision Conference 2018, BMVC 2018* (2019).

[41] H. Tjaden et al. «A Region-Based Gauss-Newton Approach to Real-Time Monocular Multiple Object Tracking». In: *IEEE Transactions on Pattern Analysis and Machine Intelligence 41(8),8565885, pp. 1797-1812* (2019).

[42] H. Wang et al. « Normalized object coordinate space for category-level 6D object pose and size estimation». In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2019-June,8953761, pp. 2637-2646* (2019).

[43] Hinterstoisser S et al. «Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes». In: *Computer VisionACCV 2012: 11th Asian Conference on Computer Vision, Daejeon* (2013).

[44] Xiang Y., Schmidt T. e Narayanan V. «PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes». In: *Proc. Robot, Sci. Syst.* (2018).

[45] Hodan T., Haluza P. e Obdrzalek S. «T-less: An RGB-D dataset for 6D pose estimation of texture-less objects». In: *Proc. IEEE Winter Conf. Appl. Comput. Vis., 2017, pp. 880888* (2017).

[46] Zeng Andy, Yu Kuan-Ting e Song Shuran. «Multi-view Self-supervised Deep Learning for 6D Pose Estimation in the Amazon Picking Challenge». In: *IEEE International Conference on Robotics and Automation (ICRA) 2017* (2017).

[47] Brégier R. e Devernay F. «Symmetry aware evaluation of 3D object detection and pose estimation in scenes of many parts in bulk». In: *ICCV, 2017* (2017).

[48] P.-C. WU e Y.-Y. LEE. «A benchmark dataset for 6dof object pose tracking.» In: *ISMAR (2017), pp. 186191* (2017).

[49] A Doumanoglou e R Kouskouridas. «Recovering 6d object pose and predicting next-best-view in the crowd.» In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 35833592.* (2016).

[50] Wang He, Sridhar Srinath e Huang Jingwei. «Normalized object coordinate space for category-level 6d object pose and size esti- mation.» In: *Proceedings ofthe IEEE Conference on Computer Vision and Pattern Recognition, pages 26422651* (2019).

[51] Armeni I., Sax S. e Zamir A.R. «Joint 2D-3D-semantic data for indoor scene understanding». In: *CoRR, abs/1702.01105* (2017).

[52] Mitash C., Boularias A. e Bekris K. «Robust 6d object pose estimation with stochastic congruent sets». In: *29th British Machine Vision Conference, BMVC 2018* (2018).

[53] Fischler M A e Bolles R C. «Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography». In: *Commun. ACM, vol. 24, no. 6, pp. 381395* (1981).