Deep Learning for Industrial Robotics

Daniele De Gregorio[1], Gianluca Palli[2] e Luigi Di Stefano[3]
DISI[1,3]-DEI[2], Università degli Studi di Bologna, 40136 Bologna.
d.degregorio@unibo.it[1], gianluca.palli@unibo.it[2], luigi.distefano@unibo.it[3]

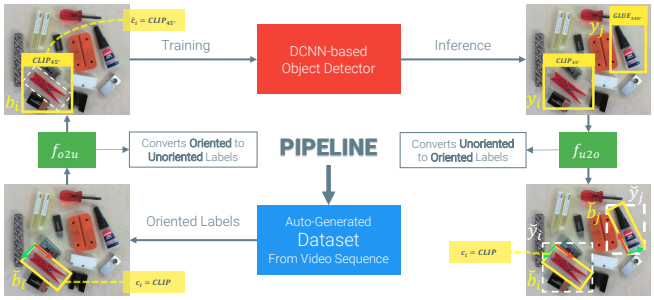

Figura 1: La pipeline complessiva del sistema proposto. Si parte dalla generazione completamente automatizzata di un dataset composto da Bounding Box Orientate. Queste ultime vengono convertite in Bounding Box Non-Orientate per addestrare un Object Detector convenzionale che codifica l'informazione di orientazione nel processo di classificazione. L' informazione di orientazione è poi sfruttata per ricostruire delle Bounding Box Orientate a partire da quelle non orientate.

Abstract

Questo contributo affronta la problematica dell'effettivo utilizzo di tecniche di Deep Learning per il riconoscimento di immagini in ambito industriale. In particolare, proponiamo un nuovo paradigma volto a estendere i moderni Object Detector 2D basati su modelli data-driven, come le CNN, ai fini dell'impiego in applicazioni di Pick&Place e Visual Servoing. In tale contesto, proponiamo anche un metodo automatizzato per generare un dataset con cui addestrare il modello. Grazie all'approccio proposto è possibile realizzare un sistema autoconfigurabile adatto ad ambienti non controllati che nei nostri esperimenti raggiunge più del 99% di precisione.

1 Introduction

L' Object Detection 2D multi-categoria è uno dei task di Computer Vision che ha raggiunto performance fino a poco tempo fa inimmaginabili grazie alle Convolutional Neural Networks (CNNs) [Redmon *et al.*, 2016]. Tuttavia, task molto rilevanti in ambito industriale come la Pose Estimation non hanno beneficiato di miglioramenti così netti nonostante i numerosi lavori basati su tecniche di Deep Learning [Sundermeyer *et al.*, 2018]. La nostra ipotesi circa questo divario in termini di prestazioni attiene la mancanza di dati di training adeguati e la difficoltà nel generarne di nuovi.

Quindi, focalizzare la ricerca su modelli data-driven, come le Deep Neural Nerworks, che siano in grado di andare oltre la semplice object detection senza fornire una soluzione concreta alla produzione dei dati con cui il modello sarà addestrato non pare una soluzione di interesse pratico in ambito industriale. Nei prossimi paragrafi descriveremo la nostra proposta, che consiste in due elementi distinti e complementari. Il primo elemento è l'estensione di un generico modello di Object Detector basato su CNN al fine di stimare anche l'orientamento e la scala degli oggetti, così che la CNN possa essere utlizzata in uno schema di controllo come modulo di Visual Servoing, per – e.g. – applicazioni di Pick&Place. Il secondo è un metodo completamente automatizzato per la generazione di un dataset per l'addestramento di un modello data-driven come quello precedentemnte descritto.

2 CNN per la Predizione dell'Orientazione

L'idea base è sfruttare a nostro vantaggio l'incapacità delle odierne CNN di essere invarianti alla rotazione. Come ampiamente dimostrato [Zhou *et al.*, 2017], questo tipo di reti impara più rappresentazioni degli oggetti quando essi appaiono ruotati all'interno del dataset di addestrameto. Noi abbiamo sfruttato questa multi-rappresentazione formulando la stima dell'angolo di un oggetto come un problema di classificazione, giacchè, di fatto, ogni versione ruotata dello stesso oggetto costituisce per la rete una categoria differente. Come mostrato in Figura 1, il dataset quindi è formato da Oriented Labels (Bounding Box orientate), ed ognuna di esse viene trasformata in una label Non-Orientata (funzione f_{o2u}) a cui viene assegnata una categoria che contiene intrinsecamente l'informazione circa l'orientazione originale (ogni oggetto viene rappresentato da k categorie che dipendono dal passo di discretizzazione angolare scelto). Con le label non-orientate è possibile addestrare un Object Detector standard che produrrà in uscita delle predizioni ugualmente non orientate. Queste ultime possono essere a loro volta riconvertite in predizioni orientate (funzione f_{u2o}) usando l'informazione angolare, condificata nel processo di classificazione stesso, per ricostruire una box orientata inscritta in quella non orientata. Implicitamente questo metodo codifica anche la scala del-

l'oggetto che può essere ricavata confrontando la dimensione delle box predette e quelle originali di riferimento.

Questo tipo di estensione può essere applicato a qualsiasi modello di CNN per object detection pre-esistente. Nei nostri esperimenti abbiamo utilizzato YOLOv3 [Redmon *et al.*, 2016], una CNN per object detection allo stato dell'arte. Nei nostri esperimenti abbiamo riscontrato che con questa tecnica ottieniamo non solo una *Mean Average Precision* superiore a 0.99 ma anche prestazioni migliori rispetto ai metodi classici attualmente utilizzati in ambito industriale, quali SIFT [Lowe, 2004] per gli oggetti *Textured* e BOLD [Tombari *et al.*, 2013] per gli oggetti *Untextured*.

3 Generazione Automatizzata del Dataset

Come introdotto in sezione 1, la disponibilità dei dati è un problema fondamentale che impedisce l'adozione massiccia del Deep Learning in ambito industriale. Per questo motivo abbiamo progettato, insieme al nostro modello di CNN per Object Detection, anche un metodo automatizzato per la generazione dei dati per addestrarlo. Nella pratica, dato che il modello proposto effettua la stima della posa 3-DoF, la situazione operativa sarà quella in cui gli oggetti, di dimensione per lo più omogenea, giaceranno su di un piano di lavoro e la camera potrà effettuare tutte le roto-traslazioni che mantengano il piano immagine parallelo a quello di lavoro. Il dataset può quindi essere generato con la stessa configurazione: acquisendo un video con il piano imamgine della camera parallelo al piano di lavoro è possibile, dopo aver generato manualmente le bounding box orientate nel primo frame della sequenza, propagarle in maniera automatizzata stimando l'omografia dei frame successivi rispetto al primo. Grazie a questa tecnica, a scopo dimostrativo, abbiamo generato un dataset completo, di 12 oggetti e 15 scene per un totale di circa 70000 immagini, con un intervento umano di poco più di 60 minuti. Il dataset è disponibile publicamente online[1] congiuntamente al software di esempio usato per la sua generazione.

4 Risultati Sperimentali

Grazie alla grande dimensione del dataset creato con il metodo descritto in precedenza, è stato possibile effettuare test estensivi del nostro approccio e comparare i risultati con quelli che si otterrebbero con metodi classici non data-driven. In Tabella 1 sono riassunte le Mean Average Precision complessive suddivise per scene *Simple* e *Hard*, per sottolineare come, intuitivamente, i metodi classici funzionino bene su situazioni semplici (Simple = scene con un background di colore uniforme) mentre per situazioni complesse (Hard = scene con un background molto complesso), laddove possa essere addestrato in maniera automatica con grandi quantità di dati, il Deep Learning costituisce oggi la soluzione migliore. La Figura 2 riporta alcuni risultati qualitiativi, sempre suddivisi in base alla loro difficoltà, ottenuti dal sistema sviluppato.

[1] https://github.com/m4nh/Loop

Model	Simple	Hard	Overall
SIFT [Lowe, 2004]	0.54	0.45	0.49
BOLD [Tombari *et al.*, 2013]	0.78	0.57	0.67
CNN - θ_5	0.96	0.96	0.96
CNN - θ_{10}	**0.99**	**0.98**	**0.99**

Tabella 1: La mean average precision (*mAP*) per i metodi classici di object detection basati su features SIFT e BOLD e per la nostra CNN con una discretizzazione dell'orientazione di 5 e 10 gradi.

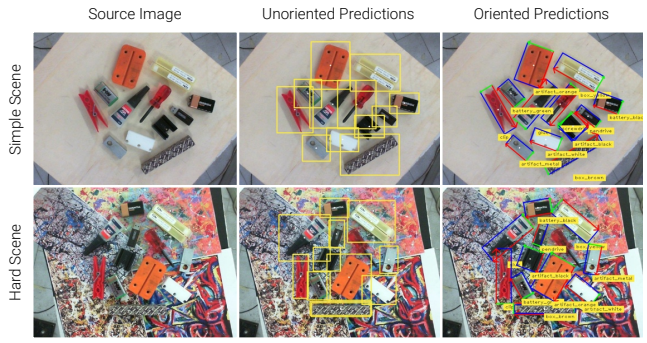

Figura 2: Risultati ottenuti con la nostra CNN. La colonna centrale mostra le predizioni non orientate dell'object detector 2D, mentre quella di destra le predizioni orientate generate grazie alle informazioni di angolo codificate nel processo di classificazione. Entrambe le scene mostrate sono state ottenute usando il nostro metodo per la generazione automatica dei dataset e non utilizzate nella fase di training.

Riferimenti bibliografici

[Lowe, 2004] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[Redmon *et al.*, 2016] Joseph Redmon, Santosh Divvala, Ross Girshick, e Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[Sundermeyer *et al.*, 2018] Martin Sundermeyer, Zoltan Marton, Maximilian Durner, e Rudolph Triebel. Implicit 3d orientation learning for 6d object detection from rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 699–715, 2018.

[Tombari *et al.*, 2013] Federico Tombari, Alessandro Franchi, e Luigi Di Stefano. Bold features to detect textureless objects. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1265–1272, 2013.

[Zhou *et al.*, 2017] Yanzhao Zhou, Qixiang Ye, Qiang Qiu, e Jianbin Jiao. Oriented response networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4961–4970. IEEE, 2017.