Lokalizacija in ocenjevanje lege predmeta v treh prostostnih stopnjah s središčnimi smernimi vektorji

Domen Tabernik, Jon Natanael Muhovič, Danijel Skočaj

Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
E-pošta: {domen.tabernik,jon.muhovic,danijel.skocaj}@fri.uni-lj.si

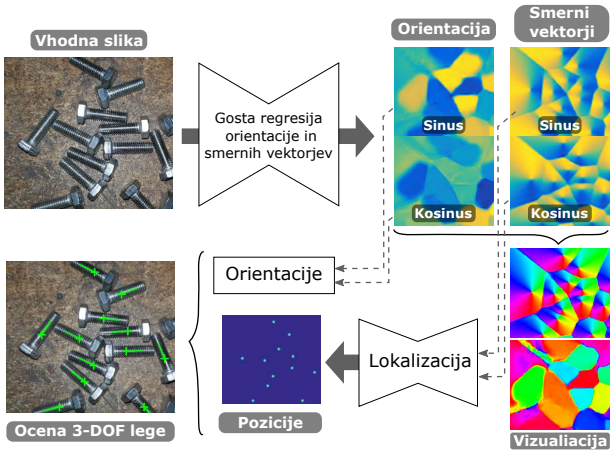
Localization and Pose Estimation of Objects in Three Degrees of Freedom using Central Directional Vector

In this paper, we propose an approach to localize and estimate the pose of objects in three degrees of freedom (3-DOF). Our method is based on point localization combined with regression of the orientation angle for each detected object. We extend existing point localization method to estimate the orientation of all detected objects in an image. The orientation regression is parameterized with trigonometric functions, similar to the direction to the object center. We evaluate our method on the proposed screw dataset, composed of a training set containing synthetic images with photorealistic appearance and a test set containing real images of screws. Compared to the state-of-the-art 6-DOF position estimation method applied to the 3-DOF problem, our approach achieves comparable results at a significantly lower computational cost.

1 Uvod

Ocenjevanje lege objekta v industrijskih aplikacijah robotskega prijemanja je ključnega pomena za manipulacijo predmetov. Zahteva natančno določanje lege predmeta v vseh šestih prostostnih stopnjah (6-DOF), kar vključuje pozicijo in orientacijo v treh dimenzijah. Vendar v industriji pogosto zadostuje ocena lege predmeta v treh prostostnih stopnjah (3-DOF), kar zajema oceno orientacije v eni dimenziji in lokalizacijo v dveh dimenzijah slike. To je dovolj za robotsko prijemanje, ko je robot omejen na eno površino s fiksnim pogledom od zgoraj navzdol, kar je pogost scenarij pri prijemanju predmetov s tekočega traku z ortogonalnim pogledom kamere. Prav tako se 3-DOF ocena izkaže za zadostno v domenah oddaljenega zaznavanja na satelitskih slikah, kjer je narava problema omejena na ortografski pogled.

V tem delu se ukvarjamo s problemom lokalizacije in ocenjevanja lege predmeta v treh prostostnih stopnjah. Trivialna rešitev tega problema bi bila uporaba najnaprednejših pristopov za ocenjevanje lege objekta v 6-DOF [9, 2], nato pa bi orientacijo projicirali samo na površino, ki nas zanima, in zanemarili dimenzije orientacije, ki niso ortogonalne pogledu kamere. Kljub fleksibilnosti, ta rešitev ni najbolj optimalna, saj je formulacija modela neposredno vezana na tri dimenzije, kar zahteva večjo kapaciteto modela za reševanje problema več dimenzij, kot



Slika 1: Predlagan pristop za lokalizacijo in oceno lege predmeta v 3-DOF.

je dejansko potrebno. Na drugi strani bi bilo oceno lege predmeta mogoče pridobiti tudi s pomočjo najnovejših pristopov za detekcijo predmetov z orientiranim očrtanim okvirjem [10, 4]. Ti pristopi definirajo predmet z lokacijo na sliki, velikostjo okvirja in kotom orientacije. Vendar je kot orientacije definiran le v območju od 0 do π, kar ni primerno za ocenjevanje 3-DOF lege predmeta, kjer zahtevamo orientacijo predmeta v celotnem krogu (od 0 do 2π).

Za reševanje problema lokalizacije in ocenjevanja lege predmeta v treh prostostnih stopnjah predlagamo pristop, ki sloni na točkovni lokalizaciji ter regresiji kota orientacije za zaznane objekte (Slika 1). Za točkovno lokalizacijo uporabimo metodo za lokalizacijo in štetje predmetov CeDiRNet [8], ki izvede lokalizacijo za vse objekte naenkrat, ter predlagamo njeno nadgradnjo v CeDiRNet-3DOF za dodatno ocenjevanje orientacije vseh zaznanih predmetov. Regresijo orientacije parametriziramo s trigonometrično funkcijo podobno kot je parametrizirana smer proti središču objekta v metodi CeDiRNet. Metodo preizkusimo na podatkovni zbirki vijakov, ki jo sestavljajo sintetične slike vijakov s foto-realističnim videzom za učenje modela, ter jo testiramo na realnih slikah vijakov. V primerjavi z najsodobnejšo metodo za ocenjevanje lege v 6-DOF aplicirano na problem 3-DOF dosežemo primerljiv rezultat pri občutno manjši računski zmogljivosti, saj ocenimo lego za vse objekte naenkrat ter lahko bolje iz-

koristimo kapaciteto modela za regresijo le potrebnih parametrov za tri prostostne stopnje.

2 Sorodna dela

Za oceno lege predmeta v 3-DOF lahko uporabimo obstoječe metode 6-DOF, med katerimi je najzmogljivejša metoda GDR-Net [9]. Metoda sloni na več-stopenjski obdelavi, kjer se v prvi fazi zazna objekte, nato pa aplicira mrežo za oceno lege za vsak zaznan objekt posebej. Za oceno lege se regresira 3D pozicijo ter parametrizirano 3D rotacijsko matriko z vmesno regresijo gostih točk korespondence. Izvedenka GDRNPP je na tekmovanju za oceno lege predmetov, ki je potekalo v okviru delavnice računalniškega vida za mešano resničnost na konferenci CVPR 2022, dosegla prvo mesto [7]. Težava teh pristopov je v več-stopenjski izvedbi, ki je odvisna od števila predmetov na sliki ter močno povečuje računsko zahtevnost metode.

V literaturi so bili raziskani tudi namenski pristopi za reševanje problema v 3-DOF. DeGregorio in sod. [1] so predlagali nadgradnjo za poljubno obstoječo metodo z očrtanim okvirjem. Predlagali so formulacijo orientiranega očrtanega okvirja z orientacijo definirano med 0 in 2π za dejansko oceno lege predmeta. Predikcijo orientacije so naslovili kot problem klasifikacije, kjer kot orientacije diskretizirajo na N delov ter jo učijo kot klasifikacijo s prečno entropijo za funkcijo izgube. Taka diskretizacija kota pa onemogoča doseganje večje natančnosti, saj se z večjo natančnostjo veča tudi dimenzija problema. Avtorji zato diskretizirajo kot le na 10°, kjer dobijo najboljše rezultate. Za ustrezno primerjavo z našo metodo, ki dosega natančnost pod 2° napake, bi tako potrebovali 10-krat bolj natančno diskretizacijo, pri čemer avtorji pokažejo, da že pri diskretizaciji na 5° uspešnost metode upade.

3 CeDiRNet-3DOF

Za reševanje problema lokalizacije in ocenjevanja lege predmeta v 3-DOF predlagamo pristop, ki temelji na gosti predikciji smernih vektorjev za lokalizacijo ter jo nadgradimo z gosto predikcijo smeri za orientacijo predmeta. Pristop se opira na naše predhodne delo za točkovno lokalizacijo in štetje predmetov z metodo CeDiRNet [8], ki jo v naslednjem razdelku povzamemo za lažje razumevanje, nato pa opišemo še predlagano nadgradnjo za predikcijo orientacije objekta. Nadgrajeno metodo poimenujemo CeDiRNet-3DOF.

3.1 Točkovna lokalizacija objekta s CeDiRNet

Točkovna lokalizacija objekta se izvede v dveh fazah. V prvi fazi izvedemo gosto regresijo smernih vektorjev središča objekta, nato pa v drugi fazi na podlagi regresiranih smernih vektorjev izvedemo lokalizacijo z manjšo nevronsko mrežo, ki je neodvisna od specifične domene.

Regresija smernih vektorjev središča objekta. Detekcija objektov poteka kot regresija smernih vektorjev, ki kažejo proti središču objekta. Smerne vektorje regresiramo v obliki goste predikcije za vsak piksel, kjer vsak piksel kaže proti središču najbližjega objekta. Regresija se izvaja tako za piksle, ki pripadajo objektu, kot tudi za piksle, ki so daleč stran od objekta. Čeprav je pričakovana informacija o poziciji središča objekta skoncentrirana ob objektu, pa je mogoče uspešno regresirati tudi smerne vektorje stran od objekta, če je dovzetno polje dovolj veliko. Smerne vektorje dodatno parametriziramo s pomočjo trigonometričnih funkcij. Za vsak piksel $\mathbf{x}_{i,j} = (i, j)$ tako regresiramo $sin(\phi)$ in $cos(\phi)$, kjer je ϕ definiran kot smer proti središču najbližjega objekta $\mathbf{y} = (n, m)$:

$$\phi_{i,j} = \tan\left(\frac{m-j}{n-i}\right). \tag{1}$$

S parametrizacijo smernega vektorja tako naslovimo problem cikličnosti kota ϕ, dodatno pa omejimo prostor vrednosti na [-1,1]. Za regresijo smernih vektorjev uporabimo arhitekturo kodirnik-dekodirnik z modeloma ConvNext [6] za kodirnik in FPN [5, 3] za dekodirnik. Model regresije učimo s funkcijo izgube L1.

Lokalizacija iz smernih vektorjev. Specifičen videz vrednosti $sin(\phi)$ in $cos(\phi)$ v samem središču objekta omogoča uporabo manjše nevronsko mrežo, ki vrača visoko verjetnost okoli centra, ter nizko verjetnost drugod. Z aplikacijo lokalnega maksimuma nato enostavno dobimo točko centra. Za nevronsko mrežo uporabimo arhitekturo peščene ure, kjer za kodirnik in dekodirnik uporabimo pri vsakem po 4 plasti konvolucije s 16 ali 32 kanali. Na zadnji plasti uporabimo več različnih faktorjev razširjanja (1, 4, 8 in 12) podobno kot na plasti ASPP v arhitekturi DeepLab. Nevronsko mrežo učimo s funkcijo izgube L1, kjer za ciljne vrednosti uporabimo Gausovo razdaljo do središča objekta. Lokalizacijsko mrežo učimo izključno na sintetično ustvarjenih podatkih, saj je mogoče vhodne podatke za učenje ustvariti povsem umetno s trigonometričnimi funkcijami neodvisno od domene.

3.2 Nadgradnja za ocenjevanje orientacije

Metodo CeDiRNet dodatno nadgradimo za regresijo orientacije objekta v dimenziji, ki je ortogonalna na pogled kamere. Orientacijo objekta Φ definiramo kot dodatno dimenzijo k smernem vektorju ϕ ter jo regresiramo v vsakem pikslu. Vsi piksli, ki imajo isti najbližji objekt, bodo regresirali iste vrednosti. Orientacijo parametriziramo s trigonometričnimi funkcijami na enak način kot smerne vektorje. Regresiramo torej dve dodatni polji $sin(\Phi)$ ter $cos(\Phi)$, ki ju učimo s funkcijo izgube L1. Razmerje med funkcijo izgube za smerne vektorje ϕ ter funkcijo izgube za orientacijo Φ dodatno utežimo:

$$\mathcal{L} = \mathcal{L}_{\phi} + \lambda \cdot \mathcal{L}_{\Phi}. \tag{2}$$

V vseh eksperimentih uporabimo $\lambda = 4$. Regresijo za orientacijo učimo le v okolici središča objekta (30×30 pikslov), kar se je izkazalo za uspešnejše od učenja na celotni sliki. V času inference sicer regresiramo orientacijo za vse piksle, vendar orientacijo za detektiran predmet preberemo le v detektirani točki centra.

Slika 2: Primeri sintetičnih učnih slik zgoraj in realnih testnih slik spodaj.

4 Eksperimentalni rezultati

Metodo smo preizkusili za lokalizacijo in ocenjevanje lege kovinskih vijakov, ter jo primerjali z najsodobnejšo metodo za ocenjevanje lege predmeta v 6-DOF, GDR-Net [9], oz. zmagajočo GDRNPP izvedenko iz tekmovanja BOP 2022 [7].

4.1 Podatkovna zbirka slik

Obstoječe javno dostopne zbirke obravnavajo izključno oceno lege v 6-DOF in tako niso primerne za vrednotenje problemov v 3-DOF. Za namen ocenjevanja lege predmeta v 3-DOF smo zato ustvarili novo podatkovno zbirko slik kovinskih vijakov.

Učna množica. Za učenje smo uporabili sintetično podatkovno množico s 1000 slikami. Slike prikazujejo pogled na kovinske vijake, ki so naključno razporejeni po površini s pogledom od zgoraj navzdol, tako da je površina z objekti pravokotna na pogled kamere. Vsaka slika vsebuje med 20 in 80 ne-prekrivajočih se objektov, ki so bili naključno postavljeni na površini v skladu s fizikalnimi zakonitostmi. Na slikah so upodobljeni objekti s foto-realističnim videzom kovine, medtem ko smo za ozadje uporabili naključno izbran foto-realističen material izmed 18 različnih tipov (les, kovina, kamen, plastika, asfalt, itd.). Primeri iz sintetične podatkovne zbirke so prikazani v zgornji vrstici na Sliki 2.

Testna množica. Za vrednotenje smo zajeli množico slik z realnimi predmeti. Množico smo ustvarili na dveh podlagah (leseni in kovinski), kjer smo naključno postavili okoli 20 kovinskih vijakov in jih zajeli z zgornjega pogleda. Vsako postavitev predmetov smo zajeli pri treh različnih osvetlitvah. Skupaj smo tako zajeli 48 slik s 836 označenimi objekti (označeni s središčem predmeta in orientacijo). Primeri zajetih slik so prikazani v spodnji vrstici na Sliki 2.

4.2 Mere vrednotenja

Uspešnost določanja pozicije in lege objekta v 3-DOF merimo z dvema tipoma napak: i) napaka pri lokalizaciji na sliki, izražena v evklidski razdalji do središča objekta v pikslih, ter ii) napaka pri orientaciji, izražena v stopinjah. Ker je izmerjena napaka lege objekta odvisna tudi od števila pravilno zaznanih predmetov, dodatno poročamo uspešnost detekcije predmeta z metodami priklica, natančnostjo, ter mero F1, ki so izračunane za posamezno sliko ter povprečene preko vseh testnih slik. Za pravilno detekcijo štejemo predikcije, ki so od dejanskega središča objekta oddaljene največ 20 pikslov.

Slika 3: Primeri lokalizacije in ocene lege v 3-DOF, kjer z modro barvo označimo zlati rez, z zeleno pravilno detekcijo ter z rdečo napačno detekcijo.

4.3 Rezultati

Rezultati so poročani v Tabeli 1. Vse eksperimente smo ponovili 10-krat in poročamo povprečne vrednosti ter standardni odklon preko različnih ponovitev. Obe preizkušeni metodi izkazujeta odlične rezultate pri ocenjevanju pozicije in orientacije na ustvarjeni podatkovni množici. Naša predlagana metoda CeDiRNet-3DOF oceni orietnacijo objekta v povprečju z napako 1,92°, medtem kot je sorodna metoda malenkost boljša z napako 1,66°. CeDiRNet-3DOF uspe lokalizirati objekt s povprečno napako 3,26 pikslov, medtem ko uspe GDR-Net+YOLOX lokalizirati objekt s povprečno napako 2,54 pikslov. Vse vrednosti so poročane pri upragovanju, ki doseže najboljšo F1 mero. Pri tem lahko opazimo, da je naša predlagana metoda boljša pri sami detekciji predmetov. CeDiRNet-3DOF doseže v povprečju F1 mero 99,73% s povprečnim priklicem in natančnostjo nad 99,6% in 99,8%. Na drugi strani ima metoda GDR-Net+YOLOX slabšo detekcijo z F1 mero nižjo za 0,75 p.t., povprečenim priklicem nižjim za 0,91 p.t., ter povprečno natančnostjo nižjo za 0,55 p.t. Nekaj primerov detekcije in ocene lege v 3-DOF je

	Detekcija predmeta			*Napaka ocena lege in pozicije*	
	Priklic [%]	*Natančnost [%]*	*Mera F1 [%]*	*Lokalizacija [px]*	*Orientacija [°]*
CeDiRNet-3DOF	**99,61 ± 0,0840**	**99,86 ± 0,103**	**99,73 ± 0,077**	3,26 ± 0,0857	1,92± 0,229
GDR-Net/GDRNPP +YOLOX [9]	98,70 ± 0,160	99,31 ± 0,162	98,98 ± 0,161	**2,54 ± 0,062**	**1,66 ± 0,102**

Tabela 1: Rezultati na podatkovni zbirki kovinskega vijaka. Poročamo vrednosti povprečene preko 10 ponovitev skupaj s standardnim odklonom.

	Čas izvajanja	
CeDiRNet-3DOF	**46,3 ms**	**21,59 FPS**
GDR-Net/GDRNPP+ YOLOX [9]	335,4 ms	2,89 FPS

Tabela 2: Primerjava časa izvajanja metode na velikosti slike 640x480.

prikazanih na Sliki 3.

V Tabeli 2 poročamo tudi primerjavo hitrosti izvajanja obeh metod preizkušenih na velikost slike 640x480 z grafično kartico NVIDIA A100 40GB. Metoda CeDiRNet-3DOF doseže poročano uspešnost detekcije pri občutno manjši računski zahtevnosti kot sorodna metoda. CeDiRNet-3DOF izvede detekcijo in oceno lege v le 46 milisekundah, medtem ko sorodna metoda potrebuje 145 milisekund za detekcijo objektov z metodo YOLOX ter dodatnih 190 milisekund za oceno lege z metodo GDR-Net. Skupaj potrebuje metoda GDR-Net+YOLOX tako 335,36 milisekund, ter lahko obdela manj kot 3 slike na sekundo medtem ko naša predlagana metoda obdela preko 21 slik na sekundo. Pri obeh metodah uporabljamo isto arhitekturo ConvNext-base, ter poročamo samo čas procesiranja globoke mreže na grafični enoti ter obdelavo rezultatov za končni izhod, brez časa potrebnega za pripravo vhodnih podatkov. Metoda GDR-Net se izkaže za časovno bolj potratno od metode CeDiRNet-3DOF, saj aplicira oceno lege za vsak objekt ločeno na izrezane dele slike. Pri tem nismo upoštevali časa za premik podatkov v pomnilniku, ki je potreben za izrez slike pri metodi GDR-Net zaradi česar se čas izvajanja še dodatno poveča.

5 Zaključek

V tem delu smo predlagali nadgradnjo metode CeDiRNet za ocenjevanje orientacije objekta, kjer metodo za lokalizacijo in štetje predmetov s predikcijo smernih vektorjev [8] razširimo s predikcijo orientacije objekta, ter omogočimo ocenjevanje lege objektov v treh prostostnih stopnjah. Predlagana metoda izvede oceno lege na učinkovit način, saj v enem koraku izvede sočasno regresijo smernih vektorjev ter orientacije za vse objekte na sliki, nato pa se vse objekte lokalizira v enem dodatnem koraku z majhno lokalizacijsko mrežo.

Metodo smo primerjali z najsodobnejšo metodo za oceno lege v šestih prostostnih stopnjah, GDR-Net oz. njeno izpeljanko GDRNPP z detektorjem YOLOX, ki je zmagovalna metoda s tekmovanja BOP 2022 [7]. Metodi smo preizkusili na novi podatkovni zbirki kovinskih vijakov. Metoda GDR-Net/YOLOX se sicer izkaže za malenkost bolj uspešno pri natančnosti lokalizacije in oceni lege, vendar naša predlagana metoda doseže boljšo uspešnost detekcije pri občutno manjši računski zahtevnosti. Metoda CeDiRNet-3DOF doseže preko 21 FPS, kar je skoraj 7-krat več kot metoda GDR-Net/YOLOX z le 3 FPS. Pomemben razlog za hitrejše delovanje je v sočasnem izračunu pozicije in lege za vse objekte na sliki, medtem ko metoda GDR-Net aplicira mrežo za oceno lege za vsak zaznan objekt posebej.

Literatura

[1] Daniele de Gregorio, Riccardo Zanella, Gianluca Palli, and Luigi Di Stefano. Effective deployment of CNNs for 3DOF pose estimation and grasping in industrial settings. *Proceedings - International Conference on Pattern Recognition*, pages 7419–7426, 2020.

[2] Yan Di, Fabian Manhardt, Gu Wang, Xiangyang Ji, Nassir Navab, and Federico Tombari. SO-Pose: Exploiting Self-Occlusion for Direct 6D Pose Estimation. *Proceedings of the IEEE International Conference on Computer Vision*, 1:12376–12385, 2021.

[3] Wei Li, Hongliang Li, Qingbo Wu, Xiaoyu Chen, and King Ngi Ngan. Simultaneously Detecting and Counting Dense Vehicles From Drone Images. *IEEE Transactions on Industrial Electronics*, 66(12):9651–9662, 12 2019.

[4] Wentong Li, Yijie Chen, Kaixuan Hu, and Jianke Zhu. Oriented RepPoints for Aerial Object Detection. In *Computer Vision and Pattern Recognition*, volume 2022-June, pages 1819–1828, 2022.

[5] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature Pyramid Networks for Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 936–944. IEEE, 7 2017.

[6] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. pages 11966–11976, 2022.

[7] Martin Sundermeyer, Tomas Hodan, Yann Labbe, Gu Wang, Eric Brachmann, Bertram Drost, Carsten Rother, and Jiri Matas. BOP Challenge 2022 on Detection, Segmentation and Pose Estimation of Specific Rigid Objects. In *CVPRW (CV4MR workshop)*, 2023.

[8] Domen Tabernik, Jon Natanael Muhovič, and Danijel Skočaj. Dense Center-Direction Regression for Object Counting and Localization with Point Supervision. In *https://prints.vicos.si/publications/files/424*, 2023.

[9] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. GDR-Net: Geometry-guided direct regression network for monocular 6D object pose estimation. In *Computer Vision and Pattern Recognition*, pages 16606–16616, 2021.

[10] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented R-CNN for Object Detection. In *International Conference on Computer Vision*, pages 3520–3529, 2021.

Zahvala: To delo je bilo delno financirano s strani projektov ARIS J2-3169 (MV4.0) in J2-4457 (RTFM) ter raziskovalnega programa Računalniški vid (P2-0214).