KU LEUVEN

FACULTEIT INDUSTRIËLE INGENIEURSWETENSCHAPPEN

CAMPUS DE NAYER SINT-KATELIJNE-WAVER

Visie voor semantische robotnavigatie in ziekhuisgangen

Olivier VAN DEN EEDE

Promotor(en): Prof. dr. ir. Toon Goedemé

Co-promotor(en): Filip Reniers

Masterproef ingediend tot het behalen van de graad van master of Science in de industriële wetenschappen: Electronica-ICT afstudeerrichting ICT

Academiejaar 2018 - 2019

Voorwoord

Het voorwoord vul je persoonlijk in met een appreciatie of dankbetuiging aan de mensen die je hebben bijgestaan tijdens het verwezenlijken van je masterproef en je hebben gesteund tijdens je studie.

Samenvatting

De (korte) samenvatting, toegankelijk voor een breed publiek, wordt in het Nederlands geschreven en bevat **maximum 3500 tekens**. Deze samenvatting moet ook verplicht opgeladen worden in KU Loket.

Abstract

Het extended abstract of de wetenschappelijke samenvatting wordt in het Engels geschreven en bevat **500 tot 1.500 woorden**. Dit abstract moet **niet** in KU Loket opgeladen worden (vanwege de beperkte beschikbare ruimte daar).

Keywords: Voeg een vijftal keywords in (bv: Latex-template, thesis, ...)

Inhoudsopgave

Lijst van figuren

Lijst van tabellen

Acroniemen

CNN Convolutional Neural Network. 2, 3, 5

EM Expectation-maximization. 4

HOG Histogram of Oriented Gradients. 2

HSI Hue Saturation Intensity. 2

OCR Optical character recognition. 1

RANSAC Random sample consensus. 2, 4

RGB Rood Groen Blauw. 1, 5

ROI Region Of Intrest. 4

SIFT Scale-invariant feature transform. 2, 4

SVM Support Vector Machine. 2, 4

YOLO You Only Look Once. 3

Hoofdstuk 1

Literatuurstudie

1.1 Indoor navigatie & visie

Op visie gebaseerde navigatie is een onderwerp dat zeer vaak onderzocht wordt. Oudere onderzoeken zoals [16] maken gebruik van een robot met een Rood Groen Blauw (RGB) camera die zonder kaart informatie. De enige informatie die gegeven wordt is een eenvoudige object beschrijving van de gang en een beschrijving van een deur met een deurnummer ernaast. Met enkel een deurnummer als doel vertrekt de robot door de gang, en houd zichzelf parallel met de muren door gebruik te maken van andere sensoren. Eens er een deur in beeld komt, worden er een aantal features(randen) herkent in het beeld. Nadat de deuren herkend worden kan er via Optical character recognition (OCR) de deurnummer herkend worden en nagegaan of het doel bereikt is. Dit is uiteraard een zeer eenvoudige techniek omdat de robot geen begrip heeft van de omgeving, en moeilijk plaatsen t.o.v elkaar kan onderscheiden.

Nieuwere technieken zoals [6] maken gebruik van RGB-D camera's zoals bijvoorbeeld een kinect waardoor ze ook over diepte informatie beschikken. Die diepte info kan dan gebruikt worden om heel de omgeving in 3d te mappen en op basis van de effectief gemeten positie navigatie te doen. Een andere manier om een 3d representatie van de omgeving te verkrijgen zoals [13] is gebruik te maken van stereo visie. Hierbij wordt de informatie van 2 RGB camera's die op een vaste afstand van elkaar staan gecombineerd om diepte informatie te verzamelen.

In dit onderzoek gaan we ons echter beperken tot een enkele RGB camera.

1.2 Object detectie

Een belangrijk aspect van dit onderzoek is het detecteren van individuele objecten in het beeld van 1 enkele RGB camera. De te detecteren objecten zijn op voorhand vastgelegd, en zijn afhankelijk van de ruimte waarin de robot zich bevindt.

In de logistieke gangen van een ziekenhuis zijn er heel wat objecten te zien die we kunnen detecteren, een kleine selectie van deze objecten zijn.

- Pictogrammen
- Brandblussers
- Deurklinken

Voor deze objecten gaan we kijken naar detectie technieken uit de traditionele beeldverwerking, en naar meer *state of the art* technieken.

1.2.1 Traditionele object detectie

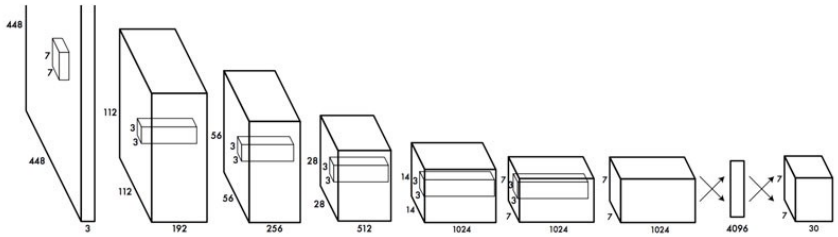
In openbare gebouwen zijn er heel wat pictogrammen te vinden zoals nooduitgang, hoogspanning en brandblusser. Deze pictogrammen hebben steeds een specifieke vorm, kleur en symbool. De literatuur leert ons weinig over pictogramdetectie, maar pictogrammen kunnen wel vergeleken worden met verkeersborden die bijna dezelfde kenmerken hebben. De aanpak van [5] is om 2 soorten features in een beeld te onderscheiden. In eerste instantie detecteren ze vormen op basis van kleur randen en anderzijds wordt de afbeelding omgezet naar Hue Saturation Intensity (HSI) waaruit enkel de hue gebruikt wordt. De hue is de belangrijkste component voor het onderscheiden van kleuren omdat er zo geen rekening wordt gehouden met de hoeveelheid licht en schaduwen. Een recenter onderzoek [17] bouwt voort op deze technieken, maar berekenen de Histogram of Oriented Gradients (HOG) features van het beeld. Vervolgens wordt er gebruik gemaakt van een Support Vector Machine (SVM) om te bepalen waar er zich een match bevindt.

De vorm en kleur features kunnen dan gecombineerd worden om de plaats voor een mogelijke match te vinden. Eens er een mogelijke boundig box gevonden is, kan er geprobeerd worden een template te matchen om het effectieve pictogram te achterhalen. Het grootste probleem bij de techniek van [5] is dat hun gebruikte template matching techniek niet robuust is voor schaal invariantie. Bij [17] maken ze voor de herkenningsfase gebruik van Scale-invariant feature transform (SIFT)[9] features en kleur informatie. Hierdoor is het probleem van schaal invariantie grotendeels opgelost. Hierbij worden de SIFT features van de kandidaat matches en de templates vergeleken, en er wordt een gemiddelde genomen van de verschillen tussen hue, saturation en value. Door middel van Random sample consensus (RANSAC) en een treshold wordt er bepaald welke matches gebruikt worden. Deze techniek zou gebruikt kunnen worden voor het detecteren van pictogrammen.

1.2.2 Convolutional neural nework

De laatste jaren in het domein van beeldverwerking wordt er steeds meer gegrepen naar deep learning technieken. Dit is komt omdat rekenkracht steeds beter en beter wordt, en de resultaten die bekomen worden de traditionele manieren overtreffen op verschillende vlakken. Een deep learning techniek die veel gebruikt wordt in de beeldverwerking is een Convolutional Neural Network (CNN).

Een CNN is een supervised deep learning techniek die gebruikt kan worden om complexere beeld-interpretatie te doen. Een CNN kan bestaan uit meerdere lagen die meestal een combinatie zijn van 'convolutional-layers' en 'fully connected-layers'. Elk van deze lagen bevat een aantal neuronen met elk een eigen set van gewichten. Het doel van een CNN is om de gewichten zodanig bij te stellen zodat data die aan de eerste laag gegeven wordt een verwacht resultaat geeft aan de laatste laag. Deze laatste laag kan men de classificatielaag noemen, en geeft een representatie van wat het netwerk denkt dat er aan de input staat. In figuur 1.1 is een voorbeeld tezien van een CNN met de verschillende soorten lagen.



Figuur 1.1: De lagen van een CNN volgens het YOLO [11] detection system.

Een 'convolutional-layer' is een laag die een convolutie operatie uitvoert op zijn input, de convolutie gebeurd d.m.v een masker dat meestal voorgesteld wordt als een tensor. Door een tensormasker te gebruiken kan de operatie uitgevoerd worden op meerdere inputdimensies tegelijkertijd, denk hierbij aan bijvoorbeeld 3 kleurkanalen.

Om uiteindelijk een classificatie te verkrijgen moet er een dimensievermindering doorgevoerd te worden, dit wordt gedaan door 'pooling layers' aan het netwerk toe te voegen na elke convolutie laag.

Een CNN kan pas gebruikt worden nadat het getraind is. Voor de training van een netwerk zijn er 2 dingen noodzakelijk, veel voorbeeld data en per voorbeeld de verwachte output (label). Bij het trainingsproces wordt alle inputdata aangelegd, en wordt er gekeken wat het netwerk aan zijn output heeft. De loss functie is een maat van hoe goed een netwerk een voorspelling kan doen van de input data, met andere woorden een vergelijking tussen de input en de output. Het doel van de training van een netwerk is het minimaliseren van deze loss functie. Dit kan gedaan worden d.m.v 'backprogagation'. Backpropagation is het steeds een klein beetje aanpassen van de gewichten in de inwendige neuronen om zo het resultaat te verbeteren en de loss functie te verkleinen. Een netwerk heeft een goede training gehad als de loss functie minimaal is.

Een voorbeeld van een CNN is het 'You Only Look Once (YOLO) detection system' [11]. Het YOLO netwerk is opgebouwd uit 24 convolutielagen en 2 fully connected layers. Dit netwerk heeft een uitgebreide training gehad op de ImageNet dataset en kan gebruikt worden om object detectie en classificatie te doen door 1 keer de input afbeelding door het netwerk te laten gaan. Door middel van een hertraining kan deze detector leren om alle objecten te detecteren en te classificeren en dus een mogelijke detector zijn voor onze toepassing.

Zoals [8] voorstelt is het niet moeilijk om het 'YOLO detection system' een hertraining te geven om deuren te herkennen. Zo kan dit ook toegevoegd worden aan de lijst met te detecteren kenmerken.

1.3 Object tracking

Object tracking of het volgen van objecten heeft als doel het bepalen van de positie van hetzelfde object over meerdere frames heen. In het geval van dit onderzoek kan het een indicatie geven van relatieve posities t.o.v. objecten die zich in de gangen bevinden. Een grote moeilijkheid bij het volgen van objecten stilstaan t.o.v. de camera is dat ze veranderen in grootte, oriëntatie en perspectief. [19] stelt voor om gebruik te maken van SIFT voor het volgen van objecten. Ze zoeken een Region Of Intrest (ROI) op het eerste frame waarop ze een kleurhistogram en SIFT features berekenen. Op het volgende frame worden dezelfde bewerkingen uitgevoerd in een regio die net iets groter is dan de originele ROI. Een overeenkomst regio wordt dan berekend door middel van een Expectation-maximization (EM) algoritme. Volgens [3] is het beter om gebruik te maken van het KLT feature algoritme [15]. Hiermee wordt er een transformatie berekend waardoor de ROI tussen de 2 frames gelijkaardig wordt. De initiële transformatieparameters worden berekend via RANSAC.

Schrijf nog over [10].

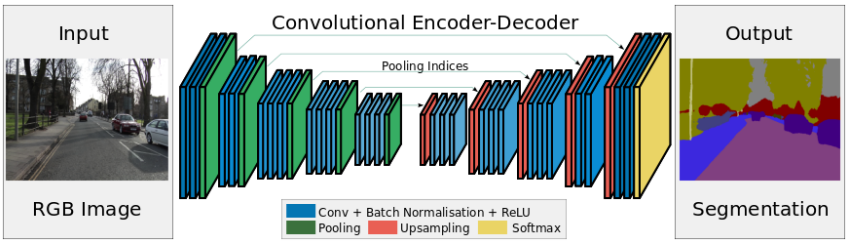
1.4 Image segmentation

Het correct segmenteren van de beelden zal een belangrijke rol spelen. Niet in elk beeld zal er een distinctief object aanwezig zijn om te detecteren. Daarom is het belangrijk om de vloer van de muren te kunnen onderscheiden. Een eenvoudige aanpak zou kunnen zijn om via K-means een verdeling van een beeld te doen en met een soort regressie de regio's te labelen. Volgens [18] werkt de K-means aanpak met een op textuur en kleur gebaseerde aanpak redelijk goed, maar wordt steeds de muur verbonden met het plafond omwille van kleur en textuur gelijkenissen. Hun regressie gebaseerde labeling techniek blijkt echter een slechte oplossing. Verder zoals [7] aangeeft zijn reflecties en overbelichting eigenschappen van indoor omgevingen die het moeilijk kunnen maken om een correcte segmentatie te doen. [7] stelt een techniek voor die begint met het detecteren van verticale en horizontale lijn segmenten. Dit doen ze door eerst een Canny edge detector[4] toe te passen en vervolgens een line fitting. Een zelf geleerde SVM classifier verdeeld alle lijnsegmenten in 2 categorieën namelijk horizontaal en verticaal. De vluchtlijnen van de gang worden hierbij onderverdeeld in de horizontale categorie. Alle lijnstukken krijgen een score via een reeks van operaties waarna enkel de beste lijnen bijgehouden worden. Op basis van de kleur van de vlakken tussen de lijnstukken kan een segmentatie gemaakt worden. Dit geeft een resultaat waarbij de vloer meestal een mooi homogeen geheel is, maar de muren worden in meerdere vlakken gesegmenteerd door eventuele kleurverschillen en objecten aan de muur.

Een andere manier om de vloer te segmenteren is voorgesteld in [12]. Zij doen een superpixel segmentatie volgens het SLIC algoritme [1], vervolgens bekijken ze de randen van de superpixels. Na observaties blijkt dat de randen van superpixels onregelmatig worden bij objectovergangen. Door het aanduiden van een paar vloerpixels kan hun algoritme superpixels aanduiden die tot de vloer behoren. Deze aanpak geeft een goede schatting van vrije ruimte op de vloer, maar is minder

bruikbaar voor segmentatie van muren.

Een meer recente technologie om afbeeldingen te segmenteren is gebruik te maken van een CNN. Het netwerk voor segmentatie is verschillend van een traditioneel CNN voor bijvoorbeeld object detectie. Een voorbeeld van een segmentatienetwerk is te zien in figuur 1.2.



Figuur 1.2: Het SegNet [2] segmentatie netwerk.

Het segmentatienetwerk SegNet [2] is een combinatie van convolutielagen en pooling layers, er zijn geen fully connected layers aanwezig zoals het geval is bij een classificatie netwerk. De bedoeling van het SegNet netwerk is om als output opnieuw een afbeelding te genereren, daarom zijn de lagen opgebouwd als een zandloper, op deze manier is de output even groot als de oorspronkelijke afbeelding. Een segmentatienetwerk wordt getraind op gelijkaardige manier aan een traditioneel CNN met als verschil dat de labeling gebeurd op pixelbasis aangezien de output even groot is als de input van het systeem. De output van het systeem is een per pixel gelabelde afbeelding afhankelijk van het aantal classen waarmee het systeem getraind is.

Het SegNet netwerk is getraind op de SUN RGB-D [14] dataset. Deze dataset bevat een groot aantal indoor scenes, waarbij er onder andere segmentatie klassen zijn voor muren, vloeren en plafonds. De training is gebeurd met enkel de RGB gegevens van de dataset. Deze trainingsdata zou uiteraard nuttig kunnen zijn voor dit onderzoek.

Hoofdstuk 2

Bibliografie

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. SÃ$\frac{1}{4}$sstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, Nov 2012.

[2] Vijay Badrinarayanan, Alex Kendall, Roberto Cipolla, and Senior Member. SegNet : A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. pages 1–14.

[3] Bhakti Baheti, Ujjwal Baid, and Sanjay Talbar. An approach to automatic object tracking system by combination of SIFT and RANSAC with mean shift and KLT. *Conference on Advances in Signal Processing, CASP 2016*, pages 254–259, 2016.

[4] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, Nov 1986.

[5] Chiung Yao Fang, Sei Wang Chen, and Chiou Shann Fuh. Road-sign detection and tracking. *IEEE Transactions on Vehicular Technology*, 52(5):1329–1341, 2003.

[6] Peter Henry, Michael Krainin, Evan Herbst, Xiaofeng Ren, and Dieter Fox. Rgb-d mapping: Using depth cameras for dense 3d modeling of indoor environments. In *In the 12th International Symposium on Experimental Robotics (ISER*, 2010.

[7] Yinxiao Li and Stanley T. Birchfield. Image-based segmentation of indoor corridor floors for a mobile robot. *IEEE/RSJ 2010 International Conference on Intelligent Robots and Systems, IROS 2010 - Conference Proceedings*, pages 837–843, 2010.

[8] Adrian Llopart, Ole Ravn, and Nils A. Andersen. Door and cabinet recognition using Convolutional Neural Nets and real-time method fusion for handle detection and grasping. *2017 3rd International Conference on Control, Automation and Robotics, ICCAR 2017*, pages 144–149, 2017.

[9] D G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157 vol.2, sep 1999.

[10] Guanghan Ning. Spatially Supervised Recurrent Convolutional Neural Networks for Visual Object Tracking. (1):1–4, 2017.

[11] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016.

[12] F. Geovani Rodríguez-Telles, L. Abril Torres-Méndez, and Edgar A. Martínez-García. A fast floor segmentation algorithm for visual-based robot navigation. *Proceedings - 2013 International Conference on Computer and Robot Vision, CRV 2013*, pages 167–173, 2013.

[13] K. Schmid, T. Tomic, F. Ruess, H. HirschmÃ$\frac{1}{4}$ller, and M. Suppa. Stereo vision based indoor/outdoor navigation for flying robots. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3955–3962, Nov 2013.

[14] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[15] Carlo Tomasi and Takeo Kanade. Detection and tracking of point features. 1991.

[16] M Tomono and S Yuta. Mobile robot navigation in indoor environments using object and character recognition. In *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No.00CH37065)*, volume 1, pages 313–320 vol.1, apr 2000.

[17] S. J. Zabihi, S. M. Zabihi, S. S. Beauchemin, and M. A. Bauer. Detection and recognition of traffic signs inside the attentional visual field of drivers. *IEEE Intelligent Vehicles Symposium, Proceedings*, (Iv):583–588, 2017.

[18] Zhong-Ju Zhang. Wall, floor, ceiling, object region identification from single image.

[19] Huiyu Zhou, Yuan Yuan, and Chunmei Shi. Object tracking using SIFT features and mean shift. *Computer Vision and Image Understanding*, 113(3):345–352, 2009.

FACULTEIT INDUSTRIËLE INGENIEURSWETENSCHAPPEN
CAMPUS DE NAYER SINT-KATELIJNE-WAVER
J. De Nayerlaan 5
2860 SINT-KATELIJNE-WAVER, België
tel. + 32 15 31 69 44
iiw.denayer@kuleuven.be
www.iiw.kuleuven.be